Towards the French Biomedical Ontology Enrichment

Juan Antonio Lossio-Ventura

Collège Doctoral du Languedoc-Roussillon

THESIS

To obtain the dregree of PHD - Doctor

Awarded by **University of Montpellier**

Prepared at **I2S**[*] Graduate School,
UMR 5506 Research Unit, ADVANSE Team,
and Laboratory of Informatics, Robotics And Microelectronics of Montpellier

Speciality: **Computer Science**

Defended by **Mr Juan Antonio LOSSIO-VENTURA**
juan.lossio@lirmm.fr

Towards the French Biomedical Ontology Enrichment

Defended on 09/11/2015 in front of a jury composed by:

Sophia ANANIADOU	Professor	Univ. of Manchester	President
Fabio CRESTANI	Professor	Univ. of Lugano	Reviewer
Pierre ZWEIGENBAUM	Professor	LIMSI-CNRS	Reviewer
Natalia GRABAR	Researcher	CNRS	Examinator
Mathieu ROCHE	Research Professor	Cirad, TETIS, LIRMM	Director
Clement JONQUET	Associate Professor	Univ. of Montpellier	Advisor
Maguelonne TEISSEIRE	Research Professor	TETIS, LIRMM	Advisor

[<Dreams have only one owner at a time. For that, dreamers tend to be alone.]

[*] **I2S:** INFORMATION, STRUCTURES AND SYSTEMS.

UNIVERSITÉ MONTPELLIER

Dedication

This thesis is lovingly
dedicated to my mother Laly Ventura.
The fact of making me stronger,
her support, encouragement, and
constant love have positively
influenced all my life.

Acknowledgments

During the past three years, I met a lot of wonderful people. They helped me without asking any response. These people contributed to this thesis as well as to my personal development.

First, I would like to thank my thesis committee: Prof. Sophia Ananiadou, Prof. Fabio Crestani, Prof. Pierre Zweigenbaum, and Dr. Natalia Grabar, for their time reading deeply my thesis, for their insightful comments and for the hard questions which allowed me to improve my research from various perspectives.

I also would like to express my sincere gratitude to my advisors Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire for the continuous support of my PhD and related research. Thanks for their patience, motivation, encouragement, knowledge and inspirational guidance during the three years of my thesis. Their enthusiasm and encouragement have been a great motivation to me. I would not change them, I could not have found better advisors and mentors for my PhD.

My sincere thanks also goes to Prof. Pascal Poncelet, who provided me the opportunity to join the ADVANSE team, giving me access to the laboratory and research facilities. His precious support made possible to conduct other related research to my PhD.

I am extremely grateful to my colleagues Lilia, Jessica, Sarah, Amine, Mike, Vijay, etc. for their help and useful discussions during my research. We had stimulating discussions, as well as several sleepless nights, which served to have fun. In particular, I am grateful to Lilia, Jessica, and Mike.

I sincerely thanks all the ADVANSE team for their moral support and encouragement during my stay in Montpellier.

Finally, I want to express my very profound gratitude to my mother and to my sisters for providing me unfailing support and continuous encouragement throughout my three years of research. This accomplishment would not have been possible without them. Thank you.

This thesis was supported in part by the French National Research Agency under the JCJC program, grant ANR-12-JS02-01001, as well as by the University of Montpellier, CNRS, IBC of Montpellier project and the FINCyT program, Peru.

Abstract

Big Data for the biomedical domain involves a major issue: the analysis of large volumes of heterogeneous data (e.g. video, audio, text, image). Ontology, i.e. conceptual models of the reality, can play a crucial role in biomedical fields for automating data processing, querying, and matching heterogeneous data. Various English resources exist, but considerably fewer are available in French and there is a substantial lack of related tools and services to exploit them. Ontologies were initially built manually. A few semi-automatic methodologies have been proposed in recent years. Semi-automatic construction/enrichment of ontologies are mostly achieved using natural language processing (NLP) techniques to assess texts. NLP methods have to take the lexical and semantic complexity of biomedical data into account: (1) lexical refers to complex phrases to take into account, (2) semantic refers to sense and context induction of the terminology.

In this thesis, we address the above-mentioned challenges by proposing methodologies for construction/enrichment of biomedical ontologies based on two main contributions. The first contribution concerns the automatic extraction of specialized biomedical terms (lexical complexity) from corpora. New ranking measures for single- and multi-word term extraction methods are proposed and evaluated. In addition, we present BioTex web and desktop application that implements the proposed measures. The second contribution concerns concept extraction and semantic linkage of extracted terminology (semantic complexity). This work seeks to induce semantic concepts of new candidate terms, and to find semantic links, i.e. relevant locations of new candidate terms, in an existing biomedical ontology. We propose a methodology that extracts new terms in MeSH ontology. Quantitative and qualitative assessments conducted by experts and non-experts on real data highlight the relevance of the contributions.

Résumé

En biomedicine, le domaine du “Big Data” (l’infobésité) pose le problème de l’analyse de gros volumes de données hétérogènes (i.e. vidéo, audio, texte, image). Les ontologies biomédicales, modèle conceptuel de la réalité, peuvent jouer un rôle important afin d’automatiser le traitement des données, les requêtes et la mise en correspondance des données hétérogènes. Il existe plusieurs ressources en anglais mais elles sont moins riches pour le français. Le manque d’outils et de services connexes pour les exploiter accentue ces lacunes. Dans un premier temps, les ontologies ont été construites manuellement. Au cours de ces dernières années, quelques méthodes semi-automatiques ont été proposées. Ces techniques semi-automatiques de construction/enrichissement d’ontologies sont principalement induites à partir de textes en utilisant des techniques du traitement automatique du langage naturel (TALN). Les méthodes de TALN permettent de prendre en compte la complexité lexicale et sémantique des données biomédicales : (1) lexicale pour faire référence aux syntagmes biomédicaux complexes à considérer et (2) sémantique pour traiter l’induction du concept et du contexte de la terminologie.

Dans cette thèse, afin de relever les défis mentionnés précédemment, nous proposons des méthodologies pour l’enrichissement/la construction d’ontologies biomédicales fondées sur deux principales contributions. La première contribution est liée à l’extraction automatique de termes biomédicaux spécialisés (complexité lexicale) à partir de corpus. De nouvelles mesures d’extraction et de classement de termes composés d’un ou plusieurs mots ont été proposées et évaluées. L’application BioTex implémente les mesures définies. La seconde contribution concerne l’extraction de concepts et le lien sémantique de la terminologie extraite (complexité sémantique). Ce travail vise à induire des concepts pour les nouveaux termes candidats et de déterminer leurs liens sémantiques, c’est-à-dire les positions les plus pertinentes au sein d’une ontologie biomédicale existante. Nous avons ainsi proposé une approche d’extraction de concepts qui intègre de nouveaux termes dans l’ontologie MeSH. Les évaluations, quantitatives et qualitatives, menées par des experts et non experts sur des données réelles, soulignent l’intérêt de ces contributions.

Research Publications

Edition of International Conference Proceedings

- **Lossio-Ventura, J. A.**, and Alatrista-Salas, H., Editors. *Proceedings of the 2nd Annual International Symposium on Information Management and Big Data - (SIMBig 2015)*, Cusco, Peru, September 2-4, 2015, volume 1478 of CEUR Workshop Proceedings. CEUR-WS.org, 2015.
- **Lossio-Ventura, J. A.**, and Alatrista-Salas, H., Editors. *Proceedings of the 1st Annual International Symposium on Information Management and Big Data - (SIMBig 2014)*, Cusco, Peru, October 8-10, 2014, volume 1318 of CEUR Workshop Proceedings. CEUR-WS.org, 2014.

International Journals

- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, Springer Netherlands, vol. 19, 1, August 2015, pp 55-99.
- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. Towards a mixed approach to extract biomedical terms from documents. *International Journal of Knowledge Discovery in Bioinformatics - (IJKDB)*, vol. 4, 1, January-March 2014, 1-15.

French Journal

- Roche, M., Fortuno, S., **Lossio-Ventura, J. A.**, Akli, A., Belkebir, S., Lounis, T., and Toure, S. Extraction automatique des mots-clés à partir de publications scientifiques pour l'indexation et l'ouverture des données en agronomie. *Cahiers Agricultures*, vol. 24, 5, Septembre-Octobre 2015. pp. 313-320.

International Conferences

- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. A Way to Automatically Enrich Biomedical Ontologies. *In Proceedings of the 19th International Conference on Extending Database Technology - Posters - (EDBT'2016).* Bordeaux, France, 2016 *(to appear).*

- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. Communication Overload Management Through Social Interactions Clustering. *In Proceedings of the 31st ACM/SIGAPP Symposium on Applied Computing - (SAC'2016).* Pisa, Italy, 2016 *(to appear).*

- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. Automatic Biomedical Term Polysemy Detection. *In Proceedings of the 10th International Language Resources and Evaluation Conference - (LREC'2016).* Portorož, Slovenia, 2016 *(to appear).*

- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. BioTex: A system for biomedical terminology extraction, ranking, and validation. *In Proceedings of the 13th International Semantic Web Conference - Posters & Demos - (ISWC'2014).* Riva del Garda, Italy, 2014, pp 157-160.

- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. Yet another ranking function for automatic multiword term extraction. *In Proceedings of the 9th International Conference on Natural Language Processing - (PolTAL'2014).* Warsaw, Poland, 2014, no. 8686 in LNAI, Springer, pp. 52-64.

- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. SIFR Project: The Semantic Indexing of French Biomedical Data Resources. *In Proceedings of the 1st International Symposium on Information Management and Big Data - (SIMBig'2014).* Cusco, Peru, September 2014, vol. 1318, CEUR Workshop, p. 58-61.

- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. Integration of linguistic and web information to improve biomedical terminology extraction. *In Proceedings of the 18th International Database Engineering & Applications Symposium - (IDEAS'14).* Porto, Portugal, July 2014, ACM, Ed., ACM, pp. 265-269.

- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. Biomedical terminology extraction: A new combination of statistical and web mining approaches. *In Proceedings of Journées internationales d'Analyse statistique des Données Textuelles (JADT'2014).* Paris, France, June 2014, pp 421-432.

- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. Combining c-value and keyword extraction methods for biomedical terms extraction. *In Proceedings of the Fifth International Symposium on Languages in Biology and Medicine (LBM'13).* Tokyo, Japan, December 2013, pp. 45-49.

- **Lossio-Ventura, J. A.**, Hacid, H., Ansiaux, A., and Maag, M. L. Conversations reconstruction in the social web. *In Proceedings of the 21st international conference companion on World Wide Web (WWW'12).* Lyon, France, April 2012, ACM, pp. 573-574.

French Conferences

- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. Prédiction de la polysémie pour un terme biomédical. *In Proceedings of the 12th COnférence en Recherche d'Information et Applications - CORIA'2015).* Paris, France, March 2015, pp 437-452.
- **Lossio-Ventura, J. A.**, Jonquet, C., Roche, M., and Teisseire, M. Extraction automatique de termes combinant différentes informations. *In Proceedings of the 21ème Traitement Automatique des Langues Naturelles - (TALN'2014).* Marseille, France, July 2014, Association pour le Traitement Automatique des Langues, pp. 407-412.

Contents

1 INTRODUCTION 1

1.1 Motivation . . . 3

1.2 Context . . . 4

1.3 Objective . . . 6

1.4 Organization . . . 6

2 STATE-OF-THE-ART 9

I Automatic Biomedical Term Extraction 13

3 Introduction 15

4 State-of-the-Art 19

4.1 Text Mining approaches . . . 19

4.1.1 Linguistic approaches . . . 19

4.1.2 Statistical methods . . . 20

4.1.3 Machine Learning . . . 22

4.1.4 Hybrid methods . . . 22

4.1.5 Terminology Extraction from Parallel and Comparable Corpora 24

4.1.6 Tools and applications for biomedical term extraction . . . 24

4.2 Graph-based approaches . . . 26

4.3 Web Mining approaches . . . 28

4.4 Discussion . . . 29

5 Methodology 33

5.1 Candidate Term Extraction (step 1) . . . 38

5.2 Ranking Measures (step 2) . . . 39

5.2.1 C-value . . . 40

5.2.2 TF-IDF and Okapi . . . 42

5.2.3 F-OCapi and F-TFIDF-C . . . 43

5.2.4 C-Okapi and C-TFIDF . . . 43

5.2.5 LIDF-value and L-value . . . 44

5.3 Re-ranking (step 3) . . . 46

5.3.1 A new graph-based ranking measure: “TeRGraph” (Terminology Ranking based on Graph information) 47
5.3.2 WebR 50
5.3.3 A new web ranking measure: WAHI (**W**eb **A**ssociation based on **H**its **I**nformation) 51

6 Data and Results 55
6.1 Data, Protocol, and Validation 55
6.1.1 Data 55
6.1.2 Protocol 56
6.1.3 Validation 56
6.2 Multilingual Comparison (LabTestsOnline) 57
6.3 Evaluation of the global process (GENIA) 62
6.3.1 Ranking Results (step 2 in Figure 5.1) 62
6.3.2 Re-ranking Results (step 3 in Figure 5.1) 67
6.3.3 Summary 70

7 Discussion and Conclusions 73
7.1 Discussion 73
7.1.1 Impact of Pattern List 73
7.1.2 Effect of Dictionary Size 74
7.1.3 Term Extraction Errors 74
7.2 Conclusions and Future Work 75

8 BioTex 77
8.1 Introduction 77
8.2 Implementation 78
8.2.1 Term extraction and ranking measures 78
8.2.2 Validation of candidate terms 79
8.3 Use-Cases 80
8.3.1 From the biomedical domain 81
8.3.2 To more general domain 82
8.4 Summary 84
8.5 Conclusions 85

II Concept Extraction and Semantic Linkage 87

9 Introduction 89

10 State-of-the-art 93
10.1 Polysemy Detection 93
10.2 Term Sense Induction 95
10.2.1 Word Sense Induction 96

10.2.2 Word Sense Disambiguation . . . 105
10.2.3 Sense Number Prediction . . . 107
10.3 Meta-learning . . . 110
10.4 Semantic Linkage . . . 113
10.5 Discussion . . . 118

11 Methodology 121
11.1 Polysemy Detection (step 1) . . . 122
11.1.1 Meta-features . . . 123
11.1.2 Machine Learning Algorithms . . . 131
11.2 Term Sense Induction (step 2) . . . 131
11.2.1 Sense Number Prediction . . . 133
11.2.2 Clustering for Concept Induction . . . 139
11.3 Semantic Linkage (step 3) . . . 139

12 Data and Results 143
12.1 Polysemy Detection . . . 144
12.1.1 Data set . . . 144
12.1.2 Polysemy Detection Results . . . 146
12.2 Term Sense Induction . . . 150
12.2.1 Data set . . . 150
12.2.2 Results of Sense Number Prediction . . . 151
12.3 Semantic Linkage . . . 159
12.3.1 Semantic LinkageData set . . . 159
12.3.2 Semantic Linkage Results . . . 162

13 Discussion and Conclusions 169
13.1 Polysemy Detection . . . 170
13.2 Term Sense Induction . . . 170
13.3 Semantic Linkage . . . 171

14 GENERAL CONCLUSIONS 173
14.1 Conclusions . . . 173
14.2 Perspectives . . . 177

References 186

Chapter

1 INTRODUCTION

Big Data is a popular term used to describe the exponential growth and availability of both structured and unstructured data. This has taken place over the last 20 years. For instance, social networks such as Facebook, Twitter and Linkedin generate masses of data, which is available to be accessed by other applications. Several domains, including biomedicine, life sciences and scientific research, have been affected by Big Data[1]. Therefore there is a need to understand and exploit this data. This process is called “Big Data Analytics”, which allows us to gain new insight through data-driven research [Madden, 2012, Embley and Liddle, 2013]. For instance, the combination of genomics and clinical health data together with Big Data Analytics will leverage personalized medicine, which will significantly improve patient care. A major problem hampering Big Data Analytics development is the need to process several types of data, such as structured, numeric and unstructured data (e.g. video, audio, text, image, etc)[2].

Semantic knowledge could be integrated to solve this problem. Semantic knowledge is generally represented by ontologies. In the context of computer and information sciences, an ontology is a description of the concepts and relationships that really or fundamentally exist for a particular domain [Gruber, 1993]. The representations are typically classes (or concepts), attributes (or properties), and relationships (or relations among concept members).

[1] By 2015 the average of data annually generated in hospitals is 665TB : http://ihealthtran.com/wordpress/2013/03/infographic-friday-the-body-as-a-source-of-big-data/.

[2] Today, 80% of data is unstructured such as images, video, and notes

Therefore, ontologies, as conceptual models of reality, can play a role in Big Data applications for data processing automation, querying, and reconciling data heterogeneity. In this context, ontologies can: (i) provide semantics to add raw data, (ii) build bridges across domains, (iii) connect data at various concept levels across domains, via their generalized concepts, and (iv) be used as authorized knowledge to analyze Big Data. The ontologies are used in information retrieval systems (e.g. in Google with the knowledge graph introduced in 2012) or decision support in several domains, for instance, life science [Sy et al., 2012], medicine [Aimé et al., 2012], low [Casellas, 2011], learning [Lundqvist et al., 2011].

In the biomedical domain, as we have already mentioned, the increasing capability and sophistication of biomedical instruments has led to the build-up of large volumes of heterogeneous data (e.g. the use of electronic health records "EHR" for storing patient information). Analysis of this biomedical data ("biomedical Big Data") is opening new avenues for developing biomedical research, but it could be hard to efficiently manage data without ontologies, support storage, query, and to perform analytic functions.

Biomedical ontologies can be integrated to manage heterogeneous data in emerging bioinformatics projects. Two ways for integrating ontology in biomedical Big Data are: (i) ontology to database mapping: mapping ontology to database classes (concepts)/instances; and (ii) adding metadata on data using the terms of an ontology. For instance, we may find a part of text like "blood pressure 200 mmHg" in a document, which could be mapped to "hyperglycemia". Some related works integrate biomedical ontologies in both ways, for instance, in the neurology disease domain [Jayapandian et al., 2014], BRAIN Project[3],etc.

In addition to the importance of ontology as previously described, biomedical ontologies are useful, for instance in: (i) medicine, by facilitating the continuity of health care; and (ii) biological research, by facilitating the sharing of experimental data among researchers [Bodenreider, 2008]. Biomedical ontologies serve to standardize the terminology, enable access to domain knowledge, verify data and facilitate integrative analysis of heterogeneous data.

Initially, ontologies were built manually. That has changed in recent years, and a few semi-automatic methodologies have been proposed. The semi-automatic construction/enrichment of ontologies are mostly induced through textual data. For instance, *Text2Onto* [Cimiano and Völker, 2005], *OntoLearn* [Velardi et al., 2007], *Sprat* [Maynard et al., 2009] were developed to construct/enrich ontologies from textual corpus, among other studies [Blomqvist, 2009, Deborah et al., 2011, Dixit et al., 2012, Mondary, 2011, Sánchez and Moreno, 2008]. These related works have usu-

[3] http://braininitiative.nih.gov/

ally been proposed in general domains, with different techniques to automate tasks. These studies have generated interesting results, but experts are always needed for the construction/enrichment methodology, as recently described in [Gherasim, 2013]. The main reason for the expert intervention is because the ontology construction/enrichment task requires a considerable amount of domain knowledge. We briefly describe studies related to ontology construction/enrichment in next chapter. For instance, in the biomedical domain, ontologies can have millions of concepts and they are always based on considerable expert help. However, because of current large-scale changes, ontology enrichment/construction requires an automatic process to reduce the participation time of experts.

Biomedical ontologies have recently started to emerge that are built on the basis of theories and methods from diverse disciplines such as information management, knowledge representation, and natural language processing (NLP) to perform or improve biomedical applications. Automatic biomedical ontology construction/enrichment from textual data requires strict in-depth studies of the linguistic structures, concepts, and relations present in the text. This study is part of Natural Language Processing (NLP), which involves methods to communicate to people and machines through natural language.

NLP for biomedical ontology enrichment/construction is faced with several challenges. First, the complexity and specialization of texts to be evaluated, i.e. biological data differs from EHR (electronic health record) data, which in turn differs from data of other domains. Second, the complexity of extracting new terminology according the specific already existing resources, for instance biomedical terms often contain numbers, e.g. *"epididymal protein 9"*, *"pargyline 10 mg"*. Third, the concept induction of newly found terms, by using associated biomedical complex texts. Fourth, unifying several disciplines to set up a general workflow. Fifth, to make this multidisciplinary aspect"user friendly" and "multilingual".

To the best of our knowledge, no studies have been carried out to address all of these challenges. As mentioned before, the proposed methodologies (see Chapter 2 for a brief description): (i) need expert intervention during their workflow to obtain good results, (ii) are not applied for different languages, (iii) are in most cases applied for general domains.

1.1 Motivation

As described previously, the volume of biomedical data is constantly increasing. Despite the widespread adoption of English in Science, a significant quantity of these data are in French. Usually, the content of resources is indexed to enable querying with keywords. However, there are obvious limits to keyword-based indexing, e.g.

use of synonyms, polysemy, lack of domain knowledge.

Biomedical data integration and semantic interoperability is necessary to enable new scientific discoveries that could be achieved by merging different available data (i.e. translational research). A key aspect in addressing semantic interoperability for life sciences is the use of terminologies and ontologies as a common denominator to structure biomedical data and make them interoperable. The community has turned especially toward ontologies to design semantic indexes of data that leverage medical knowledge for better information mining and retrieval.

However, besides the existence of various English tools, there are considerably fewer ontologies available in French [Névéol et al., 2014] and there is a marked lack of related tools and services to exploit them. This shortcoming is out of line with the huge amount of biomedical data produced in French, especially in the clinical world (e.g. electronic health records).

The SIFR project was proposed to overcome this problem. It seeks to annotate biomedical resources and enrich/build new biomedical ontologies. This project is described in next section.

1.2 Context

The Semantic Indexing of French Biomedical Data Resources (SIFR)[4] project proposes to investigate scientific and technical challenges in building ontology-based services to leverage biomedical ontologies and terminologies in indexing, mining and retrieval of French biomedical data. The main goal of SIFR is to semantically index all possible French biomedical resources in order to enable straightforward use of ontologies, thus enabling health researchers to deal with knowledge engineering issues and to concentrate on the biological and medical challenges.

The SIFR project involves a cyclic process to address this situation. The process consists of indexing the resources and automatically enriching the terminologies/ontologies used for the indexation. We show a simple lifecycle of an SIFR project, where we included only two major processes: A) Annotation, and B) Enrichment. The Annotation process involves semantically annotating all possible French biomedical resources, such as scientific articles, electronic health records, doctor's notes, etc., using existing biomedical ontologies/terminologies. SIFR has thus created the French BioPortal Annotator[5], which processes text submitted by users, recognizes relevant ontology terms in the text and returns the annotations to the user. The enrichment process involves new data reported by researchers, for instance new scientific

[4] http://www.lirmm.fr/sifr/

[5] http://bioportal.lirmm.fr/annotator

biomedical articles. So the objective is to enrich existing biomedical ontologies with new terms that appear in the biomedical literature. Hence, these new terms are added to the ontologies and are reused in the future annotation process. Figure 1.1 illustrates the processes of the SIFR Lyfecicle project.

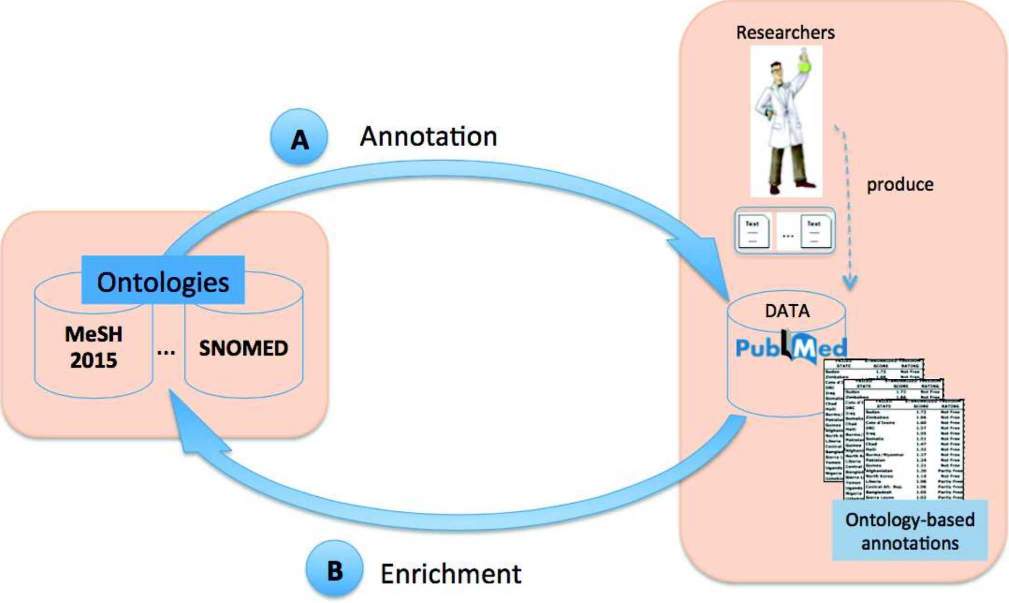


Figure 1.1: Processes of the SIFR Lyfecicle.

This thesis tackles process B in the SIFR Lyfecicle project, which is named Enrichment in the Figure 1.1. We extended and called this process "Towards the French Biomedical Ontology Enrichment", which seeks to enrich biomedical ontologies from textual data. In the next chapters, we explain the proposed workflow to tackle this process.

There are approaches to create ontologies and approaches to enrich already existing ontologies, as explained in [Gherasim, 2013]. Several approaches take different resources as input, such as textual documents, databases, taxonomies, thesauri, internet, etc. In this thesis, we focus on ontology enrichment by using textual documents as resources.

The design and creation of ontologies is initially performed by experts. In the last 10 years, because of the corpora size and the dissemination of ontologies, automatic approaches have been developed to create ontologies. These automatic approaches are also called as automatic construction, semi-automatic construction and/or learning-based construction. To the best of our knowledge, there are no existing studies that propose a complete automatic ontology enrichment methodology, and expert intervention is generally needed in the workflow.

1.3 Objective

As we have already discussed, this thesis seeks to enrich biomedical ontologies from textual data, especially French biomedical ontologies, to address the abovementioned challenges related to ontology enrichment. These challenges are pooled in two main groups: (i) lexical complexity, and (ii) semantic complexity. The lexical complexity involves the first and the second challenge. This means extraction of new complex biomedical terminology from a specialized text corpus. The semantic complexity is related to the third challenge. This means concept induction and semantic linkage of new terminology. Both implicitly use several disciplines to get to the final solution, so it is "user friendly" and "multilingual".

Thesis Contribution

To fulfill our biomedical ontology enrichment objective, we address (i) the lexical complexity of the terms, and (ii) their semantic complexity. More precisely, these two aspects will be detailed into two main parts:

1. **Automatic Biomedical Term Extraction:** (lexical complexity) which aims at automatically extracting technical biomedical terminology from a specialized text corpus. In this case, we focus on terms that do not exist in an ontology/terminology. These are called new biomedical candidate terms. The approach proposed here is based on linguistic, statistic, graph, and web features to improve the ranking of new biomedical candidate terms.

2. **Concept Extraction and Semantic Linkage:** (semantic complexity) which seeks to extract the concepts and to find the semantic links, i.e. a position semantically close to a biomedical ontology, of these new candidate terms before extraction. We propose to perform this in three steps. First, we believe that it is important to detect if a new candidate term could be polysemic. Second, we propose to identify the possible senses or concepts of terms. Third, we would like to find the semantic links that can have new candidate terms in an already defined biomedical ontology, i.e. to find a position in a biomedical ontology to add new candidate terms.

1.4 Organization

The rest of this thesis is organized as follows.

The first part of this thesis, "Automatic Biomedical Term Extraction", is composed of the Chapters 3, 4, 5, 6, 7, 8, in which, first, we study how to extract relevant new candidate terms from textual data. Second, we present the different term extraction methods and several related existing works. Then, we detail our proposal to extract biomedical candidate terms based on linguistic, statistic, graph, and web features. We present the results of experiments. Then, we conclude and discuss the results. We finalize by presenting the application called BioTex, in Chapter 8, which implements the proposed methodology, and the different industrial/research applications of our approach.

The second part of this thesis, called “Concept Extraction and Semantic Linkage”, proposes three approaches to fulfil the final biomedical ontology enrichment objective: (i) Polysemy Detection, (ii) Term Sense Induction, and (iii) Semantic Linkage. The Chapters 9, 10, 11, 12, 13 are devoted to presenting the acquisition of term concepts and to adding them to an existing ontology. We first present the motivations and introduce the problems. We study existing work regarding these three issues. We subsequently present the three approaches. Then we describe the several experiments we performed, and finally, we present the results of our experiments.

Finally, in Chapter 14, we conclude and summarize the contribution of this thesis, and we discuss the prospects and propose some future research directions.

Chapter

2

STATE-OF-THE-ART

This chapter describes briefly the related methodologies to build and/or enrich ontologies, this state-of-the-art is a short resume described in [Gherasim, 2013].

The construction of ontologies manually is a long and complex process. Several methodologies haven been defined [Fernández-López et al., 1997, Pinto et al., 2004, Pérez et al., 2008]. We describe briefly the methodologies to create manually ontologies. The most well-known methodology is called **Methontology** [Corcho et al., 2005, Fernández-López et al., 1997, Gómez-Pérez et al., 2007], which is considered the most complete methodology [Pérez et al., 2008]. It is based in three process: i) Process of Management, ii) Process of Development, and iii) Process of Maintenance.

On-To-Knowledge [Staab et al., 2001] is another methodology, built to be applied to a specific domain. This means that the ontologies created following this methodology are specialized and dependent of a domain. This methodology has five process: i) Feasibility study, ii) Ontology kickoff, iii) Refinement, iv) Evaluation, and v) Maintenance.

A methodology proposing a novel principle is **DILIGENT** [Pinto et al., 2004], which introduces the term of collaborative construction of ontologies. This means the ontology is built by the collaboration of several communities related to the ontology, for instance user, engineers, editors, etc. This methodology contains five process: i) Construction, ii) Local adaptation, iii) Analyze, iv) Revision, and iv) Local update.

The most recent methodology, to our knowledge is **NeOn** [Pérez et al., 2008], this has a novel paradigm, which is the development of ontologies based on the reutilization, modification, restructure, and the adaptation of already existent ontologies.

As we could in the previous chapter, the conceptualization of a ontology is the central activity for the ontology construction. In fact, the ontology construction is a recent domain, and to our knowledge there no exist a study proposing a complete automatic methodology to built an ontology.

For instance, there exists **Terminae** [Aussenac-Gilles et al., 2000, Aussenac-Gilles et al., 2008], an approach which proposes a tool to assist automatically users to build ontologies, using functions based on natural language processing to extract relevant terms for the ontology, then it builds the concept structure. Indeed, the user identifies and structures the ontology, choosing the concepts, relations and instances. For instance, the user associates the term and the concept.

Text-To-Onto [Maedche and Staab, 2000, Maedche and Volz, 2001] and **Text2Onto** [Cimiano and Völker, 2005], *Text2Onto* is the improved version of *Text-To-Onto*. *Text-To-Onto* proposes algorithms to extract concepts, instances, and taxonomic relations. The relations are extracted according to patterns created by the user using WordNet. The user also must validate the concepts and relations. *Text2Onto* proposed new algorithms and the option of combining the results of using several algorithms. The last methodology uses the GATE architecture [Cunningham et al., 2002].

Another approach proposes to add concepts and instances according to a set of predefined patterns by the user is **Sprat** [Maynard et al., 2009]. A disadvantage of this approach is that user need a corpus containing the most patterns previously defined. A close approach called **Asium** [Faure and Nédellec, 1998, Faure et al., 1998, Faure and Nedellec, 1999] also exploits patterns to extract terms and relations. This approach has as output a hierarchy of concepts.

OntoLearn [Navigli et al., 2003, Velardi et al., 2007] is an approach less structured than the other three before mentioned. It does not offer a tool integrating all the possible functionalities. In contrast, *OntoLearn* offers a set of independent tools for each step, such as term extraction, concept extraction and relation extraction. For instance, the authors use *TermExtractor* for extracting terms; *GlossExtractor* to associate a definition for the extracted term; and *WordNet* to validate the relations.

There also exist an approach called **OntoGen** [Fortuna et al., 2006, Fortuna et al., 2007], which allows the construction semi-automatic of small ontologies from collections of documents. So, the created ontologies represent the “topic ontologies”, where each concept of the ontology is related to one topic of the collectionof documents.

We have to mention as well **OntoLT** [Buitelaar et al., 2004], which precisely is not an approach to build ontologies. It allows to manually define patterns to rely elements from the text to elements from the ontology. We mentioned **OntoLT** because that is usually cited in the literature as an ontology learning tool.

Part I

Automatic Biomedical Term Extraction

Chapter 3

Introduction

The huge amount of biomedical data available today often consists of plain text fields, e.g. clinical trial descriptions, adverse event reports, electronic health records, emails or notes expressed by patients within forums [Murdoch and Detsky, 2013]. These texts are often written using a specific language (expressions and terms) of the associated community. Therefore, there is a need for formalization and cataloging of these technical terms or concepts via the construction of terminologies and ontologies [Rubin et al., 2008]. These technical terms are also important for Information Retrieval (IR), for instance when indexing documents or formulating queries. However, as the task of manually extracting terms of a domain is very long and cumbersome, researchers have striving to design automatic methods to assist knowledge experts in the process of cataloging the terms and concepts of a domain under the form of vocabularies, thesauri, terminologies or ontologies.

Automatic Term Extraction (ATE), or Automatic Term Recognition (ATR), is a domain which aims to automatically extract technical terminology from a given text corpus. We define technical terminology as the set of terms used in a domain. Term extraction is an essential task in domain knowledge acquisition because the technical terminology can be used for lexicon updating, domain ontology construction, summarization, named entity recognition or, as previously mentioned, IR.

In the biomedical domain, there is a substantial difference between existing resources (hereafter called *terminologies* or *ontologies*) in English, French, and Spanish. In English, there are about 9 919 000 terms associated with about 8 864 000 concepts such as those in UMLS[1] or BioPortal [Noy et al., 2009]. Whereas in French there are only about 330 000 terms associated with about 160 000 concepts [Névéol et al., 2014], and in Spanish 1 172 000 terms associated with about 1 140 000 concepts. Note the strong difference in the number of ontologies and terminologies available in French or Spanish. This makes ATE even more important for these languages.

In biomedical ontologies, different terms may be linked to the same concept and are semantically similar with different writing, for instance *"neoplasm"* and *"cancer"* in MeSH or SNOMED-CT. Ontologies also contain terms with morphosyntaxic variants, for instance plurals like '*'external fistula"* and *"external fistulas"*, and this group of variants is linked to a preferred term. As one of our goals is to extract new terms to enrich ontologies, our approach does not normalize variant terms, mainly because normalization would lead to penalization in extracting new variant terms. Technical terms are useful to gain further insight into the conceptual structure of a domain. These may be: (i) single-word terms (simple), or (ii) multi-word terms (complex). The proposed study focuses on both cases.
Term extraction methods usually involve two main steps. The first step extracts candidate terms by unithood calculation to qualify a string as a valid term, while the second step verifies them through termhood measures to validate their domain specificity. Formally, unithood refers to the degree of strength or stability of syntagmatic combinations and collocations, and termhood is defined as the degree to which a linguistic unit is related to domain-specific concepts [Kageura and Umino, 1996]. ATE has been applied to several domains, e.g. biomedical [Lossio-Ventura et al., 2014d, Frantzi et al., 2000, Zhang et al., 2008, Newman et al., 2012], ecological [Conrado et al., 2013], mathematical, [Stoykova and Petkova, 2012], social networks [Lossio-Ventura et al., 2012], banking [Dobrov and Loukachevitch, 2011], natural sciences [Dobrov and Loukachevitch, 2011], information technology [Newman et al., 2012, Yang et al., 2009], legal [Yang et al., 2009], as well as post-graduate school websites [Qureshi et al., 2012].

The main issues in ATE are: (i) extraction of non-valid terms (noise) or omission of terms with low frequency (silence), (ii) extraction of multi-word terms having various complex various structures, (iii) manual validation efforts of the candidate terms [Conrado et al., 2013], and (iv) management of large-scale corpora. Inspired by our previously published results and in response to the above issues, we propose a cutting edge methodology to extract biomedical terms. We propose new measures and some modifications of existing baseline measures. Those measures are divided into: 1) ranking measures, and 2) re-ranking measures. Our ranking measures are statistical- and linguistic-based and address issues i), ii) and iv). Our two re-ranking measures – the first one called *TeRGraph* – is a graph-based measure which deals with issues i), ii) and iii). The second one, called *WAHI*, is a web-based measure

[1] `http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html`

which also deals with issues i), ii) and iii). The novelty of the *WAHI* measure is that it is web-based which has, to the best of our knowledge, never been applied within ATE approaches.

The main contributions of the methodology of this first part are: (1) enhanced consideration of the term unithood, by computing a degree of quality for the term unithood, and, (2) consideration of the term dependence in the ATE process. The quality of the proposed methodology is highlighted by comparing the results obtained with the most commonly used baseline measures. Our evaluation experiments were conducted despite difficulties in comparing ATE measures, mainly because of the size of the corpora used and the lack of available libraries associated with previous studies. Our three measures improve the process of automatic extraction of domain-specific terms from text collections that do not offer reliable statistical evidence (i.e. low frequency).

This first part is organized as follows. We first discuss related work in Chapter 4. Then the methodology to extract biomedical terms is detailed in Chapter 5. The results are presented in Chapter 6, followed by discussions in Chapter 7.1. Finally, the conclusions in Chapter 7.2, and the application created with this methodology in Chapter 8.

Chapter

4

State-of-the-Art

Recent studies have focused on multi-word (n-grams) and single-word (unigrams) term extraction. Term extraction techniques can be divided into four broad categories: (i) *Linguistic*, (ii) *Statistical*, (iii) *Machine Learning*, and (iv) *Hybrid*. All of these techniques are encompassed in Text Mining approaches. To our knowledge, graph-based approaches have not yet been applied to ATE, although they have been successively adopted in other Information Retrieval fields and could be suitable for our applications. Existing web techniques have not been applied to ATE but, as we will see, these techniques can be adapted for such purposes.

In this chapter, we describe the state-of-the-art divided into three types of approaches: (i) Text Mining approaches, (ii) Graph-based approaches, and (iii) Web Mining approaches. These techniques are defined and associated studies are reported in each respective section.

4.1 Text Mining approaches

4.1.1 Linguistic approaches

These techniques attempt to recover terms via linguistic pattern formation. This involves building rules to describe naming structures for different classes based on orthographic, lexical, or morphosyntactic characteristics, e.g. [Gaizauskas et al., 2000]. The main approach is to develop rules (typically manually) describing common naming structures for certain term classes using orthographic or lexical clues, or more complex morpho-syntactic features. Moreover, in many cases, dictionaries of typical term constituents (e.g. terminological heads, affixes, and specific acronyms) are used to facilitate term recognition [Krauthammer and Nenadic, 2004].

In the general domain, several systems only use linguistic patterns, such us Termino [David and Plante, 1990, Faraj et al., 1996], which uses syntactic analysis to extract nominal terms. It allows extraction of groups of nouns, adjectives, verbs, and prepositions. These systems are devoted to French applications.

Another well-known application is Lexter [Bourigault, 1993, Bourigault and Jacquemin, 1999, Aussenac-Gilles and Bourigault, 2000], which aims at extracting maximal nominal phrases. Locating noun phrase boundaries is the idea underlying the Lexter design. Rather than exploiting knowledge on possible grammatical structures of complex terms, the authors use grammatical configurations that are known not to be parts of terms. First, the authors split the text by locating these potential boundaries (i.e. verbs, pronouns, conjunctions), between which noun phrases likely to be occurrences of terms are isolated. This process is also used in other related works, such as [Drouin, 2003, Turenne and Barbier, 2004]. The second process involves decomposition of maximal nominal groups according heads and arguments. And the third one is the presentation of terms as a network, taking the *head* of terms into account as a criterion to group terms sharing the same *head*.

There are applications to update terminology, such as Fastr (FAst Syntactic Term Recognizer) [Jacquemin, 1996, Jacquemin, 1999], which is a rule-based system geared mainly towards extracting variants of target terms. It takes a list of reference terms and a corpus as input and then it enriches the reference terms with their variants.

A recent study on biomedical term extraction [Golik et al., 2013] (BioYaTeA system) is based on linguistic patterns plus additional context-based rules to extract candidate terms, which are not scored and the authors leave the term relevance decision to experts.

In this section, we describe the state-of-the-art of linguistic approaches to extract terminology, and in the following section we describe the statistical approaches.

4.1.2 Statistical methods

Statistical techniques chiefly rely on external evidence presented through surrounding (contextual) information. Such approaches are mainly focused on the recognition of general terms [Van Eck et al., 2010]. The most basic measures are based on frequency. For instance, *term frequency (tf)* counts the frequency of a term in the corpus, *document frequency (df)* counts the number of documents where a term occurs, and *average term frequency (atf)*, which is $\frac{tf}{df}$.

is Mantex [Sebag, 2002] is a methodology based on the frequency criterion according to the number of segment occurrences in the corpus. This system does not extract single-word terms. There are also studies that use mutual information to extract terminology, as in [Church and Hanks, 1990], where the authors compute the dependence between two words composing a term. Another study based on mutual information measures is focused on cubic mutual information [Daille, 1994], which gives more importance to the most frequent phrases (i.e. words composing a term). These methodologies are applied to binary terms.
Mutual information and co-occurrence measures are associative measures. They can be used for terminology ranking but also for other tasks like the identification of semantic proximity (see Section 4.3).

Another system that only focuses on the statistical aspect is Ana (Automatic Natural Acquisition) [Enguehard et al., 1993, Enguehard and Pantera, 1995], which is applied for English and French. This system is inspired by human learning of the mother tongue. It has two main steps: (i) familiarization, which builds the first list of words, and (ii) discovery, which incrementally builds the list of domain terms taking previously provided words into account.

A similar research topic, called Automatic Keyword Extraction (AKE), proposes to extract the most relevant words or phrases in a document using automatic indexation. Keywords, which we define as a sequence of one or more words, provide a compact representation of a document's content. Such measures can be adapted to extract terms from a corpus as well as ATE measures.

We take two popular AKE measures as baselines measures, i.e. *Term Frequency Inverse Document Frequency (TF-IDF)* [Salton and Buckley, 1988], and *Okapi BM25* (hereafter *Okapi*) [Robertson et al., 1999], these weight the word frequency according to their distribution along the corpus. *Residual inverse document frequency (RIDF)* compares the document frequency to another chance model where terms with a particular term frequency are distributed randomly throughout the collection, while *Chi-square* [Matsuo and Ishizuka, 2004] assesses how selectively words and phrases co-occur within the same sentences as a particular subset of frequent terms in the document text. This is applied to determine the bias of word co-occurrences in the document text, which is then used to rank words and phrases as keywords of the document; *RAKE* [Rose et al., 2010] hypothesises that keywords usually consist of multiple words and do not contain punctuation or stop words. It uses word co-occurrence information to determine the keywords.

In next section, we list the few existing studies based on machine learning methods, which could be considered as part of statistical approaches. A difference is that all the approaches described in this section are unsupervised methods.

4.1.3 Machine Learning

Machine Learning (ML) or supervised techniques. These systems are often designed for specific entity classes and thus integrate term extraction and term classification. Machine Learning systems use training data to learn features useful for term extraction and classification. But the avaibility of reliable training resources is one of the main problems. Some proposed ATE approaches use machine learning [Conrado et al., 2013, Zhang et al., 2010, Newman et al., 2012]. However, ML may also generate noise and silence. The main challenge is how to select a set of discriminating features that can be used for accurate recognition (and classification) of term instances. Another challenge concerns the detection of term boundaries, which are the most difficult to learn.

As we mentioned before, there are applications to extract variants of target terms. For instance, there is a method [Loginova Clouet, 2014, Clouet and Daille, 2014] for recognizing and splitting to align variants with the original terms through multiword term extraction. Combining dependent and independent characteristics of the language, and using a probabilistic method. It is validated in Germanic, Slavic, and Romance languages. These studies are more related to building bilingual terminology lexicons according to comparable corpora.

4.1.4 Hybrid methods

Most approaches combine several methods (typically linguistic and statistically based) for the term extraction task.

GlossEx [Kozakov et al., 2007] considers the probability of a word in the domain corpus divided by the probability of the appearance of the same word in a general corpus. Moreover, the importance of the word is increased according to its frequency in the domain corpus.
Weirdness [Ahmad et al., 1999] considers that the distribution of words in a specific domain corpus differs from that in a general corpus.

Among such approaches, we can mention the studies of [Smadja, 1993], who developed a set of statistical techniques for retrieving and identifying collocations from large textual corpora. These techniques have been implemented in a system called XTRACT, which involves three stages. The first stage is based on a statistical technique for identifying word pairs involved in a syntactic relation. The words can appear in the text in any order and can be separated by an arbitrary number of other words. The second stage is based on a technique to extract n-word collocations (or n-grams). A third stage is then applied to the output of stage one and applies parsing techniques to sentences involving a given word pair in order to identify the proper syntactic relation between the two words.

Several studies combining linguistic and statistic aspects are focused on the TERMIGHT system [Dagan and Church, 1997], with the aim of building bilingual glossaries. TERMIGHT consists of two stages: (a) preparing a monolingual list of all technical terms in a source-language document, and (b) finding translations for these terms in parallel source–target documents. As a first step (in each component), the tool automatically extracts candidate terms and candidate translations based on term-extraction and word-alignment algorithms.

ACABIT is also a system that combines linguistic and statistic information for French and English. This system has two main processes: (i) extraction of candidate terms, which consists of extracting terms according to a syntactic structure [Daille, 1994, Daille, 1996], in most cases a binary structure, and (ii) ranking candidate terms, where terms are ranked according to a statistic measure [Daille, 1998].

In this section, we describe the SYNTEX system [Bourigault and Fabre, 2000, Bourigault et al., 2005], which is an extension of the LEXTER system. The main difference is that SYNTEX takes nominal groups containing verbs, i.e. verb syntagms, into account. It performs a syntactic analysis of sentences in the corpus, and yields a dependency network of word and syntagms. This proximity is based on the identification of shared syntactic contexts.

C/NC-value [Frantzi et al., 2000] combines statistical and linguistic information for the extraction of multi-word and nested terms. This is the most well-known measure in the literature. While most studies address specific types of entities, *C/NC-value* is a domain-independent method. It has also been used for recognizing terms in the biomedical literature [Hliaoutakis et al., 2009, Hamon et al., 2014]. In [Zhang et al., 2008], the authors showed that *C-value* obtains the best results compared to the other measures cited above. *C-value* has been also modified to extract single-word terms [Nakagawa and Mori, 2002], and in this work the authors extract only terms composed of nouns.

Moreover, *C-value* has also been applied to different languages other than English, e.g. Japanese, Serbian, Slovenian, Polish, Chinese [Ji et al., 2007], Spanish [Barrón-Cedeño et al., 2009], Arabic, and French. We have thus chosen *C-value* as one of our baseline measure. Those baseline measures will be modified and evaluated with the new proposed measures.

4.1.5 Terminology Extraction from Parallel and Comparable Corpora

Another kind of approach suggests that terminology may be extracted from parallel and/or comparable corpora. Parallel corpora contain texts and their translation into one or more languages, but such corpora are scarce [Bowker and Pearson, 2002]. Thus parallel corpora are scarce for specialized domains. Comparable corpora are those which select similar texts in more than one language or variety [Déjean and Gaussier, 2002]. Comparable corpora are built more easily than parallel corpora. They are often used for machine translation and their approaches are based on linguistics, statistics, machine learning, and hybrid methods. The main objective of these approaches is to extract translation pairs from parallel/comparable corpora. Different studies propose translation of biomedical terms for English-French by alignment techniques [Deléger et al., 2009]. English-Greek and English-Romanian bilingual medical dictionaries are also constructed with a hybrid approach that combines semantic information and term alignments [Kontonatsios et al., 2014b]. Other approaches are applied for single- and multi-word terms with English-French comparable corpora [Daille and Morin, 2005]. The authors use statistical methods to align elements by exploiting contextual information. Another study proposes to use graph-based label propagation [Tamura et al., 2012]. This approach is based on a graph for each language (English and Japanese) and the application of a similarity calculus between two words in each graph. Moreover, some machine learning algorithms can be used, e.g. the logistic regression classifier [Kontonatsios et al., 2014a]. There are also approaches that combine both corpora [Morin and Prochasson, 2011] (i.e. parallel and comparable) in an approach to reinforce extraction. Note that our corpora are not parallel and are far of being comparable because of the difference in their size. Therefore these approaches are not evaluated in our study.

4.1.6 Tools and applications for biomedical term extraction

There are several applications implementing some measures previously mentioned, especially *C-value*, which is a domain independent measure, but frequently used for biomedical term extraction. The study of related tools revealed that most existing systems that especially implement statistical methods are made to extract keywords and, to a lesser extent, to extract terminology from a text corpus. Indeed, most systems take a single text document as input, not a set of documents (as corpus), for which the *IDF* can be computed. Most systems are available only in English and the most relevant for the biomedical domain are:

- *TerMine*[1], developed by the authors of the *C-value* method, only for English term extraction;

[1] http://www.nactem.ac.uk/software/termine/

- *Java Automatic Term Extraction*[2] [Zhang et al., 2008], a toolkit which implements several extraction methods including *C-value*, GlossEx, TermEx and offer other measures such as frequency, average term frequency, *IDF*, *TF-IDF*, *RIDF*;

- *FlexiTerm*[3] [Spasic et al., 2013], a tool explicitly evaluated on biomedical copora and which offer more flexibility than *C-value* when comparing term candidates (treating them as bag of words and ignoring the word order);

- *BioYaTeA* [4] [Golik et al., 2013], is a version of the YaTeA term extractor [Aubin and Hamon, 2006], both are available as a Perl module. It is a biomedical term extractor. The method used is based only on linguistic aspects.

There also exist applications for Automatic Term Extraction (ATE) and for Automatic Keyword Extraction (AKE) used in general domains, including the biomedical domain. The major applications are based on statistic approaches, as well as some of them integrate the linguistic approaches. For instance, we list some application in the following paragraphs:

- *Maui-indexer*[5] [Medelyan et al., 2009], an open-source software program and a library for identification of main "topics" in text documents. These topics are tags, keywords, keyphrases, vocabulary terms, descriptors, index terms or titles of Wikipedia articles. It uses both, linguistic and statistic features. For statistical features, it uses *frequency, occurrence positions*;

- *KEA*[6] [Medelyan and Witten, 2006], (Keyphrase Extraction Algorithm), is an algorithm for extracting keyphrases from text documents. It can be either used for free indexing or for indexing with a controlled vocabulary. Some statistic measure are used, such as: *TF-IDF, first occurrence of the phrase, length of the phrase, number of phrases related to that phrase*;

- *Exit*[Roche et al., 2004], based on maximum likelihood estimation and mutual information methods, for French and English text;

- *TermExtractor*[Sclano and Velardi, 2007], which implements two entropy-based measures, domain consensus (terms which are consensually referred to throughout the corpus) and domain relevance (terms which are relevant to the domain of interest);

[2] `https://code.google.com/p/jatetoolkit/`

[3] `http://users.cs.cf.ac.uk/I.Spasic/flexiterm/`

[4] `http://search.cpan.org/~bibliome/Lingua-BioYaTeA/`

[5] `https://code.google.com/p/maui-indexer/`

[6] `http://www.nzdl.org/Kea/`

- *Whatizit*[7] [Rebholz-Schuhmann et al., 2008], which offers a set of web services for biomedical text processing, including recognition of specific terms and matching to corresponding entries in bioinformatics databases;

- *Araya*[8], a licensed terminology extraction tool, where terms are extracted only from databases supported by Araya, and also offers bilingual extraction. Based on a TMX file all possible relevant term pairs are computed. This is based on a statistical approach which determines the frequency of terms;

- *FiveFilters*[9], a web application available for English term extraction, it allows to extract just 50 terms. Terms can be returned in a variety of formats as HTML, JSON, XML. The application is intended to be a simple, free alternative to Yahoo's Term Extraction service.

- *Yahoo Content Analysis*[10], a web service API which provides a service that extracts terms from a piece of content using the Yahoo search index. It detects entities/concepts, categories, and relationships within unstructured content. It ranks those detected entities/concepts by their overall relevance, using Wikipedia pages;

- *Topia Termextract*[11], a Python library which uses linguistic and statistical analysis for English term extraction. This application extracts terms by using Part-Of-Speech (POS) tagging algorithm.

A shown, most existing systems implementing statistical methods are made to extract keywords and, to a lesser extent, to extract terminology from a text corpus. Indeed, most systems take a single text document as input, not a set of documents (as corpus), for which the *IDF* can be computed. Finally, most systems are available only in English.

This section explained the measures based especially on linguistic and statistic features. Next section explains the graph-based approaches.

4.2 Graph-based approaches

Graph modeling is an alternative for representing information, which clearly highlights relationships of nodes among vertices. It also groups related information in a specific way, and a centrality algorithm can be applied to enhance their efficiency. Centrality in a graph is the identification of the most important vertices within a

[7] http://www.ebi.ac.uk/webservices/whatizit/info.jsf

[8] http://www.heartsome.de/en/araya.php

[9] http://fivefilters.org/term-extraction/

[10] https://developer.yahoo.com/search/content/V2/contentAnalysis.html

[11] https://pypi.python.org/pypi/topia.termextract/1.1.0

graph. A host of measures have been proposed to analyze complex networks, especially in the social network domain [Borgatti, 2005, Borgatti et al., 2009, Banerjee et al., 2014]. Freeman [Freeman, 1979], formalized three different measures of node centrality: degree, closeness and betweenness. Degree is the number of neighbors that a node is connected to. Closeness is the inverse sum of shortest distances to all other neighbor nodes. Betweenness is the number of shortest paths from all vertices to all others that pass through that node.

One study proposes to take the number of edges and their weights into account [Opsahl et al., 2010], since the three last measures do not do this. Another well known measure is PageRank [Page et al., 1999], which ranks websites. Boldi [Boldi and Vigna, 2014], evaluated the behavior of ten measures, and associated the centrality to the node with largest degree. Our approach proposes the opposite, i.e. we focus on nodes with a lower degree. An increasingly popular recent application of graph approaches to IR concerns social or collaborative networks and recommender systems [Noh et al., 2009, Banerjee et al., 2014].

Graph representations of text and scoring function definition are two widely explored research topics, but few studies have focused on graph-based IR in terms of both document representation and weighting models [Rousseau and Vazirgiannis, 2013]. First, text is modeled as a graph where nodes represent words and edges represent relations between words, defined on the basis of any meaningful statistical or linguistic relation [Blanco and Lioma, 2012]. In [Blanco and Lioma, 2012], the authors developed a graph-based word weighting model that represents each document as a graph. The importance of a word within a document is estimated by the number of related words and their importance, in the same way that PageRank [Page et al., 1999] estimates the importance of a page via the pages that are linked to it. Another study [Rousseau and Vazirgiannis, 2013] introduces a different representation of document that captures relationships between words by using an unweighted directed graph-of-words with a novel scoring function called *tw-idf*. Another recent study [Rousseau and Vazirgiannis, 2015] proposes to apply the *k-degenerate* graph on the graph-of-words to extract keywords from a single document.

In the above approaches, graphs are used to measure the influence of words in documents like automatic keyword extraction methods (AKE), while ranking documents against queries. These approaches differ from ours as they use graphs focused on the extraction of relevant words in a document and computing relations between words. In our proposal, a graph is built such that the vertices are multi-word terms and the edges are relations between multi-word terms. Moreover, we focus especially on a scoring function of relevant multi-word terms in a domain rather than in a document.

4.3 Web Mining approaches

Different web mining studies focus on semantic similarity, semantic relatedness. This means quantifying the degree to which some words are related, considering not only similarity but also any possible semantic relationship among them. The word association measures can be divided into three categories [Chaudhari et al., 2011]: (i) *Co-occurrence measures* that rely on co-occurrence frequencies of both words in a corpus, (ii) *Distributional similarity-based measures* that characterize a word by the distribution of other words around it, and (iii) *Knowledge-based measures* that use knowledge-sources like thesauri, semantic networks, or taxonomies [Harispe et al., 2014].

In this section, we focus on co-occurrence measures because our goal is to extract multi-word terms and we suggest computing a degree of association between words composing a term. Word association measures are used in several domains like ecology, psychology, medicine, and language processing, and were recently studied in [Pantel et al., 2009, Zadeh and Goel, 2013], such as *Dice*, *Jaccard*, *Overlap*, *Cosine*. Another measure to compute the association between words using web search engines results is the Normalized Google Distance [Cilibrasi and Vitanyi, 2007], which relies on the number of times words co-occur in the document indexed by an information retrieval system. In this study, experimental results with our web-based measure will be compared with the basic measures (*Dice*, *Jaccard*, *Overlap*, *Cosine*).

There are also measures that use the web to extract synonym terms as in [Turney, 2001], the author presents a simple unsupervised learning algorithm for recognizing synonyms based on statistical data acquired by querying a Web search engine. The algorithm, called PMI-IR, uses Pointwise Mutual Information (PMI) and Information Retrieval (IR) to measure the similarity of pairs of words. Another related task where the Web is used is in Named Entity Recognition (NER). [Roche and Prince, 2010] addresses a particular case of NER, i.e. acronym expansion when this does not appear in the document (i.e. expansion). In this paper, nine quality measures are provided regarding relevant definition prediction based on mutual information (MI), cubic MI (MI^3), and Dice's coefficient. It is applied in the biomedical domain, where acronyms are numerous.

These co-occurrence measures are corpus-based association measures. As the Web is the largest resource of data, therefore, these measures can be applied to the Web. The Web contains many corpus-based and/or corpus-driven dictionaries of several domains.

Nowadays is relatively easy to collect a large corpus from the Web using search engines or web crawlers. A first problem is that we know little about the domains of texts in the collected corpus [Sharoff, 2011]. Even if we collect domain-specific corpora and we are sure that all texts are about this specific domain, i.e. the amount of research papers, technical texts, newspapers, webpages, tutorials, social media, etc., is still unknown. A second problem is if this corpus is very big, we might not be able to access it from a personal computer. Therefore, the use of the Web is beneficial, it allowd access to all kinds of data and domains. For this, we can use search engines for querying the Web, thus taking advantage of its indexation and retrieval algorithms.

As we have seen, this chapter introduced the state-of-the-art of terminology extraction tasks. This chapter has highlighted the most relevant studies regarding this task. The most well-known approaches are based on linguistic and statistical features. Another important task related to this is Structuring and Grouping of Terms, which aims to gather terms considered related. For instance, [Habert et al., 1996] outlines an approach to identify essential concepts and relationships in a corpus. This approach is based on a symbolic method, it uses parse trees to classify similar words. A more recent study related to the biomedical domain improves the signal in pharmacovigilance [Dupuch et al., 2011]. Pharmacovigilance is related to the collection, analysis and prevention of adverse drug-induced reactions (ADR). In this study, the authors proposed to use methods designed for term structuring (detection of synonymous and hierarchical relations) for the generation of these groupings.

There are also systems, such as Camaleon [Séguéla and Aussenac-Gilles, 1999], Lexiclass [Assadi, 1997], Asium [Faure and Nédellec, 1998]. For instance, the Lexiclass system is used after the Syntex system to structure similar terms.

4.4 Discussion

We have outlined the most relevant studies related to Automatic Term Extraction. We have described methodologies applied to the general domain, as well as methodologies applied to a more specific domain, i.e. the biomedical domain.

We divided the state-of-the art of Automatic Term Extraction into three kinds of approach: (I) Text Mining approaches, (II) Graph-based approaches, and (III) Web-based approaches. Text Mining studies are classified into 4 categories: (i) Linguistic, (ii) Statistical, (iii) Machine Learning, and (iv) Hybrid. Among these approaches, we have outlined that the major works are based on unsupervised techniques, while a few methodologies are also based on supervised techniques due to the lack of annotated terminologies to evaluate the results for several languages.

Table 4.1 sums up the most relevant methodologies and the features they take into account (i.e. linguistic, statistical, graph, web). In this table, the *knowledge* aspect is added for terminology extraction. In the state-of-the-art, approaches based on knowledge bases are not mentioned because we did not identify studies related to knowledge-based methods. This table also contains the most relevant AKE methods that have shown good results in other tasks such as Information Retrieval, etc., and might be used for our objective. In this thesis, these methods will be adapted for terminology extraction. Major works have been built for extract terms in general domain. Few of these have been applied to specific domains such as biomedicine, implying in few cases a soft change in its methodology.

As noted in previous sections, linguistic and statistical methods are frequently proposed to extract terminology. The linguistic aspect is important because it allows extraction of complex terms for a specific domain. In [Daille, 1994, Daille, 1996], the authors use general patterns to extract terminology, while focusing more on 2-gram terms. These approaches can also extract longer terms according to the extraction of variations. A complicated variation extraction step involves the identification of “head” and ‘argument”, so errors obtained during the recognizing task can add more problems to the extraction process.

As we mentioned in previous sections, to our knowledge, graphs and the Web have never been applied to terminology extraction. However, graphs have recently been used for keyword extraction tasks, as shown in [Rousseau and Vazirgiannis, 2013].

In our context, we use all of the previously mentioned aspects to automatically extract biomedical terms. In our workflow, first we apply a linguistic filter, then we propose a ranking of terms according to statistical measures. Finally, we propose a re-ranking of terms via graph- and Web-based measures. The knowledge aspect appears when knowledge bases are used to generate a list of linguistic patterns. We hence propose a complete automatic workflow to achieve automatic term extraction, which has been divided into three steps: (i) Candidate term extraction, (ii) Ranking of candidate terms, and (iii) Re-ranking. This workflow is applied in particular to the biomedical domain.

Our objectives for using these five aspects are:

- We take the linguistic aspect into account because we work specifically in the biomedical domain. For this, we propose biomedicine-oriented linguistic patterns. A major difference between the linguistic patterns used in [Daille, 1994, Daille, 1996, Frantzi et al., 2000] and our patterns is that the before mentioned study uses patterns focused on the general domain. In [Daille, 1994, Daille, 1996], we can observe that patterns are frequently oriented towards extracting 2-gram terms, in several cases it extracts 3-gram and 4+gram terms because of the extraction of variants. In [Frantzi et al., 2000], the upper limit is 7-gram terms. Our linguistic patterns can extract up to 12-gram terms. Another important difference in comparison to these studies and the before mentioned studies is that our linguistic patterns are focused on the biomedical domain, allowing us to extract specialized terms, for instance terms containing numbers in its structure. In addition, in our methodology, each linguistic pattern is associated with a probability. Syntex [Bourigault and Fabre, 2000, Bourigault et al., 2005] is a system that takes complex linguistic patterns into account. One difference in comparison to our methodology is that we do not take verbs into account for term extraction. According the UMLS and BioPortal statistic over biomedical terms, they do not contain verbs in their structure. Verbs can be used to extract a kind of relationship towards the biomedical enrichment. In a recent study related to the biomedical domain, BioYaTeA [Golik et al., 2013], the authors use biomedical linguistic patterns proposed by a human specialist and then they apply some rules and propose a list of terms to experts. A possible inconvenience is that the experts have to validate the term relevance.

- We use knowledge bases to build linguistic patterns. These knowledge bases are specialized in the biomedical domain, e.g. UMLS, MeSH, SNOMED, etc. They allow us to create complex linguistic patterns for the extraction of rare terms.

- As the linguistic aspect identifies a high number of terms, the statistic aspect is used to rank them according to a score of importance within the corpus, thus avoiding potential noise. Without the statistical aspect, term selection becomes a hard task if done by humans. We take as base *C-value* and we also adapt AKE measures that have shown good results in the literature to extract biomedical terms.

- Graphs allow to visualize important information that is hidden as simple text, as showed in [Rousseau and Vazirgiannis, 2013]. Therefore, we use graphs to represent our corpus for improving the term extraction task. In [Rousseau and Vazirgiannis, 2013], the authors create a graph where each vertex represents a word, in our case each vertex represents a term (single- or multi- word term).

- The web is the largest resource of data. The main reason of using the web is that it contains a large vast of domains which might give an idea of the general use of biomedical terms in similar domains and less related domains.

Therefore, our methodology for Automatic Biomedical Term Extraction involves: (i) the use of specialized linguistic patterns, (ii) the creation of new ranking measures based on already existing measures, (iii) the use of graphs, and (iv) the use of the Web.

In the following chapter, we describe the use of these aspect to achieve the automatic biomedical terminology extraction.

	Text Mining			*Graph*	*Web*	*Biomedical*	*ATE*	*AKE*
	Linguistic	Statistic	ML					
[Krauthammer and Nenadic, 2004]	✓						✓	
Termino [David and Plante, 1990, Faraj et al., 1996]	✓						✓	
Lexter [Bourigault, 1993, Bourigault and Jacquemin, 1999, Aussenac-Gilles and Bourigault, 2000]	✓						✓	
[Drouin, 2003]	✓						✓	
[Turenne and Barbier, 2004]	✓						✓	
Fastr [Jacquemin, 1996, Jacquemin, 1999]	✓							
BioYaTeA [Golik et al., 2013]	✓					✓	✓	
Mantex [Sebag, 2002]		✓					✓	
Mutual Information [Church and Hanks, 1990]		✓					✓	
Cubic Mutual Information [Daille, 1994]		✓					✓	
Ana [Enguehard et al., 1993, Enguehard and Pantera, 1995]		✓					✓	
TF-IDF [Salton and Buckley, 1988]		✓						✓
Okapi BM25 [Robertson et al., 1999]		✓						✓
RIDF [Sebag, 2002]		✓						✓
Chi-square [Matsuo and Ishizuka, 2004]		✓						✓
RAKE [Rose et al., 2010]		✓						✓
[Conrado et al., 2013]			✓				✓	
[Newman et al., 2012]			✓				✓	
[Zhang et al., 2010]			✓				✓	
GlossEx [Kozakov et al., 2007]	✓	✓					✓	
Weirdness [Ahmad et al., 1999]	✓	✓					✓	
Xtract [Smadja, 1993]	✓	✓					✓	
Termight [Dagan and Church, 1997]	✓	✓					✓	
Acabit [Daille, 1994, Daille, 1996]	✓	✓					✓	
Syntex [Bourigault and Fabre, 2000, Bourigault et al., 2005]	✓	✓				✓	✓	
C/NC-value [Frantzi et al., 2000]	✓	✓				✓	✓	
[Rousseau and Vazirgiannis, 2013]		✓		✓				✓
[Blanco and Lioma, 2012]		✓		✓				✓
[Roche and Prince, 2010]		✓			✓	✓	✓	
[Cilibrasi and Vitanyi, 2007]		✓			✓			
[Turney, 2001]		✓			✓		✓	
Ranking + Re-ranking [Lossio-Ventura et al., 2015]	✓	✓	✓	✓	✓	✓	✓	

Table 4.1: Summarization of Methodologies for Automatic Term Extraction.

Chapter

5

Methodology

In the previous chapter we described the state-of-the-art for automatic term extraction taking into account five aspects, i.e. linguistic, knowledge, statistic, graph and web. We also positioned our methodology in comparison to those related works. As well as, we announced the use of baselines measures for this task.

Therefore, this chapter describes the baseline measures, their modifications as well as new measures that we propose for the biomedical term extraction task. The principle of our approach is to assign a weight to a term, which represents the appropriateness of being a relevant biomedical term. This allows to give as output a list ranked by their appropriateness.

We define "relevant biomedical terms" as the set of most important terms that might belong to the biomedical domain. Terms belonging to the biomedical domain will form the biomedical technical terminology.

Our methodology for automatic term extraction has three main steps plus an additional step (a), described in Figure 5.1, and in the sections hereafter:

- (a) Pattern Construction,
- (1) Candidate Term Extraction,
- (2) Ranking of Candidate Terms,
- (3) Re-ranking.

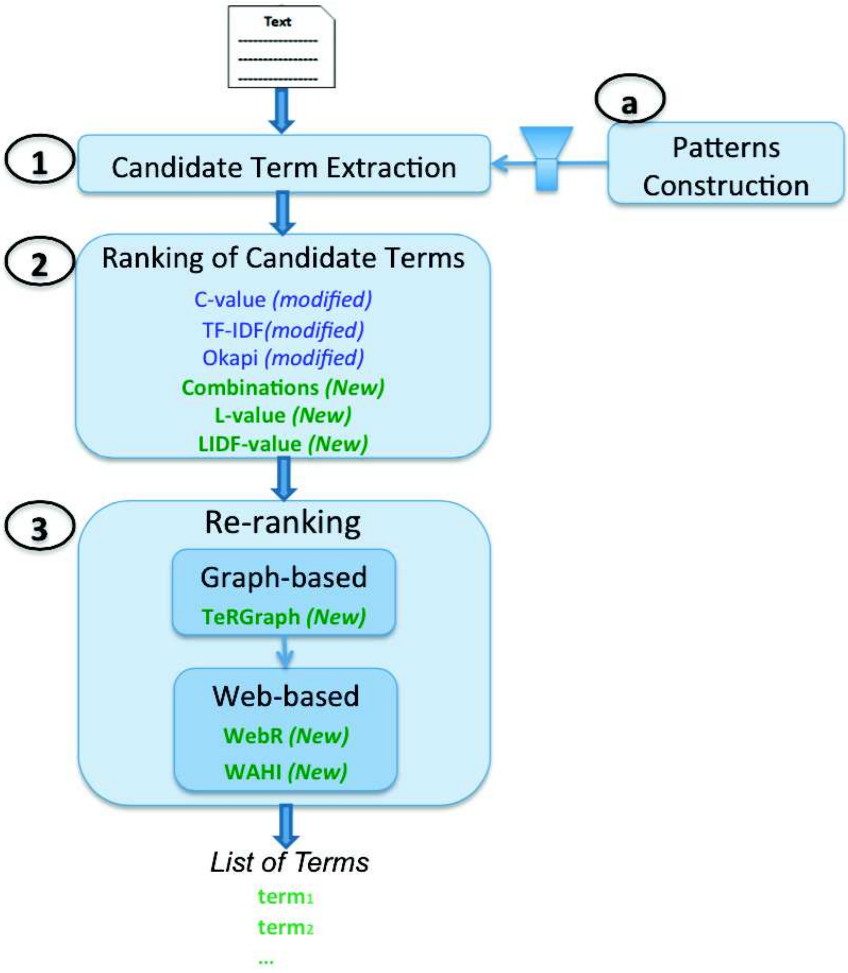


Figure 5.1: Workflow Methodology for Biomedical Term Extraction.

Patterns Construction

As previously cited, we supposed that biomedical terms have a similar syntactic structure (linguistic aspect). Therefore, we built a list of the most common linguistic patterns according to the syntactic structure of terms present in the UMLS[1] (for English and Spanish), and the French version of MeSH[2], SNOMED International and the rest of the French content in the UMLS.

Part-of-Speech (POS) tagging is the process of assigning each word in a text to its grammatical category (e.g. noun, adjective). This process is performed based on the definition of the word or on the context in which it appears. This is highly time-consuming, so we conducted automatic part-of-speech tagging.

[1] http://www.nlm.nih.gov/research/umls

[2] http://mesh.inserm.fr/mesh/

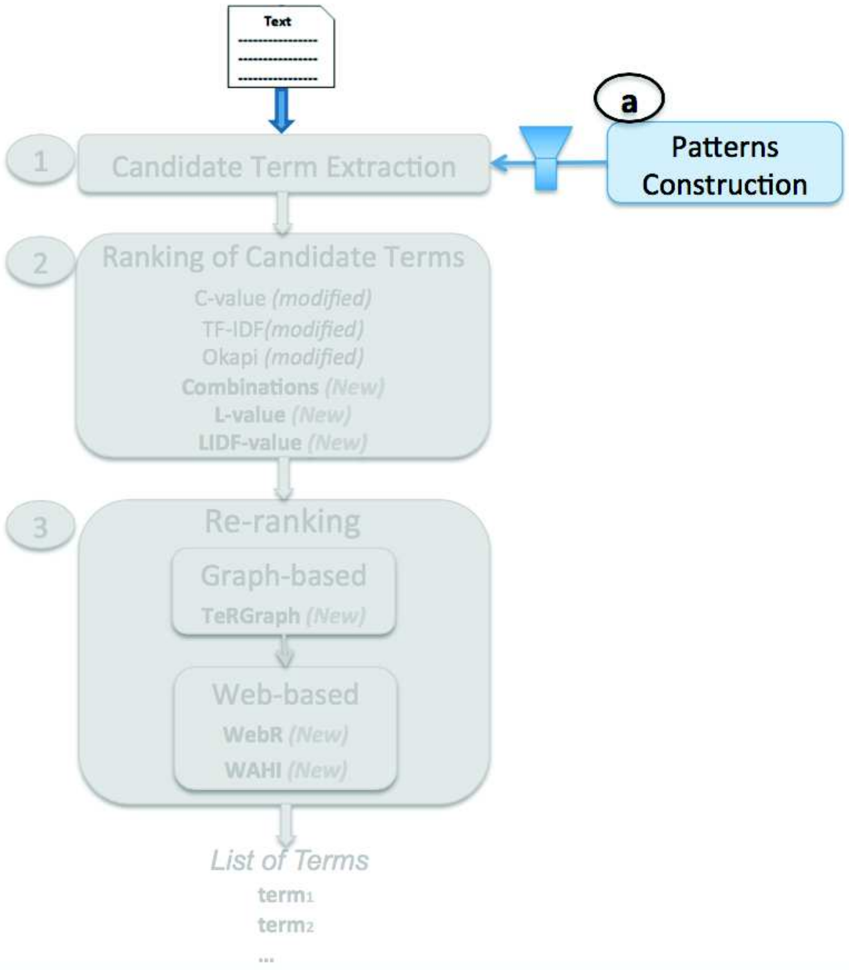


Figure 5.2: Section to Extract Patterns.

We evaluated three tools (TreeTagger[3], Stanford Tagger[4] and Brill's rules[5]). This evaluation was carried out throughout the entire workflow with the three tools and we assessed the precision of the extracted terms. We noted that in general TreeTagger gave the best results for Spanish and French. Meanwhile, for English, the Stanford tagger and TreeTagger gave similar results. We finally chose TreeTagger, which gave better results and may be used for English, French and Spanish. Moreover, our choice was validated with regard to a recent comparison study [Tian and Lo, 2015], wherein the authors showed that TreeTagger generally gives the best results, particularly for nouns and verbs.

Therefore, we carried out automatic part-of-speech tagging of the biomedical terms

[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[4] http://nlp.stanford.edu/software/tagger.shtml

[5] http://en.wikipedia.org/wiki/Brill_tagger

using TreeTagger, and then computed the frequency of the syntactic structures. Patterns among the 200 highest frequencies were selected to build the list of patterns for each language. From this list, we also computed the weight (probability) associated with each pattern, i.e. the frequency of the pattern over the sum of frequencies (see Algorithm 1). This weight represents the probability to find this pattern among the existing terms in the used terminology. This weight will be used for two measures.

The number of terms used to build these lists of patterns was 3 000 000 for English, 300 000 for French, and 500 000 for Spanish, taken from the previously mentioned terminologies. Table 5.2 illustrates the computation of the linguistic patterns and their weights for English.

Different terminology extraction studies are based on the use of regular expressions or linguistic patterns to extract candidate terms, for instance [Daille, 1994, Frantzi et al., 2000]. Generally these regular expressions are manually built for a specific language and/or domain [Daille et al., 1994]. In our setting, we prefer to (i) construct and (ii) apply patterns in order to extract terms in texts.

These patterns have the advantage of being generic because they are based on defined PoS tags, for instance few linguistic patterns are covered by the linguistic patterns defined in [Daille, 1994, Frantzi et al., 2000]. Moreover, they are very specific because they are (automatically) built with specialized biomedicine resources. Concerning this last point, we can consider we are close to the use of regular expressions.

There are three main reasons that we use specific linguistic patterns. First, we would like to restrict the patterns to the biomedical domain. For instance, biomedical terms often contain numbers in their syntactic structure, and this is very specific to the biomedical domain, e.g. *"epididymal protein 9"*, *"pargyline 10 mg"*. General patterns do not enable extraction of such terms as those used in the studies before mentioned. Table 5.1 shows several examples of terms having complex structures being extracted by our methodology. Note that these terms can not be identified by classical approaches. Our methodology is based on 200 significant patterns for English, French, or Spanish, yet different for each language. For instance, there are 55 patterns for English that contain numbers in the linguistic structure.

The second reason is in [Daille, 1994, Daille, 1996] patterns frequently are oriented to extract 2-gram terms, in several cases 3-gram and 4-gram terms because of the extraction of variants. In [Frantzi et al., 2000] the upper limit is 7-gram terms. Our linguistic patterns can extract till 12-gram terms. Thus, this kind of pattern seems quite relevant for this domain.

The third reason for using lexical patterns is that we assign a probability of occurrence to each pattern, which would not be possible with classical patterns and regular expressions.

Term ***Linguistic Pattern***
virus type 1
NN NN CD
class ii genes
NN CD NN
class ii transactivator
NN CD NN
interleukin 2 receptor
NN CD NN
polymerase ii transcription
NN CD NN
virus type 1 tax
NN NN CD NN
type 1 long terminal repeat
NN CD JJ JJ NN
fmol million cells
NN CD NN
class ii antigen
NN CD NN
class ii gene expression
NN CD NN NN
histocompatibility complex class i molecules
NN NN NN CD NN
histocompatibility complex class ii antigens
NN NN NN CD NN histocompatibility complex class ii molecules
NN NN NN CD NN
immunodeficiency virus type 1
NN NN NN CD
enhancer-binding protein 2
JJ NN CD
role for nf-kappab in human cd34 bone
NN IN NN IN JJ NN NN
mononuclear leukocytes of patients with atopic dermatitis
JJ NN IN NN IN JJ NN
leukocytes in patients with chronic pulmonary heart
NN IN NN IN JJ JJ NN
absence in lymphocytes from untreated cml patients
NN IN NN IN JJ NN NN
fresh leukemic cells of adult t-cell leukemia
JJ JJ NN IN JJ NN NN
transcription-dependent surface expression of different endothelial cell
JJ NN NN IN JJ JJ NN

Table 5.1: Example of with complex linguistic pattern (where *NN* is a noun, *IN* a preposition or subordinating conjunction, *JJ* an adjective, and *CD* a cardinal number)

```
Algorithm 1: ComputePatterns (Dictionary, np)
  Data: $Dictionary$ = dictionary of a domain, $np$ = number of patterns to use
  Result: $HT_{patterns}(pattern, probability)$ = Hashtable of the first $np$ linguistic
          patterns with its probability
  begin
      $HT_{patterns} \longleftarrow \emptyset$;
      $HT_{aux}(tag, freq) \longleftarrow \emptyset$ // Hashtable of the tag of each term with its
      frequency ;
      $sizeHT \longleftarrow$ number of terms in $Dictionary$;
      $freq_{total} \longleftarrow 0$ ;
      $probability \longleftarrow 0.0$ ;
      Tag of the $Dictionary$;
      for $tag$ of each term $\in Dictionary$ do
          if $tag \in HT_{aux}$ then
              update $HT_{aux}(tag, freq + 1)$;
          else
              add $HT_{aux}(tag, 1)$;
          end
      end
      Rank $HT_{aux}(tag, freq)$ by the $freq$;
      $freq_{total} \longleftarrow \sum_{i=1}^{np} freq(HT_{aux}(i))$;
      for $i = 1; i \leq np; i{+}{+}$ do
          $probability \longleftarrow \frac{freq(HT_{aux}(i))}{freq_{total}}$ ;
          add $HT_{patterns}(tag(HT_{aux}(i)), probability)$;
      end
  end
```

Pattern	Frequency	Probability
NN IN JJ NN IN JJ NN	3006	$3006/4113 = 0.73$
NN CD NN NN NN	1107	$1107/4113 = 0.27$
	4113	1.00

Table 5.2: Example of pattern construction (where *NN* is a noun, *IN* a preposition or subordinating conjunction, *JJ* an adjective, and *CD* a cardinal number)

5.1 Candidate Term Extraction (step 1)

The first main step is to extract the candidate terms. So we apply part-of-speech to the whole corpus using TreeTagger. Then we filter out the content of our input corpus using previously computed patterns. We select only terms whose syntactic structure is in the patterns list. The pattern filtering is specifically done on a per-

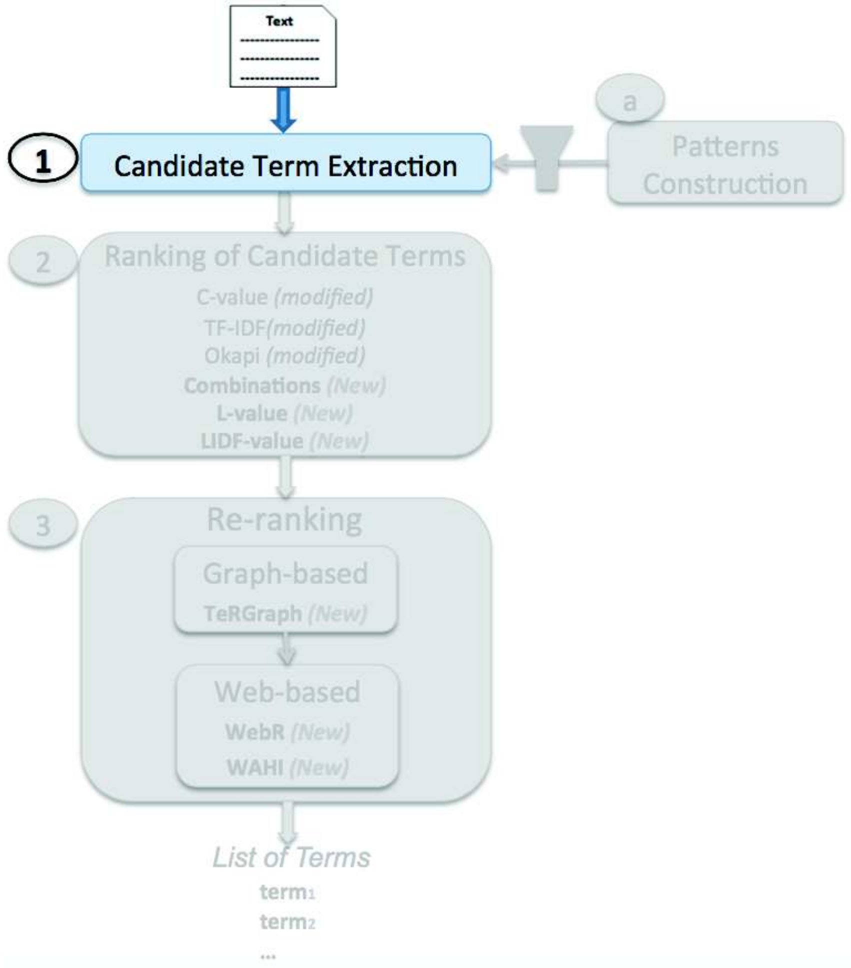


Figure 5.3: Section to Extract Candidate Terms.

language basis (i.e. when the text is in French, only the French list of patterns is used).

5.2 Ranking Measures (step 2)

We need to select the most appropriate terms for the biomedical domain. Candidate term ranking is therefore essential. For this purpose, several measures are proposed and Figure 5.4(2) shows the set of available measures.

We propose some modifications of the most known measures in the literature (i.e., *C-value*, *TF-IDF*, *Okapi*), and propose new ones (i.e., *F-TFIDF-C*, *F-OCapi*, *LIDF-value*, *L-value*). Those measures are linguistic- and statistic- based, they are also not very time-consuming. In this step, only one measure will be selected to perform the ranking.

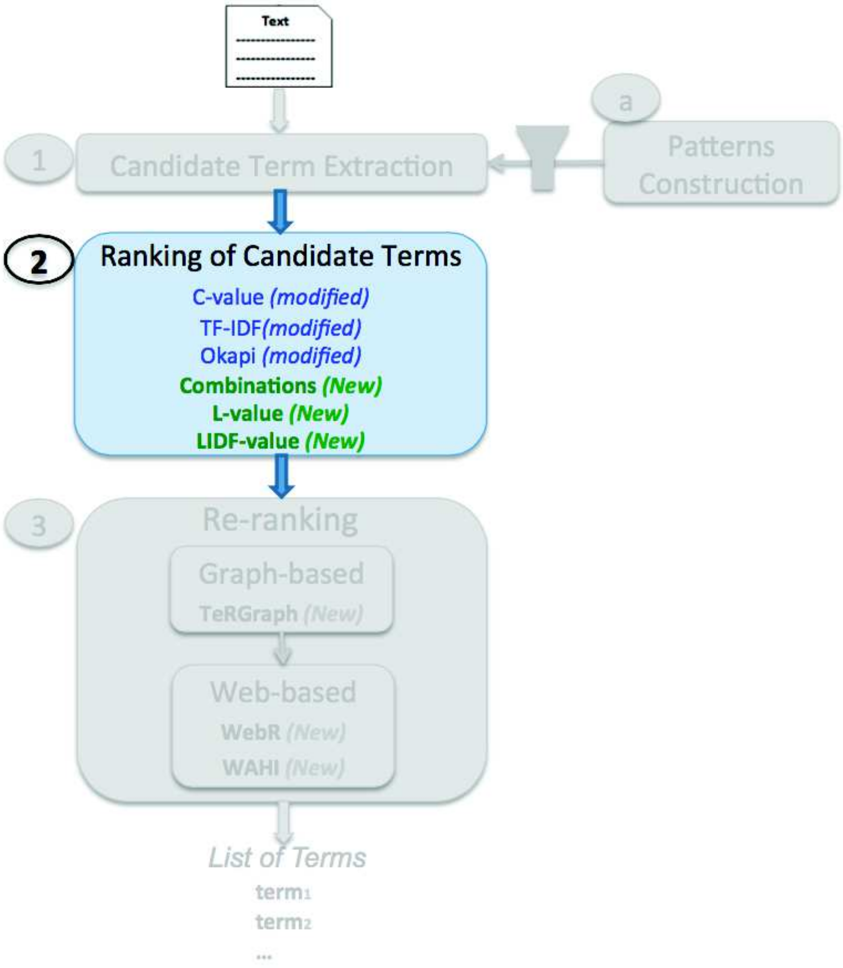


Figure 5.4: Section Describing the Ranking Measures.

The measures of this section take a list of candidate terms previously filtered by linguistic patterns as input, which makes it possible to assess less invalid terms while dealing with the noise problem. In addition to the use of linguistic patterns to alleviate the problem of the extraction of multi-word terms having various complex structures. Moreover, the frequency decreases the number of invalid terms to evaluate (noise). The measures mentioned above are effective on large amounts of data [Lv and Zhai, 2011b, Lv and Zhai, 2011a, Singhal et al., 1996], which overcomes the problem of large-scale corpora. Hereafter we describe all measures.

5.2.1 C-value

The *C-value* method combines linguistic and statistical information [Frantzi et al., 2000]. Linguistic information is the use of a general regular expression as linguistic patterns, and the statistical information is the value assigned with the *C-value* measure based on the frequency of terms to compute the *termhood* (i.e. the association strength of a term to domain concepts). The *C-value* method aims to improve the extraction of long terms, and it was specially built for extracting multi-word terms.

The principle of this measure is to privilege the extraction of the longest terms, penalizing the nested terms, i.e. terms appearing in longer terms. In specialized domains, it is really important to extract complex terms, *C-value* works well extracting complex terminology. For instance, in a ophthalmology corpus, the authors found "soft contact lens" more relevant than "soft contact" which is not a term of this domain.

$$C\text{-}value(A) = \begin{cases} w(A) \times f(A) & \text{if } A \notin nested \\ w(A) \times \left(f(A) - \frac{1}{|S_A|} \times \sum_{b \in S_A} f(b) \right) & \\ \quad \text{otherwise} & \end{cases} \tag{5.1}$$

Where A is the candidate term, $w(A) = \log_2(|A|)$, $|A|$ the number of words in A, $f(A)$ the frequency of A in the unique document, S_A the set of terms that contain A and $|S_A|$ the number of terms in S_A. In a nutshell, *C-value* uses either the frequency of the term if the term is not included in other terms (first line), or decreases this frequency if the term appears in other terms, based on the frequency of those other terms (second line).

We modified the measure in order to extract all terms (single-word + multi-words terms), as also suggested in [Barrón-Cedeño et al., 2009], but in a different manner. The original *C-value* defines $w(A) = \log_2(|A|)$, and we modified $w(A) = \log_2(|A|+1)$ in order to avoid null values for single-word terms, as illustrated in Table 5.3. Note that we do not use a stop word list or a frequency threshold as was originally proposed.

	Original *C-value*	**Modified *C-value***
	$w(A) = \log_2(\|A\|)$	$w(A) = \log_2(\|A\| + 1)$
antiphospholipid antibodies	$\log_2(2) = 1$	$\log_2(2+1) = 1,6$
white blood	$\log_2(2) = 1$	$\log_2(2+1) = 1,6$
platelet	$\log_2(1) = 0$	$\log_2(1+1) = 1$

Table 5.3: Calculation of $w(A)$

5.2.2 TF-IDF and Okapi

These measures are used to associate a weight to each term in a document [Salton and Buckley, 1988]. This weight represents the term relevance for the document. The output is a ranked list of terms for each document, which is often used in information retrieval so as to order documents by their importance for a given query [Robertson et al., 1999]. *Okapi* can be seen as an improvement of the *TF-IDF* measure, while taking the document length into account.

$$TF\text{-}IDF(A,d,D) = tf(A,d) \times idf(A,d) \tag{5.2}$$

$$tf(A,d) = \frac{f(A,d)}{max\{f(A,d) : w \in d\}}$$
$$idf(A,d) = \log \frac{|D|}{|\{d \in D : A \in d\}|}$$

$$Okapi(A,d,D) = tf_{BM25}(A,d) \times idf_{BM25}(A,d) \tag{5.3}$$

$$tf_{BM25}(A,d) = \frac{tf(A,d) \times (k_1 + 1)}{tf(A,d) + k_1 \times (1 - b + b \times \frac{dl(d)}{dl_{avg}}))}$$
$$idf_{BM25}(A,d) = \log \frac{|D| - dc(A) + 0.5}{dc(A) + 0.5}$$

Where A is a term, considering d a document, D the collection of documents, $f(A,d)$ the frequency of A in d, $tf(A,d)$ the term frequency of A in d, $idf(A,D)$ the inverse document frequency of A in D, $dc(t)$ the number of documents containing term A, this means: $|\{d \in D : t \in d\}|$, $dl(d)$ the length of the document d in number of words, dl_{avg} the average document length of the collection.

As the output is a ranked list of terms per document, we could find the same term in different documents, with different weights in each document. So we need to merge the term into a single list. For this, we propose to merge them according to three functions, which respectively calculate the sum(S), max(M) and average(A) of the weights of a term. At the end of this task, we have three lists from *Okapi* and three lists from *TF-IDF*. The notation for these lists are $Okapi_X(A)$ and $TF\text{-}IDF_X(A)$, where A is the term, and X the factor $\in \{M, S, A\}$. For example, $Okapi_M(A)$ is the value obtained by taking the maximum *Okapi* value for a term A in the whole corpus. Figure 5.5 shows the merging process.

With aim of improving the term extraction precision, we designed two new combined measures, while taking the values obtained in the above steps into account. Both are based on harmonic means of two values.

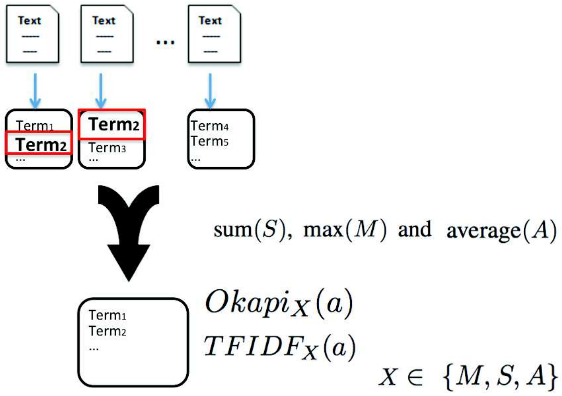


Figure 5.5: Merging lists.

5.2.3 F-OCapi and F-TFIDF-C

Considered as the harmonic mean of the two used values, this method has the advantage of using all values of the distribution. These measures were inspired by the *F-measure*, for this reason they start with the *F* letter.

The arithmetic mean is closer to the highest of the two values obtained by the two measures for a term, which is not a good representation of both. On the other hand, harmonic mean is closer to minimum between two values, which is a better representation between these two. Therefore, the principle of these measures is to have two high and close values to get a good ranking for a term.

$$F\text{-}OCapi_X(A) = 2 \times \frac{Okapi_X(A) \times C\text{-}value(A)}{Okapi_X(A) + C\text{-}value(A)} \tag{5.4}$$

$$F\text{-}TFIDF\text{-}C_X(A) = 2 \times \frac{TFIDF_X(A) \times C\text{-}value(A)}{TFIDF_X(A) + C\text{-}value(A)} \tag{5.5}$$

5.2.4 C-Okapi and C-TFIDF

Our assumption is that *C-value* might be more representative if the term frequency (in equation (5.1)) of the terms is replaced with a more significant value, in this case the *Okapi's* and *TF-IDF's* values of the terms over the whole corpus.

$$C\text{-}wm(a) = \begin{cases} w(a) \times wm_X(a) & \text{if } \ a \notin nested \\ \\ w(a) \times \left(wm(a) - \frac{1}{|S_a|} \times \sum_{b \in S_a} wm(a) \right) \\ \quad \text{otherwise} \end{cases} \tag{5.6}$$

Where $wm(a)$ is a weighting measure $= \{Okapi_X, TFIDF_X\}$.

5.2.5 LIDF-value and L-value

In this section we present two new measures. The first one, called *LIDF-value* (**L**inguisitic patterns, **IDF**, and C-**value** information). *LIDF-value* is partially presented in [Lossio-Ventura et al., 2014d]. This is a new ranking measure based on linguistic and statistical information.

Our method *LIDF-value* is aimed at computing the termhood for each term, using the *linguistic* information calculated as described below, the *idf*, and the *C-value* of each term. The *linguistic* information gives greater importance to the term unithood in order to detect low frequency terms. So, we associate the pattern weight (see Table 5.2) with the candidate term *probability*. In our hypothesis this weight represents the *probability* of a candidate term of being a relevant biomedical term. The *probability* is associated only if the syntactic structure of the term appears in the linguistic pattern list.

The inverse document frequency *(idf)* is a measure indicating the extent to which a term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then by taking the logarithm of that quotient.

The *probability* and *idf* improve low frequency term extraction. The objective of these two components is to tackle the silence problem, allowing extraction of discriminant terms, for instance, in a biomedical corpus, *"virus production"* with low frequency being better ranked than *"human monocytic cell"*, which has a higher frequency. This means that for a low frequency candidate term, its score can be favored if its linguistic pattern is associated with a high probability and/or its *idf* value is also high. The *C-value* measure is based on the term frequency. The *C-value* (see formula 5.1) measure favors a candidate term that does not often appear in a longer term. For instance, in a specialized corpus (Ophthalmology), the authors of [Frantzi et al., 2000] found the irrelevant term *"soft contact"* while the frequent and longer term *"soft contact lens"* is relevant.

As an example, we implement the Algorithm 2, which describes the applied process. These different statistical information items (i.e. *probability* of linguisitic patterns, *C-value*, *idf*) are combined to define the global ranking measure *LIDF-value* (see formula 5.7); where $\mathrm{P}(A_{LP})$ is the probability of a term A which has the same linguistic structure pattern LP, i.e. the weight of the linguistic pattern LP computed in Section *Pattern Construction*.

$$LIDF\text{-}value(A) = \mathrm{P}(A_{LP}) \times idf(A) \times C\text{-}value(A) \tag{5.7}$$

Note that *LIDF-value* works only for a set of documents, mainly because the *idf* measure can only be computed on a set of documents (see formula 5.2). Therefore,

Algorithm 2: ComputeLIDF-value *(Corpus, Patterns, min_{freq}, num_{terms})*

```
Data: $Corpus$ = set of documents of a specific-domain;
$Patterns = HT_{patterns}(pattern, probability)$ //Hashtable of linguistic patterns
with its probability;
$min_{freq}$ = frequency threshold for candidate terms;
$num_{terms}$ = number of terms to take as output
Result: $L_{terms}$ = List of ranked terms
begin
    Tag the $Corpus$;
    Take the $lemma$ of each tagged word;
    Extract candidate terms $A$ by filtering with $Patterns$;
    Remove candidate terms $A$ below $min_{freq}$;
    for each candidate term $A \in Corpus$ do
        $LIDF\text{-}value(A) = \mathrm{P}(A_{LP}) \times idf(A) \times C\text{-}value(A)$;
        add $A$ to $L_{terms}$;
    end
    Rank $L_{terms}$ by the value obtained with $LIDF\text{-}value$;
    Select the first $num_{terms}$ terms of $L_{terms}$ ;
end
```

for datasets composed of one document, we propose a new measure, *L-value*, as explained in the following paragraphs.

L-value is a variant of *LIDF-value*, focused on one document with the goal of benefiting from the *probability* of linguisitic patterns computed for *LIDF-value*. This measure does not contain the *idf* (see formula 5.8). *L-value* is interesting to highlight the more representative terms of a single corpus without considering the discriminative aspects, e.g. *idf*. This measure gives another point of view and is complementary to those based on the *idf* weighting.

A single document can be considered as a free text without delimitation. For instance, a scientist article, a book, a document created with titles/abstracts from a library database. *L-value* becomes interesting when it does not exist a considerable amount of data for a new subject, i.e. an emergent term in the community. For instance, the “Ataxia Neuropathy Spectrum” term appears only in 4 titles/abstracts of scientist articles from PubMed[6] between 2009 and 2015. PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics.

$$L\text{-}value(A) = \mathrm{P}(A_{LP}) \times C\text{-}value(A) \tag{5.8}$$

[6] `http://www.ncbi.nlm.nih.gov/pubmed`

5.3 Re-ranking (step 3)

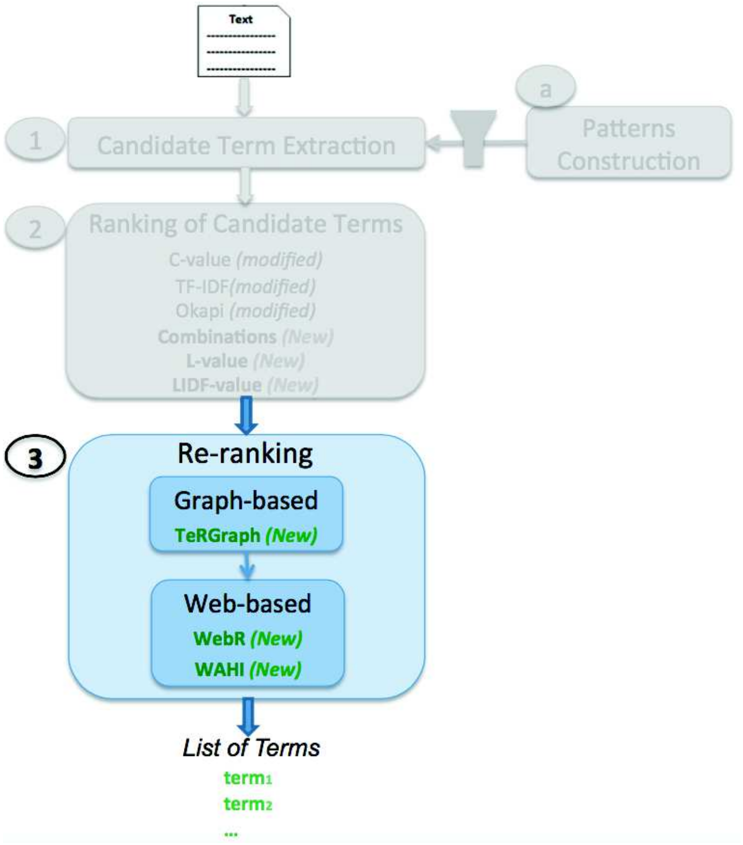


Figure 5.6: Section Describing the Re-ranking Measures.

After the term extraction, we propose new measures to re-rank the candidate terms in order to increase the top k term precision. The re-ranking measures aim to improve the term extraction results of ranking measures. This involves positioning the most relevant biomedical terms at the top of the list. That provides more confidence that the terms appearing at the top of this list are true biomedical terms.

These re-ranking functions represent an extension of the measures presented in [Lossio-Ventura et al., 2014c]. Therefore, as improvements, we propose to take graph-theoretic information into account to highlight relevant terms, as well as web information, as explained in the following subsections. These measures can be executed separately, but the graph construction is time consuming, and the number of search engine queries is limited. Therefore, we just apply these measures for a group of selected terms given by a ranking measure. Because the ranking measures have proved to be more efficient applied before than *TeRGraph* and *web*-based measures.

As these measures are applied to the list of terms obtained with a ranking measure, which tackles noise, silence and multi-word term extraction problems, so they also take into account those problems. As mentioned, the objective of re-raking measures is to re-rank terms, so the manual validation efforts of the candidate terms decrease because the relevant biomedical term is allocated at the top of the list.

5.3.1 A new graph-based ranking measure: "TeRGraph" (Terminology Ranking based on Graph information)

This approach aims to improve the ranking (and therefore the precision results) of extracted terms. As mentioned above, in contrast to the above-cited study, the graph is built with a list of terms obtained according to a measure described in Section 5.2, where vertices denote terms linked by their co-occurrence in sentences in the corpus. Moreover, we make the hypothesis that the term representativeness in a graph, for a specific-domain, depends on its number of neighbors, and the number of neighbors of its neighbors. We assume that a term with more neighbors is less representative of the specific domain. This means that this term is used in the general domain. Figure 5.7 illustrates our hypothesis.

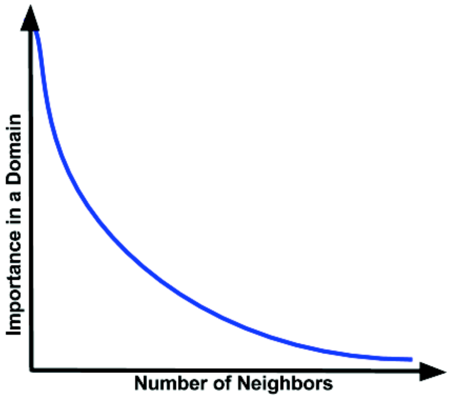


Figure 5.7: Importance of a term in a domain

The graph-based approach is divided into two steps:

(i) **Graph construction:** a graph (see Figure 5.9) is built where vertices denote terms, and edges denote co-occurrence relations between terms, co-occurrences between terms are measured as the weight of the relation in the initial corpus. This approach is statistical because it links all co-occurring terms without considering their meaning or function in the text. This graph is undirected as the edges imply that terms simply co-occur, without any further distinction regarding their role. We take the *Dice coefficient*, a basic measure to compute the co-occurrence between two terms x and y, as defined by the following formula:

$$D(x,y) = \frac{2 \times P(x,y)}{P(x)+P(y)} \tag{5.9}$$

In [Rousseau and Vazirgiannis, 2015], the authors build a graph in a similar way: an undirected graph, only nouns and and adjectives are considered as vertices, each vertex represents a word, and stemming each word. In our case vertices can be multi-word terms, which depend directly of the linguistic patterns and of the used ranking measure. Stemming is not considered in our work. Another work building [Habert et al., 1996] a co-occurrence graph in a similar manner is proposed to cluster words of a same class.

(ii) **Representativeness computations on the term graph:** a principled graph-based measure to compute term weights (representativeness) is defined. The aim of this new graph-based ranking measure, *TeRGraph*, see Equation 5.10, is to derive these weights for each vertex, (i.e. multi-word term weight), in order to re-rank the list of extracted terms.

$$TeRGraph(A) = \log_2\left(k + \frac{1}{1 + |\mathrm{N}(A)| + \displaystyle\sum_{T_i \in \mathrm{N}(A)} |\mathrm{N}(T_i)|}\right) \tag{5.10}$$

Where A represents a vertex (term), $\mathrm{N}(A)$ the neighborhood of A, $|\mathrm{N}(A)|$ the number of neighbors of A, T_i the neighbor i of A, and k a constant. The intuition for Equation 5.10 is as follows: the more a term A has neighbors (directly with $\mathrm{N}(A)$ or by transitivity with $\mathrm{N}(T_i)$), the more the weight decreases. Indeed, a term A having a lot of neighbors is considered too general for the domain (i.e. this term is not salient), so it has to be penalized via the associated score.

The k constant affects the *TeRGraph* value, i.e. the set of values that *TeRGraph* takes when k changes. For instance, when $k = 0.5$, the set of values for *TeRGraph* is between -1 and 0, (i.e., $TeRGraph \in [-1, 0]$), and when $k = 1$, $TeRGraph \in [0, 0.6]$. As the values taken by *TeRGraph* are different, then the slope of the curve is also different. Figure 5.8 shows the behavior of *TeRGraph* when k changes. According the experiments, we have chosen $k = 1.5$. The main reason is that the slope of the curve is low, and the set of values for *TeRGraph* ranges from 0.6 to 1.

See Algorithm 3 for more details, it describes the entire process: (1) co-occurrence graph construction, (2) computation of the representativeness of each vertex.

Figure 5.9 shows an example to calculate the value of *TeRGraph* for a term in different graphs. These graphs are built with different co-occurrence thresholds (i.e. Dice's value between two terms). In this example, A_1 and A_2 represent the term *chloramphenicol acetyltransferase reporter* in Graphs 1 and 2, respectively. Note that the term *chloramphenicol acetyltransferase reporter* in Graph 1 has a lower value than in Graph 2, this is due the term in Graph 1 contains more neighbors

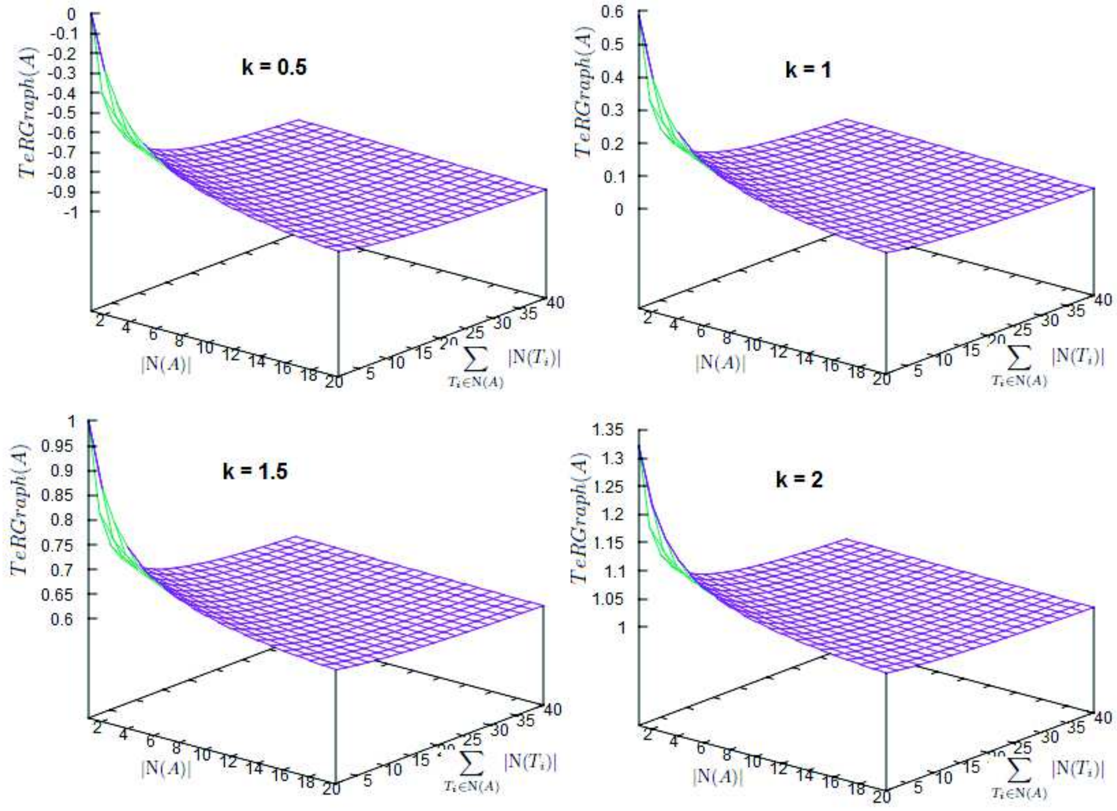
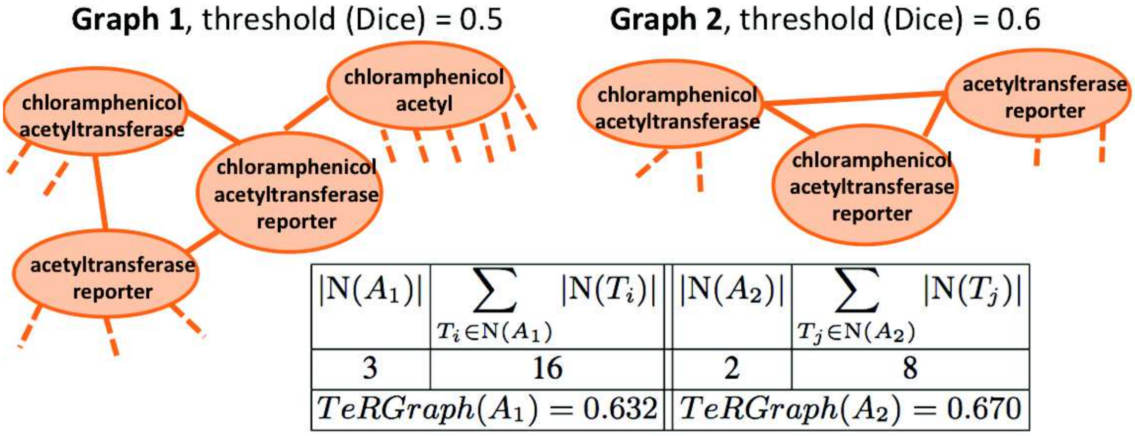


Figure 5.8: TeRGrpah's value for $k = \{0.5; 1; 1.5; 2\}$

than in Graph 2.


Graph 1, threshold (Dice) = 0.5
Graph 2, threshold (Dice) = 0.6
chloramphenicol acetyltransferase
chloramphenicol acetyl
chloramphenicol acetyltransferase reporter
acetyltransferase reporter
chloramphenicol acetyltransferase
acetyltransferase reporter
chloramphenicol acetyltransferase reporter

$\lvert N(A_1)\rvert$	$\sum_{T_i \in N(A_1)} \lvert N(T_i)\rvert$	$\lvert N(A_2)\rvert$	$\sum_{T_j \in N(A_2)} \lvert N(T_j)\rvert$
3	16	2	8
$TeRGraph(A_1) = 0.632$		$TeRGraph(A_2) = 0.670$	

Figure 5.9: *TeRGraph*'s value for ***chloramphenicol acetyltransferase reporter***

Algorithm 3: ComputeTeRGraph *(L_{terms}, num_{terms}, δ, k)*

```
Data: $L_{terms}$ = List of ranked terms;
$num_{terms}$ = number of terms to be evaluated;
$\delta$ = threshold to create an edge between two terms;
$k$ = constant;
Result: $RRL_{terms}$ = Re-Ranked List of terms
begin
    Select all possible pairs of terms of $L_{terms}$ to compute $D(x,y)$ // in total $C^2_{num_{terms}} = \frac{num_{terms}!}{2!\,(num_{terms}-2)!}$ possibilities ;
    Select pairs which $D(x,y) \geq \delta$ for creating an edge ;
    Select all terms of $L_{terms}$ to compute $TeRGraph$ ;
    for each term $A \in L_{terms}$ do
        $\mathrm{N}(A) \longleftarrow$ neighborhood of A;
        $|\mathrm{N}(A)| \longleftarrow$ number of neighbors of A;
        $TeRGraph(A) = \log_2 \left( k + \frac{1}{1+|\mathrm{N}(A)|+ \sum_{T_i \in \mathrm{N}(A)} |\mathrm{N}(T_i)|} \right)$;
        add $A$ to $RRL_{terms}$;
    end
    Rank $RRL_{terms}$ by the value obtained with $TeRGraph$;
end
```

5.3.2 WebR

The aim of our web-based measure, to predict with a better confidence if a candidate term is a valid biomedical term or not. It is appropriated for *multi-word* terms, as it computes the dependence between the words of a term. In our case, we compute a "strict" dependence, which means the proximity of words of terms (i.e. neighboring words) is calculated with a strict restriction. In comparison to other web-based measures [Cilibrasi and Vitanyi, 2007], *WebR* reduces the number of pages to consider by taking only web pages containing all words of the terms into account. In addition, our measure can be easily adopted for all types of multi-word terms.

$$WebR(A) = \frac{nb(\text{``}A\text{''})}{nb(A)} \tag{5.11}$$

Where A = multi-word term, $a_i \in A$ and $a_i = \{noun,\ adjective,\ foreign\ word\}$.

Where A is the candidate term, nb("A") the number of hits returned by a web search engine with exact match only with multi-word term A (query with quotation marks

"A"), $nb(A)$ the number of documents returned by the search engine, including not exact matches (query A without quotation marks), i.e. whole documents containing words of the multi-word term A. For example, the multi-word term *treponema pallidum*, will generate two queries, the first nb("*treponema pallidum*") which returns with Yahoo 1 100 000 documents, and the second query nb(*treponema pallidum*) which returns 1 300 000 documents, then *WebR(treponema pallidum)*= $\frac{1100000}{1300000}$ = 0.85.

As we mentioned in Section 4.3, our web-based measures can be applied to larger corpora. The likely inconvenient with this are: (i) the lack of knowledge about the corpora, and (ii) the inability to access it from a personal computer. Therefore, the use of web is beneficial, it will allow to access all kind of data of a lot of domains. For this, we can use search engines for querying the web taking advantage of its indexation and of its retrieving algorithms. Therefore, in our workflow, we tested Yahoo and Bing search engines. *WebR* re-ranks the list of candidate terms returned by the combined measures.

5.3.3 A new web ranking measure: WAHI (Web Association based on Hits Information)

Previous studies of web mining approaches query the web via search engines to measure word associations. This enables measurement of the association of words composing a term (e.g. *soft*, *contact*, and *lens* that compose the relevant term *soft contact lens*). To measure this association, our web-mining approach takes the number of pages provided by search engines into account (i.e.number of hits).

Our web-based measure re-ranks the list obtained previously with *TeRGraph*. We will show that this improves the precision of the k first terms extracted (see Chapter 6) and that it is specially appropriate for multi-word term extraction.

Formula 5.9 leads directly to formula 5.12[7]. The *nb* function used in formula 5.12 represents the number of pages returned by search engines (i.e. Yahoo and Bing). With this measure, we compute a *strict* dependence (i.e. neighboring words by using the operator ' " ' of search engines). For instance, x might represent the word *soft* and y the word *contact* in order to calculate the association measure of the *soft contact* term.

$$Dice(x, y) = \frac{2 \times nb(\text{``}x\ y\text{''})}{nb(x) + nb(y)} \tag{5.12}$$

[7] by writing $P(x) = \frac{nb(x)}{nb\_total}$, $P(y) = \frac{nb(y)}{nb\_total}$, $P(x, y) = \frac{nb(x,y)}{nb\_total}$

Then we extend this formula to n elements as follows:

$$Dice(a_1, ..., a_n) \quad = \quad \frac{n \times nb(\text{``}a_1\ ...\ a_n\text{''})}{nb(a_1) + ... + nb(a_n)} = \frac{n \times nb(\text{``}A\text{''})}{\displaystyle\sum_{i=1}^{n} nb(a_i)} \tag{5.13}$$

This measure enables us to calculate a score for all multi-word terms, such as *soft contact lens*.

To obtain *WAHI*, we propose to associate Dice criteria with *WebR* (see formula 5.11). This only takes the number of web pages containing all the words of the terms into account by using operators *“ ”* and *AND*.

For example, *soft contact lens*, the numerator corresponds to the number of web pages with the query *“soft contact lens”*, and for the denominator, we consider the query *soft AND contact AND lens*.

Finally, the global ranking approach combining *Dice* and *WebR* is given by *WAHI* measure (**W**eb **A**ssociation based on **H**its **I**nformation):

$$WAHI(A) = \frac{n \times nb(\text{``}A\text{''})}{\displaystyle\sum_{i=1}^{n} nb(a_i)} \times \frac{nb(\text{``}A\text{''})}{nb(A)} \tag{5.14}$$

The main difference between *WAHI* and *WebR* is that *WebR* computes the dependence between the words of a term. *WAHI* introduces the degree of association between the words composing a term.

Algorithm 4 details the global web mining process to rank terms. We show in the next chapter that open-domain (general) resources, such as the web, can be tapped to support domain-specific term extraction. They can thus be used to compensate for the unavailability of domain-specific resources.

Algorithm 4: ComputeWAHI *(L_{terms}, num_{terms}, LC)*

```
Data: $L_{terms}$ = List of ranked terms;
$num_{terms}$ = number of terms to be evaluated;
$LC = \{noun, adjective, foreign\ word\}$ // linguistic categories
Result: $RRL_{terms}$ = Re-Ranked List of terms
begin
    Select the first $num_{terms}$ terms of $L_{terms}$ to compute $WAHI$;
    for each term $A \in L_{terms}$ do
        for all words $a_i$ of $A \in LC$ do
            $n \longleftarrow$ number of words in A;
            $WAHI(A) \longleftarrow n \times \frac{num\text{-}hits(\text{“}A\text{”})}{\displaystyle\sum_{i=1}^{n} num\text{-}hits(a_i)} \times \frac{num\text{-}hits(\text{“}A\text{”})}{num\text{-}hits(A)}$;
        end
        add $A$ to $RRL_{terms}$;
    end
    Rank $RRL_{terms}$ by the value obtained with $WAHI$;
end
```

Chapter

6

Data and Results

This chapter reports the data set and experiments of our proposal for Automatic Biomedical Term Extraction. This chapter presents experiments for multilingual term extraction in Section 6.2, as well a detailed study of the entire process for multi-word term extraction in Section 6.3.

6.1 Data, Protocol, and Validation

In this section we describe in detail the different data sets and the protocol used for our experiments, as well as the method to evaluate our results.

6.1.1 Data

We used two corpora for our experiments. The first one is a set of biological laboratory tests, extracted from LabTestsOnline[1]. This website provides information in several languages to patients or family caregivers about clinical lab tests. Each test includes the *formal lab test name*, some *synonyms* and possible *alternate names* as well as a description of the test. To reduce bias in our results only descriptions are used in our data set. The LabTestsOnline website was extracted totally for English, French, and Spanish with a crawler created specifically for this purpose. These documents are available online[2]. The choice of this corpus is based mainly in

[1] `http://labtestsonline.org/`

[2] `http://www.lirmm.fr/~lossio/labtestsonline.zip`

the multilingualism, it also exists for several other languages. We have also done experiments in other corpus such as PubMed, Medline Plus, Cochrane, etc. Table 6.1 shows the details of LabTestsOnline data set for different languages.

	Number of Clinical Tests	**Number of Words**
English	235	377 000 words
French	137	174 000 words
Spanish	238	396 000 words

Table 6.1: Details of LabTestsOnline data set.

The second data set is GENIA[3], which is made up of 2 000 titles and abstracts of journal articles that were culled from the Medline database, with more than 400 000 words in English. The GENIA data set contains linguistic expressions referring to entities of interest in molecular biology, such as proteins, genes and cells. GENIA is an annotated data set, in which technical term annotation covers the identification of physical biological entities as well as other important terms. This is our *gold standard data set*.

Whereas the Medline indexes a broad range of academic articles covering the general or specific domains of life sciences, GENIA is intended to cover a smaller subject domain: biological reactions concerning transcription factors in human blood cells.

6.1.2 Protocol

As the measures described in step 2 of our workflow (i.e. *Ranking the Candidate Terms*) are not very time-consuming, and as they are easily applicable for large corpora, they were evaluated over the LabTestsOnline data set for English, French, and Spanish, and over the gold standard data set, GENIA. In contrast, as the measures described in step 3 (i.e. *Re-ranking*) are highly time-consuming, and they are used at the end of the process, to enhance the performance of the results, we evaluate them only over the GENIA data set.

6.1.3 Validation

In order to automatically validate and cover medical terms, we use UMLS for English and Spanish, and the French version of MeSH, SNOMED International and the rest of the French content in the UMLS. For instance, if an extracted candidate term is found in the UMLS dictionary, this term will be automatically validated. The results are evaluated in terms of *precision* obtained over the top k extracted terms ($P@k$). The upper limit of k is 20 000, which allows to compensate the absence of recall value.

[3] http://www.nactem.ac.uk/genia/genia-corpus/term-corpus

Biomedical terminologies or ontologies (e.g. UMLS, SNOMED, MeSH), contain terms composed of signs. Therefore, we cleaned these terminologies by eliminating all terms containing (; , ? ! : { } [ ]), and we only took terms without signs. Table 6.2 shows the distribution in n-gram (i.e. n-gram is a term of n words, with $n \geq 1$) of biomedical resources for three languages, as well as the number of terms that we took after the cleaning task. For instance, the first cell means that 13.73% of terms are composed of one word (1-gram) in UMLS for English.

	1-gram	**2-gram**	**3-gram**	**4+ gram**	**Number of Terms**
English	13.73 %	27.65 %	14.44 %	44.18 %	3 006 946
French	13.17 %	25.82 %	17.08 %	43.93 %	304 644
Spanish	8.39 %	19.31 %	16.33 %	55.97 %	534 110

Table 6.2: Details of Available Resources for Validation.

6.2 Multilingual Comparison (LabTestsOnline)

In this section, we show results obtained only with all the ranking measures, i.e. step 2 (ranking) in Figure 5.1. In addition, we tested the measures for single- plus multi-word terms, or just for multi-word terms in English, French and Spanish. Table 6.3, 6.4, 6.5 show the results in English, French and Spanish, respectively. At the top of each table, the single-word + multi-word term extraction results are presented, while the multi-word term extraction results are presented at the bottom of the table.

These tables show that *LIDF-value* and *L-value* obtain the best results for both extraction cases and for the three languages. The combined measures based on the harmonic mean, and on the *SUM* and *MAX* (i.e. *F-TFIDF-*C_M, *F-TFIDF-*C_S), also give interesting results.

The single-word + multi-word term extraction results are better than just the multi-word term extraction results. The main reason for this is that the extraction of single-word terms is more efficient due to their syntactic structure (linguistic structure), i.e. usually a *noun*. In addition, this syntactic structure has fewer variations. The results are lower as compared to multi-word term extraction, which is more complicated and involves more variations.

We observe that *LIDF-value* and *L-value* obtain very close results. In most cases *LIDF-value* performs better than *L-value*. These two measures show that the probability associated with the linguistic patterns helps to improve the term extraction results. Note that the *idf* influences *LIDF-value*, for this reason *LIDF-value* has better results than *L-value*.

	Single- and Multi- Word Terms													
	P@100	P@200	P@300	P@400	P@500	P@600	P@700	P@800	P@900	P@1000	P@2000	P@5000	P@10000	P@20000
C-value	0.930	0.935	0.927	0.943	0.938	0.930	0.920	0.916	0.904	0.892	0.802	0.629	0.480	0.318
TF-IDF$_A$	0.7	0.715	0.697	0.663	0.636	0.637	0.616	0.603	0.6	0.588	0.515	0.421	0.350	0.322
TF-IDF$_M$	0.910	0.920	0.917	0.898	0.868	0.843	0.824	0.811	0.794	0.781	0.688	0.542	0.448	0.358
TF-IDF$_S$	0.970	0.955	0.960	0.960	0.960	0.950	0.943	0.936	0.917	0.906	0.822	0.659	0.511	0.370
Okapi$_A$	0.570	0.390	0.4	0.378	0.366	0.347	0.341	0.329	0.336	0.335	0.295	0.314	0.310	0.326
Okapi$_M$	0.910	0.915	0.917	0.878	0.824	0.793	0.711	0.685	0.677	0.692	0.613	0.513	0.438	0.355
Okapi$_S$	0.940	0.945	0.927	0.930	0.926	0.920	0.909	0.9	0.882	0.873	0.810	0.656	0.513	0.363
F-OCapi$_A$	0.690	0.645	0.603	0.570	0.546	0.545	0.527	0.519	0.522	0.527	0.509	0.462	0.414	0.333
F-OCapi$_M$	0.970	0.955	0.920	0.895	0.9	0.888	0.874	0.859	0.849	0.842	0.757	0.615	0.465	0.340
F-OCapi$_S$	0.940	0.945	0.930	0.935	0.922	0.923	0.917	0.896	0.887	0.879	0.807	0.653	0.515	0.345
F-TFIDF-C$_A$	0.710	0.715	0.7	0.670	0.642	0.638	0.626	0.613	0.596	0.596	0.532	0.421	0.346	0.323
F-TFIDF-C$_M$	0.970	0.960	0.917	0.898	0.866	0.845	0.826	0.811	0.8	0.787	0.692	0.545	0.449	0.357
F-TFIDF-C$_S$	0.970	0.960	0.960	0.960	0.960	0.948	0.943	0.931	0.914	0.906	0.823	0.659	0.510	0.369
L-value	0.960	0.975	**0.970**	**0.965**	0.960	**0.955**	**0.951**	**0.943**	**0.934**	**0.933**	**0.849**	0.707	**0.597**	**0.431**
LIDF-value	**1.000**	**0.980**	**0.970**	**0.965**	**0.962**	**0.955**	0.950	**0.943**	**0.934**	0.925	**0.849**	**0.716**	**0.597**	**0.431**
	Multi-Word Terms													
	P@100	P@200	P@300	P@400	P@500	P@600	P@700	P@800	P@900	P@1000	P@2000	P@5000	P@10000	P@20000
C-value	0.810	0.790	0.757	0.715	0.686	0.668	0.646	0.633	0.623	0.621	0.527	0.395	0.284	0.189
TF-IDF$_A$	0.390	0.4	0.393	0.405	0.424	0.415	0.406	0.401	0.393	0.384	0.315	0.265	0.230	0.204
TF-IDF$_M$	0.570	0.6	0.603	0.585	0.578	0.555	0.541	0.526	0.528	0.535	0.456	0.336	0.261	0.209
TF-IDF$_S$	0.820	0.765	0.760	0.723	0.700	0.692	0.671	0.639	0.628	0.608	0.526	0.387	0.288	0.211
Okapi$_A$	0.4	0.410	0.397	0.373	0.344	0.350	0.343	0.330	0.309	0.292	0.269	0.222	0.214	0.204
Okapi$_M$	0.550	0.580	0.580	0.565	0.544	0.545	0.531	0.511	0.510	0.485	0.396	0.325	0.264	0.208
Okapi$_S$	0.740	0.680	0.663	0.648	0.644	0.627	0.623	0.623	0.606	0.592	0.497	0.394	0.287	0.209
F-OCapi$_A$	0.440	0.435	0.443	0.420	0.422	0.423	0.413	0.399	0.396	0.393	0.338	0.3	0.251	0.206
F-OCapi$_M$	0.810	0.720	0.627	0.630	0.606	0.573	0.570	0.571	0.562	0.549	0.495	0.372	0.272	0.208
F-OCapi$_S$	0.730	0.705	0.670	0.663	0.646	0.633	0.631	0.624	0.626	0.609	0.503	0.397	0.285	0.208
F-TFIDF-C$_A$	0.460	0.410	0.433	0.413	0.430	0.433	0.416	0.406	0.4	0.386	0.331	0.276	0.233	0.204
F-TFIDF-C$_M$	0.820	0.735	0.630	0.608	0.590	0.565	0.561	0.539	0.529	0.535	0.462	0.349	0.262	0.209
F-TFIDF-C$_S$	0.820	0.760	**0.763**	0.723	0.7	**0.692**	0.673	0.651	0.641	0.616	0.525	0.390	0.288	0.211
L-value	**0.860**	0.760	0.760	0.725	0.704	**0.692**	**0.687**	0.651	**0.654**	0.625	0.536	**0.428**	0.326	0.235
LIDF-value	**0.860**	**0.795**	0.757	**0.743**	**0.718**	0.682	0.667	**0.653**	0.637	**0.626**	**0.537**	**0.428**	**0.327**	**0.235**

Table 6.3: Biomedical Term Extraction for English

	Single- and Multi- Word Terms													
	P@100	P@200	P@300	P@400	P@500	P@600	P@700	P@800	P@900	P@1000	P@2000	P@5000	P@10000	P@20000
C-value	0.560	0.610	0.607	0.605	0.594	0.595	0.589	0.584	0.567	0.565	0.469	0.302	0.198	0.121
TF-IDF$_A$	0.630	0.575	0.550	0.525	0.486	0.430	0.413	0.395	0.394	0.388	0.291	0.199	0.163	0.145
TF-IDF$_M$	0.8	0.745	0.703	0.648	0.626	0.603	0.583	0.566	0.540	0.507	0.419	0.260	0.195	0.156
TF-IDF$_S$	0.810	0.780	0.723	0.698	0.662	0.650	0.637	0.625	0.613	0.606	0.510	0.334	0.226	0.161
Okapi$_A$	0.580	0.415	0.383	0.315	0.270	0.242	0.229	0.205	0.207	0.220	0.190	0.145	0.146	0.150
Okapi$_M$	0.8	0.740	0.683	0.645	0.542	0.532	0.527	0.479	0.432	0.399	0.344	0.256	0.198	0.156
Okapi$_S$	0.530	0.455	0.523	0.530	0.558	0.547	0.564	0.574	0.564	0.566	0.5	0.338	0.230	0.159
F-OCapi$_A$	0.6	0.525	0.457	0.418	0.386	0.345	0.324	0.308	0.296	0.272	0.214	0.158	0.155	0.149
F-OCapi$_M$	**0.880**	0.735	0.703	0.668	0.654	0.618	0.593	0.574	0.559	0.532	0.408	0.274	0.193	0.153
F-OCapi$_S$	0.520	0.470	0.527	0.550	0.550	0.553	0.573	0.568	0.578	0.566	0.5	0.342	0.217	0.153
F-TFIDF-C$_A$	0.640	0.575	0.557	0.528	0.486	0.453	0.417	0.404	0.396	0.388	0.298	0.199	0.160	0.144
F-TFIDF-C$_M$	**0.880**	0.750	0.703	0.650	0.628	0.603	0.584	0.573	0.546	0.522	0.420	0.261	0.196	0.156
F-TFIDF-C$_S$	0.820	0.775	0.720	0.693	0.666	0.647	0.639	0.619	0.613	0.603	0.510	0.334	0.224	0.160
L-value	0.630	0.650	0.643	0.650	0.654	0.640	0.643	0.646	0.634	**0.628**	0.543	**0.409**	**0.320**	**0.187**
LIDF-value	0.860	**0.780**	**0.733**	**0.705**	**0.680**	**0.670**	**0.654**	**0.651**	**0.643**	**0.628**	**0.550**	**0.409**	**0.320**	**0.187**

	Multi-Word Terms													
	P@100	P@200	P@300	P@400	P@500	P@600	P@700	P@800	P@900	P@1000	P@2000	P@5000	P@10000	P@20000
C-value	0.450	0.470	0.460	0.425	0.398	0.377	0.359	0.353	0.338	0.315	0.233	0.168	0.091	0.062
TF-IDF$_A$	0.330	0.280	0.240	0.245	0.250	0.235	0.209	0.205	0.2	0.195	0.133	0.103	0.087	0.074
TF-IDF$_M$	0.460	0.430	0.4	0.340	0.328	0.333	0.314	0.310	0.282	0.258	0.184	0.120	0.098	0.076
TF-IDF$_S$	0.610	0.495	0.480	0.438	0.410	0.397	0.386	0.358	0.342	0.333	0.240	0.150	0.106	0.076
Okapi$_A$	0.320	0.270	0.230	0.203	0.196	0.173	0.174	0.165	0.152	0.145	0.120	0.095	0.082	0.074
Okapi$_M$	0.430	0.420	0.383	0.330	0.328	0.312	0.304	0.275	0.248	0.242	0.168	0.120	0.095	0.075
Okapi$_S$	0.420	0.435	0.437	0.405	0.392	0.370	0.350	0.346	0.334	0.326	0.248	0.150	0.103	0.075
F-OCapi$_A$	0.330	0.290	0.250	0.258	0.252	0.250	0.229	0.213	0.201	0.196	0.137	0.101	0.085	0.074
F-OCapi$_M$	0.560	0.410	0.403	0.395	0.356	0.337	0.326	0.309	0.294	0.281	0.2	0.127	0.096	0.074
F-OCapi$_S$	0.420	0.445	0.440	0.415	0.390	0.380	0.364	0.344	0.342	0.334	0.251	0.149	0.103	0.074
F-TFIDF-C$_A$	0.330	0.295	0.257	0.255	0.248	0.250	0.233	0.221	0.216	0.209	0.136	0.102	0.087	0.074
F-TFIDF-C$_M$	0.540	0.425	0.403	0.353	0.328	0.342	0.317	0.313	0.293	0.278	0.184	0.123	0.096	0.076
F-TFIDF-C$_S$	0.610	0.475	0.483	0.445	0.422	0.393	0.387	0.368	0.350	0.330	0.242	0.151	0.106	0.076
L-value	0.620	0.620	0.557	**0.515**	**0.480**	**0.460**	0.442	0.425	0.407	**0.401**	0.314	0.211	**0.138**	**0.083**
LIDF-value	**0.660**	**0.640**	**0.563**	**0.515**	**0.480**	**0.460**	**0.443**	**0.429**	**0.413**	0.396	**0.315**	**0.212**	**0.138**	**0.083**

Table 6.4: Biomedical Term Extraction for French

	Single- and Multi- Word Terms													
	P@100	P@200	P@300	P@400	P@500	P@600	P@700	P@800	P@900	P@1000	P@2000	P@5000	P@10000	P@20000
C-value	0.630	0.650	0.657	0.625	0.618	0.620	0.609	0.598	0.581	0.570	0.463	0.315	0.216	0.140
TF-IDF$_A$	0.340	0.3	0.3	0.325	0.328	0.330	0.323	0.299	0.288	0.283	0.235	0.183	0.151	0.131
TF-IDF$_M$	0.740	0.690	0.633	0.575	0.538	0.498	0.496	0.493	0.463	0.462	0.371	0.274	0.208	0.155
TF-IDF$_S$	0.810	0.735	**0.740**	0.718	**0.706**	0.675	0.651	0.633	0.621	0.599	0.491	0.337	0.239	0.165
Okapi$_A$	0.210	0.270	0.210	0.203	0.196	0.182	0.177	0.173	0.171	0.169	0.140	0.116	0.115	0.123
Okapi$_M$	0.580	0.6	0.540	0.548	0.530	0.493	0.436	0.429	0.413	0.411	0.326	0.248	0.195	0.150
Okapi$_S$	0.560	0.570	0.597	0.615	0.6	0.595	0.586	0.583	0.580	0.580	0.5	0.346	0.238	0.161
F-OCapi$_A$	0.250	0.275	0.227	0.245	0.248	0.252	0.249	0.234	0.228	0.223	0.158	0.124	0.122	0.131
F-OCapi$_M$	0.810	0.695	0.587	0.548	0.528	0.522	0.471	0.449	0.439	0.448	0.414	0.275	0.199	0.149
F-OCapi$_S$	0.560	0.570	0.613	0.615	0.602	0.595	0.586	0.586	0.578	0.576	0.5	0.343	0.231	0.158
F-TFIDF-C$_A$	0.350	0.330	0.303	0.333	0.328	0.330	0.321	0.301	0.288	0.284	0.235	0.186	0.150	0.131
F-TFIDF-C$_M$	**0.820**	0.705	0.640	0.583	0.552	0.497	0.497	0.494	0.473	0.467	0.375	0.274	0.210	0.155
F-TFIDF-C$_S$	0.810	0.745	**0.740**	**0.720**	0.702	0.675	0.660	0.630	0.612	0.602	0.491	0.338	0.238	0.165
L-value	0.660	0.660	0.623	0.630	0.610	0.608	0.597	0.590	0.573	0.557	0.467	0.339	0.250	0.176
LIDF-value	0.810	**0.755**	0.730	0.710	0.696	**0.682**	**0.677**	**0.663**	**0.653**	**0.645**	**0.512**	**0.436**	**0.324**	**0.248**

	Multi-Word Terms													
	P@100	P@200	P@300	P@400	P@500	P@600	P@700	P@800	P@900	P@1000	P@2000	P@5000	P@10000	P@20000
C-value	0.420	0.435	0.417	0.378	0.368	0.352	0.340	0.321	0.306	0.294	0.225	0.157	0.106	0.068
TF-IDF$_A$	0.150	0.160	0.173	0.140	0.128	0.147	0.151	0.156	0.143	0.140	0.119	0.098	0.081	0.068
TF-IDF$_M$	0.350	0.350	0.343	0.290	0.272	0.242	0.217	0.226	0.221	0.216	0.175	0.132	0.101	0.075
TF-IDF$_S$	0.570	0.470	0.430	0.405	0.380	0.362	0.340	0.333	0.323	0.304	0.228	0.156	0.110	0.080
Okapi$_A$	0.110	0.135	0.120	0.123	0.116	0.125	0.131	0.134	0.123	0.115	0.094	0.080	0.076	0.069
Okapi$_M$	0.310	0.280	0.317	0.288	0.246	0.230	0.214	0.208	0.209	0.205	0.158	0.120	0.096	0.077
Okapi$_S$	0.420	0.415	0.393	0.393	0.366	0.342	0.323	0.328	0.323	0.305	0.238	0.153	0.107	0.080
F-OCapi$_A$	0.110	0.150	0.130	0.128	0.132	0.138	0.134	0.136	0.147	0.144	0.104	0.085	0.079	0.070
F-OCapi$_M$	0.460	0.325	0.333	0.295	0.260	0.240	0.223	0.223	0.218	0.220	0.193	0.122	0.098	0.077
F-OCapi$_S$	0.410	0.410	0.403	0.393	0.372	0.340	0.331	0.328	0.322	0.315	0.239	0.153	0.108	0.079
F-TFIDF-C$_A$	0.160	0.170	0.177	0.148	0.134	0.148	0.157	0.159	0.157	0.152	0.128	0.099	0.082	0.068
F-TFIDF-C$_M$	0.480	0.375	0.347	0.298	0.280	0.245	0.221	0.228	0.223	0.223	0.188	0.133	0.1	0.075
F-TFIDF-C$_S$	**0.570**	0.455	0.430	0.408	0.392	0.367	0.344	0.335	0.327	0.314	0.230	0.157	0.111	0.080
L-value	0.470	0.490	**0.460**	**0.438**	0.410	0.387	**0.370**	0.359	**0.349**	**0.337**	0.266	**0.189**	**0.144**	**0.090**
LIDF-value	0.530	**0.510**	**0.460**	**0.438**	**0.418**	**0.392**	**0.370**	**0.368**	**0.349**	**0.337**	**0.274**	**0.189**	**0.144**	**0.090**

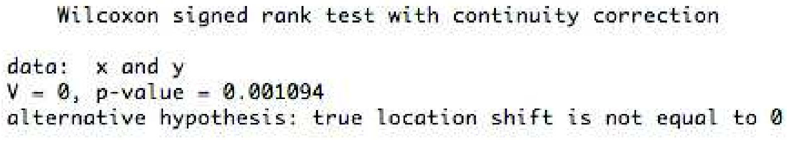
Table 6.5: Biomedical Term Extraction for Spanish

As we can see, the *TF-IDF$_S$* measure obtains high precision values approaching to those obtained by using *LIDF-value*. This behavior is repeated for the three languages, i.e. English, French, and Spanish, and for the extraction of "multi-word terms" and "multi- and single-word terms". In several cases the difference between these two values is small. Therefore, it would be interesting to find out if a a complex measure as *LIDF-value* and a simpler measure as *TF-IDF$_S$* do not have a statistically significant difference in their results.

In statistics, statistical significance is attained when a *p-value* is less than the significance level, in general cases is *p-value* < 0.05. For this, we use the "Wilcoxon signed rank", which is a non-parametric statistical hypothesis test used when comparing two related samples. Non-parametric statistics are not based on probability distributions, which is ideal to measure the obtained precision with the measures before mentioned. We considered *p-value* < 0.05 as statistically significant. All analyzes were performed with the statistical software "R". The following is the R code used in this evaluation:

```
x <- c(0.970, 0.955, 0.960, 0.960, 0.960, 0.950, 0.943, 0.936,
0.917, 0.906, 0.822, 0.659, 0.511, 0.370)
y <- c(1.000, 0.980, 0.970, 0.965, 0.962, 0.955, 0.950, 0.943,
0.934, 0.925, 0.849, 0.716, 0.597, 0.431)
wilcox.test(x, y, paired = TRUE, alternative = "two.sided")
```

We can see the results of the previous code in Figure 6.1.



```
        Wilcoxon signed rank test with continuity correction

data:  x and y
V = 0, p-value = 0.001094
alternative hypothesis: true location shift is not equal to 0
```

Figure 6.1: resultats with R

Table 6.6 shows the results of *p-value* obtained by using the previous code in "R". In this table, we can see that in all the cases *p-value* < 0.05. Therefore, there exists a statistical significance between the *LIDF* and *TF-IDF$_S$* measures.

	Single- and Multi- Word Terms	**Multi-Word Terms**
English	0.0010	0.0047
French	0.0016	0.0010
Spanish	0.0118	0.0120

Table 6.6: *p-value* between *LIDF* and *TF-IDF$_S$*.

6.3 Evaluation of the global process (GENIA)

Since GENIA is the gold standard data set, we conduct a detailed assessment of the experiments in this section. We evaluated the entire workflow of our methodology, i.e. steps 2 (ranking) and 3 (re-ranking) in Figure 5.1.

As noted earlier, the multi-word term extraction results are influenced by the syntactic structure and their variations. So, our experimentation in this section is focused only on multi-word term extraction.

In the following paragraphs, we also narrow down the presented results by keeping only the first 8 000 extracted terms for the graph-based measure and the first 1000 extracted terms for the web-based measure.

6.3.1 Ranking Results (step 2 in Figure 5.1)

Table 6.7 presents and compares the multi-word term extraction results with the best ranking measures, as shown earlier, i.e. *C-value*, *F-TFIDF-*C_S, and *LIDF-value*. The best results were obtained with *LIDF-value* with an 11% improvement in precision for the first hundred extracted multi-word terms.

These precision results are also shown in Figure 6.2. The precision of *LIDF-value* will be further improved with *TeRGraph*.

	C-value	*F-TFIDF-*C_S	***LIDF-value***
P@100	0.690	0.715	**0.820**
P@200	0.690	0.715	**0.770**
P@300	0.697	0.710	**0.750**
P@400	0.665	0.690	**0.738**
P@500	0.642	0.678	**0.718**
P@600	0.638	0.668	**0.723**
P@700	0.627	0.669	**0.717**
P@800	0.611	0.650	**0.710**
P@900	0.612	0.629	**0.714**
P@1000	0.605	0.618	**0.697**
P@2000	0.570	0.557	**0.662**
P@5000	0.498	0.482	**0.575**
P@10000	0.428	0.412	**0.526**
P@20000	0.353	0.314	**0.377**

Table 6.7: Precision comparison of *LIDF-value* with baseline measures

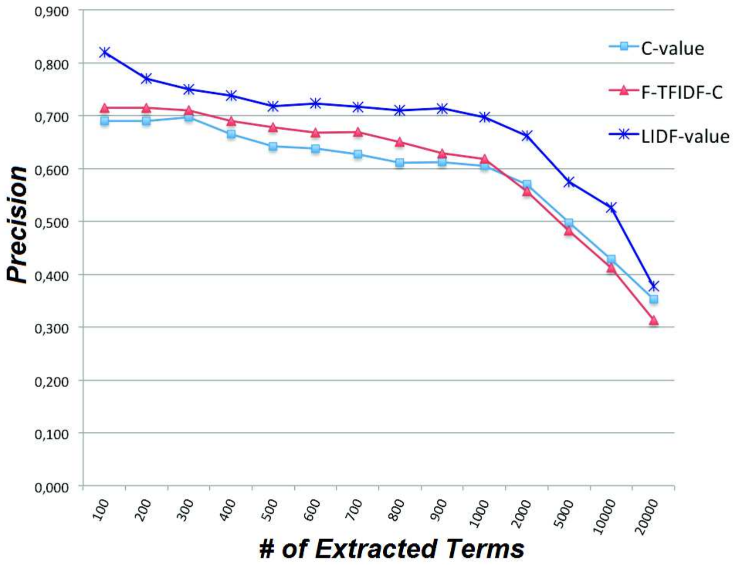


Figure 6.2: Precision comparison with *LIDF-value* and baseline measures

Results of n-gram Terms

We also evaluated *C-value*, *F-TFIDF-*C_S, and *LIDF-value* in a sequence of n-gram terms (i.e. n-gram term is a multi-word term of n words), for this we require an index term to be a n-gram terms of length $n \geq 2$. We tested the performance of *LIDF-value* on the n-gram term extraction taking the first 1 000 n-gram terms ($n \geq 2$).

Table 6.8 shows the precision comparison for the 2-gram, 3-gram and 4+ gram term extracted with *C-value*, *F-TFIDF-*C_S, and *LIDF-value*. We can see that *LIDF-value* obtains the best results for all intervals for any $n \geq 2$. These precision results are also shown in Figure 6.3 for the 2-gram terms, Figure 6.4 for the 3-gram terms, and finally Figure 6.5 for the 4+ gram terms.

Table 6.9 shows the top-20 ranked 2-gram terms extracted with the baseline measures and *LIDF-value*. *C-value* obtained 3 irrelevant terms, *F-TFIDF-C* obtained 5 irrelevant terms while *LIDF-value* obtained only 2 irrelevant terms for the top-20 ranked 2-gram terms.

Similarly, Table 6.10 shows top-10 ranked 3-gram terms extracted with the baseline measures and *LIDF-value*. Finally, Table 6.11 shows the top-10 ranked 4+ gram terms extracted with the baseline measures and *LIDF-value*.

Note that in this context, “irrelevant” means that the terms are not in the above mentioned resources. These candidate terms might be interesting for ontology extension or population, however they must pass through polysemy detection in order to identify the possible meanings.

	2-gram terms		
	C-value	*F-TFIDF-C*	***LIDF-value***
P@100	0.770	0.760	**0.830**
P@200	0.755	0.755	**0.805**
P@300	0.710	0.743	**0.790**
P@400	0.695	0.725	**0.768**
P@500	0.692	0.736	**0.752**
P@600	0.683	0.733	**0.763**
P@700	0.670	0.714	**0.757**
P@800	0.669	0.703	**0.749**
P@900	0.654	0.692	**0.749**
P@1000	0.648	0.684	**0.743**
	3-gram terms		
	C-value	*F-TFIDF-C*	***LIDF-value***
P@100	0.670	0.530	**0.820**
P@200	0.590	0.450	**0.795**
P@300	0.577	0.430	**0.777**
P@400	0.560	0.425	**0.755**
P@500	0.548	0.398	**0.744**
P@600	0.520	0.378	**0.720**
P@700	0.499	0.370	**0.706**
P@800	0.488	0.379	**0.691**
P@900	0.482	0.399	**0.667**
P@1000	0.475	0.401	**0.660**
	4+ gram terms		
	C-value	*F-TFIDF-C*	***LIDF-value***
P@100	0.510	0.370	**0.640**
P@200	0.455	0.330	**0.520**
P@300	0.387	0.273	**0.477**
P@400	0.393	0.270	**0.463**
P@500	0.378	0.266	**0.418**
P@600	0.348	0.253	**0.419**
P@700	0.346	0.249	**0.390**
P@800	0.323	0.248	**0.395**
P@900	0.323	0.240	**0.364**
P@1000	0.312	0.232	**0.354**

Table 6.8: Precision comparison of 2-gram terms, 3-gram terms, and 4+ gram terms

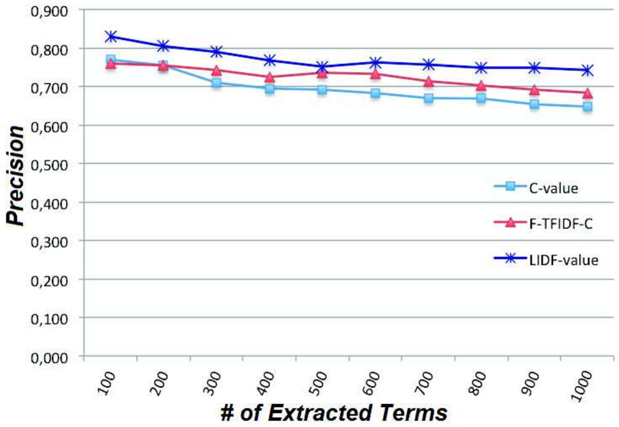


Figure 6.3: Precision comparison of 2-gram terms

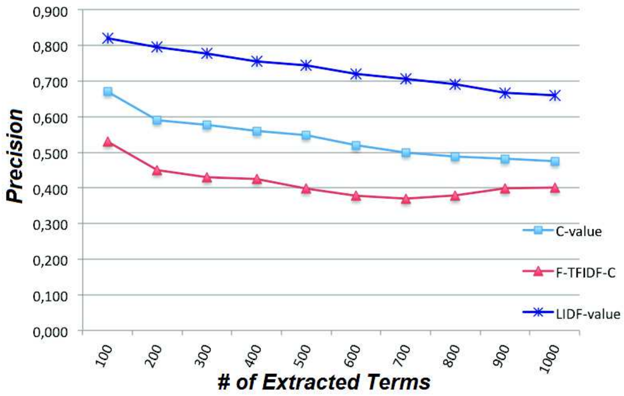


Figure 6.4: Precision comparison of 3-gram terms

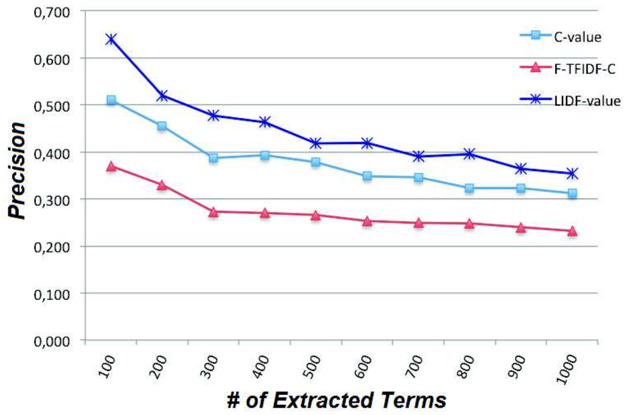


Figure 6.5: Precision comparison of 4+ gram terms

	C-value	_F-TFIDF-C_	**_LIDF-value_**
1	t cell	t cell	t cell
2	nf-kappa b	nf-kappa b	transcription factor
3	transcription factor	kappa b	nf-kappa b
4	gene expression	b cell	cell line
5	kappa b	class ii	b cell
6	cell line	glucocorticoid receptor	gene expression
7	b cell	_b activation_ *	kappa b
8	peripheral blood	_b alpha_ *	t lymphocyte
9	t lymphocyte	reporter gene	dna binding
10	nuclear factor	endothelial cell	_i kappa_ *
11	protein kinase	cell cycle	binding site
12	class ii	b lymphocyte	protein kinase
13	_b activation_ *	_nf kappa_ *	glucocorticoid receptor
14	human t	nf-kappab activation	tumor necrosis
15	tyrosine phosphorylation	u937 cell	binding activity
16	dna binding	_mhc class_ *	tyrosine phosphorylation
17	_human immunodeficiency_ *	_c ebp_*	_shift assay_ *
18	binding site	il-2 promoter	immunodeficiency virus
19	_necrosis factor_ *	monocytic cell	signal transduction
20	mobility shift	t-cell leukemia	mobility shift

Table 6.9: Comparison of top-20 ranked 2-gram terms (irrelevant terms are italicized and marked with *).

	C-value	_F-TFIDF-C_
1	human immunodeficiency virus	_kappa b alpha_*
2	_kappa b alpha_*	nf kappa b
3	tumor necrosis factor	jurkat t cell
4	electrophoretic mobility shift	human t cell
5	nf-kappa b activation	mhc class ii
6	_virus type 1_*	cd4+ t cell
7	protein kinase c	_c-fos and c-jun_*
8	long terminal repeat	peripheral blood monocyte
9	nf kappa b	t cell proliferation
10	jurkat t cell	_transcription factor nf-kappa_*

	LIDF-value
1	i kappa b
2	human immunodeficiency virus
3	electrophoretic mobility shift
4	human t cell
5	mobility shift assay
6	_kappa b alpha_*
7	tumor necrosis factor
8	nf-kappa b activation
9	protein kinase c
10	jurkat t cell

Table 6.10: Comparison of the top-10 ranked 3-gram terms (irrelevant terms are italicized and marked with *).

	C-value	*F-TFIDF-C*
1	human immunodeficiency virus type 1	transcription factor nf-kappa b
2	*human immunodeficiency virus type* *	*expression of nf-kappa b* *
3	*immunodeficiency virus type 1* *	tumor necrosis factor alpha
4	activation of nf-kappa b	normal human t cell
5	nuclear factor kappa b	primary human t cell
6	tumor necrosis factor alpha	germline c epsilon transcription
7	*human t-cell leukemia viru* *	gm-csf receptor alpha promoter
8	*human t-cell leukemia virus type* *	il-2 receptor alpha chain
9	*t-cell leukemia virus type* *	*transcription from the gm-csf* *
10	electrophoretic mobility shift assay	*translocation of nf-kappa b* *

	LIDF-value
1	i kappa b alpha
2	electrophoretic mobility shift assay
3	*human immunodeficiency virus type* *
4	human t-cell leukemia virus
5	nuclear factor kappa b
6	tumor necrosis factor alpha
7	*t-cell leukemia virus type* *
8	activation of nf-kappa b
9	peripheral blood t cell
10	major histocompatibility complex class

Table 6.11: Comparison of the top-10 ranked 4+ gram terms (irrelevant terms are italicized and marked with *).

6.3.2 Re-ranking Results (step 3 in Figure 5.1)

In this section, we will evaluate the graph-based and the web)based re-ranking measures.

6.3.2.1 Graph-based Results

Our graph-based approach is applied to the first 8 000 terms extracted by the best ranking measure. The objective is to re-rank the 8 000 terms while trying to improve the precision by intervals.

One parameter is involved in the computation of graph-based term weights, i.e. the *threshold* of Dice value which represents the relation when building the term graph. This involves linking terms whose *Dice value* of the relation is higher than *threshold*. We vary *threshold* (δ) within $\delta = [0.25, 0.35, 0.50, 0.60, 0.70]$ and report the precision performance for each of these values.

Table 6.12 gives the precision performance obtained by *TeRGraph* and shows that it is well adapted for ATE.

	TeRGraph				
	$\delta \geq 0.25$	$\delta \geq 0.35$	$\delta \geq 0.50$	$\delta \geq 0.60$	$\delta \geq 0.70$
P@100	0.840	0.860	0.910	**0.930**	0.900
P@200	0.800	0.790	0.850	**0.855**	**0.855**
P@300	0.803	0.773	**0.833**	0.830	0.820
P@400	0.780	0.732	**0.820**	**0.820**	0.815
P@500	0.774	0.712	0.798	**0.810**	0.806
P@600	0.773	0.675	0.797	**0.807**	0.792
P@700	0.760	0.647	0.769	**0.796**	0.787
P@800	0.756	0.619	0.748	**0.784**	0.779
P@900	0.748	0.584	0.724	0.773	**0.777**
P@1000	0.751	0.578	0.720	0.766	**0.769**
P@2000	0.689	0.476	0.601	0.657	**0.694**
P@3000	0.642	0.522	0.535	0.605	**0.644**
P@4000	**0.612**	0.540	0.543	0.559	0.593
P@5000	**0.574**	0.546	0.544	0.554	0.562
P@6000	0.558	0.539	0.540	0.549	**0.561**
P@7000	**0.556**	0.540	0.540	0.545	0.552
P@8000	**0.546**	**0.546**	**0.546**	**0.546**	**0.546**

Table 6.12: Precision performance of *TeRGraph* when varying δ (*threshold* parameter for Dice)

6.3.2.2 Web-based Results

Our web-based approach is applied at the end of the process, with only the first 1 000 terms extracted during the previous linguistic, statistic and graph measures. For space reasons, we show only the results obtained with *WAHI*, which are higher than *WebR*.

We took the list obtained with *TeRGraph* and $\delta \geq 0.60$. The main reason for this limitation is the limited number of automatic queries possible in search engines. At this step, the aim is to re-rank the 1 000 terms to try to improve the precision by intervals. Each measure listed in Table 6.13 and Table 6.14 shows the precision obtained after re-ranking. We tested *WAHI* with *Yahoo* and *Bing* search engines.

Table 6.13 and Table 6.14 prove that *WAHI* (either using *Yahoo* or *Bing*) is well adapted for ATE and this measure obtains better precision results than the baselines measures for word association. So our measures obtain real terms of our dictionary with a better ranking.

	WAHI	*Dice*	*Jaccard*	*Cosine*	*Overlap*
P@100	**0.960**	0.720	0.720	0.760	0.730
P@200	**0.950**	0.785	0.770	0.740	0.765
P@300	**0.900**	0.783	0.780	0.767	0.753
P@400	**0.900**	0.770	0.765	0.770	0.740
P@500	**0.920**	0.764	0.754	0.762	0.738
P@600	**0.850**	0.748	0.740	0.765	0.748
P@700	**0.817**	0.747	0.744	0.747	0.757
P@800	**0.875**	0.752	0.746	0.740	0.760
P@900	**0.870**	0.749	0.747	0.749	0.747
P@1000	**0.766**	**0.766**	**0.766**	**0.766**	**0.766**

Table 6.13: Precision comparison of *WAHI with YAHOO* and word association measures

	WAHI	*Dice*	*Jaccard*	*Cosine*	*Overlap*
P@100	**0.900**	0.740	0.730	0.680	0.650
P@200	**0.900**	0.775	0.775	0.735	0.705
P@300	**0.900**	0.770	0.763	0.740	0.713
P@400	**0.900**	0.765	0.765	0.752	0.712
P@500	**0.900**	0.760	0.762	0.758	0.726
P@600	**0.917**	0.753	0.752	0.753	0.743
P@700	**0.914**	0.751	0.751	0.733	0.749
P@800	**0.875**	0.745	0.747	0.741	0.754
P@900	**0.878**	0.747	0.748	0.742	0.748
P@1000	**0.766**	**0.766**	**0.766**	**0.766**	**0.766**

Table 6.14: Precision comparison of *WAHI with BING* and word association measures

As we mentioned in Section 4.3, the web is the largest available corpora. We mentioned the benefits by using the web. Such as, the access to all kind of data of a lot of domains. Search engines are used for querying the web. We evaluate the time in seconds of their indexation and retrieving algorithms. Table 6.15 presents the time consumed by Bing and Yahoo search engines. We can see that in general Yahoo takes double time than Bing for retrieving information. For instance, for the first 10 terms (N@10) Bing takes about 15 seconds to retrieve the number of hits. Due to the large number of existing domains in the web, and the time taken to evaluate the terms, we believe it is useful to use Bing.

	Bing	Yahoo
N@10	**15.40**	33.19
N@50	**89.31**	170.31
N@100	**177.54**	334.45
N@200	**360.35**	656.16
N@500	**907.26**	1703.07
N@1000	**1739.99**	3345.48

Table 6.15: Time execution comparison between Bing and Yahoo in seconds.

6.3.3 Summary

LIDF-value obtains the best precision results for multi-word term extraction, for each index term extraction (n-gram) and for intervals.

Table 6.16 presents a precision comparison of *LIDF-value* and *TeRGraph* measures. In terms of overall precision, our experiments produce consistent results from the GENIA data set. In most cases, *TeRGraph* obtains better precision with a δ of 0.60 and 0.70 (i.e. better precision in most $P@k$ intervals), which is very good because it helps alleviate the problem of manual validation of candidate terms. These precisions are also illustrated in Figure 6.6.

The performance of our graph-based measure somewhat depends on the value of the co-occurrence relation between terms. Specifically, the value of the co-occurrence relation affects how the graph is built (whose edges are taken), and hence it is critical for computation of the graph-based term weight. Another performance factor of our graph-based measure is the quality of the results obtained with *LIDF-value* due to the fact that the list of terms extracted with *LIDF-value* is required as input to re-rank *TeRGraph* in order to construct the graph, where nodes denote terms, and edges denote co-occurrence relations.

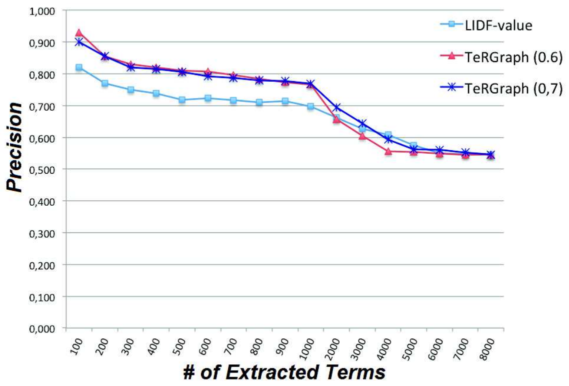


Figure 6.6: Precision comparison of *LIDF-value* and *TeRGraph*

	LIDF-value	*TeRGraph* ($\delta \geq 0.60$)	*TeRGraph* ($\delta \geq 0.70$)
P@100	0.820	**0.930**	0.900
P@200	0.770	**0.855**	**0.855**
P@300	0.750	**0.830**	0.820
P@400	0.738	**0.820**	0.815
P@500	0.718	**0.810**	0.806
P@600	0.723	**0.807**	0.792
P@700	0.717	**0.796**	0.787
P@800	0.710	**0.784**	0.779
P@900	0.714	0.773	**0.777**
P@1000	0.697	0.766	**0.769**
P@2000	0.662	0.657	**0.694**
P@3000	0.627	0.605	**0.644**
P@4000	**0.608**	0.5585	0.593
P@5000	**0.575**	0.5538	0.562
P@6000	0.550	0.549	**0.561**
P@7000	0.547	0.545	**0.552**
P@8000	**0.546**	**0.546**	**0.546**

Table 6.16: Precision comparison of *LIDF-value* and *TeRGraph*

Table 6.17 presents the precision comparison of our three measures.

WAHI based on Yahoo obtains better precision for the first *P*@100 extracted terms with 96% precision whereas, in comparison, *WAHI* based on Bing obtains 90 precision. For the other interval, Table 6.17 shows that *WAHI* based on Bing generally gives the best results. This is very encouraging because it also helps alleviate the problem of manual validation of candidate terms.

The performance of *WAHI* depends on the search engine because algorithms designed for searching information on the web are different, so the number of hits returned will differ in all cases. Another performance factor is the quality of the re-ranked list obtained with *TeRGraph*, because this list is required as input.

Moreover, Table 6.17 highlights that re-ranking with *WAHI* enables us to increase the precision of *TeRGraph*. For all cases, our re-ranking methods improve the precision obtained with *LIDF-value*. The purpose for which this web-mining measure was designed has thus been fulfilled.

Note that these measures do not normalize the possible variants. This could be a limitation for researchers looking for a preferred term for a group of variants.

	LIDF-value	*TeRGraph* ($\delta \geq 0.60$)	*WAHI* *(Bing)*	*WAHI* *(Yahoo)*
P@100	0.820	0.930	0.900	**0.960**
P@200	0.770	0.855	0.900	**0.950**
P@300	0.750	0.830	**0.900**	**0.900**
P@400	0.738	0.820	**0.900**	**0.900**
P@500	0.718	0.810	0.900	**0.920**
P@600	0.723	0.807	**0.917**	0.850
P@700	0.717	0.796	**0.914**	0.817
P@800	0.710	0.784	**0.875**	**0.875**
P@900	0.714	0.773	**0.878**	0.870
P@1000	0.697	**0.766**	**0.766**	**0.766**

Table 6.17: Precision comparison *LIDF-value*, *TeRGraph*, and *WAHI*

In this chapter we have shown the experiments of the proposed approach to extract biomedical candidate term. The approach is based on linguistic, statistic, graph, and web features. This approach obtained the best results in comparison to the baseline measures.

In next chapter, we discuss the effects of some parameters of our approach. As well as, we conclude and present the perspectives of the first part of this thesis.

Chapter

7

Discussion and Conclusions

As we mentioned before, this chapter illustrates the effects of some parameters of our workflow (see Section 7.1) and concludes and presents the perspectives of the automatic biomedical term extraction approach (see Section 7.2).

7.1 Discussion

This section presents the effects of the impacts of biomedical pattern lists, size of dictionaries, and the extraction errors.

7.1.1 Impact of Pattern List

In our methodology, we have shown that biomedical patterns directly affect the term extraction results. For instance, we can see that *L-value*, which is a combination of *C-value* and the probability of pattern lists, gives better results than *C-value* for the three languages, and *LIDF-value* outperforms *L-value* in major cases. These pattern lists work specifically for the biomedical domain. If we use these biomedical patterns in another domain instead of using specific patterns of that domain, they will impact the term extraction results. To prove this, we have extracted terms from an agronomic corpus for English and French while taking biomedical patterns and agronomic patterns into account. We built the agronomic patterns using AGROVOC[1], which is an agronomic dictionary. AGROVOC contains 39 542 and 37 382 English and French terms, respectively. Our corpus consists of titles plus abstracts extracted

[1] `http://aims.fao.org/agrovoc`

from the list of Cirad publications (French Agricultural Research Centre for International Development). Table 7.1 shows the details of the corpus formed for this comparison.

Table 7.2 presents a term extraction comparison while taking patterns built from two different domains into account.
Again we note that *LIDF-value* obtains the best results. We also see that the results of terms extracted with agronomic patterns gives better results than when using biomedical patterns for English and French.

Note that even if the term extraction results obtained using agronomic patterns are higher than using biomedical patterns, these results are a bit close. The main reason is that the biomedical and agronomic terms overlap. It means that identical patterns exist in both domains. The results could be improved by using patterns of two completely different domains.

	Number of Titles/Abstracts	**Number of Words**
English	156	29 740 words
French	84	14 850 words

Table 7.1: Details of Cirad corpus.

7.1.2 Effect of Dictionary Size

Dictionaries play an important role in term extraction, specifically during the construction of pattern lists. Table 7.2 shows that a reduction in dictionary size degrades the performance of the precision results in comparison to Tables 6.3, 6.4, 6.7. For instance, for the agronomic and biomedical domain, Table 7.2 and Table 6.3 show the P@100 of 0.92 and 1.00 respectively, and this difference increases as the number of extracted terms increases (i.e. P@k).

7.1.3 Term Extraction Errors

As explained in Section 5 (step a), the term extraction results are influenced by the Part-of-Speech (PoS) tagging tools, which have different results for different languages. Briefly, the tool *"A"* can perform very well for English, while for French the tool *"B"* gives the best results. For instance, the sentence *"Red blood cells increase with ..."* was tagged with the Stanford tool as *"adjective noun noun **verb** preposition ..."*, whereas the TreeTagger tool tagged it as *"adjective noun noun **noun** preposition ..."*. Therefore, in order to show the generality of our approach, we choose a uniform PoS tool, i.e. TreeTagger, as a trade-off for three languages (English,

	English (Single- and Multi- Word Terms)							
	With Agronomic Patterns				**With Biomedical Patterns**			
	P@100	P@200	P@1000	P@5000	P@100	P@200	P@1000	P@5000
C-value	0.910	0.825	0.631	0.255	0.870	0.790	0.527	0.223
TF-IDF$_S$	0.900	0.830	0.667	0.335	0.810	0.845	0.587	0.284
Okapi$_S$	0.910	0.865	0.680	0.331	0.870	0.845	0.625	0.281
F-OCapi$_M$	0.640	0.600	0.419	0.273	0.660	0.605	0.403	0.252
F-OCapi$_S$	0.900	0.845	0.672	0.304	0.870	0.840	0.612	0.260
F-TFIDF-C$_M$	0.740	0.610	0.412	0.261	0.760	0.610	0.402	0.270
F-TFIDF-C$_S$	0.900	0.835	0.664	0.323	0.810	0.845	0.600	0.272
L-value	0.700	0.660	0.542	0.338	0.840	0.795	**0.688**	**0.320**
LIDF-value	**0.920**	**0.875**	**0.766**	**0.340**	**0.880**	**0.855**	0.682	**0.320**

	French (Single- and Multi- Word Terms)							
	With Agronomic Patterns				**With Biomedical Patterns**			
	P@100	P@200	P@1000	P@5000	P@100	P@200	P@1000	P@5000
C-value	0.400	0.360	0.210	0.086	0.450	0.455	0.223	0.084
TF-IDF$_S$	0.430	0.380	0.248	0.114	0.500	0.450	0.293	0.119
Okapi$_S$	0.390	0.360	0.256	0.115	0.490	0.450	0.300	0.120
F-OCapi$_M$	0.310	0.225	0.154	0.100	0.340	0.245	0.167	0.115
F-OCapi$_S$	0.400	0.355	0.248	0.106	0.480	0.465	0.269	0.115
F-TFIDF-C$_M$	0.350	0.240	0.163	0.099	0.380	0.295	0.170	0.118
F-TFIDF-C$_S$	0.350	0.240	0.163	0.099	0.500	0.475	0.268	0.119
L-value	0.550	0.510	**0.367**	**0.135**	**0.520**	0.480	0.333	**0.130**
LIDF-value	**0.560**	**0.535**	**0.367**	**0.135**	0.510	**0.510**	**0.336**	**0.130**

Table 7.2: Precision comparison of Term Extraction with Agronomic and Biomedical Patterns

French, and Spanish), while understanding that it will penalize the results for the three languages.

7.2 Conclusions and Future Work

This first part of this thesis defines and evaluates several measures for automatic single-word term, multi-word term extraction. These measures are classified as *ranking measures*, and *re-ranking measures*. The measures are based on the linguistic, statistical, graph and web information. We modified some baseline measures (i.e. *C-value*, *TF-IDF*, *Okapi*) and we proposed new measures.

All the ranking measures are linguistic- and statistic-based. The best ranking measure is *LIDF-value*, which overcomes the lack of frequency information with the *linguistic pattern probability* and *idf* values.

We experimentally showed that *LIDF-value* applied in the biomedical domain, over two corpora (i.e. LabTestsOnline, GENIA), outperformed a state-of-the-art baseline for extracting terms (i.e. *C-value*), while obtaining the best precision results in all intervals (i.e. $P@k$). And with three languages the *LIDF-value* trends were similar.

We have shown that multi-word term extraction is more complex than single-word term extraction. We detailed an evaluation over the GENIA corpus for multi-word term extraction. Moreover, in that case, *LIDF-value* improved the automatic term extraction precision in comparison to the most popular term extraction measure.

We also evaluated the re-ranking measures. The first re-ranking measure, *TeRGraph*, is a graph-based measure. It decreases the human effort required to validate candidate terms. The graph-based measure has never been applied for automatic term extraction. *TeRGraph* takes the neighborhood to compute the term representativeness in a specific domain into account.

The other re-ranking measures are web-based. The best one, called *WAHI*, takes the list of terms obtained with *TeRGraph* as input. *WAHI* enables us to further reduce the huge human effort required for validating candidate terms.

Our experimental evaluations revealed that *TeRGraph* had better precision than *LIDF-value* for all intervals. Moreover, our experimental assessments revealed that *WAHI* improved the results given with *TeRGraph* for all intervals.

As a future extension of this work, we intend to use the relation value within *TeRGraph*. We plan to include the use of other graph ranking computations, e.g. PageRank, adapted for automatic term extraction. Moreover, a future work consists of using the web to extract more terms than those extracted.

One prospect could be the creation of a regular expression for the biomedical domain from the linguistic pattern list. We plan to modify our measures in order to normalize the possible variants, looking towards for a preferred term for those variants.

Next chapter presents the BioTex application, which implements the measures defined in the first part of this thesis.

Chapter

8

BioTex

8.1 Introduction

As we said, term extraction is an essential task in domain knowledge acquisition. Although hundreds of terminologies and ontologies exist in the biomedical domain, for other languages there exist a need to enrich currently available terminologies and ontologies. Automatic term extraction and keyword extraction measures are widely used in biomedical text mining applications. In this chapter we present BioTex [Lossio-Ventura et al., 2014b], an application that implements several state-of-the-art measures and those proposed in Section 5.2, for English, French, and Spanish.

In a corpus, there exist different information to be expressed by a community. For this purpose, a specific terminology can be used, which does not exist for several domains. For instance, in the biomedical domain, there is a lack of terminologies. We thus intend to offer to users an opportunity to automatically extract biomedical terms and use them for any natural language, indexing, knowledge extraction, or annotation purpose. Extracted terms can also be used to enrich biomedical ontologies or terminologies by offering new terms or synonyms to attach to existing defined classes. Automatic Term Extraction (ATE) methods are designed to automatically extract relevant terms from a given corpus. Note again that we refer to ATE when extracted terms are not previously defined in existing standard ontologies or terminologies. We refer to "semantic annotation" when an extracted term can match to an existing class (URI) such as in [Jonquet et al., 2009].
Relevant terms are useful to gain further insight into the conceptual structure of a domain. These may be: (i) single-word terms (simple to extract), or (ii) multi-word

terms (hard). In the biomedical domain, there is a substantial difference between existing resources (ontologies) in English, French, and Spanish (see Chapter 3). In general, our project involves two main stages: (i) Biomedical term extraction, and (ii) Concept extraction and semantic linkage, in order to populate ontologies with the extracted terms.

In this chapter, we present BioTex, an application to extract biomedical terms. Given a text corpus, it extracts and ranks biomedical terms according to a selected extraction measure. In addition, BioTex automatically validates terms that already exist in UMLS terminology. We have presented different measures in Section 5.2, and performed comparative assessments in Chapter 6 as well as in other publications [Lossio-Ventura et al., 2014a, Lossio-Ventura et al., 2014d]. In next sections, we present the interfaces supported by BioTex; its use-cases (i.e. projects using the application); as well as, a small comparison of BioTex with other biomedical applications, such as *TerMine, FlexiTerm* (see Section 4.1.6.

8.2 Implementation

BioTex is an application for biomedical terminology extraction which offers several baselines and new measures to rank candidate terms for a given text corpus. BioTex can be used either as: (i) a Web application taking a text file as input, or (ii) as a Java library. When used as a Web application, it produces a file with a maximum of 1200 ranked candidate terms. Used as a Java library, it produces four files with ranked candidate terms found in the corpus, respectively, unigram, bigram, 3-gram and all the 4+ gram terms. BioTex has two main interfaces:

8.2.1 Term extraction and ranking measures

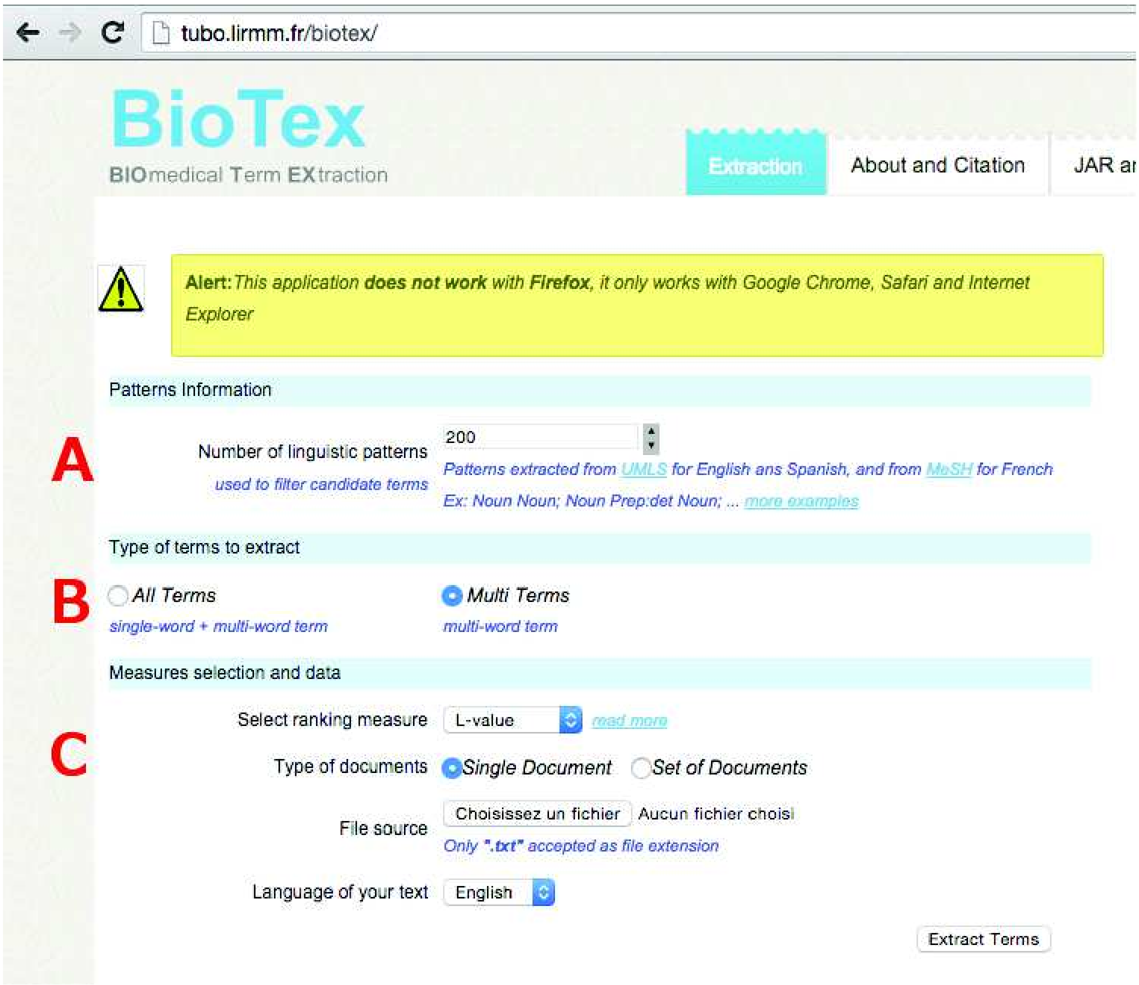
As illustrated by the Web application interface, Figure 8.1, BioTex users can customize the workflow by changing the following parameters:

- (A) Select a number of *patterns* to filter out the candidate terms (200 by default according to our experiments, see Chapter 6). Those reference patterns (e.g., noun-noun, noun-prep-noun, etc.) were built with terms taken from UMLS for English and Spanish, and MeSH for French (see Section 5). They are ranked by their frequency.

- (B) Select the *type of terms* to extract: **all terms** (i.e., single-word and multi-word terms) or **multi-terms**, which means the multi-word terms only. In our work, we also call *n-gram* terms. Where $n \geq 1$ for **all terms**, $n \geq 2$ and **multi-terms**.

- (C) Here, the user has to select 3 parameters:

- Select one of the *ranking measure* to apply. It is necessary to take into account the type of data set (single document or a set of documents). Only the measures, such as *C-value* and *L-value* are allowed to work with both single document and a set of documents. The measures using *idf* only with a set of documents.
- Select your file in *TXT* format.
- Choose the corpus *language* (i.e., English, French, or Spanish).



Alert: *This application* ***does not work*** *with* ***Firefox***, *it only works with Google Chrome, Safari and Internet Explorer*

Patterns Information

Number of linguistic patterns used to filter candidate terms: 200

Patterns extracted from UMLS for English ans Spanish, and from MeSH for French

Ex: Noun Noun; Noun Prep:det Noun; ... more examples

Type of terms to extract

○ *All Terms* — *single-word + multi-word term*

● *Multi Terms* — *multi-word term*

Measures selection and data

Select ranking measure: L-value *read more*

Type of documents: ● *Single Document* ○ *Set of Documents*

File source: Choisissez un fichier Aucun fichier choisi

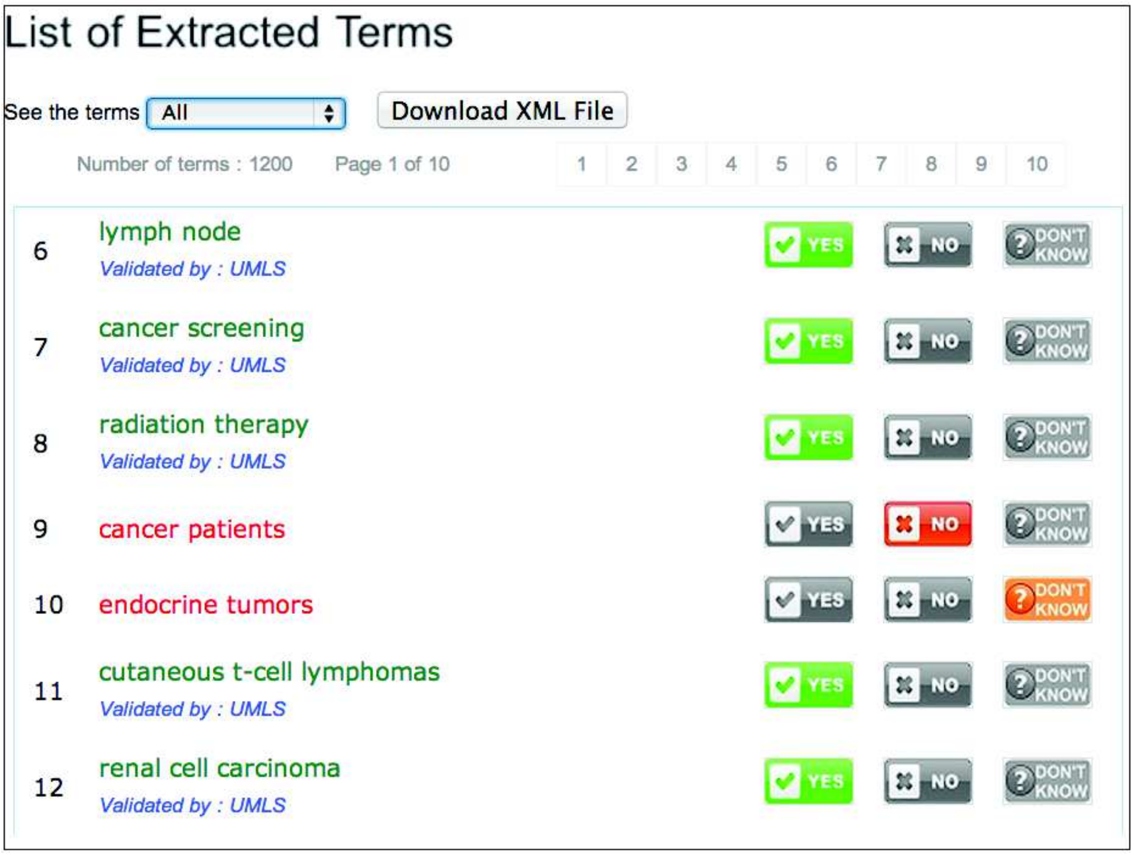
Only ".txt" accepted as file extension

Language of your text: English

Figure 8.1: BioTex Term extraction interface.

8.2.2 Validation of candidate terms

After the extraction process, BioTex automatically validates the extracted terms by using UMLS (English, Spanish) & MeSH-fr, SNOMED and the rest of UMLS (French). As illustrated in Figure 8.2, these validated terms are displayed in green, specifying the used knowledge source and the others in red. Therefore, BioTex allows someone to easily distinguish the classes annotating the original corpus (in green) from the terms that maybe also considered relevant for their data, but need to be curated (in red). The last ones may be considered candidates for ontology enrichment. Users can also manually validate/invalidate the remaining red terms for their own purposes, while potentially enhancing the BioTex validation dictionary.



List of Extracted Terms

See the terms All

Number of terms : 1200 Page 1 of 10

6	lymph node *Validated by : UMLS*
7	cancer screening *Validated by : UMLS*
8	radiation therapy *Validated by : UMLS*
9	cancer patients
10	endocrine tumors
11	cutaneous t-cell lymphomas *Validated by : UMLS*
12	renal cell carcinoma *Validated by : UMLS*

Figure 8.2: BioTex Term validation interface.

8.3 Use-Cases

In this section, we briefly describe the different use-cases of the *BioTex* application. Defined basically for biomedical texts, BioTex has appeared to be efficient for other context as explained in the following sections. In the following paragraphs we describe the biomedical and the general usages.

8.3.1 From the biomedical domain

In this section, we only describe the projects and task where BioTex has been used. In the major cases, our application has been used to build terminologies.

8.3.1.1 Ontologos use-case

Ontologos[1] is a company, which provides innovative products corresponding to customers' needs, facilitating working organization by using relevant software tools and an adequate methodology. In the biomedical domain, Ontologos works with VARAPP[2] association. VARAPP has as objective to promote the improvement of professional practices in the health sector and other related sectors. VARAPP provided us its French health documents through Ontologos. We used these health document to extract the most relevant terms to propose only those not belonging to a biomedical terminology. So, we extracted the first 500 French terms[3] that are not present in a terminology. The next step was the validation by the doctors of VARAPP association. They selected the relevant biomedical terms (i.e. those which can be added to a biomedical terminology) and the irrelevant biomedical terms. Table 8.1 illustrates the results of the validation done by the experts (i.e. doctors). For instance, the first row shows a precision of 74.6%, this means that 373 out of 500 terms have been found as relevant terms. This means, that these 373 could be evaluated to be added to a terminology. The other 25.4 % (127 terms) have been considered as irrelevant in the medical domain.

	Precision
True Biomedial Terms	74.6 %
False Biomedial Terms	25.4 %

Table 8.1: Results of validation in percentage.

8.3.1.2 Psychiatric Ontology use-case

The project Covalmo was developed in partnership between the hospital Sainte-Anne and the Knowledge Engineering Laboratory in e-Health of the University Paris 6. Covalmo [Richard et al., 2015] is a project of Computer Sciences and Medicine. Its main objective is to discover all the possible factors of psychiatric illness in order to help the development of a consensus about the descriptive categories of psychiatric disorders. All this, through tools and methods of Knowledge Engineering [Aimé, 2015]. Hence, Covalmo aims at developing tools for a: (1) better description of the

[1] http://www.ontologos-corp.com/corporate/index.php

[2] http://www.varapp.org/

[3] The proposed terms are available here: http://tubo.lirmm.fr:8080/ontologos/

diagnoses and procedures, as well as a (2) better indexation of patient records. For this project, they used BioTex to extract relevant terms for ONTOPSYCHIA, an ontology for psychiatry. We are not able to show results for confidentiality reasons.

8.3.1.3 Patient Vocabulary use-case

This is a joint project between LIRMM and I3M[4] laboratories, which seeks to create a patient vocabulary. The authors describe the construction of a lexical resource that aligns the patient vocabulary with that of health professionals [Nzali et al., 2015]. The project describes that social media is increasingly used by patients and health professionals. Patients usually use slang words, abbreviations, and a vocabulary of their own during their exchanges. To automatically analyze the texts of social networks, the acquisition of this specific vocabulary is required. This work will allow to improve the information retrieval in health forums, as well as, to facilitate the development of statistical studies based on the information extracted from these forums. To build this vocabulary, the candidate terms have been extracted with BioTex software. Table 8.2 shows a sample of the alignment between biomedical terms and patient terms.

Biomedical Term	Patient Term
oncologue	onco
chimiothérapie	chimio
mammographie	mammo

Table 8.2: Alignment of Biomedical Terms and Patient Terms.

8.3.2 To more general domain

As before mentioned, basically was developed for biomedical domain. Recently, this application was applied to other domains; such as agricultural, epidemiological; showing very good results. The main reason is that BioTex application contains several identical linguistic patterns also used in other domains. In this section, we describe the use-cases of our application in other domains.

8.3.2.1 Epidemiological use-case

This application was applied for a work of the CMAEE laboratory[5]. In recent years methodologies are proposed for epidemiological surveillance on the web. In this context, services need tools that could refine the search and detect the relevant information. In the face of many diseases and even more symptoms, the analysts

[4] Institut de Mathématiques et de Modélisation de Montpellier

[5] `http://umr-cmaee.cirad.fr/`

tackle a difficult challenge: How to identify appropriate queries for Internet disease surveillance? In order to address this issue, the work detailed in [Arsevska et al., 2014] consists of using terms extracted with BioTex in order to propose adapted queries for surveillance tasks. The selection of query terms is based on the sources where the terms are extracted (i.e. Google, PubMed). The authors identified 66 terms extracted to characterize Schmallenberg virus (SBV) emergence.

8.3.2.2 Agricultural use-case

To conduct a study with thematic expertise of our approach, two researchers from CIRAD[6], who are members of TETIS laboratory, were solicited. So, in the context of large amounts of textual data related to agriculture now available, indexing becomes a crucial issue for research organizations. One way to index documents consists of extracting terminology. The authors in [Roche and Fortuno, 2014, Roche et al., 2015] investigates the use and combination of Text Mining methodologies to highlight and publish in Open Data systems the most appropriate terms extracted with BioTex (In French and in English). Moreover these terms are used to match heterogeneous data of agricultural domain.

8.3.2.3 Geographical use-case

This work was carried out during a study by researchers of TETIS laboratory [7]. In recent study, text mining approaches are used to analyze data set in French related to land-use planning. In this use case, the relevant information concerns sentiments, spatial information, and everything else related to the territory (i.e. Land-use Planning). For extracting topics associated with land-use planning, BioTex is applied [Roche and Teissire, 2015]. The authors classified the extracted terms in five categories, which convey information associated with the territory triptych (spatial entities and themes), actors and sentiments.

8.3.2.4 Publicis Groupe

Publicis Groupe[8] is a global leader in marketing, communication, and business transformation, it is present in 108 countries. This company accompanies clients through their business transformation, offering a full range of services and expertise across digital, technology, consulting, creative, corporate communications and public affairs, media strategy, planning and buying, healthcare communications, and brand asset production. They have used BioTex as an experiment for one of their real time marketing projects with Orange[9]. The application was mainly used to find

[6] `http://www.cirad.fr/`

[7] `https://tetis.teledetection.fr/index.php/fr/`

[8] `http://www.publicisgroupe.com/`

[9] `http://www.orange.com/fr/accueil`

insights about communities of games/e-sports specialists. The objective is to understand what kind of problems do they encounter with respect to their internet connections. BioTex has also used for topic labelisation. This work is in process, the qualitatif results should be produced soon. The researchers conclude that the tool was useful for them to extract insights for their client's brands and also for other research purposes.

8.4 Summary

As shown in the detailed study done in Section 4.1.6, most existing systems implementing statistical methods are made to extract keywords and, to a lesser extent, to extract technical terms from a text corpus. Certainly, most systems take a single text document as input, not a set of documents (as corpus), for which the *IDF* can be computed. Most systems are available only in English. Table 8.3 shows a quick comparison of the most important characteristics of typical systems. This comparison has been done between BioTex, *TerMine*[10] (*C-value*), the most commonly used application, and *FlexiTerm*[11], the most recent one. To know more about these applications and the used methods see Section 4.1.6.

	BioTex	***TerMine***	*FlexiTerm*
Languages	en/fr/es	en	en
Type of Application	Desktop/Web	Web	Desktop
License	Open	Open	Open
Processing Capacity	No Limits / < 6 MB	< 2 MB	No Limits
Possibility to save results	XML	-	CSV
POS tool	TreeTagger	Genia/TreeTagger	Stanford POS
# of Implemented Measures	7	1	1

Table 8.3: Brief comparison of biomedical terminology extraction applications.

In general, Table 8.3 shows that BioTex integrates more characteristics than the other systems. BioTex allows to be added as a library for any independent application. It also allows to validate manually the candidate terms that do not belong to a terminology. This application has been done for three languages.

In previous sections, we explained in detail BioTex. Next section concludes and list the most important aspects of our application. As well as, the possible perspectives to be integrated to become more useful BioTex.

[10] http://www.nactem.ac.uk/software/termine/

[11] http://users.cs.cf.ac.uk/I.Spasic/flexiterm/

8.5 Conclusions

In this chapter, we presented the BioTex application for biomedical terminology extraction, which is the implementation of all the ranking measures detailed in Section 5.2. It is available for online testing and evaluation but can also be used in any program as a Java library (POS tagger not included). In contrast to other existing systems, our system allows us to analyze French and Spanish corpus, manually validate extracted terms and, export the list of extracted terms.

BioTex starts to be a valuable tool for the biomedical community, as well as for related domains. It is currently used in other independent projects. In addition, it is integrated in a couple of test-beds within the SIFR project[12]. The application is available at `http://tubo.lirmm.fr/biotex/` along with a video demonstration `http://www.youtube.com/watch?v=EBbkZj7HcL8`. For our future validations, we will enrich the validation dictionaries with BioPortal [Noy et al., 2009] terms for English and CISMeF [Darmoni et al., 2009] terms for French.

To sum up, this first part, “Automatic Term Extraction”, described the lexical complexity to extract new biomedical terms. For this, we proposed several measures that overcomes the baselines. These measures have been based on linguistic, statistic, graph, and web features showing good results. Finally this first part is implemented as an application called BioTex, which is starting to be used for several biomedical projects.

An interesting perspective is to add the context visualization of each candidate term. For instance, to show the most relevant sentences containing the candidate term, as done in [Fischl and Scharl, 2014, Hullman et al., 2015]. It will allow to users to perform a better validation of candidate terms. This context may also allow recognize manually if a candidate term is ambiguous (or polysemic). Likewise, the application could show the documents in which these candidate terms are relevant, as done in [Chuang et al., 2012] forassessing topic model quality.

Another perspective is to present a graph of term co-occurrence, showing the relatedness between the candidate terms. Note that the candidate terms are terms presents in a terminology (green terms), as well as new candidate terms (red terms). Therefore, this will give an appreciation of the possible position for a new biomedical term in an ontology.

In the future, it would be interesting to offer a disambiguation application using the context of each term in order to populate biomedical ontologies with the new extracted terms (red terms), while looking into the possibility of extracting relations [Abacha and Zweigenbaum, 2011] between new terms and already known terms.

[12] `http://www.lirmm.fr/sifr`

As previously mentioned, adding more functionalities to BioTex, users could be able to recognize "manually" if a candidate term is ambiguous, the possible senses according to the contexts, and to figure out a good position for new terms in a biomedical ontology. Note that the objective of this thesis is to automatically enrich biomedical ontologies. So, for this, the process to evaluate if a new biomedical term is ambiguous, and to evaluate its right position in an ontology has to be done automatically.

Hence, next part called "Concept Extraction and Semantic Linkage", seeks to induce the possible sense or senses for new candidate terms, and to propose where these new candidate terms could be added in an ontology.

Part II

Concept Extraction and Semantic Linkage

Chapter 9

Introduction

As a reminder to readers, the Web is by far the largest information archive available worldwide getting larger every day with the input of new user content. This vast pool of text expressions and terms contains important information about several domains. This is the case for biomedicine that, accompanied by recent advances in research, has accelerated the rate of publishing electronic biomedical documents [El-Rab et al., 2013]. Such growth makes it difficult to keep track of recent developments [Stevenson et al., 2008], so the use of automated methods to analyze and process data generated by users is thus mandatory. Extracting and enriching ontologies/vocabularies with new terms is challenging because of the ambiguity of natural language.

Therefore, to accomplish the general objective, the first part of this manuscript proposed a methodology to extract new biomedical candidate terms from textual data derived from several resources. Specifically, a methodology consisting of different measures based on linguistic, statistic, graph and Web features. At the end of this part, the new potential candidate terms are identified without further knowledge about them.

Hence, the second part seeks to achieve ontology/vocabulary enrichment and requires a methodology to figure out the concepts of the new candidate terms. Then, to seek the relation of new candidate terms with those appearing in an already existing biomedical ontology, the correct position of new candidate terms in a biomedical ontology must be determined.

Consequently, this second part, called *Concept Extraction and Semantic Linkage*, as we mentioned, seeks to extract the concepts and find the semantic links (i.e. a position semantically close) in a biomedical ontology of these new candidate terms, as mentioned before. Therefore, for this purpose, we believe it is important and necessary to perform three steps. First, it is essential to determine if a new candidate term is polysemic. Second, after having predicted the polysemy, we need to identify the possible senses or concepts of each term, this is the well-known Word Sense Induction (WSI) domain in which just the detected polysemic candidate terms will be evaluated by conducting an exhaustive search of their several concepts. Third, we have to be capable of finding semantic links that could have new candidate terms in an already defined biomedical ontology, i.e. to find a position in a biomedical ontology to add new candidate terms. Figure 9.1 illustrates the three steps necessary to reach the final objective.

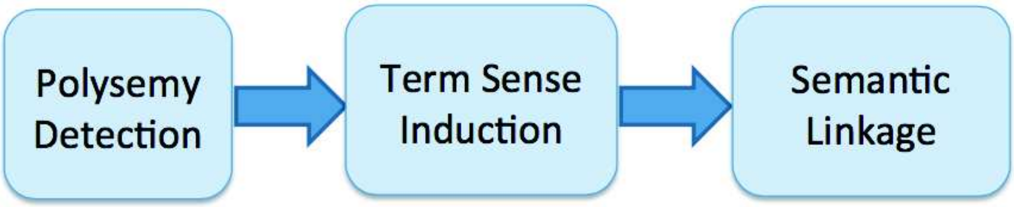


Figure 9.1: Workflow for Concept Extraction and Semantic Linkage.

Polysemy Detection aims at identifying, as *true* or *false*, when a new candidate term could have multiple meanings. This is very important because polysemic candidate terms must be evaluated in a different way than those that are not polysemic. In fact, polysemy is a major problem in linguistic semantics and it allows more accurate reading and better language acquisition [Crossley et al., 2010].

Term Sense Induction (TSI) aims to assign possible senses to these new biomedical candidate terms. In TSI, the time needed to detect the set of senses is an inconvenience, including the unnecessary time spent for non-polysemic terms with a single sense. This disadvantage is affected directly by the major problem in TSI, i.e. determination of the number of senses (i.e. number of clusters) of a term. This is because there is no a priori knowledge about the new candidate terms. In several domains, determining the exact cluster number becomes impossible, so the clustering algorithms often give poor performance results [Dehkordi et al., 2009]. For instance, to predict the number of senses, the non-parametric Bayesian method [Lau et al., 2012] uses Hierarchical Dirichlet Processes (HDP). All the approaches tend to induce a larger number of word senses compared to the gold standard per ambiguous word on the SEMEVAL-2010 WSI dataset [Lau et al., 2012]. Although many algorithms have been suggested, there is no convincing acceptable solution to the best number of clusters problem [Mirkin, 2011].

To achieve Polysemy Detection and Term Sense Induction, many techniques have been proposed, mainly based on supervised classification and unsupervised classification. Using words, the context of words [Navigli, 2012] and the graph of words [Agirre and Soroa, 2009, Agirre et al., 2010] as representation.

A related field, i.e. meta-learning, which aims to assign a good classifier for a specific dataset, gives very good classification results. Meta-learning uses several dataset characterization approaches. Meta-feature extraction is the first step of meta-learning, which exploits meta-knowledge to select the best method for the classification task. General techniques are used to extract meta-features, which can be transformed to extract even more meta-features. This enhances the efficiency of the classification process. Meta-learning has been applied in different domains but never for polysemy detection to our knowledge. Therefore, we take advantage of meta-features to meet the challenge of Polysemy Detection. In this context, we propose a novel approach to detect if a term is polysemic by defining new meta-features extracted directly from the textual dataset and from an induced co-occurrence graph. In turn, these meta-features use two dictionaries from two different domains (i.e. medical and agronomy), thus allowing us to determine the use of a same term in different domains. To the best of our knowledge, the properties of a co-occurrence graph have never been used to define meta-features. The main idea is to capture the dataset characteristics from the structural shape and size of the graph induced from the dataset.

To solve the major problem of TSI, we propose new indexes to evaluate the cluster quality. Moreover, to solve the problem related to TSI, we also use the proposed meta-features extracted for Polysemy Detection. So only terms detected as polysemic have to be evaluated to identify the number of senses. Then we will extract the set of senses for polysemic terms and the unique sense of the non-polysemic terms.

Semantic Linkage seeks to find a semantic relation between two entities. This is part of the well-known Relation Extraction domain. A lot of applications in information extraction, natural language processing, require an understanding of the semantic relations between entities. Relation Extraction approaches are generally categorized in supervised and unsupervised techniques [Bach and Badaskar, 2007]. In the unsupervised paradigms, contextual features are used. The distributional hypothesis theory [Harris, 1954] indicates that words that occur in the same context tend to have similar meanings. Accordingly, it is assumed that pairs of terms that occur in similar contexts tend to have similar relations. These unsupervised techniques are what we call Semantic Linkage. Relation extraction can also be divided into two tasks: (i) To find couples of terms which are semantically close, and (ii) To identify the kind of relations. The first task is usually tackled with unsupervised techniques, which aims to find the closest position of a new candidate term in a biomedical ontology. In our case, we only tackle the first task: (i) with unsupervised techniques.

As previously mentioned, this task is called semantic linkage, which in our work is essentially based on the context of the terms. This is, if two biomedical context of two terms are close, so they likely have a semantic link between them.

All of these steps are done automatically in our project. Therefore, the final workflow of our methodology is shown in Figure 9.2. In this figure, we can see that the polysemic terms follow a different approach to extract possible concepts in comparison to non-polysemic terms.

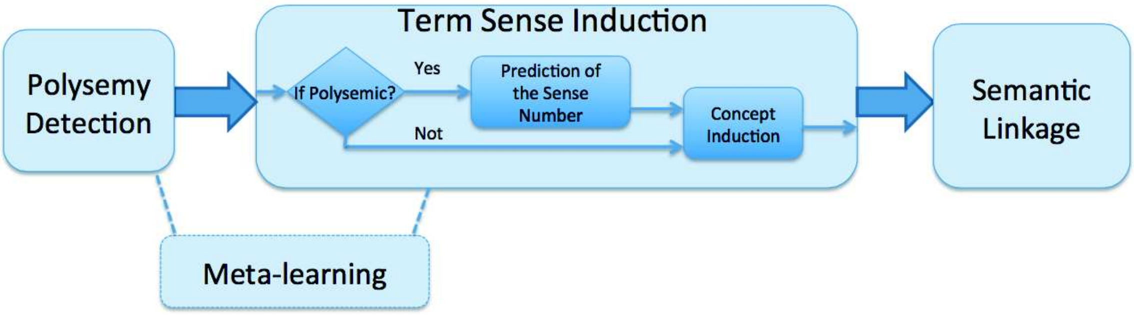


Figure 9.2: Final Proposed Workflow for Concept Extraction and Semantic Linkage.

The rest of this part is organized as follows. A detailed study on related works and their relevance in the ontology enrichment context is discussed in Chapter 10. Then, in Chapter 11, we present the main phases of our methodology. Results of experiments are presented in Chapter 12. A discussion and conclusions are addressed in Chapter 13.

Chapter

10

State-of-the-art

As we mentioned in the previous chapter, we propose a methodology aimed at enriching biomedical ontologies/terminologies. For this purpose, we split this part into three steps, as illustrated in Figure 10.1. In this chapter, we will describe the state-of-the-art for each step: (i) Polysemy Detection, (ii) Term Sense Induction, and (iii) Semantic Linkage. These domains are defined and the associated studies are reported in each step respectively.

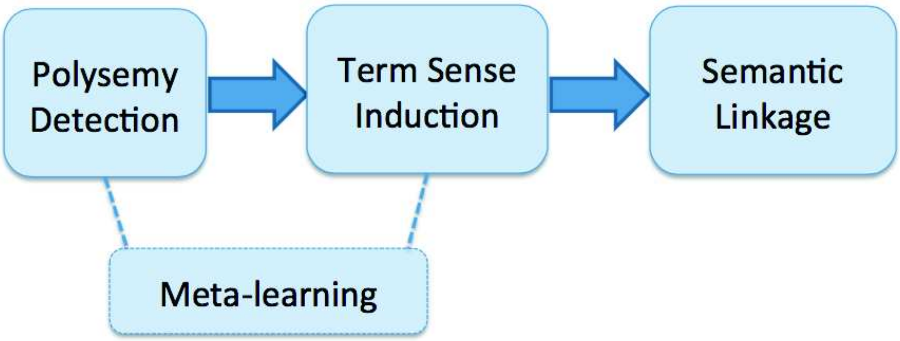


Figure 10.1: Workflow for Concept Extraction and Semantic Linkage.

10.1 Polysemy Detection

Polysemy detection seeks to detect if a term is polysemic given the context. The term “polysemy detection” is used in two different contexts. The first one, as in major studies, polysemy detection is used to detect the set of senses for a target word, this is the well-known Word Sense Induction domain. The second context seeks only to detect if a term is polysemic, i.e. true or false as result. In this section, we only describe studies belonging to the second context.

A task related to polysemy detection is term ambiguity detection (TAD), as shown in [Baldwin et al., 2013] which, given a term and a corresponding topic domain, determines whether the term uniquely references a member of that topic domain. For instance, given a term such as *Brave* and a category such as *film*, the task is to make a binary decision as to whether all instances of *Brave* reference a film by that name. They use a hybrid approach consisting of three modules. The first module is primarily designed to detect non-referential ambiguity. This means terms appearing in non-named entity contexts are likely to be non-referential, and terms that can be non-referential are ambiguous. The authors compute the non-referentiality as a probability based on the lower-cased form of words composing the term. Their idea is if the term is in lower form, it does not reference a named entity. The second module employs ontologies to detect across-domain ambiguity. Two ontologies were examined. Wiktionary, where terms having multiple senses were labeled as ambiguous. The second ontology used was Wikipedia disambiguation pages. All terms that had a disambiguation page were marked as ambiguous. The final module attempts to detect both non-referential and across-domain ambiguity by clustering the contexts in which words appear. They utilized the popular Latent Dirichlet Allocation (LDA) [Blei et al., 2003] topic modeling method. LDA represents a document as a distribution of topics, and each topic as a distribution of words. In this work, some terms can be already indexed in Wiktionary and/or Wikipedia ontologies. In such cases, their framework is nonviable to enrich ontologies.

Another related study proposes a measure to decide if a preposition is polysemous or not, with the final objective being to determine the preposition senses [Köper and im Walde, 2014]. The authors propose a rank-based distance measure to explore the vector-spatial properties of the ambiguous terms, thus solving two issues: (i) to distinguish polysemous from monosemous prepositions in vector space; and (ii) to determine salient vector-space features for the classification of preposition senses. The rank-based measure predicts the polysemy vs. monosemy of prepositions with up to 88% precision, and suggests that noun-based features are better than verb-based features to predict the sense of prepositions. Their rank-based distance measure is computed in three parts: (i) For a set of 49 German prepositions, they compute pair-wise distances for each pair of prepositions, which are represented by high-dimensional vectors, and they use the standard cosine measure for calculating vector similarities/distances; (ii) They rank according to the distance values, i.e. they determine, for each preposition, the most similar preposition, the second most similar preposition, etc. and; (iii) An overall rank is calculated for each preposition in order to distinguish the polysemy and monosemy. This rank corresponds to the mean position of a specific preposition in the distance-based sorted lists across all prepositions. In this case, the prepositions are already known.

Polysemy detection is similar to the well studied issues of named entity disambiguation (NED) and word sense disambiguation (WSD). These tasks assume that the number of senses of a word is given. This makes these tasks inapplicable in enriching terminology tasks because the number of senses of a new candidate term is not known. One task that requires polysemy detection (true or false as output) is word sense induction (WSI), which attempts to both figure out the number of senses of a word, and what they are (see section 10.2).

In our methodology, we are interested in polysemy detection in order to enrich terminologies/ontologies with new biomedical candidate terms (i.e. terms that are not indexed). Therefore, we try to predict when a new candidate biomedical term might be polysemic (i.e. true or false). Our approach will be detailed in Section 11.1. In general, the approaches mentioned before characterize the dataset as a vector of features, with the most well-known being "bag-of-words", then they use a learning algorithm (supervised/unsupervised) to capture the polysemy and senses of a word. An area that involves a similar process is meta-learning, which seeks first to characterize the dataset, and second, to assign it a highly performant classification algorithm. Meta-learning has proved to perform weel for classification tasks. An overview of meta-learning-based methods is proposed hereafter.

10.2 Term Sense Induction

Word Sense Induction is a natural language processing area which aims at the automatic identification of the senses of a word (i.e. meanings). As the output of word sense induction is a set of senses (i.e. sense inventory) for a word, obtained without any knowledge resource, so this task is related and really important for enriching terminologies and ontologies. In our methodology, we talk of technical terms, which means terms that could be added to a terminology in a domain. In our case, we extract new biomedical candidate terms and our aim is to use them for increasing terminologies and ontologies. These are single-word or multi-word terms. For this reason, we call this task "Term Sense Induction" (TSI) instead of Word Sense Induction, which aims, in major cases, to find senses for single-word terms. To our knowledge, few studies have been conducted related to word sense induction applied to the biomedical domain. In contrast, several studies are related to word sense disambiguation related to the biomedical domain.

TSI are always based on clustering algorithms. Two major problems must be addressed to automatically determine the senses of a term: (i) Determine the number of senses, i.e. to induce the number of meanings from the dataset, which is usually taken as a prior in most clustering algorithms; and (ii) Choose the best clustering algorithm for our dataset, i.e. in our case a textual dataset.

10.2.1 Word Sense Induction

WSI uses unsupervised techniques to automatically identify the set of senses denoted by a word. In general, most WSI research is based on: (a) Distributional Hypothesis [Harris, 1954], which indicates that words surrounded with similar contexts tend to have similar meanings, and (b) Topic Modeling Methods, which can discover latent topic structures from contexts without involving feature engineering.

More specifically, the main WSI approaches proposed are categorized in four types [Navigli, 2012, Wang et al., 2015b]: (i) *Context clustering*, (ii) *Word clustering*, (iii) *Co-occurrence Graphs*, and (iv) *Probabilistic*. The types (i), (ii), and (iii) are based on (a). And, the type (iv) is based on (b). Hereafter, we present the main WSI approaches.

- **Context Clustering**: The main idea underlying this approach is that the distributional profile of words implicitly expresses their semantics. This means that the word sense can be derived from its context. Each occurrence of a target word in a corpus is represented here as a context vector. These context vectors are composed of linguistic features and can be extracted i two manners: (i) Direct representation of the context (a context window), or (ii) The representation of contexts of the target word containing words that co-occur together. The vectors are then clustered into groups, with each identifying a sense of the target word [Nasiruddin, 2013].

  More specifically, this idea is based on the word space model [Schutze, 1992], where dimensions are words. A word w in a corpus can be represented as a vector whose jth component counts the number of times that word w_j co-occurs with w within a fixed context (a sentence or a larger context). The hypothesis underlying this model is that the distributional profile of words implicitly expresses their semantics. The similarity between two words can then be calculated, for example, with the cosine measure between the corresponding vectors of each word.

  The set of vectors for each word in the corpus creates a co-occurrence matrix. This might involve a large number of dimensions. Therefore, latent semantic analysis (LSA) can be applied to reduce the dimensionality of the resulting multidimensional space via singular value decomposition (SVD) [Golub and Van Loan, 1989]. SVD finds the major axes of variation in the word space. Dimensionality reduction take the set of word vectors in the high-dimensional space and represents them in a lower-dimensional space: As a result, dimensions associated with terms that have similar meanings are expected to be merged. After the reduction, contextual similarity between two words can also be computed with the cosine measure.

The aim is then to cluster context vectors, i.e. vectors which represent the context of specific occurrences of a target word. A context vector is built as the centroid (i.e. the normalized average) of vectors of the words occurring in the target context, which can be seen as an approximation of its semantic context [Schutze, 1992, Schütze, 1998]. These context vectors are second order vectors (composed words), in that they do not directly represent the context at hand. In contrast to this representation, another work [Pedersen and Bruce, 1997] models the target context directly as a first-order vector of several features.

Finally, sense induction can be performed by grouping the context vectors of a target word using a clustering algorithm. [Schütze, 1998] proposed an algorithm, called context-group discrimination, which gathers occurrences of an ambiguous word into clusters of senses based on the contextual similarity between occurrences. Contextual similarity is calculated as described above, whereas clustering is performed with the Expectation Maximization algorithm, i.e. an iterative maximum likelihood estimation procedure of a probabilistic model. A different clustering approach consists of agglomerative clustering [Pedersen and Bruce, 1997]. Initially, each instance constitutes a singleton cluster. Next, agglomerative clustering merges the most similar pair of clusters, and continues with successively less similar pairs until a stopping threshold is reached. The performance of the agglomerative clustering of context vectors was assessed in an unconstrained setting and in the biomedical domain [Savova and Pedersen, 2005].

There are methods that are based on the previously cited works. For instance, [Purandare and Pedersen, 2004b] makes a systematic comparison of WSI methods, with the first one being [Pedersen and Bruce, 1997] and the second one [Schutze, 1992, Schütze, 1998]. They cluster instances of a target word that occur in raw text using both vector and similarity spaces. The context of each instance is represented as a vector in a high dimensional feature space. Discrimination is achieved by clustering context vectors directly in the vector space and also by finding pairwise similarities among the vectors and then clustering in similarity space. The authors employ two different representations of the context in which a target word occurs. First order and second order.

In this context, [Udani et al., 2005] present an approach to noun sense induction using an on-the-fly unsupervised clustering algorithm operating on Web search results. The main idea is that the Web search appears to be a good indicator of the type of corpora available on the Web.

In [Bordag, 2004], the authors presented the "Triplet-based algorithm", which represents an instantiation of the "one sense per collocation" observation [Gale et al., 1992]. That essentially means that whenever a pair of words co-occurs significantly often in a corpus (collocation), the concept referenced by that pair is unambiguous. This approach differs from the others, which use clustering of word co-occurrences, in that it enhances the effect of the one sense per collocation observation by using triplets of words instead of pairs. The authors implement a two step clustering process using sentence co-occurrences as features. Moreover, a novel likewise automatic and unsupervised evaluation method inspired from [Schutze, 1992] is used.

SenseClusters[1] [Purandare and Pedersen, 2004a] is a package of (mostly) Perl programs that allows a user to cluster similar contexts together using unsupervised knowledge-lean methods. These techniques have been applied to word sense discrimination, email categorization, and name discrimination. In [Pedersen, 2007, Pedersen, 2010], SenseClusters was configured to construct representations of the instances. These instances use second order co–occurrence vectors. These are constructed by first identifying word co–occurrences, and then replacing each word in an instance to be clustered with its co-occurrence vector. Then all the vectors that make up an instance are averaged together to represent that instance.

A recent work [Van de Cruys and Apidianaki, 2011], presented a unified model for the automatic induction of senses, and the subsequent disambiguation of particular word instances using the automatically extracted sense inventory. The induction step and the disambiguation step are based on the same principle: words and contexts are mapped to a limited number of topical dimensions in a latent semantic word space. The intuition is that a particular sense is associated with a particular topic, so that different senses can be discriminated through their association with particular topical dimensions; in a similar way, a particular instance of a word can be disambiguated by determining its most important topical dimensions.

A related study applied to the biomedical domain is presented in [Savova and Pedersen, 2005], which aims to find semantic ambiguities in the biomedical domain. This method uses first and second order representations of context

[1] http://www.d.umn.edu/~tpederse/senseclusters.html

and it is evaluated on the National Library of Medicine Word Sense Disambiguation Corpus. The authors showed that the method of clustering second order contexts in similarity space is especially effective on such domain-specific corpora. The goal of this method is to divide contexts that contain a particular target word into clusters, where each cluster represents a different meaning of that target word. Each cluster is made up of similar contexts, and they presume that a target word used in similar contexts will have the same or very similar meaning.

Multilingual context vectors are also used to determine word senses [Ide and Erjavec, 2001, Albano et al., 2014, Albano et al., 2015]. In this setting, a word occurrence in a multilingual corpus is represented as a context vector which includes all the possible lexical translations of the target word.

- **Word Clustering**: Before we represented word senses as first- or second-order context vectors. A different approach to the induction of word senses consists of word clustering techniques, such approaches seek to cluster words which are semantically similar and can hence find a specific sense.

  A well-known approach to word clustering [Lin, 1998a] consists of identifying words $W = (w_1, \dots, w_k)$ similar (possibly synonymous) to a target word w_0. The similarity between w_0 and w_i is determined based on the information content of their single features, given by the syntactic dependencies which occur in a corpus (e.g. subject-verb, verb-object, adjective-noun, etc.). The more dependencies the two words share, the higher the information content [Salton and Buckley, 1988]. However, as for context vectors, the words in W will cover all senses of w_0. A word clustering algorithm is applied to discriminate between the senses. Let W be the list of similar words ordered by degree of similarity to w_0. A similarity tree T is initially created which consists of a single node w_0. Next, for each $i \in \{1, ..., k\}$, $w_i \in W$ is added as a child of w_j in the tree T such that w_j is the most similar word to w_i among $\{w_0, ..., w_{i-1}\}$. After a pruning step, each subtree rooted at w_0 is considered as a distinct sense of w_0.

  The clustering by committee (CBC) algorithm [Pantel and Lin, 2002] also uses syntactic contexts. For each target word, a set of similar words is computed as above. To calculate the similarity, again, each word is represented as a feature vector, where each feature is the expression of a syntactic context in which the word occurs. Given a set of target words (e.g. all those occurring in a corpus), a similarity matrix S is built such that S_{ij} contains the pairwise similarity between words w_i and w_j. As a second step, given a set of words E, a recursive procedure is applied to determine sets of clusters, called committees, of the words in E. To this end, a standard clustering technique, i.e.

average-link clustering, is employed. In each step, residue words not covered by any committee (i.e. not similar enough to the centroid of each committee) are identified. Recursive attempts are made to discover more committees from residue words. Note that, as above, committees conflate senses as each word belongs to a single committee. Finally, as a sense discrimination step, each target word $w \in E$, again represented as a feature vector, is iteratively assigned to its most similar cluster based on its similarity to the centroid of each committee. After a word w is assigned to a committee c, the intersecting features between w and elements in c are removed from the representation of w, so as to allow for the identification of less frequent senses of the same word at a later iteration.

The authors described the ASIUM system in [Faure and Nédellec, 1998], which learns subcategorization frames of verbs and ontologies from syntactic parsing of technical texts. The selection restrictions in the subcategorization frames are filled by the ontology concepts. ASIUM takes syntactic parsing of texts, which are subcategorization examples and basic clusters formed by head words that occur with the same verb after the same preposition (or with the same syntactical role). ASIUM successively aggregates the clusters to form new concepts in the form of a generality graph that represents the ontology of the domain. The ASIUM method is based on conceptual clustering.

Some recent studies take advantage of syntactic relations between words [Chen et al., 2009, Van de Cruys and Apidianaki, 2011] to conduct context modeling. In [Van de Cruys and Apidianaki, 2011], the authors combine two kinds of approaches: Context and Word clustering. The induction is based on words and contexts are mapped to a limited number of topical dimensions in a latent semantic word space. The key idea is that the model combines tight, synonym-like similarity (based on dependency relations) with broad, topical similarity (based on a large "bag of words" context window).

Another work [Pinto et al., 2007] uses a an information theory based co-occurrence measure, named pointwise Mutual Information (MI), which is fully discussed in [Manning and Schütze, 1999], and its applications for finding collocations are analyzed by determining the co-occurrence degree among two terms. This may be done by calculating the ratio between the number of times that both terms appear together (in the same context and not necessarily in the same order) and the product of the number of times that each term occurs alone. The objective is to construct a co-occurrence list for performing self-term expansion in order to improve the usability of limited, narrow-domain corpora. The self term expansion method consists of replacing terms of a document by a set of correlated terms. The goal is to improve natural language processing tasks such as clustering narrow-domain short texts.

In [Niu et al., 2007], the authors used three types of features to capture contextual information: part-of-speech of neighboring words (no more than three word distance) with position information, unordered single words in a topical context (all the contextual sentences), and local collocations (including 11 collocations). The feature set used in this work was the feature set used in [Lee and Ng, 2002], except that they did not use syntactic relations.

- **Co-occurrence Graphs**: These techniques assume that the semantics of a word can be reached by building and analyzing a word co-occurrence graph. This allows identification of the set of senses using graph-based clustering [Navigli and Crisafulli, 2010]. These techniques are related to word clustering methods, where co-occurrences between words can be obtained on the basis of grammatical or collocation relations.

For this, a graph is built, $G = (V, E)$ whose vertices V correspond to words in a text and edges E connect pairs of words which co-occur in a syntactic relation, in the same paragraph, or in a larger context. The construction of a co-occurrence graph based on grammatical relations between words in context was described in [Widdows and Dorow, 2002]. Given a target ambiguous word w, a local graph G_w is built around w. By normalizing the adjacency matrix associated with G_w, we can interpret the graph as a Markov chain. The Markov clustering algorithm [Van Dongen, 2001] is then applied to determine word senses, based on an expansion and an inflation step, aiming, respectively, at inspecting new more distant neighbors and supporting more popular nodes.

Another work [Véronis, 2004] proposed an ad hoc approach called HyperLex. First, a co-occurrence graph is built such that nodes are words occurring in the paragraphs of a text corpus in which a target word occurs, and an edge between a pair of words is added to the graph if they co-occur in the same paragraph. Each edge is assigned a weight according to the relative co-occurrence frequency of the two words connected by the edge. As a second step, an iterative algorithm is applied to the co-occurrence graph: At each iteration, the node with the highest relative degree in the graph is selected as a hub (based on the experimental finding that a node's degree and its frequency in the original text are highly correlated). As a result, all its neighbors are no longer eligible as hub candidates. The algorithm stops when the relative frequency of the word corresponding to the selected hub is below a fixed threshold. The entire set of hubs selected is said to represent the senses of the word of interest. Hubs are then linked to the target word with zero-weight edges and the minimum spanning tree (MST) of the entire graph is calculated. Finally, MST is used to disambiguate specific instances of our target word.

“PageRank”is an alternative graph-based algorithm for inducing word senses [Brin and Page, 1998]. PageRank is a well-known algorithm developed for computing the ranking of web pages, and is the main component of the Google search engine. It has been employed in several research areas for determining the importance of entities whose relations can be represented in terms of a graph. Further experiments on HyperLex and PageRank have been performed in [Agirre and Edmonds, 2007], who tuned a number of parameters of the former algorithm, such as the number of adjacent vertices in a hub, the minimum frequencies of edges, vertices, hubs, etc.

In [Dorow and Widdows, 2003], they extracted only noun neighbors that appeared in conjunctions or disjunctions with the target word. Additionally, they extracted second-order co-occurrences. Nouns are represented as vertices, while edges between vertices are drawn, if their associated nouns co-occur in conjunctions or disjunctions more than a given number of times. This co-occurrence frequency is also used to weight the edges. The resulting graph is then pruned by removing the target word and vertices with a low degree. Finally, the MCL algorithm [Dongen, 2000] is used to cluster the graph and produce a set of clusters (senses), each one consisting of a set of contextually related words.

Another work using a graph model [Agirre and Soroa, 2007b] proposed a system for performing two-stage graph based clustering where a co-occurrence graph is first clustered to compute similarities against contexts. The context similarity matrix is pruned and the resulting associated graph is clustered again using a random walk type algorithm.

An alternative method [Klapaftis and Manandhar, 2008] creates a graph, where each vertex corresponds to a collocation that co-occurs with the target word, and edges between vertices are weighted based on the co-occurrence frequency of their associated collocations. A smoothing technique is applied to identify more edges between vertices and the resulting graph is then clustered.

In [Korkontzelos and Manandhar, 2010], the authors presented a graph-based approach for word sense induction and disambiguation. This approach represented an unambiguous unit as a graph vertex: (a) a single word, if it is considered unambiguous, or (b) a pair of words, otherwise. The co-occurrences of the content of the vertices that they join were modeled by graph edges. Then hard-clustering on the graph was done. To disambiguate a test instance, the authors assigned it to the induced sense whose vertex contents occurred mostly in the instance.

A hierarchical structure (binary tree) of a graph was inferred [Klapaftis and Manandhar, 2010], in which vertices were contexts of a polysemous word and edges represented the similarity between contexts. The method they used to infer that hierarchical structure is the Hierarchical Random Graphs (HRGs) algorithm [Clauset et al., 2007].

These studies of WSI are required for Information Retrieval (IE) as well [Navigli and Crisafulli, 2010, Di Marco and Navigli, 2011, Di Marco and Navigli, 2013, Lau et al., 2013]. In [Di Marco and Navigli, 2011], the authors presented an approach to Web search result clustering based on the automatic discovery of word senses. First, the senses are acquired from a query by means of a graph-based clustering algorithm that exploits cycles (triangles and squares) in the co-occurrence graph of the query. Then they clustered the search results based on their semantic similarity to the induced word senses.

In such approaches, few works related to the biomedical domain have been proposed. For instance, in [Noh et al., 2010] a network representation of co-occurrence data is first defined to represent both word senses and word contexts. The representation expresses the textual context observed around a certain term as a network, where nodes are terms and edges are the number of co-occurrences between connected terms. A graph kernel is adopted as a similarity measure between terms and senses represented in networks. Candidate senses and ambiguous contexts are then compared directly in the representation space to resolve the word sense. They conducted experiments in the biomedical domain and found, according to the recall results, that the method outperformed a baseline vector representation method. No precision, accuracy, and F-measures were computed.

An additional related work in the biomedical domain [Duan et al., 2009] presents a an efficient graph-based algorithm to cluster words into groups. That algorithm follows the principle of finding a maximum margin between clusters, determining data splits that maximize the minimum distance between pairs of data points belonging to two different clusters.

- **Probabilistic Clustering**: Another option is to adopt some probabilistic approaches. Advanced Bayesian methods have been explored in recent years, the main reason is that the methods can discover latent topic structures from contexts without involving feature engineering. First, a distribution of senses is drawn for each ambiguous word. Then, context words are generated according to this distribution. Different senses can thus be obtained which have different word distributions. Hence, in this kind of approach, the contexts of ambiguous instances are regarded as pseudo documents and their induced topic distributions are considered as sense distributions.

A Bayesian framework [Brody and Lapata, 2009] that uses parametric LDA [Blei et al., 2003], and formalize WSI in a generative model. They proposed a topic model that uses a weighted combination of separate LDA models based on different feature sets (e.g. word tokens, parts of speech, and dependency relations). They only used smaller units of text surrounding the ambiguous word, while discarding the global context of each instance.

In [Yao and Van Durme, 2011], the authors proposed a model based on a hierarchical Dirichlet process (HDP) [Teh et al., 2006] to learn the sense distributions. The advantage of this method is that it can automatically learn the number of word senses for each ambiguous word, as compared to LDA which needs to be pre-assigned a topic number in advance. Experimental results have shown that the HDP model is superior to the standard LDA model. [Lau et al., 2012] also showed improvements in supervised F-scores after incorporating position features in the HDP model. [Choe and Charniak, 2013] extended the naive Bayes model based on the idea that the closer a word is to the target word, the more relevant this word will be in WSI.

As we can see, LDA has been used for WSI tasks in recent years. [Tang et al., 2014] propose a joint model to automatically induce document topics and word senses simultaneously. Instead of using some predefined word sense resources, the word sense information is captured via a latent variable and directly induced in a fully unsupervised manner from the corpora.

Recently, [Wang et al., 2015b] presented a sense-topic model also based on LDA, which treats sense and topic as two separate latent variables to be inferred jointly. Topics are informed by the entire document, while senses are informed by the local context surrounding the ambiguous word.

In the biomedical domain, a comparison between graph-based approaches and topic model approaches was carried out [Chasin et al., 2014]. The objective of this work was to evaluate the state-of-the-art of approaches for the clinical domain. In particular, to compare graph-based approaches relying on a clinical knowledge base with bottom-up topic-modeling-based approaches such as LDA. Topic-modeling methods achieved better accuracy on a subset of Mayo Clinic data than graph-based methods and than the most-frequent-sense baseline. For word sense disambiguation, there are also several approaches that use probabilistic methods, and topic models [Cai et al., 2007, Boyd-Graber and Blei, 2007, Boyd-Graber et al., 2007, Agirre et al., 2014]. Other hybrid approaches [Huang et al., 2015] construct multi-granularity semantic spaces to learn representations of ambiguous instances, in order to capture richer semantic knowledge during context modeling. In particular, the authors consider the semantic space of words,the semantic space of word clusters, as well as topics. To circumvent the difficulty of selecting the number of word senses, they adapted a rival penalized competitive learning method to determine the number of word senses automatically via gradually repelling the redundant sense clusters.

As previously mentioned, there are few studies related to biomedical WSI in comparison to the number of studies in biomedical WSD. An overview of studies related to WSD are proposed in section 10.2.2.

In WSI as in the WSD (see section 10.2.2), the major studies characterize the dataset with a word vector (bag-of-words) or word graph. A related field of computational intelligence is meta-learning, where classifications are done using several dataset characterization strategies. This field has shown very good classification results. An overview of types of dataset characterization are described in section 10.3.

We also mentioned that a small number of these approaches seek to automatically detect the number of senses for a target word. That is a major issue, which has a performance effect on the results of unsupervised algorithms. To detect the number of senses, several indices have been proposed, which compute the quality, and they opt for the best solution based to this quality evaluation. An overview of studies related to the determination of number of clusters are proposed in section 10.2.3.

10.2.2 Word Sense Disambiguation

One domain related to WSI is WSD, which has aims at determining what sense of a word (i.e. meaning) is used in a sentence, when the word has multiple senses. For this task, the words are already known and they are also indexed in an ontology/terminology. WSD is generally considered to be the first brick on the road to automated text understanding, as well as for information retrieval. Since our work is not based on WSD, hereafter we briefly describe some approaches, especially those related to biomedicine. There are three main approaches to WSD [Navigli, 2009]:

- **Supervised Approaches:** These approaches use supervised classification algorithms to learn a classifier for a target word from an annotated training dataset. The training dataset is a set of examples represented as vectors of features. In this vector there is a special element representing the appropriate sense (or class). In this kind of approach, the best have proved to be Super Vector Machines (SVM) and memory-based learning [Hoste et al., 2002, Decadt et al., 2004, Mihalcea and Faruque, 2004, Grozea, 2004, Chan et al., 2007, Zhong and Ng, 2010].

- **Knowledge-based Approaches:** These approaches aim at inferring the sense of a word in a context primarily using dictionaries, thesauri, glossaries, ontologies, and lexical knowledge bases, without using any corpus evidence. They have the advantage of a wider coverage, thanks to the use of large amounts of structured knowledge. Generally resources, such as WordNet, BabelNet and UMLS are used in this kind of approach in order to benefit from the graph structure to assign the correct sense of a word, for instance, Degree [Navigli and Lapata, 2010, Ponzetto and Navigli, 2010] or Personalized PageRank [Agirre and Soroa, 2009]. Other [Agirre et al., 2014].
- **Unsupervised Approaches:** These methods work directly with raw unannotated datasets to discover senses automatically. These methods are also called Word Sense Induction (see Section 10.2.1).

In biomedical contexts, knowledge-based approaches are generally proposed, with the most well-known resource used being UMLS[2] (Unified Medical Language System) [Bodenreider, 2004]. That is a large multi-purpose and multi-lingual thesaurus that contains millions of biomedical and health related concepts, their synonymous names, and their relationships. The Metathesaurus includes over 150 electronic versions of classifications, code sets, thesauri, and lists of controlled terms in the biomedical domain. Their uses include: patient care, health services billing; public health statistics; indexing and cataloging of biomedical literature; basic, clinical, and health services research. The major studies in this area rely on UMLS, for instance, [McInnes, 2008, Stevenson et al., 2008, McInnes and Pedersen, 2013, an tSaoir, 2014, McInnes and Stevenson, 2014].

In this context, there are fewer examples of unsupervised than supervised (including knowledge-based) approaches. Most unsupervised approaches use UMLS to some extent [Jimeno Yepes and Aronson, 2012] and Information Content Similarity [McInnes et al., 2011]. In [Schuemie et al., 2005], the authors believe that combining unsupervised learning and established knowledge is most effective.

The commonly cited example of a supervised approach that consistently outperforms known unsupervised approaches is Naive Bayes [Jimeno Yepes and Aronson, 2012, McInnes et al., 2011]. Several recent approaches to biomedical WSD use structured knowledge in the form of a graph. Examples range from the use of cooccurrence data from a domain-specific corpus [Agirre et al., 2006], to variations

[2] `http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html`

of PageRank [Agirre et al., 2010], to the graph representation of an ontology, or portion of an ontology [El-Rab et al., 2013].

WSD studies in the clinical context is growing, and is also based on UMLS, for instance [Savova et al., 2008, Moon et al., 2015].

10.2.3 Sense Number Prediction

In WSI, the sense number (cluster number) of ambiguous words cannot be appropriately determined. Specifically, in cluster analysis a major problem is to determine the best number of clusters, which has a performance effect on the clustering results.

In many practical WSI applications, it becomes impossible to know the exact cluster number in advance, so these clustering algorithms often result in poor performance [Dehkordi et al., 2009]. More recently, the non-parametric Bayesian method [Lau et al., 2012] uses Hierarchical Dirichlet Processes (HDP) [Teh et al., 2006] to learn the number of word senses automatically. Moreover, in [Klapaftis and Manandhar, 2010], the authors use HDP and one of their objectives is to predict the number of senses of a target word.

Another related work [Niu et al., 2007], applies a cluster validation method to estimate the number of senses of a target word in untagged data, and then grouped the instances of this target word into the estimated number of clusters. Specifically, this approach presents an index to predict the sense number of a word. They set $k_{min} = 2$, and $k_{max} = 5$ in their system.

However, all these approaches tend to induce larger numbers of word sense comparing to the gold standard per ambiguous word on SEMEVAL-2010 WSI dataset [Lau et al., 2012]. Hence, exploring a word sense clustering algorithm to learn appropriate sense numbers for ambiguous words is also crucial for WSI tasks.

In general, many popular clustering methods, such as the *k-means* algorithm, require the cluster number to be precisely predefined. A limitation in current applications is that there is no convincing acceptable solution to the best number of clusters problem [Mirkin, 2011]. That is due to the high complexity of real datasets.

Many strategies for estimating the optimal number of clusters have been proposed. A very extensive comparative evaluation was conducted [Milligan and Cooper, 1985], where they compared 30 proposed methods to estimate the true number of clusters when applying hierarchical clustering algorithms to simulated data with well-separated clusters. From this work, in [Anderson, 2001], it has been proved that *Calinski and Harabasz's* index is the most effective. This is followed by *Duda and Hart's* method [Javed et al., 2008], and the C-index. Although many algorithms have been suggested to solve this problem, there does not appear to be one most reliable method. In addition, several are unknown, consequently never used.

Algorithms to determine the cluster number can be categorized according to the data set type application, as: (i) Algorithms for numerical datasets, (ii) Algorithms for categorical datasets, and (iii) Algorithms for mixed datasets [Liang et al., 2012]. They can also be categorized in two categories [Gordon, 1999]: (i) Global methods, and (ii) Local methods. With global methods, the quality of clustering, given a specific number of clusters, is measured by a criterion, and the optimal estimate is obtained by comparing the values of the criterion. Local methods are intended to test the hypothesis that a pair of clusters should be amalgamated. They are only suitable for assessing hierarchically-nested partitions.

A method [Kolesnikov et al., 2015] has recently been proposed for determining the optimal number of clusters in a dataset. The method is based on parametric modeling of the quantification error. The model parameter is treated as the effective dimensionality of the dataset. This method was applied for numerical datasets and was tested with artificial and real numerical datasets.

Another study [Yu et al., 2014] proposes an automatic method by extending the decision-theoretic rough set model to clustering. They also propose a new clustering validity evaluation function based on the risk calculated by loss functions and possibilities. Then a hierarchical clustering algorithm is proposed, named *ACA-DTRS*, which stops automatically when the function is optimized for predicting the number of clusters.

In [Liang et al., 2012], to tackle the problem of cluster number prediction, the authors propose a generalized mechanism by integrating Renyi entropy and complement entropy. The mechanism characterizes within-cluster entropy and between-cluster entropy to identify the worst cluster in a mixed dataset. An index is also defined to evaluate the clustering results for mixed data.

MCS (maximum clustering similarity) [Albatineh and Niewiadomska-Bugaj, 2011] was proposed to solve this problem by studying the behavior of similarity indices comparing two (of several) clustering methods. The similarity between the two clustering methods is calculated with the same number of clusters. Finally, the number of clusters at which the index attains its maximum is a candidate for the optimal number of clusters.

Two methods [Yan, 2005] for estimating the number of clusters are proposed. The first one uses the weighted within-clusters sum of errors, i.e. a robust measure of the within-cluster homogeneity. The second one is a "multi-layer" analytic approach, which is particularly useful in detecting the nested cluster structure of data. The methods are applicable when the input data contain only continuous measurements and are partitioned based on any clustering method. Both are based on the GAP method.

Currently, there is a package in R named *NbClust* which was developed for that purpose. It provides 30 indices which determine the number of clusters in a dataset and it also offers clustering schemes from different results to the user. In this package, there are also several indices described in [Milligan and Cooper, 1985], with the following being the most well-known for clustering evaluation:

- CH [Caliński and Harabasz, 1974]
- CCC [Sarle, 1983]
- Pseudot2 [Duda et al., 1973]
- KL [Krzanowski and Lai, 1988]
- Gamma [Baker and Hubert, 1975]
- Gap [Tibshirani et al., 2001]
- Silhouette [Rousseeuw, 1987]
- Hartigan [Hartigan, 1975]
- Cindex [Hubert and Levin, 1976]
- DB [Davies and Bouldin, 1979]
- Ratkowsky [Ratkowsky and Lance, 1978]
- Scott [Scott and Symons, 1971]
- Marriot [Marriott, 1971]
- Ball [Ball and Hall, 1965]
- Trcovw [Milligan and Cooper, 1985]
- Tracew [Milligan and Cooper, 1985]
- Friedman [Friedman and Rubin, 1967]
- Rubin [Friedman and Rubin, 1967]
- Dunn [Dunn, 1974]

As previously mentioned, in general for polysemy detection problems, WSI, WSD, cluster number prediction, the major works tend to use words to characterize datasets. So, taking as base words, they are grouped in a vector, or the well-known "bag-of-words". They are also used to build graphs of words. Then supervised and unsupervised algorithms are applied for performing clustering and classification tasks. As we mentioned, meta-learning is a related computational intelligence field, which conducts classification using several dataset characterization strategies and gives very good classification results. We propose an overview of these types of dataset characterization in section 10.3, with the aim of using these types for performing the task of polysemy detection, WSI, and cluster number prediction.

10.3 Meta-learning

Meta-learning was originally described by [Maudsley, 1979] as “the process by which learners become aware of and are increasingly in control of habits of perception, inquiry, learning, and growth that they have internalized”.

In classification, various approaches are available to solve algorithm selection problems and a lot of research carried out in that direction. In [Bhatt et al., 2012], meta-learning is defined as a hot topic in machine learning research, which has emerged from the need to improve the generalization ability and stability of learned models and support data mining automation in issues related to algorithm and parameter selection. It is the process of generating knowledge that relates the performance of machine learning algorithms to the characteristics of the problem (i.e. characteristics of its datasets). Therefore, meta-learning is the study of principled methods that exploit meta-data to obtain efficient models and solutions by adapting machine learning and data mining processes [Bhatt et al., 2013].

Meta-learning has many aspects, but its final goal is to automatically discover many interesting models for given data [Duch et al., 2011].
In [Peng et al., 2002], the authors mention two basic tasks in meta-learning: the description of learning tasks (datasets), and the correlation between the task description and the optimal learning algorithm. The first task is to characterize datasets with meta-features, which constitutes the meta-data for meta-learning. The second is learning at the meta-level, which develops meta-knowledge for selecting appropriate algorithms in classification.

The meta-learner is a learning system that receives a set of such meta-examples as input and then acquires knowledge used to predict the algorithm’s performance for new problems being solved. The meta-features are, in general, statistics describing the training dataset of the problem, such as the number of training examples, number of attributes, correlation between attributes, class entropy, etc. [Lemke et al., 2013].

As in any learning task, the characterization of instances (meta-features) plays a crucial role in enabling learning. In particular, the meta-features used must have some predictive power. To date, authors have categorized meta-features in several categories. Major studies divide these meta-features in three categories, for instance [Giraud-Carrier, 2008, Bhatt et al., 2012]:

- **Direct meta-features:** In general simple, statistical and information-theoretic meta-features. The simplest and most widely used functions. The objective is to extract a number of statistical and information-theoretic measures. Typical measures include the number of features, number of classes, ratio of examples to features, degree of correlation between features and the target concept, average class entropy, class-conditional entropy, skewness, kurtosis, and signal to noise ratio. The idea here is that learning algorithms are sensitive to the underlying structure of the data on which they operate, so it may be hoped that it could be possible to map structures to algorithms. Several experiments suggest that the size of the training set and the size of the input space play a crucial role in determining the difference between algorithms. Since, in practice, these are usually different, one may expect to capture sufficient information from these and other measures to discriminate among learners. Empirical results do seem to confirm this intuition.

- **Model-based meta-features:** A different form to exploit properties according to an induced model. This approach has some advantages:

  (i) The dataset is summarized in a data structure that can embed the complexity and performance of the induced hypothesis, and thus is not limited to the example distribution.

  (ii) The resulting representation can serve as a basis to explain the reasons behind the performance of the learning algorithm. Currently, only decision trees have been considered, and they consist of some extracted properties, such as nodes per meta-feature, maximum tree depth, shape, and tree imbalance.

- **Landmarking meta-features:** Another source of characterization falls under the landmarking concept. Each learner has a class of tasks on which it performs particularly well, under a reasonable measure of performance. We call this class the area of expertise of a learner. The basic idea of the landmarking approach is that the performance of a learner on a task uncovers information about the nature of the task. Hence, a task can be described by the collection of areas of expertise to which it belongs. We call a landmark learner, or simply a landmarker, a learning mechanism whose performance is used to describe a task. Landmarking is the use of these learners to locate the task in the expertise space, the space of all areas of expertise. Landmarking thus views meta-learning as intending to find locations of tasks in the expertise space. While other approaches rely on peripheral information (e.g. statistical characteristics, model-based properties, etc.), landmarking uses an expertise map as its main source of information.

As previously mentioned, the most frequent measures to characterize datasets, are frequency, mean, standard deviation, etc. For instance, new measures are presented in [Peng et al., 2002] based on an induced decision tree to characterize datasets in order to select appropriate learning algorithms. The main idea is to capture the dataset characteristics from the structural shape and size of the decision tree induced from the dataset. The authors extracted 15 meta-features of three types: general meta-features, statistical meta-features, information-theoretic meta-features.

Additional meta-features have been proposed, as transformations of existing ones [Castiello et al., 2005], and some guidelines have been set to select the most informative ones. In their work, 9 new meta-features have been proposed of the three types above mentioned.

Other statistic meta-features have been presented in [Reif et al., 2012b], where the authors present a novel approach for constructing more informative meta-features using a two-stage method based on traditional meta-features. The proposed meta-features are able to describe differences over datasets that are not accessible using the standard meta-measures method. Moreover, they added an additional meta-feature selection method in order to automatically select the most useful measures. A similar work [Reif et al., 2012a] presented a new function, which is a novel data generator for creating datasets with specific characteristics that can be used for the development and evaluation of meta-learning systems.

As we saw, induced decision trees are generally used to extract meta-features, but there is no meta-feature extraction approach based on graph-based models. The properties of graphs make them a very useful structure. In addition, graphs have been shown to achieve state-of-the-art performances in standard evaluation tasks. That will be detailed in the next section.

In [Brazdil et al., 2008], the authors consider the number of classes as meta-features, the ratio of examples to meta-features, the degree of correlation between meta-features and target concept and average class entropy. In comparison with other areas, meta-features can look completely different, as for example summarized in [Lemke et al., 2009] for time series forecasting, where meta-features can include, for example, length, seasonality, autocorrelation, standard deviation and trends of the series.

Other measures are proposed [Vilalta and Drissi, 2009], they include class variation from the dataset, probability of variation, distance measure, cohesiveness, and density of the example distribution in the training set. In a similar approach, [Köpf and Iglezakis, 2002] suggests comparing observations with each other and extracts case base properties, which assess the dataset quality using measures such as redundancy, for example induced by data records that are exactly the same, or incoherency, which, for example occurs if data records have the same meta-features but different class labels.

Several meta-features can be extracted by using a model that is fast to build and train to take advantage of its properties. In this spirit, [Bensusan et al., 2000] suggests building a decision tree for a classification problem and using properties of the tree such as nodes per meta-feature, tree depth or shape to characterize it. Another approach wherby the landmarking type is extracted as meta-feature as proposed in

[Pfahringer et al., 2000], using the performance of simple algorithms to describe a problem and correlating this information with the performance of more advanced learning algorithms. A list of landmarking algorithms can be found in [Vanschoren, 2010]. An empirical evaluation of different categories of meta-features can be found in [Reif et al., 2014], where the authors distinguish 5 such categories of meta-features, i.e. simple, statistical, information-theoretic, landmarking and model-based, which corresponds to the general categorization evident from the literature.

10.4 Semantic Linkage

The aim of Semantic Linkage is to find a semantic relation between two entities. This belongs to the Relation Extraction (RE) domain. RE is one of the most important topics in NLP. RE is the task of determining semantic relations between entities mentioned in a text. RE is an essential part of information extraction and is useful for question answering [Ravichandran and Hovy, 2002], textual entailment [Marelli et al., 2014] and many other applications. Many applications in information extraction, natural language understanding and information retrieval require an understanding of the semantic relations between entities. Several relation extraction approaches in different domains have been proposed.

Relations are extracted after the new candidate terms have been found. Several approaches use identified terms to locate relations between them. These approaches have been categorized in several different categories, where one division could be [Bach and Badaskar, 2007]: (i) unsupervised relation discovery, and (ii) supervised classification.

In unsupervised paradigms, contextual features are used, this is also called semantic linkage. Distributional hypothesis theory [Harris, 1954] indicates that words that occur in the same context tend to have similar meanings. Accordingly, it is assumed that pairs of terms that occur in similar contexts tend to have similar relations.

In [Hasegawa et al., 2004], the authors adopted a hierarchical clustering method to cluster the contexts of terms and simply select the most frequent words in the contexts to represent the relation between the terms. This study involves context based clustering of pairs of entities. The assumption is that pairs of entities occurring in similar context can be clustered and that each pair in a cluster is an instance of the same relation. In cases where contexts linking a pair of entities express multiple relations, the pair of entities either would not be clustered at all, or would be placed in a cluster corresponding to its most frequently expressed relation.

Another work [Chen et al., 2005] proposed a novel unsupervised method based on model order selection and discriminative label identification to address this problem.

In recent works, the main idea is that the relation extractor simultaneously discovers facts expressed in natural language, and the ontology into which they are assigned [Banko and Etzioni, 2008, Lin and Pantel, 2001, Bollegala et al., 2010, Yao et al., 2011].

In the supervised paradigm, relation classification is considered a multi-classification problem, and researchers concentrate on extracting more complex features. Generally, these methods can be categorized into two types: feature-based and kernel-based. In feature-based methods, a diverse set of strategies have been exploited to convert the classification clues (such as sequences and parse trees) into feature vectors [Kambhatla, 2004, Suchanek et al., 2006]. Feature-based methods are hampered by the problem of selecting a suitable feature set when converting the structured representation into feature vectors. Kernel-based methods provide a natural alternative to exploit rich representations of the input classification clues, such as syntactic parse trees. Kernel-based methods allow the use of a large set of features without explicitly extracting the features. Various kernels, such as the convolution tree kernel [Qian et al., 2008], sub-sequence kernel and dependency tree kernel [Bunescu and Mooney, 2005], have been proposed to solve the relation classification problem. However, the methods mentioned above suffer from a lack of sufficient labeled data for training.

In the biomedical domain, there are several approaches available to extract relations, by matching linguistic patterns including those first described in [Liu et al., 2011] and those used in [Charlet et al., 2006, Baneyx et al., 2007]. These patterns include the relations “is a” for hierarchy and “also known as” for synonymy [Liu et al., 2011]. Other methods include hierarchical clustering [Kuo et al., 2007] rules to extract synonyms as well [Henriksson et al., 2014, Wang et al., 2015a].

In [Abacha and Zweigenbaum, 2011], the authors presented MeTAE (Medical Text Annotation and Exploration). MeTAE is used to extract and annotate medical entities and relationships from medical texts and to semantically explore the produced RDF annotations.
A recent work [Doing-Harris et al., 2015] presents a system to extract concepts and relationships from clinical and biomedical documents called SEAM. This system features a natural language processing pipeline for information extraction. Synonym and hierarchical groups are identified using corpus-based semantics and lexico-syntactic patterns.

A very closely related task is automatic thesaurus building from a corpus, i.e. a so-called distributional thesaurus. Given an input word, the thesaurus semantically identifies similar words based on the assumption that they share a distribution similar to that of the input word. It is usually ordered in descending values of similarity to the input word. Distributional thesauri are useful for several tasks, such as the extraction of relationships [Min et al., 2012] and syntactic analysis [Anguiano, 2013]. Generally this construction is based on distributional similarity, which has been widely discussed for several years, as in [Sparck Jones, 1986, Grefenstette, 1994, Lin, 1998b, Curran and Moens, 2002, Weeds, 2003, Heylen et al., 2008].

Distributional similarity establishes the fact that two words are closely related, i.e. a semantic relation, if they share similar contexts. These contexts are typically co-occurrent words in a limited window around the target word, or syntactically related words [Claveau et al., 2014].

There are two kinds of semantic relations: (i) paradigmatic relations, such as hyperonymy or synonymy, and (ii) syntagmatic relations, or so-called collocation relations [Halliday and Hasan, 1976] in the context of lexical cohesion or "non-classical relations" by [Morris and Hirst, 2004, Budanitsky and Hirst, 2006a, Adam et al., 2013]. In [Ferret, 2013], the author states that the difference between these two kinds of relations depends basically on the difference between semantic similarity and semantic relatedness. For instance, the studies carried out in [Budanitsky and Hirst, 2006b, Zesch and Gurevych, 2010]. As mentioned in [Ferret, 2013], in existing studies it is hard to determine the limit between these two notions of semantic similarity and semantic relatedness, so they are usually interchanged. Generally, semantic similarity is considered to be included in semantic relatedness.

One way to improve a distributional thesaurus is focused on the content of distributional contexts of words, for instance the filtering of components [Padró et al., 2014, Polajnar and Clark, 2014]; the weighting of distributional contexts [Ferret, 2015]; and looking for better algorithmic efficiency [Rychlý and Kilgarriff, 2007].

A quality evaluation of a distributional thesaurus [Adam et al., 2013] can be carried out by: (i) comparing the thesaurus to reference lexicons, also called direct evaluation, and (ii) in a specific task, which is called indirect evaluation, e.g. replacing a word by one of its neighbors so as not to alter the meaning of the sentence.

In a recent study [Claveau and Kijak, 2015], the authors use IR tools and concepts to build a thesaurus, they directly assess the results with reference lexicons, and indirectly in an IR task. In this study, only 25 000 nouns are considered, with a context window of +2-. WordNet and Moby thesauri are used for direct evaluation, and indirect evaluation is conducted through an RI task. The corpus used is AQUAINT-2[3]. WordNet[4] provides 3 neighbors on average for 10 473 nouns found in AQUAINT-2, and Moby[5] provides 50 neighbors on average for 9 216 nouns. Combined, these two ressources cover 12 243 nouns of the corpus with 38 neighbors on

[3] https://catalog.ldc.upenn.edu/LDC2008T25

[4] https://wordnet.princeton.edu/

[5] http://moby-thesaurus.org/

average.

Recent studies of Ferret range from the creation of distributional thesauri to reranking of distributional thesauri. In [Ferret, 2015], the authors propose a new criterion for improving distributional thesauri from a bootstrapping perspective. It is geared towards compound and single-word terms. The authors use AQUAINT-2 as a corpus to create the single-word term thesaurus (A2ST) and the compound thesaurus (A2ST-comp). The proposed methodologies generally involve a context window of +3-. This study is a follow-up to the methodologies proposed in [Ferret, 2010, Ferret, 2012, Ferret, 2013].

In the above-mentioned studies, the authors tend to extract paradigmatic relations such as hyperonymy or synonymy. In many studies it is not clear if just the synonyms are taken into account, or synonyms and hyperonyms. Most methodologies are based on single words, but none are based on hyponymy, which is very important in ontologies. This is where the semantic field of a term is included in another, i.e. its hyperonym. In the previously mentioned studies, the context of a word is generally represented by a windows of +2,3- neighbors. All of these approaches rank similar words without a linked concept.

Our objective is to enrich biomedical ontologies, so in this thesis we focus on the extraction of paradigmatic relations such as hyperonymy (fathers), synonymy, and hyponymy (sons). We use conventional semantic similarity approaches in a different manner, where each term has its own context. For us, single- or mutli-word terms are possible. Each context is represented by the most important features occurring in the paragraphs that contain the term, forming a bag-of-words (vector). Then a measure is applied to compute the similarity between the two vectors. For each new term, a concept is induced, which is not offered by the distributional thesaurus methodologies. The corpora used in other works are different from ours, which is an inconvenience fir comparing our approach.

Another close task is Entity Linking, which seeks to map an entity mention appearing in a text document to an entity in a knowledge base. Entity linking goes beyond NER tasks. The challenges [D'Souza and Ng, 2015] in this task are: (i) same words or phrases can be used to refer to different entities, (ii) same entities can be referred to by different words or phrases, and (iii) many mentioned entities may not appear in a knowledge base. In other words, entity linking is challenging due to name variations and entity ambiguity [Shen et al., 2015]. A named entity may have multiple surface forms, such as its full name, partial names, aliases, abbreviations, and alternate spellings.

One open-domain based approach [Rao et al., 2013] involves linking organizations, geo-political entities, and persons to entities in a Wikipedia-derived knowledge base, utilizing heuristics for matching mention strings with candidate concept phrases.

In the biomedical domain, this task is also known as normalization of entities or of concepts. These tasks are geared towards mapping a word or phrase in a document to an unique concept in an ontology. The disease normalization task consists of finding disease mentions and assigning a unique identifier to each. A recent study [D'Souza and Ng, 2015] presents a multi-pass sieve approach to the under-studied task of normalizing disorder mentions in the biomedical domain. It takes documents from two genres, i.e. clinical reports and biomedical abstracts.

In [Ghiasvand and Kate, 2014], the methodology first generates variations in a given disorder word/phrase based on a set of learned edit distance patterns for converting one word/phrase into another, and then attempts to normalize these query phrase variations by performing exact matches with a training disorder mention or a concept term.

In this kind of task, we can cite DNorm[6] [Leaman et al., 2013], which is an automated method for determining which diseases are mentioned in biomedical texts and this task is called disease normalization. DNorm learning similarities between mentions and concept names directly from training data. DNorm is the first technique to use machine learning to normalize disease names and also the first method that uses pairwise learning to rank in a normalization task. DNorm achieved the best performance in the 2013 ShARe/CLEF shared task on disease normalization in clinical notes [Leaman et al., 2011].

MetaMap[7] aims to map biomedical texts to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in texts. Briefly, given a textual passage, MetaMap identifies candidate UMLS concepts and the corresponding spans of mentions. MetaMap uses a knowledge-intensive approach based on symbolic, natural-language processing (NLP) and computational-linguistic techniques. MetaMap is a highly configurable system for biomedical named entity recognition and UMLS normalization. According to our MetaMap use experience, in all cases we obtained entities already existing in the UMLS metathesaurus. We did not find any new entity variants, i.e. not existing in UMLS.

A key difference between entity linking and our methodology is that the terms already exist for entity linking. In our methodology, we seek to find new terms or new term forms for an entity. We find the best position in an ontology for these new terms. Experts must validate if this is a new term and concept, or if it is a synonym of another already existing term.

[6] http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/DNorm.html

[7] https://metamap.nlm.nih.gov/

10.5 Discussion

We have outlined works related to several domains. These domains are inscribed in the workflow that we designed for this part of the thesis. Concerning polysemy detection, few works seek to predict polysemy as true or false. Most studies seek to identify the senses as well.

Regarding the term sense induction, studies are classified in four categories: (i) Context clustering, ii) Word clustering, iii) Co-occurrence Graphs, and iv) Probabilistic clustering. Most works are based on the first three categories and for general domains. This leads to characterization of the dataset using bags-of-words and graphs of words, i.e. the most commonly used. For this problem, other kinds of dataset characterization have never been adopted, such as those proposed in the meta-learning field.

A major problem in term sense induction is the determination of the number of senses of a word. A second problem is the best choice of a clustering algorithm for textual datasets.

The few studies solving the determination of the number of senses problem have given poor results. This task relies on determination of the number of clusters for the clustering task. Many measures have been proposed to induce the number of clusters, several proposed measures are not used, because of the dataset type and because there is not a consensus to determine which ones are the best. The choice of clustering algorithm becomes easier to solve when the number of senses is a priori known.

As we mentioned previously, we focus on the extraction of paradigmatic relations as hyperonymy (fathers), synonymy, and hyponymy (sons). We do not identify the type of relation. This part is based on conventional approaches of semantic similarity between two biomedical terms. For us, a term can be single- or multi-word term. Each context is represented by the most important features occurring in the paragraphs that contain the term, forming a bag-of-words (vector). Then, a measure is applied to compute the similarity between the two vectors. For each new term, a concept is induced, which is not offered by the distributional thesaurus methodologies.

In our context, in order to automatically enrich biomedical ontologies/terminologies, we tackle all the identified steps without human intervention. In our workflow, the human operator (i.e. expert) can participate at the end of the entire workflow, suggesting and validating the options to enrich biomedical ontologies. To sum up, we take the new candidate terms extracted in the first part of this thesis. We hence propose a complete automatic workflow to achieve automatic enrichment, which has been divided in three steps: (i) Polysemy Detection, (ii) Term Sense Induction, and (iii) Semantic Linkage. Our approach enables identification of different places where the new candidate term could be added in an ontology. By using our approach, we overcome polysemy detection for new terms by using a new type of dataset characterization involving meta-learning. We also induce the sense of a term using bag-of-words and graph-based approaches, as well as semantic linkage. This workflow is applied in particular to the biomedical domain.

At the same time, we deal with special problems such as the prediction of the number of senses for new candidate terms using the same characterization offered by meta-learning.

In the following chapter, we describe each step of our workflow to achieve automatic biomedical ontology enrichment.

Chapter

11

Methodology

In the first part of this thesis, *Automatic Term Extraction*, we extracted new biomedical candidate terms as potential candidates for ontology enrichment. In this chapter, we describe the proposed methodology in order to enrich biomedical ontologies. The main objective of this second part is to induce concepts and find the correct position in an already existing biomedical ontology. The meaning or sense of a particular candidate term is induced according to the context in which those terms appear. A clustering task is done over the contexts, and the most important features of clusters are used to give a semantic orientation, hereafter called concept or sense. The semantic orientation helps determine the appropriate position of terms in an ontology. Our methodology for enriching biomedical ontologies involves three main steps, described in Figure 11.1, and in the sections hereafter: (1) Polysemy Detection, (2) Term Sense Induction, and (3) Semantic Linkage.

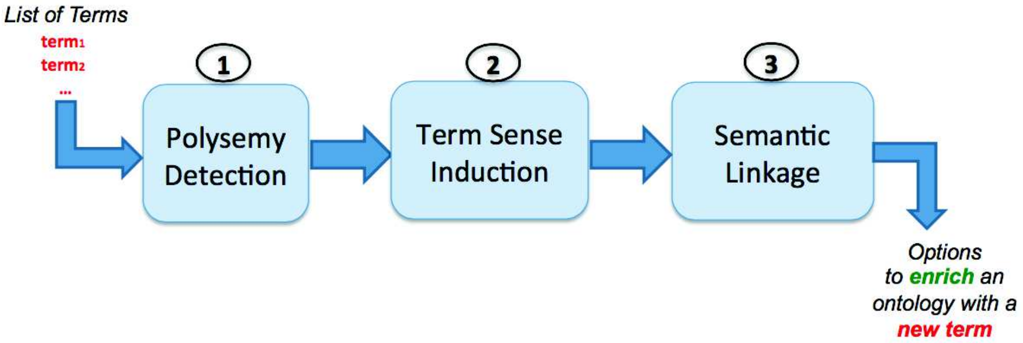


Figure 11.1: Workflow for Concept Extraction and Semantic Linkage.

11.1 Polysemy Detection (step 1)

In this section, we present the methodology proposed to determine if a biomedical term is polysemic. Polysemy is one of the most important issues of recent linguistic semantics, since the analysis of the polysemy process is essential for accurate reading, language acquisition, computational linguistics and similar tasks [Crossley et al., 2010]. In our case, polysemy is important for language acquisition in the biomedical domain via the enrichment of biomedical ontologies.

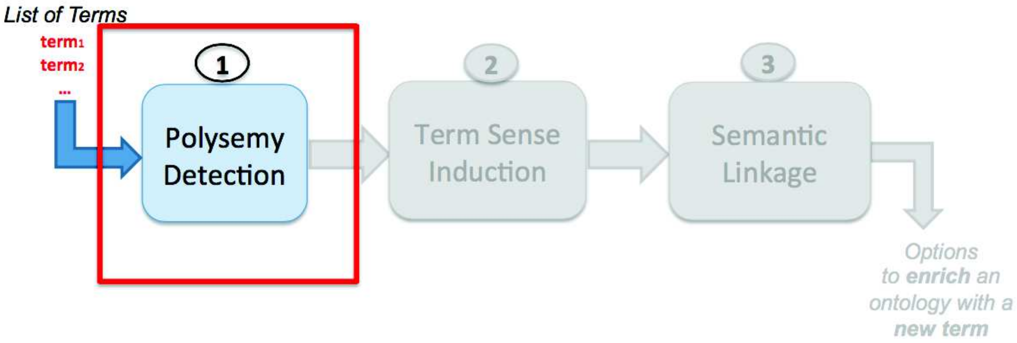


Figure 11.2: Solving the Polysemy Detection issue

As we mentioned in the previous chapter, most studies addressing polysemy detection and those tackling the word sense disambiguation/induction problem generally use the most well-known representation of the dataset based on bags-of-words or graphs-of-words. Meta-learning offers another way to characterize the dataset with the aim of performing a classification process. In this thesis, this characterization is called meta-features, which allows us to represent different types of datasets, i.e. to abstract into a single manner several kind of data and to apply many types of classification algorithms. In our case, these meta-features are used to detect biomedical polysemy via the additional information they can offer, i.e. the meta-data. For instance, the number of biomedical terms in a dataset can suggest the domain of the dataset (i.e. biomedicine).

Therefore, we opted to focus on meta-features with the objective of representing our dataset for the polysemy detection task. First, we present the main meta-features that serve to characterize the dataset. The dataset is the context of the new biomedical candidate terms. To create these new meta-features, we apply some statistical measures and we use UMLS[1] and AGROVOC[2] dictionaries, which are respectively a biomedical and an agronomic thesaurus. Our intuition behind the use of two different and related dictionaries at the same time, is to determine if a term appears in two different contexts. In this case, we can suppose this term could

[1] http://www.nlm.nih.gov/research/umls/

[2] http://aims.fao.org/agrovoc

be polysemic. Figure 11.3 shows the workflow of our approach, described hereafter.

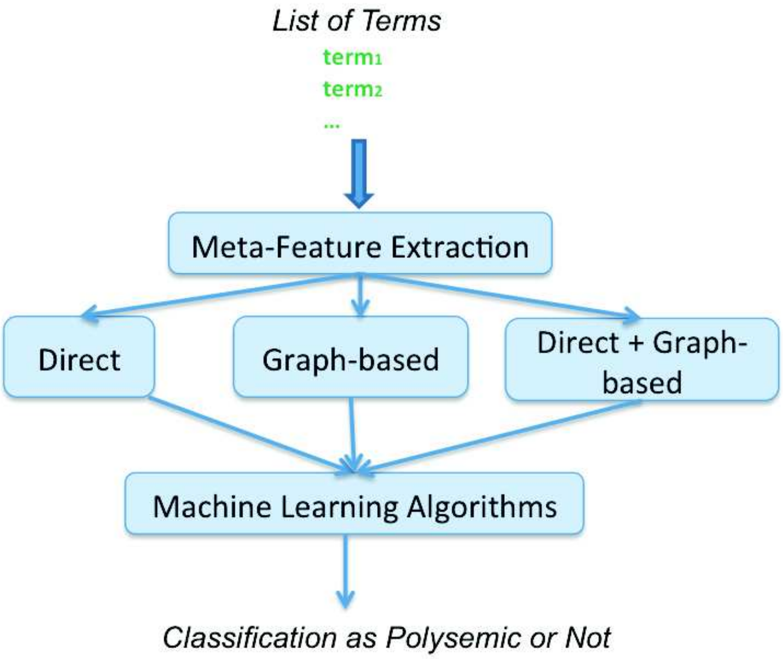


Figure 11.3: Workflow Methodology for Polysemy Prediction.

11.1.1 Meta-features

As cited in Section 10.3 there are two basic tasks in meta-learning: (i) the dataset characterization task (datasets), and (ii) the correlation between the dataset characterization and the optimal learning algorithm. The first task is to characterize datasets with meta-features, and the second is selecting appropriate algorithms for classification [Peng et al., 2002].

We present new meta-features based on statistical measures to characterize our biomedical dataset [Lossio-Ventura et al., 2016a]. They are extracted directly from the data and from a graph induced by these data. We select appropriate learning algorithms to determine if a term is polysemic. The main idea, as mentioned before, is to capture the characteristics of datasets from the structural shape and size of the graph induced from the data. Graphs are usually beneficial because they summarize and display information in a manner that is easy for most people to comprehend. Graphs are used in many domains, such as maths, social sciences, etc. Graphs are also used to concisely and clearly summarize data.
A total of 23 measures are proposed: 11 direct and 12 from the induced graph. Their effectiveness is illustrated by comparing the results obtained by different machine

learning algorithms.

Notation: for each term t, there is a set A_t of titles/abstracts extracted from Medline, $a \in A_t$, is a title/abstract associated with single or multiple senses for polysemic terms.

11.1.1.1 Direct Meta-features

These are extracted directly from the text dataset without any model construction (i.e. a tree model, a graph). They consist of statistic measures based mainly on counting some terms from UMLS and AGROVOC, i.e. the two dictionaries used.

1. **Number of Words:** represented as $nWords(t)$, is the number of words that contains the term t. For instance *nWords(Lung cancer)*$= 2$.

2. **Number of UMLS Terms:** represented by $termsU(t)$, i.e. the number of UMLS terms contained in the set of abstracts A_t.

3. **Minimum of UMLS Terms:** denoted as $minU(t)$, represents the minimum number of UMLS terms contained for each a of A_t.

   $$minU(t) = min(termsU(a_1), termsU(a_2), ...)$$

4. **Maximum of UMLS Terms:** denoted as $maxU(t)$, represents the maximum number of UMLS terms contained for each a of A_t.

   $$maxU(t) = max(termsU(a_1), termsU(a_2), ...)$$

5. **Mean of UMLS terms:** denoted as $meanU(t)$, represents the mean number of UMLS terms for each a of A_t.

   $$meanU(t) = \frac{1}{n} \times \sum_{i=1}^{n} termsU(a_i)$$

6. **Standard deviation of UMLS Terms:** denoted as $sdU(t)$, represents the standard deviation of the number of UMLS terms contained for each a of A.

   $$sdU(t) = \frac{1}{n-1} \times \sqrt{\sum_{i=1}^{n} (termsU(a_i) - meanU(t))^2}$$

7. **Number of AGROVOC Terms:** denoted as $termsA(t)$, represents the number of AGROVOC terms contained in the set of abstracts A_t of t.

8. **Minimum of AGROVOC Terms:** denoted as $minA(t)$, is the minimum number of AGROVOC terms contained in each a of A_t.

   $$minA(t) = min(termsA(a_1), termsA(a_2), ...)$$

9. **Maximum of AGROVOC Terms:** denoted as $maxA(t)$, is the maximum number of AGROVOC terms contained in each a of A_t.

$$maxA(t) = min(termsA(a_1), termsA(a_2), ...)$$

10. **Mean of AGROVOC Terms:** denoted as $meanA(t)$, represents the mean number of AGROVOC terms for each a of A_t.

$$meanA(t) = \frac{1}{n} \times \sum_{i=1}^{n} termsA(a_i)$$

11. **Standard deviation of AGROVOC Terms:** denoted as $sdA(t)$, represents the standard deviation of the number of AGROVOC terms contained for each a of A.

$$sdA(t) = \frac{1}{n-1} \times \sqrt{\sum_{i=1}^{n} (termsA(a_i) - meanA(t))^2}$$

The 11 meta-features defined previously are computed directly from the data. After that, for each candidate term context, we build a co-occurrence graph of terms. We thus represent the dataset via a graph, and compute new meta-features from these induced graphs. The meta-features based on graphs are described in the next section.

11.1.1.2 Graph-based Meta-features

As previously mentioned, we decided to use graphs to characterize a dataset. In fact, graphs are useful for summarizing and displaying information that is not highlighted by a bag-of-words characterization. We thus take advantage of the graph properties, such as, the neighborhood, the edge weights and size. We decided to represent the context of each candidate term as a graph. Hence, we built a graph for the context of each biomedical candidate term, with each graph being independent of the others.

Graph construction: A graph (see Figure 11.4) for each biomedical term is built (done in a similar way as defined in Section 5.3.1). Vertices denote terms, and edges denote co-occurrence relations between terms. Co-occurrences between terms are measured as the weight of the relation in the initial dataset. This relation is statistic-based by linking all co-occurring terms without considering their meaning or function in the text.

Each graph is built with the first 1 000 terms extracted with the BioTex application (see Chapter 8). The graph is undirected as the edges imply that terms simply co-occur, without any further distinction regarding their role. We take the *Dice coefficient* because, as explained in Section 5.3.1, this coefficient less penalizes the similarity value between two objects. We take this basic measure to compute the co-occurrence between two terms x and y, defined by the following formula:

$$D(x, y) = \frac{2 \times |x \cap y|}{|x| + |y|} \tag{11.1}$$

Where $|x|$ and $|y|$ are the number of abstracts in which we find x and y, respectively, and $|x \cap y|$ is the number of abstracts shared by the two terms; $D(x, y)$ is the similarity quotient that ranges from 0 to 1.

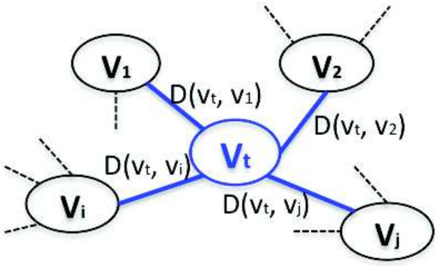


Figure 11.4: Graph created for the t term.

In Figure 5.9, v_t represents the vertex with the t term, v_i represents a vertex i in the graph (term), N(v_i) the neighborhood of v_i, $|\mathrm{N}(v_i)|$ the number of neighbors of v_i, r_j the neighbor j of v_i, $weight(v, r_j)$ the edge weight between v_i and its neighbor r_j, so $weight(v_i, r_j) = D(v_i, r_j)$.
Hereafter, we define the new proposed meta-features based on graphs and show how to compute them.

1. **Number of Neighbors:** represented as *ng*, is the number of neighbors of vertex v_t in the induced graph.

   $$ng(v_t) = |\mathrm{N}(v_t)|$$

2. **Sum of Edge Weights:** denoted *sum*, represents the sum of edge weights specifically for the vertex v_t in the graph created for t.

   $$sum(v_t) = \sum_{j=1}^{ng(v_t)} weight(v_t, r_j)$$

3. **Minimum Number of Neighbors:** denoted *minNG*, represents the minimum number of neighbors of all v_i in the graph created for t.

   $$minNG(t) = min(ng(v_1), ng(v_2), ...)$$

4. **Maximum Number of Neighbors:** denoted *maxNG*, represents the maximal number of neighbors of all v_i in the graph created for t.

   $$maxNG(t) = max(ng(v_1), ng(v_2), ...)$$

5. **Mean Number of Neighbors:** denoted *meanNG*, represents the mean number of neighbors of all v_i in the graph created for t.

$$meanNG(t) = \frac{\sum_{i=1}^{1000} ng(v_i)}{1000}$$

6. **Standard deviation of the Number of Neighbors:** denoted *sdNG*, represents the standard deviation of the number of neighbors of all v_i in the graph created for t.

$$sdNG(t) = \frac{\sqrt{\sum_{i=1}^{1000} (ng(v_i) - meanNG(t))^2}}{1000-1}$$

7. **Min Sum of Edge Weights:** denoted *minSUM*, represents the minimum sum of edge weights of all v_i on the graph created for t.

$$minSUM(t) = min(sum(v_1), sum(v_2), ...)$$

8. **Max Sum of Edge Weights:** denoted *maxSUM*, represents the maximum sum of edge weights of all v_i on the graph created for t.

$$maxSUM(t) = max(sum(v_1), sum(v_2), ...)$$

9. **Mean Sum of Edge Weights:** denoted *meanSUM*, represents the mean sum of edge weights of all v_i in the graph created for t.

$$meanSUM(t) = \frac{\sum_{i=1}^{1000} sum(v_i)}{1000}$$

10. **Standard deviation of the Sum of Edge Weights:** denoted *sdSUM*, represents the standard deviation of the sum of edge weights of all v_i in the graph created for t.

$$sdSUM(t) = \frac{\sqrt{\sum_{i=1}^{1000} (sum(v_i) - meanSUM(t))^2}}{1000-1}$$

11. **Number of Neighbors in UMLS:** represented as *ngUMLS*, is the number of neighbor terms with the vertex v_t in the graph and in turn in UMLS.

$$ngUMLS(v_t) = |\mathrm{N}(v_t)|_{r_j \in UMLS}$$

12. **Sum of Edge Weights in UMLS:** as *sumUMLS*, represents the sum of edge weights for v_t with its neighbors that are in UMLS.

$$sumUMLS(v_t) = \sum_{j=1}^{ngUMLS(v_t)} weight(v_t, r_j)$$

To illustrate the new proposed meta-features, we show an example of the entire process, step by step to compute the direct and graph-based meta-features in the next section (see Section 11.1.1.3).

11.1.1.3 Example

As mentioned, we illustrate how to compute measures to characterize a dataset. Table 11.1 shows a small extract of our dataset, a set of three titles/abstract for the term *yellow fever*. The UMLS terms found in that text are in blue and AGROVOC terms are in red. Table 11.2 presents the values obtained with the direct meta-features. Note that the term *yellow fever* exists in both UMLS and AGROVOC.

Figure 11.5 shows a subgraph created with the entire set of titles/abstracts for the term *yellow fever*. This subgraph shows the UMLS terms encircled in blue. Table 11.3 presents the values obtained with the graph-based meta-features.

Id	Title/Abstract
a_1	"***Risks*** of ***travel***, benefits of a ***specialist*** consult. If ***patients*** are planning to ***travel*** to developing countries, their ***primary care*** ***physicians*** can advise them on various medical ***risks***, especially traveler's ***diarrhea***, and offer to update their ***immunizations***. However, travelers to areas where there is a ***risk*** of ***malaria***, <*t*>***yellow fever***</*t*>, or other tropical ***diseases*** should be referred to a ***specialist***."
a_2	"***Herpes zoster*** after <*t*>***yellow fever***</*t*> ***vaccination***. An immunocompetent 64-year-old ***women*** presented with brachial ***herpes zoster*** (HZ) ***infection*** 3 days after ***vaccination*** against ***yellow fever*** (YF). The lesions disappeared after antiviral ***treatment***. There are very few ***reports*** of a possible ***association*** between YF ***vaccination*** and HZ ***infection***. This case highlights the importance of continuing ***surveillance*** of vaccine adverse events."
a_3	"Broadening the horizons for <*t*>***yellow fever***</*t*>: new uses for an old vaccine. The vaccine against ***yellow fever*** is one of the safest and most effective ever developed. With an outstanding record in ***humans***, has this live attenuated vaccine been overlooked as a promising vector for the ***development*** of ***vaccines*** against pathogens outside its own genus? Recent studies, including a ***report*** by Tao et al. on page 201 of this issue, have sparked renewed interest."

Table 11.1: An Extract of the Titles/Abstracts dataset for the term **Yellow Fever**

Terms found in UMLS (27): association, development, diarrhea, diseases, fever, herpes zoster, humans, immunizations, infection, malaria, patients, physicians, primary care, primary care physicians, report, reports, risk, risks, specialist, surveillance, travel, treatment, vaccination, vaccines, women, yellow fever, zoster.

Terms found in AGROVOC (22): countries, developing countries, development, diseases, events, fever, genus, humans, infection, lesions, malaria, pathogens, patients, physicians, planning, reports, risk, uses, vaccination, vaccines, women, yellow fever.

Item	Measure	Comment
1	$nWords(t) = 2$	"yellow fever" contains two words.
2	$termsU(t) = 27$	The total number of distinct UMLS terms found in the set of title/abstracts
3	$minU(t) = 6$	The minimum number of distinct UMLS terms found for each abstract a_i of t. That means $min(termsU(a_1), termsU(a_2), termsU(a_3))$ $= min(13, 11, 6) = 6$
4	$maxU(t) = 13$	The maximum number of distinct UMLS terms found for each abstract a_i of t. That means $max(termsU(a_1), termsU(a_2), termsU(a_3))$ $= max(13, 11, 6) = 13$
5	$meanU(t) = 12$	$\frac{\sum_{i=1}^{n} termsU(a_i)}{n} = \frac{\sum_{i=1}^{3} termsU(a_i)}{3} = \frac{13+11+6}{3} = 10$
6	$sdU(t) = 2.55$	$\frac{\sqrt{\sum_{i=1}^{n} (termsU(a_i) - meanU(t))^2}}{n-1} = \frac{\sqrt{\sum_{i=1}^{3} (termsU(a_i) - 10)^2}}{3-1} = \frac{\sqrt{(13-10)^2+(11-10)^2+(6-10)^2}}{2} = 2.55$
7	$termsA(t) = 22$	The total number of distinct AGROVOC terms found in the set of titles/abstracts
8	$minA(t) = 8$	The minimum number of distinct AGROVOC terms found for each abstract a_i of t. That means $min(termsA(a_1), termsA(a_2), termsA(a_3)) = min(10, 8, 8) = 8$
9	$maxA(t) = 10$	The minimum number of distinct AGROVOC terms found for each abstract a_i of t. That means $max(termsA(a_1), termsA(a_2), termsA(a_3)) = max(10, 8, 8) = 10$
10	$meanA(t) = 8.67$	$\frac{\sum_{i=1}^{n} termsA(a_i)}{n} = \frac{\sum_{i=1}^{3} termsA(a_i)}{3} = \frac{10+8+8}{3} = 8.67$
11	$sdA(t) = 0.82$	$\frac{\sqrt{\sum_{i=1}^{n} (termsA(a_i) - meanA(t))^2}}{n-1} = \frac{\sqrt{\sum_{i=1}^{3} (termsA(a_i) - 8.67)^2}}{3-1} = \frac{\sqrt{(10-8.67)^2+(8-8.67)^2+(8-8.67)^2}}{2} = 0.82$

Table 11.2: Direct Meta-features for t = **Yellow Fever**

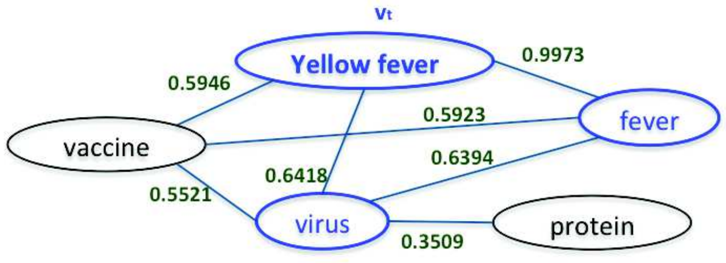


Figure 11.5: Subgraph created for the term t = **Yellow Fever**

Item	Measure	Comment
1	$ng(v_t) = 3$	the vertex v_t “yellow fever” has three neighbors.
2	$sum(v_t) = 2.2337$	$\sum_{j=1}^{ng(v_t)} weight(v_t, r_j) = \sum_{j=1}^{3} weight(v_t, r_j) = weight(\text{yellow fever, vaccine}) + weight(\text{yellow fever, fever}) + weight(\text{yellow fever, virus}) = 0.5946 + 0.9973 + 0.6418 = 2.2337$
3	$minNG(t) = 1$	The minimum number of neighbors of all v_i in the graph created for t. That means $min(ng(v_t), ng(v_1), ng(v_2), ng(v_3), ng(v_4)) =$ $min(ng(\text{yellow fever}), ng(\text{vaccine}), ng(\text{fever}), ng(\text{virus}),$ $ng(\text{protein})) =$ $min(3, 3, 3, 4, 1) = 1$
4	$maxNG(t) = 4$	The maximum number of neighbors of all v_i in the graph created for t. That means $max(ng(v_t), ng(v_1), ng(v_2), ng(v_3), ng(v_4)) =$ $max(ng(\text{yellow fever}), ng(\text{vaccine}), ng(\text{fever}), ng(\text{virus}),$ $ng(\text{protein})) =$ $max(3, 3, 3, 4, 1) = 4$
5	$meanNG(t) = 2.8$	$\frac{\sum_{i=1}^{5} ng(v_i)}{5} =$ // in this case 5 instead of 1000, because the subgraph contains only 5 vertices. $\frac{\sum_{i=1}^{5} ng(v_i)}{5} = \frac{3+3+3+4+1}{5} = 2.8$
6	$sdNG(t) = 0.55$	$\frac{\sqrt{\sum_{i=1}^{5} (ng(v_i) - meanNG(t))^2}}{5-1} =$ // in this case 5 instead of 1000, because the subgraph contains only 5 vertices. $\frac{\sqrt{\sum_{i=1}^{5} (ng(v_i) - 2.8)^2}}{5-1} =$ $\frac{\sqrt{(3-2.8)^2+(3-2.8)^2+(3-2.8)^2+(4-2.8)^2+(1-2.8)^2}}{4} = 0.55$
7	$minSUM(t) = 0.3509$	The minimum sum of edge weights of all v_i in the graph created for t. That means $min(sum(v_t), sum(v_1), sum(v_2), sum(v_3), sum(v_4)) =$ $min(sum(\text{yellow fever}), sum(\text{vaccine}), sum(\text{fever}), sum(\text{virus}),$ $sum(\text{protein})) =$ $min(2.2337, 1.739, 2.229, 2.1842, 0.3509) = 0.3509$
8	$maxSUM(t) = 2.2337$	The minimum sum of edge weights of all v_i in the graph created for t. That means $max(sum(v_t), sum(v_1), sum(v_2), sum(v_3), sum(v_4)) =$ $max(sum(\text{yellow fever}), sum(\text{vaccine}), sum(\text{fever}), sum(\text{virus}),$ $sum(\text{protein})) =$ $max(2.2337, 1.739, 2.229, 2.1842, 0.3509) = 2.2337$
9	$meanSUM(t) = 1.7474$	$\frac{\sum_{i=1}^{5} sum(v_i)}{5} =$ // in this case 5 instead of 1000, because the subgraph contains only 5 vertices. $\frac{\sum_{i=1}^{5} sum(v_i)}{5} = \frac{2.2337+1.739+2.229+2.1842+0.3509}{5} = 1.7474$
10	$sdSUM(t) = 0.40$	$\frac{\sqrt{\sum_{i=1}^{5} (sum(v_i) - meanSUM(t))^2}}{5-1} =$ // in this case 5 instead of 1000, because the subgraph contains only 5 vertices. $\frac{\sqrt{\sum_{i=1}^{5} (sum(v_i) - 1.7474)^2}}{5-1} =$ $= 0.40$
11	$ngUMLS(v_t) = 2$	the vertex v_t “yellow fever” has two neighbors (fever and virus) that are in UMLS.
12	$sumUMLS(v_t) = 1.6391$	$\sum_{j=1}^{ngUMLS(v_t)} weight(v_t, r_j) = \sum_{j=1}^{2} weight(v_t, r_j) =$ $weight(weight(\text{yellow fever, fever}) + weight(\text{yellow fever, virus}) = 0.9973 + 0.6418 = 1.6391$

Table 11.3: Graph-based Meta-features for t = **Yellow Fever**

We have just seen how to compute the new meta-features in Tables 11.2, 11.3. The next step is to apply some machine learning algorithms to classify the new candidate terms as polysemic or not. We describe the algorithm used on these meta-features in the next section (see Section 11.1.2). Experiments are reported in Section 12.1.2.

11.1.2 Machine Learning Algorithms

We use some well-known supervised algorithms, the input data for these algorithms is called training data and has a known label or result in our case polysemic/not-polysemic. A model is prepared through a training process where it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. The algorithms that we use are implemented in Weka[3] software with default values for each algorithm, such as:

- Naive Bayes (NB) [John and Langley, 1995]
- AdaBoost (AB)
- Tree Decision (TD) [Quinlan, 1993]
- Support Vector Machine (SVM) [Platt, 1999]
- Meta Bagging (MB)
- M5P Tree (M5P)
- Multilayer Perceptron (NN)
- MultiClassClassifier Logistic (MCC)

Interested readers may refer to [Hall et al., 2009] for further details on these machine learning approaches.

To sum up, this section describes the predictive process if a new biomedical candidate term is polysemic or not. The next step is to induce a possible sense or senses for these new candidate terms. Note that the workflow to induce possible senses for a polysemic term differs from the workflow to induce a single sense of a non-polysemic term. The following section describes the process to induce the sense(s) (see Section 11.2).

11.2 Term Sense Induction (step 2)

The objective of this step is to induce multiple or single sense(s) (concept) of a polysemic and not polysemic new biomedical candidate term, respectively, in order to conceptualize them. As we mentioned previously, the concepts or senses are extracted according the context of the terms. A clustering task is applied on the context, and the most important features of clusters are used to give a semantic orientation.

For this, we carry out two tasks. First, (i) *Number of senses prediction:* This task is performed only for the candidate terms predicted as polysemic in the previous section. As these candidate terms have been predicted as polysemic, it is necessary to determine the exact number of senses these terms may have according to the corpus context. Then, (ii) *Clustering for concept induction:* This task carries out

[3] `http://www.cs.waikato.ac.nz/ml/weka/`

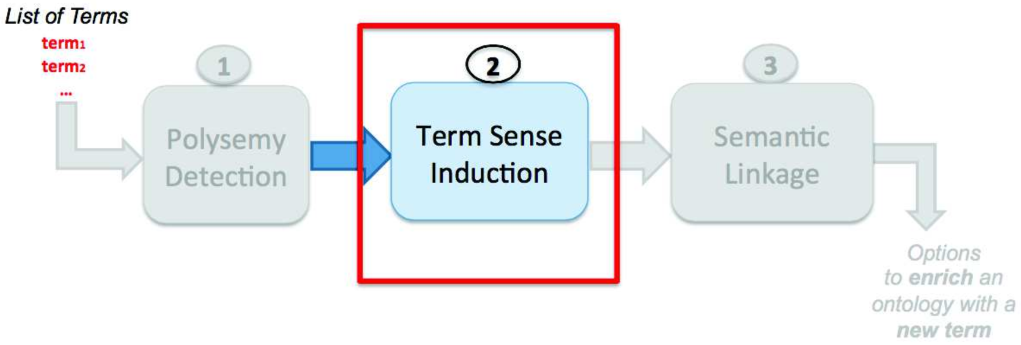


Figure 11.6: Solving the Term Sense Induction.

a clustering algorithm taking the predicted k as input, and then for each cluster it selects the most important features, which represent the formed concept. Note that $k = 1$ when the candidate term is not polysemic. Figure 11.7 shows the workflow of our approach, described hereafter.

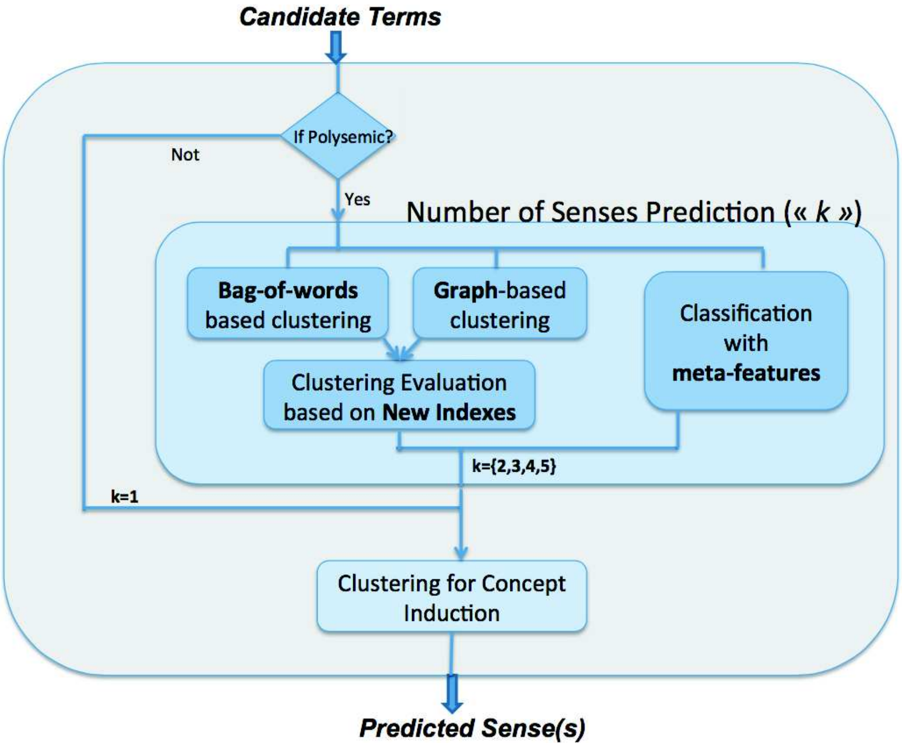

Candidate Terms
If Polysemic?
Not
Yes
Number of Senses Prediction (« k »)
Bag-of-words based clustering
Graph-based clustering
Classification with meta-features
Clustering Evaluation based on New Indexes
k={2,3,4,5}
k=1
Clustering for Concept Induction
Predicted Sense(s)

Figure 11.7: Term Sense Induction Workflow.

11.2.1 Sense Number Prediction

The prediction of the sense number of a term falls directly under clustering-based issues. In clustering tasks, one of the most difficult problems is to determine the number of clusters k, which is a basic input parameter for most clustering algorithms. Many algorithms have been proposed to solve this problem, but mostly for the general domain. One limitation of these approaches is that they tend to predict a high k value. In contrast, in the biomedical domain, according to the statistics on UMLS, polysemic terms tend to be linked to only 2 and 5 senses (2 and 5 clusters). Therefore, as we aim to identify possible senses for a new biomedical candidate term, we will limit the number of senses to between 2 and 5.

Table 11.4 shows the statistical details of polysemic terms in UMLS for English, French, and Spanish. The English version of UMLS contains about 9 919 000 distinct terms, about 54 257 of which are polysemic. This means that for approximately 200 biomedical terms there is just 1 polysemic term. So these statistics confirm that in the biomedical domain there are more non-polysemic terms (monosemous) than polysemic terms for the three languages (English, French, and Spanish). Table 11.5 shows the statistical details of polysemic terms in MeSH. The analysis is similar.

# of Senses	***English***	***French***	***Spanish***
2	54 257	1 292	10 906
3	7 770	36	414
4	1 842	1	56
5+	1 677	1	18

Table 11.4: Details of Polysemic Terms in UMLS.

# of Senses	***English***	***French***	***Spanish***
2	178	11	0
3	1	0	0
4	0	0	0
5+	0	0	0

Table 11.5: Details of Polysemic Terms in MeSH.

Three different ways are proposed to determine the number of senses: (i) Executing clustering algorithms varying k over the bag-of-words representation of the data and evaluating the quality indexes, (ii) Executing clustering algorithms varying k over the graph created as defined in Section 11.1.1.2 and evaluating the quality indexes, and (iii) Using the meta-features proposed in Section 11.1.1 and applying supervised algorithms for classification.

For (i) and (ii), we use the CLUTO[4] application, which is a software program for clustering of high-dimensional datasets, from which we select 5 clustering algorithms, included in *partitional, agglomerative,* and *graph-partitioning* types, and we implement them by varying the number of k clusters from between 2 and 5. We then evaluate the quality of clustering solutions, to finally take the best k. This process allows us to determine the best k value. To evaluate the quality of clustering solutions, we propose new indexes computed on the internal and external similarity of clusters obtained by CLUTO's clustering algorithm using an objective function. Objective functions for clustering are introduced below.

Objective function rates the **global** quality of a clustering [Booth et al., 2008], and is called a quality measure. We can obtain the optimal clustering by optimizing (i.e. maximize/minimize) this objective function. Each of the measures have strengths and weaknesses. Optimizing each of the measures is known to be an NP-hard problem. Hence, many efficient algorithms that have been claimed to solve the optimal problem with polynomial-time complexity yield sub-optimal clustering. The objective functions used with algorithms are as follows:

$$I1 = maximize \sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{v,u \in S_i} sim(v,u) \right)$$

$$I2 = maximize \sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} sim(v,u)}$$

$$E1 = minimize \sum_{i=1}^{k} n_i \frac{\sum_{v \in S_i, u \in S} sim(v,u)}{\sqrt{\sum_{v,u \in S_i} sim(v,u)}}$$

$$H1 = maximize \frac{I1}{E1}$$

$$I2 = maximize \frac{I2}{E1}$$

Where k is the number of clusters, S the total number of objects to be clustered, S_i the set of objects assigned to the i^{th} cluster, n_i the number of objects in the i^{th} cluster, v and u represents two objects, and $sim(v,u)$ the similarity between two objects.

Clustering algorithms generated from the given objective functions are shown, with a number of examples of widely used approaches discussed in Section 12.

[4] `http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview`

11.2.1.1 Definition of New Internal indexes

To evaluate the clustering solutions, there are several indexes, which are also called validity indexes. These indexes are categorized in two classes: external and internal. External indexes use pre-labelled datasets with "known" cluster configurations and measure how well clustering techniques perform with respect to these known clusters. Internal indexes are used to evaluate the "goodness" of a cluster configuration without any prior knowledge of the nature of the clusters.

We use the following measures: (i) the intra-cluser similarity (*ISIM*), and (ii) the inter-cluster similarity (*ESIM*), in order to create new indexes [Lossio-Ventura et al., 2016c].

These new internal indexes are computed from the results obtained by using an objective function. They focus on choosing the minimum or maximum value of a criterion. This gives us an idea as to whether the obtained clusters are homogeneous.

New internal indexes are described below:

Notation: $\mid S_i \mid$ is the number of objects assigned to the i_{th} cluster, $OF = I1, I2, E1, H1, H2$ the objective function used by the clustering algorithm.

1. **Average ISIM:** represented as $a_{k,OF}$, is the average of the *ISIM* value of each cluster of a solution clustering with number of clusters = k.

$$a_{k,OF} = \frac{\sum_{i=1}^{k} ISIM_i}{k}$$

2. **Average ESIM:** represented as $b_{k,OF}$, is the average of the *ESIM* value of each cluster of a solution clustering with number of clusters = k.

$$b_{k,OF} = \frac{\sum_{i=1}^{k} ESIM_i}{k}$$

3. **Average of the difference between ISIM and ESIM:** represented as $c_{k,OF}$, is the average of the difference between *ISIM* and *ESIM* multiplied by the number of objects in such clusters $\mid S_i \mid$.

$$c_{k,OF} = \frac{1}{k} \sum_{i=1}^{k} \mid S_i \mid \times (ISIM_i - ESIM_k)$$

4. **Division between the ISIM sum and ESIM sum:** represented as $e_{k,OF}$, is the division between the sum of *ISIM* multiplied by the number of objects in such clusters $\mid S_i \mid$, and the sum of *ESIM* multiplied by the number of objects in such cluster.

$$e_{k,OF} = \frac{\sum_{i=1}^{k} \mid S_i \mid \times ISIM_k}{\sum_{i=1}^{k} \mid S_i \mid \times ESIM_i}$$

5. **Global objective function divided by the logarithm:** represented as $f_{k,OF}$, is the division between the value of the objective function (OF) and the logarithm of k to the base 10.

$$f_{k,OF} = \frac{OF}{\log_{10}(k)}$$

For each clustering solution, we are able to predict the number of senses, with different types of dataset representation. Note that these indexes can only be applied to the clustering solutions. So we evaluated these indexes on two kinds of representation: i) Bag-of-words representation (see Section 11.2.1.3), and ii) Graph-of-words representation (see Section 11.2.1.4).

11.2.1.2 Clustering Algorithms to Evaluate New Internal Indexes

We evaluate the prediction of the number of clusters with the proposed internal indexes based on five clustering algorithms of classes *partitional, agglomerative,* and *graph-partitioning based*. For this, as mentioned previously, we characterize our dataset with the bag-of-words and graph-of-words representation. Then we compute the values of our new internal indexes.

We use five well-known clustering algorithms implemented in CLUTO[5] software, such as:

- **rb:** In this method, the desired *k-way* clustering solution is computed by performing a sequence of $k - 1$ repeated bisections. In this approach, the objects are first clustered into two groups, then one of these groups is selected and further bisected. This process continuous until the desired number of clusters is found. During each step, the cluster is bisected so that the resulting *2-way* clustering solution optimizes a particular clustering criterion function. Note that this approach ensures that the criterion function is locally optimized within each bisection, but in general is not globally optimized.
- **rbr:** In this method, the desired *k-way* clustering solution is computed in a fashion similar to the repeated-bisecting method but ultimately the overall solution is globally optimized.
- **direct:** In this method, the desired *k-way* clustering solution is computed by simultaneously finding all k clusters. In general, computing a *k-way* clustering directly is slower than clustering via repeated bisections. In terms of quality, for reasonably small k values (usually less than 10–20), the direct approach leads to better clusters than those obtained via repeated bisections. However, as k increases, the repeated-bisecting approach tends to be better than direct clustering.

[5] `http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview`

- **agglo:** In this method, the desired *k-way* clustering solution is computed using the agglomerative paradigm whose goal is to locally optimize (minimize or maximize) a particular clustering criterion function. The solution is obtained by stopping the agglomeration process when k clusters are left. In this kind of algorithm, we find the *k-means* algorithm.
- **graph:** In this method, the desired *k-way* clustering solution is computed by first modeling the objects using a nearest-neighbor graph (each object becomes a vertex, and each object is connected to its most similar other objects), and then the graph is split into *k-clusters* using a *min-cut* graph partitioning algorithm.

11.2.1.3 Prediction of *k* with Bag-of-words Representation

We represent our dataset as a "bag-of-words or vector of terms", we prefer to call them terms because these terms are composed of one or more words. Therefore, to select these terms we used the BioTex application (see Chapter 8), while selecting only the first 3 000 terms and using the *LIDF-value* measure to rank the terms. Then we apply clustering algorithms with *cosine* as similarity measures, a well-known similarity used on textual datasets. Given two vectors of terms p and q, the cosine measure, is represented using a dot product and magnitude as:

$$sim(p,q) = \cos(\mathbf{p},\mathbf{q}) =$$
$$\frac{\mathbf{pq}}{\|\mathbf{p}\|\|\mathbf{q}\|} = \frac{\sum_{i=1}^{n}\mathbf{p}_i\mathbf{q}_i}{\sqrt{\sum_{i=1}^{n}(\mathbf{p}_i)^2}\sqrt{\sum_{i=1}^{n}(\mathbf{q}_i)^2}}$$

Finally, to choose the number of senses, we select k of the maximum or minimum value of each index computed previously. Note that the internal similarity (*ISIM*) must be maximal, and the external similarity (*ESIM*) must be minimal. Thus, for $k = 2, 3, 4, 5$, we do:

1. $max(a_{k,OF})$**:** we choose the maximal value of the *ISIM* average of all clusters. The main idea behind this decision is that the clustering solution should have a high average internal similarity value.

$$max(a_{k,OF}) = max(a_{2,OF}, a_{3,OF}, a_{4,OF}, a_{5,OF})$$

2. $min(b_{k,OF})$**:** we choose the minimal value of the *ESIM* average of all clusters, which means that the clustering solution should have a low average external similarity value.

$$min(b_{k,OF}) = min(b_{2,OF}, b_{3,OF}, b_{4,OF}, b_{5,OF})$$

3. $max(c_{k,OF})$**:** we choose the maximal value for the average result of the subtraction between *ISIM* and *ESIM* of each cluster. The main reason for this

decision is that the clustering solution should have a high difference between *ISIM* and *ESIM*, showing that each cluster is compacter and the clusters are well separated.

$$max(c_{k,OF}) = max(c_{2,OF}, c_{3,OF}, c_{4,OF}, c_{5,OF})$$

4. $max(e_{k,OF})$**:** we choose the maximal average value of the division between *ISIM* and *ESIM* of all clusters. The main reason for this decision is that the clustering solution should show a high quotient between *ISIM* and *ESIM*, showing that each cluster is compacter and the clusters are well separated.

$$max(e_{k,OF}) = max(e_{2,OF}, e_{3,OF}, e_{4,OF}, e_{5,OF})$$

5. $max(f_{k,OF})$**:** we choose the maximal value of the division between the objective function and the logarithm of k. The hypothesis underlying this index is that the objective function is stronger when more clusters are included for the solution, so we try to reduce this drawback via the logarithm of the number of clusters.

$$max(f_{k,OF}) = max(f_{2,OF}, f_{3,OF}, f_{4,OF}, f_{5,OF})$$

11.2.1.4 Prediction of *k* with Graph Representation

As we previously mentioned, we represent our dataset as a graph of co-occurrences between terms. To predict the number of senses, we use the graph created previously, as defined in Section 11.1.1.2. Then we apply the clustering algorithms with *Dice coefficient* as a similarity measure (measure used to create the co-occurrence graph). Finally, to choose the number of senses: we select k of the maximum or minimum value of each previously computed index, the analysis is the same as in the previous section (see Section 11.2.1.3). Thus, for $k = 2, 3, 4, 5$:

$$max(a_{k,OF}) = max(a_{2,OF}, a_{3,OF}, a_{4,OF}, a_{5,OF})$$

$$min(b_{k,OF}) = min(b_{2,OF}, b_{3,OF}, b_{4,OF}, b_{5,OF})$$

$$max(c_{k,OF}) = max(c_{2,OF}, c_{3,OF}, c_{4,OF}, c_{5,OF})$$

$$max(e_{k,OF}) = max(e_{2,OF}, e_{3,OF}, e_{4,OF}, e_{5,OF})$$

$$max(f_{k,OF}) = max(f_{2,OF}, f_{3,OF}, f_{4,OF}, f_{5,OF})$$

11.2.1.5 Prediction of *k* with Meta-feature Representation

As previously carried out for polysemy detection, we represent our dataset with meta-features. Based on the meta-features computed as explained in Section 11.1.1, we apply the supervised classification algorithms to predict the number of senses.

The result of Section 11.2.1 is the prediction of the number of possible senses "k" for a polysemic candidate term. The next step is to induce the senses of the candidate term. The next section (see Section 11.2.2) describes how to induce that concept with the predicted k. Note that $k = 1$ when the candidate term is not polysemic. These new internal indexes have also been used to evaluate tweet clustering [Lossio-Ventura et al., 2016b].

11.2.2 Clustering for Concept Induction

When using cluster analysis on a dataset to gather similar cases, it is necessary to choose among a large number of clustering methods and distance measures. Sometimes, one choice might influence another, but there are many possible combinations of methods. A real problem is to find the number of clusters k, which is tackled only for the biomedical domain in Section 11.2.1.

This concerns the final task to induce possible senses for a candidate term. Therefore, for this task, we select the *k-means* clustering algorithm that proved to perform well for textual data [Jing et al., 2006, Aggarwal and Zhai, 2012, Kinnunen et al., 2011], while taking the number of sense(s) "k" computed in the preceding section as input. Then we extract the most relevant features of each cluster, which represents the formed concept. From the set of clusters (concepts) we take the cluster containing the highest number of biomedical terms to continue to next step. This is the simplest methodology and we intend to improve this process in a future study.

If the candidate term is not polysemic, then $k = 1$. Therefore it is not possible to apply a clustering algorithm, so we then extract the most relevant features of the only context of the candidate term, which represents the formed concept.

At the end of this step, we have the candidate term with its induced concepts. The next step is to find the position in a biomedical ontology where the candidate term could be added. The induced concept (one or several) is useful to determine the position, because it will be used to evaluate the semantic linkage with other terms of an ontology. The next section (see Section 11.3) describes the methodology proposed to add a candidate term to a biomedical ontology.

11.3 Semantic Linkage (step 3)

This section introduces how a new biomedical candidate term could be added in an existing biomedical ontology, i.e. how to find the correct position in the ontology. This is a first way to extract the type of relation between a new biomedical candidate term and an already existing term from an ontology.

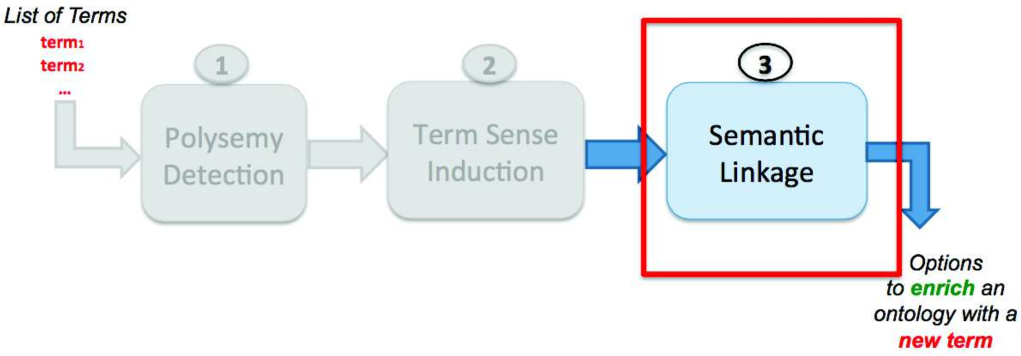
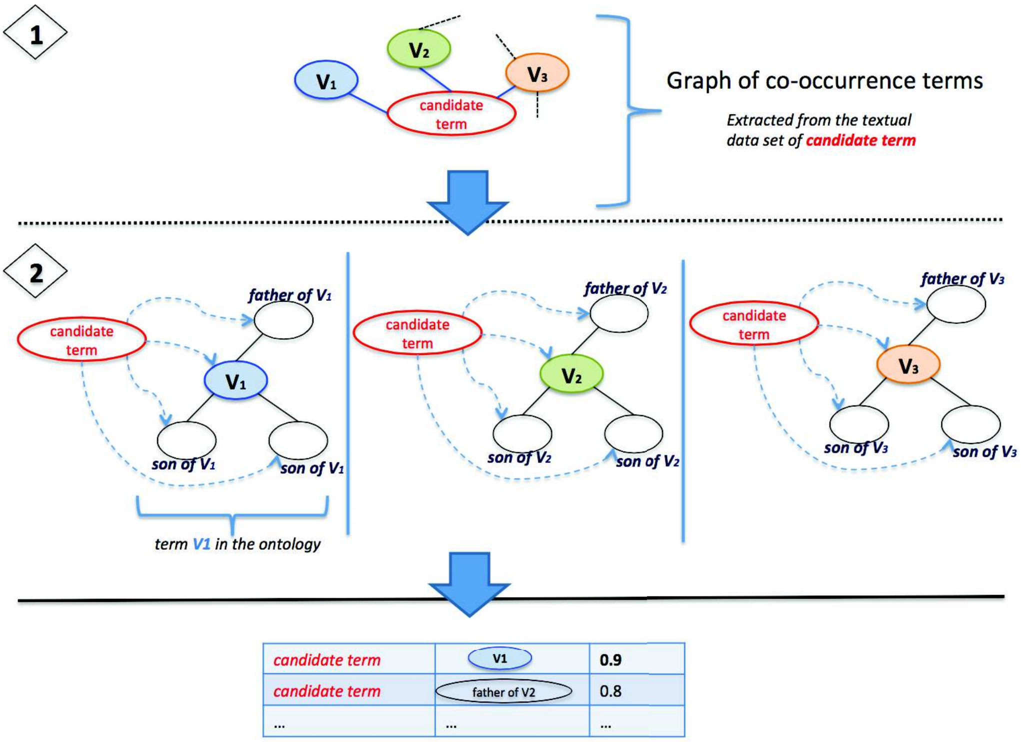


Figure 11.8: Semantic Linkage of a Candidate Term.

The idea underlying this process is, given a candidate term and its graphs of term co-occurrence (as defined in Section 11.1.1.2), (1) we select the MeSH neighborhood, then (2) we evaluate the semantic similarity of the candidate term with: (i) its MeSH neighbors, and, (ii) the fathers/sons of those neighbors in the MeSH ontology. Finally, a list of terms is proposed where the new biomedical candidate term could be positioned. Figure 11.9 illustrates the principle to semantically link the candidate terms.



candidate term	V1	0.9
candidate term	father of V2	0.8
...	...	...

Figure 11.9: Semantic Linkage Workflow.

The semantic linkage is based essentially on context similarity between the new biomedical candidate term and those appearing in an ontology. The following paragraphs describe in detail the steps shown in the proposed workflow for the semantic linkage task. As mentioned, the objective of this approach is to propose a list of terms where the candidate term could be added. This list is ranked by their cosine similarity between contexts.

Our methodology for semantic linkage follows two main steps:

1. **Selecting the MeSH neighborhood:** First, we take the neighbors of the new biomedical candidate term from the co-occurrence graph of terms. These graphs are built as defined in Section 11.1.1.2. Second, we just choose neighbors belonging to a biomedical ontology, this is the ontology we want to enrich, i.e. the MeSH ontology in our case. Figure 11.10 illustrates this first step.

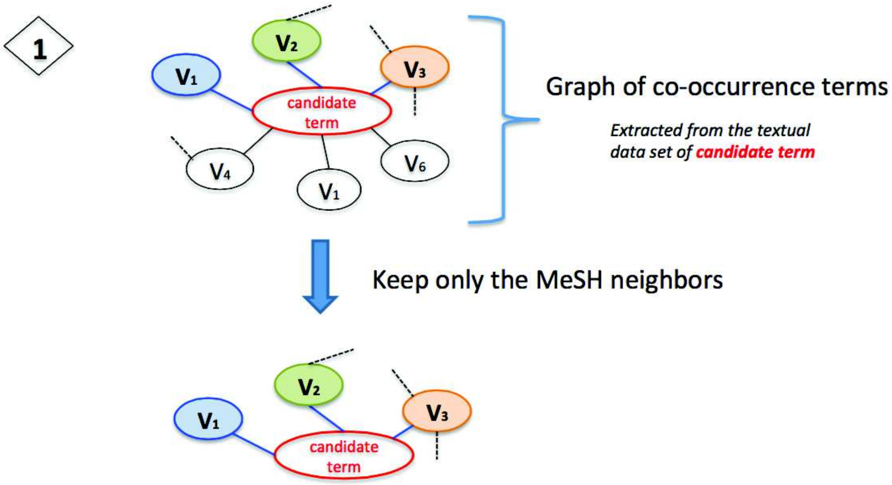

1
V₂
V₁
V₃
candidate term
V₄
V₁
V₆
Graph of co-occurrence terms
Extracted from the textual data set of candidate term
Keep only the MeSH neighbors
V₂
V₁
V₃
candidate term

Figure 11.10: First Semantic Linkage step.

2. **Evaluating the semantic similarity:** To compare the semantic similarity between the candidate terms and the MeSH terms, we need to extract the neighborhood context. For this, we use a database such as PubMed to extract the context of each MeSH neighbor term. Then we apply the cosine measure to compute the semantic similarity. We also take the subclasses and super classes in MeSH or Hyponymy/hypernymy of the neighbors of the candidate term. Figure 11.11 illustrates this second step.

   Finally, the output of this methodology is an option list, ranked by the cosine similarity. This list provides MeSH terms that are the most semantically related to the candidate term.

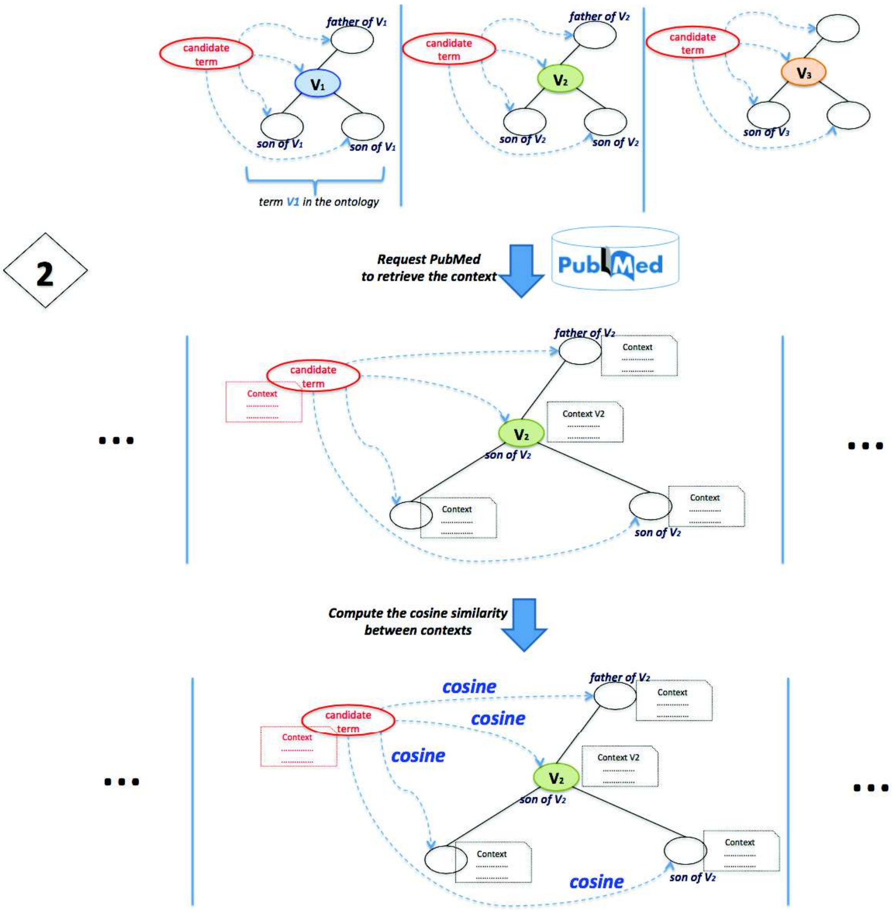


Figure 11.11: Second Semantic Linkage step.

This chapter proposed a workflow to extract concepts of new biomedical terms, and to add these terms to an already existing biomedical ontology, which represents the second part of this thesis. There were three steps in this workflow, with an associated approach for each one. An approach to: i) predict the polysemy of new biomedical terms, ii) induce senses of these new terms, and finally iii) semantically link terms to an already existing biomedical ontology.

The results of the entire second part “Concept Extraction and Semantic Linkage” are provided in next chapter (see Chapter 12).

Chapter 12

Data and Results

This chapter outlines the data sets and experiments of our proposal for: (i) Polysemy Detection, (ii) Term Sense Induction, and (iii) Semantic Linkage. It is divided into three sections, with each one first describing the data set, and then the results of each step of the proposed methodology. Sections 12.1, 12.2, 12.3, present the data sets and results for polysemy detection, term sense induction, and semantic linkage, respectively, as shown in Figure 12.1.

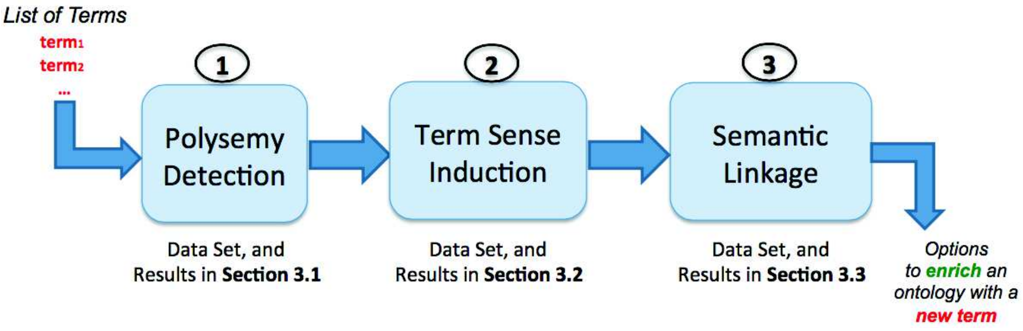

List of Terms
term1
term2
...
1
Polysemy Detection
2
Term Sense Induction
3
Semantic Linkage
Data Set, and Results in Section 3.1
Data Set, and Results in Section 3.2
Data Set, and Results in Section 3.3
Options to enrich an ontology with a new term

Figure 12.1: Data sets and Results of the Proposed Methodology.

The data sets used in this thesis consisted of textual data collections, i.e. specifically biomedical data. We used a different biomedical data set for each step because there was no single annotated data set that could be used for the three proposed approaches. Most commonly, the major data sets are annotated, i.e. *gold standard corpora*, while the others are extracted from the Web, especially from PubMed. Each data set and the results obtained are described in the next sections.

12.1 Polysemy Detection

This section describes the data set used and the results of experiments conducted to detect the polysemy of a candidate term. First, we describe the data set, then the results.

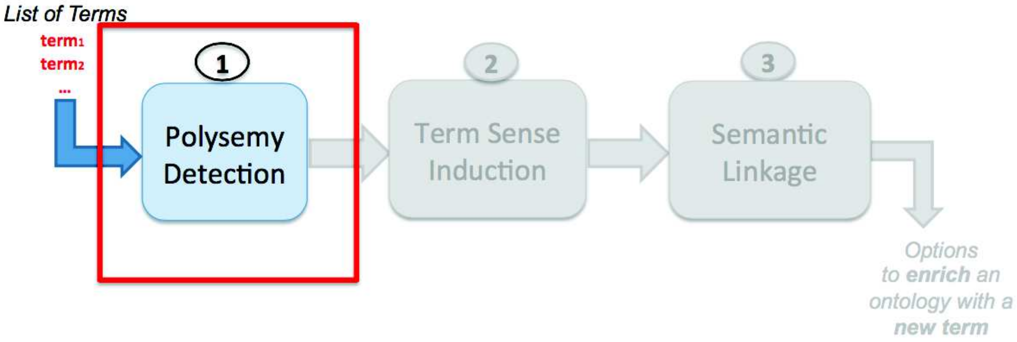


Figure 12.2: Evaluation of Polysemy Detection

12.1.1 Data set

The data set used to evaluate the polysemy detection approach must contain polysemic and non-polysemic terms, hereafter polysemic is also called ambiguous. This data set will allow to evaluate if a new biomedical term is polysemic or not (yes or no). Therefore, our data set is composed of a polysemic data set and a non-polysemic data set. The first one already exists, and the second one was built for this purpose. In next sections, we describe the polysemic data set (see Section 12.1.1.1) and non-polysemic data set (see Section 12.1.1.2).

12.1.1.1 Polysemic Data set

This data set consists of 203 ambiguous entities (polysemic) in English. These entities have been extracted from the MSH WSD[1] [Jimeno-Yepes et al., 2011] data set, which consists of 106 ambiguous abbreviations, 88 ambiguous terms, and 9 which are a combination of both. For each ambiguous term/abbreviation, on average the data set contains 180 instances (i.e. titles/abstracts) obtained from MEDLINE. This data set is well-known in Word Sense Disambiguation literature applied to the biomedical domain.

12.1.1.2 Construction of the Non-polysemic Data set

For the experiments on our Polysemy Detection approach, we required both ambiguous and non-ambiguous biomedical terms. The ambiguous data set already exists in

[1] `http://wsd.nlm.nih.gov/`

the literature. To our knowledge, there are no non-ambiguous data sets, so we constructed a non-polysemic data set using the UMLS metathesaurus and the manual MeSH indexing of MEDLINE. This non-polysemic data set was built via two steps: (i) selecting non-polysemic terms from MeSH/UMLS, and (ii) extracting a set of titles and abstracts containing those terms from PubMed.

Hence, the MSH WSD data set contains, for each term, on average 180 titles/abstracts. Actually this number differs for each term. For instance, the term "Cold" contains 260 titles/abstracts, while "Yellow Fever" contains 183. To avoid bias in the data set, we thus created our non-ambiguous data set following the same methodology used for the ambiguous data set creation. For instance, if the term *Cold* has 260 titles/abstracts, for the non-ambiguous data set, we select the non-ambiguous term *Term*$_1$ and extract the same number of titles/abstracts, i.e. again 260. Table 12.1 illustrates this principle.

Polysemic Term	**Number of Titles/Abstracts**	**Non-polysemic Term**	**Number of Titles/Abstracts**
Cold	260	Term$_1$	260
Cortical	297	Term$_2$	297
PCA	491	Term$_3$	491
Yellow Fever	183	Term$_4$	183
...	...	...	...

Table 12.1: Principle of Non-polysemic Data set Creation.

The steps for creation of the non-polysemic data set are:

- First, we select all terms contained in MeSH. We use MeSH because PubMed only indexes articles with MeSH terms and we use PubMed to build the non-polysemic data set.
- Then we screen UMLS to identify non-ambiguous terms (those associated to only one concept). We filter this list, only taking those that are non-polysemic in UMLS.
- UMLS terms contain multiple signs, so we clean them by eliminating all terms containing (; , ? ! : { } [ ]).
- Then we randomly choose 203 non-polysemic terms (in order to have a balanced data set of positive and negative examples, i.e. polysemic and non-polysemic terms), and we extract their content from PubMed. PubMed is a free resource that provides access to MEDLINE. MEDLINE is a bibliographic database of life sciences and biomedical information. It includes bibliographic information for articles from academic journals covering various medical domains.

- Finally, the non-polysemic data set is built and available for research.

Table 12.2 shows an extract of the polysemic/non-polysemic data set. Table 12.3 summarizes the details of our polysemic/non-polysemic data set.

Polysemic Term	**Number of Titles/Abstracts**	**Non-polysemic Term**	**Number of Titles/Abstracts**
Cold	260	Decision Making	260
Cortical	297	Superovulation	297
PCA	491	Policy Making	491
Yellow Fever	183	Zymosan	183
...	...	...	...

Table 12.2: Extract of Polysemic and Non-polysemic Data set.

Description	**data set**
Nb of Entities	406
Nb of Ambiguous Entities	203
Nb of Non-ambiguous Entities	203
Nb of Tokens of the Context of Ambiguous Entities	7 597 337
Nb of Tokens of the Context of Non-ambiguous Entities	8 294 378
Mean number of Tokens for each Ambiguous Entity	37 425
Mean number of Tokens for each Non-ambiguous Entity	40 859

Table 12.3: Details of our Polysemic/Non-polysemic Data set

At this stage, we finally count with a data set composed of polysemic terms and non-polysemic terms. This allows us to evaluate our methodology to detect polysemy.

12.1.2 Polysemy Detection Results

The data set built in the previous section serves for experiments and evaluation of our polysemy detection methodology, which is based on the definition of new meta-features (see Section 11.1) computed directly from the corpus and from an induced graph, to finally apply a set of machine learning algorithms. Our reason for learning a classifier is to produce the best estimation of whether a biomedical term is polysemic or not. Therefore, this section describes the experiments carried out to evaluate the performance of the new proposed meta-features (total of 23). The previously cited algorithms were applied for this analysis (see Section 11.1.2) with a 10-cross-validation. The results are provided in terms of *Accuracy (A)*, *Precision (P)*, *Recall (R)*, and *F-Measure (F)* over the data set. In Section 12.1.2.1, experiments were carried out with direct and graph-based meta-features separately. We also wanted to explore the performance of the meta-features by combining the 11 direct meta-features with the 12 graph-based meta-features and these results are presented in Section 12.1.2.3. As major studies deal with the identification of the correct meaning of a term, we cannot provide a comparison of our approach with others. To the best of our knowledge, there are no studies focused on the detection of polysemy with binary output (i.e. true or false).

Experimental Protocol 1

Data set: Polysemic and Non-polysemic Data set

Polysemic Term	**Class**
Ca	P
Cold	P
Cortical	P
Yellow Fever	P
...	...
Non-polysemic Term	**Class**
Decision Making	NP
Superovulation	NP
Policy Making	NP
Zymosan	NP
...	...

Algorithms: 8 algorithms mentioned in Section 11.1.2: NB, AB, TD, SVM, MB, M5P, NN, MCC

Input: Meta-features extracted from our Data set

Output: Polysemy Detection P (polysemic) or NP (non-polysemic)

12.1.2.1 Direct Meta-features

Table 12.4 shows the results obtained on the previously defined data set (Polysemic and Non-polysemic Data set) with the 11 direct meta-features. Note that the M5P Model Tree (M5P) gets the best results, with an *Accuracy* of 92.1%. This means that the supervised algorithms on our direct meta-features have correctly classified 92 instances (polysemic or not). The worst result is obtained with Naive Bayes (NB), with an *Accuracy* of 86%.

12.1.2.2 Graph-based Meta-features

Table 12.5 shows the results obtained with only graph-based meta-features. Note that Meta Bagging gets the best results, with an *accuracy* of 92.1%. The results obtained with the supervised algorithms differed for the two types of meta-features. This is because the meta-features and their values are different. The worst result is also obtained with Naive Bayes (NB), with an *Accuracy* of 86%.

	A	*P*	*R*	*F*
Zero Rule (ZeroR)	0.493	0.491	0.493	0.472
One Rule (OneR)	0.887	0.901	0.887	0.886
Naive Bayes (NB)	0.860	0.863	0.860	0.859
AdaBoost (AB)	0.897	0.903	0.897	0.896
Tree Decision (TD)	0.879	0.882	0.879	0.879
Support Vector Machine (SVM)	0.919	0.922	0.919	0.919
Meta Bagging (MB)	0.892	0.896	0.892	0.891
M5P Tree (M5P)	**0.921**	**0.925**	**0.921**	**0.921**
Multilayer Perceptron (NN)	0.906	0.907	0.921	0.906
MultiClassClassifier Logistic (MCC)	0.914	0.915	0.914	0.914

Table 12.4: Direct Meta-features

	A	*P*	*R*	*F*
Zero Rule (ZeroR)	0.493	0.491	0.493	0.472
One Rule (OneR)	0.847	0.850	0.847	0.847
Naive Bayes (NB)	0.860	0.863	0.860	0.859
AdaBoost (AB)	0.899	0.900	0.899	0.899
Tree Decision (TD)	0.882	0.884	0.882	0.882
Support Vector Machine (SVM)	0.874	0.875	0.874	0.874
Meta Bagging (MB)	**0.921**	**0.922**	**0.921**	**0.921**
M5P Tree (M5P)	0.884	0.885	0.884	0.884
Multilayer Perceptron (NN)	0.906	0.907	0.906	0.906
MultiClassClassifier Logistic (MCC)	0.914	0.914	0.914	0.914

Table 12.5: Graph-based Meta-features

12.1.2.3 Combining two kinds of meta-features

We study the effect of feature mixing, i.e. direct plus graph-based meta-features. These two types of meta-features are combined and Table 12.6 reports the results. We can see that the Neural Network model (Multilayer Perceptron) gets excellent results, with an *accuracy* (A) of 97.8%. This table also illustrates that the minimal accuracy performance is 95.3%. We can prove that the combination of two kinds of meta-features gives the best results.

12.1.2.4 Discussion

We evaluate the informativeness of the meta-features in detail. For this purpose, from Table 12.6, we take the created decision tree (TD), in order to discuss the types of meta-features highlighted by this algorithm, which obtains 97% of the F-measure.

	A	*P*	*R*	*F*
Zero Rule (ZeroR)	0.493	0.491	0.493	0.472
One Rule (OneR)	0.837	0.837	0.837	0.837
Naive Bayes (NB)	0.956	0.956	0.956	0.956
AdaBoost (AB)	0.975	0.976	0.975	0.975
Tree Decision (TD)	0.970	0.970	0.970	0.970
Support Vector Machine (SVM)	0.966	0.966	0.966	0.966
Meta Bagging (MB)	0.970	0.970	0.970	0.970
M5P Tree (M5P)	0.963	0.963	0.963	0.963
Multilayer Perceptron (NN)	**0.980**	**0.980**	**0.980**	**0.980**
MultiClassClassifier Logistic (MCC)	0.953	0.953	0.953	0.953

Table 12.6: Combining two kinds of meta-features

Figure 12.3 shows the associated decision tree. We can see that only 4 of the 23 meta-features have been taken into account for classification. Two direct ($minU(t)$, $sdA(t)$) and two graph-based ($sum(v_t)$, $ngUMLS(v_t)$) meta-features. The two direct meta-features are extracted with UMLS ($minU(t)$) and AGROVOC ($sdA(t)$), which confirms that overlapping between the two dictionaries is useful to detect biomedical term polysemy. Figure 12.3 shows that the combination of $minU(t)$ and $sum(v_t)$ allows us to classify the most *non-polysemic* terms, i.e. 199 out of 203, while $minU(t)$ and $ngUMLS(v_t)$ allows us to classify the most *polysemic* terms, i.e. 161 out of 203.

Table 12.7 presents the confusion matrix, where each column represents the instances in a predicted class, while each row represents the instances in an actual class, corresponding to an Accuracy (A) of 0.97 (see Table 12.6, column *A*, row *TD*). This table shows us that the prediction is balanced. The system has correctly classified 198 polysemic terms from a total of 203, and similar for non-polysemic terms, where it has correctly classified 196 terms from 203.

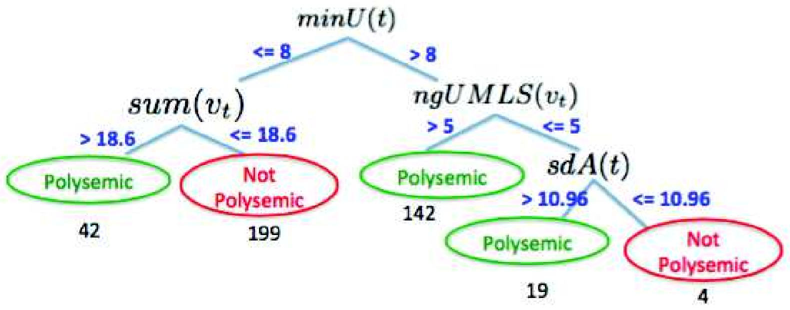


Figure 12.3: Decision Tree obtained from the Polysemic and Non-polysemic Data set.

	Pol	*Not Pol*	← Classified as
Pol	198	5	203
Not Pol	7	196	203

Table 12.7: Confusion Matrix based on the Polysemic and Non-polysemic Data set.

Finally, as pointed out before, to our knowledge, no studies have focused on polysemy detection as in our case by answering yes or no. Therefore, there is no baseline comparison with related works.

12.2 Term Sense Induction

This section describes the data set used and results of our approach to induce possible sense(s) of a candidate term. First we describe the data set and then the results.

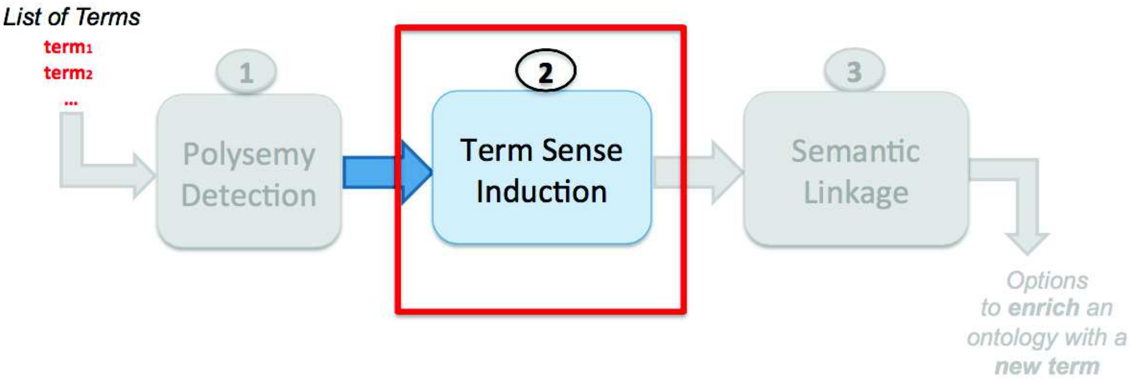


Figure 12.4: Evaluation of Term Sense Induction

12.2.1 Data set

As mentioned in Section 12.1.1.1, this data set is composed of 203 ambiguous entities (polysemic). These entities have been extracted from the MSH WSD[2] [Jimeno-Yepes et al., 2011] data set, which consists of 106 ambiguous abbreviations, 88 ambiguous terms, and 9 that are a combination of both. For each ambiguous term/abbreviation, the data set contains on average 180 instances (i.e. titles/abstracts) obtained from MEDLINE. Table 12.8 shows the details of the polysemic data set used for this experiment. Table 12.9 illustrates an extract of our data set with its respective number of senses. This additional information, number of senses, which were not

[2] http://wsd.nlm.nih.gov/

used in the previous section, is the input for our experiments.

Description	**Details**
Nb of Ambiguous Entities	203
Nb of Tokens of the Context of Ambiguous Entities	7 597 337
Mean number of Tokens for each Ambiguous Entity	37 425

Table 12.8: Sense Number per Term.

Term	**Sense Number**
Ca	4
Cold	3
Cortical	3
Yellow Fever	2
...	...

Table 12.9: Details of our Polysemic Data set

As we mentioned previously, for this methodology, the polysemic data set and its number of associated senses is all that is necessary. The next section describes the results obtained with our methodology on this data set.

12.2.2 Results of Sense Number Prediction

Readers should note that this methodology is only applied to terms which have been classified as polysemic terms. So in this section we evaluate our approach to induce the possible senses of a term. As previously cited, a major problem in TSI (Term Sense Induction) is to determine the number of senses, which represents a clustering task. Therefore, we evaluate the results of determining the number of clusters (number of senses for a new term) according to three aspects. The first one, applying clustering algorithms with the bag-of-words representation and computing the values of our new internal indexes. The second one, applying clustering algorithms with the Graph representation and computing the values of our new internal indexes. And the third one, we evaluate our meta-features with supervised algorithms.

Note that the new indexes are only used to evaluate the clustering done with the bag-of-words representation and the graph representation. We then evaluate our meta-features with supervised algorithms to determine the number of clusters.

12.2.2.1 Results With Bag of Words Representation

Experimental Protocol 2

Data set: Polysemic Data set:

Polysemic Term	**Sense Number**
Ca	4
Cold	3
Cortical	3
Yellow Fever	2
...	...

Algorithms: 5 clustering algorithms: *rb, rbr, direct, agglo, graph*
Input: Bag-of-words of our Data set, 3 000 meta-features
Output: Prediction of the Number of Clusters (2,3,4 or 5)

First, we recall the basic notation related to the concepts explained in Section 11.2.1.1: (i) intra-cluster similarity (*ISIM*), and (ii) inter-cluster similarity (*ESIM*). We opted by taking the **cosine** as distance similarity for all our clustering with a bag-of-words representation. The selected number of meta-features was 3 000 terms, extracted with the BioTex application.

Table 12.10 illustrates the process for determining the number of clusters (number of possible senses for a term) according to the value of indices $a_{k,I2}$ and $c_{k,I2}$ (see Section 11.2.1.1 and Section11.2.1.3).

Table 12.10 shows the values computed for the term "Yellow Fever", where the objective function is *I2* and the algorithm clustering is *Partitional.* In our data set, "Yellow Fever" has 2 senses. So, the correct predicted number of clusters would be 2 as well. In this table, the first column represents the number of clusters, which we vary between 2 and 5, and then we apply the clustering algorithms to compute the new index values. For instance, the first row for $k = 2$, shows that the first cluster, *Cluster-1* has 110 objects ($\mid S_i \mid= 110$), the intra-cluster similarity of this cluster is 0.058 ($ISIM = 0.058$), and the inter-cluster similarity is 0.025 ($ESIM = 0.025$). For the two clusters formed when ($k = 2$), we compute $a_{2,I2}$ and $c_{2,I2}$.

The last two rows show, according the index values, how we choose the number of clusters. For instance, if we take $\boldsymbol{max(c_{k,I2})}$, then the maximum value is 2.655 when $\boldsymbol{k = 2}$. Therefore, the $\boldsymbol{max(c_{k,I2})}$ index predicts that the number of clusters is **2** when the objective function is *I2* and the algorithm clustering is *Partitional.*

Note that this process is performed for 5 objective functions, for 5 clustering algorithms, for the 5 new indexes, and for each instance of our data set (see Section 11.2.1). Table 12.11 summarizes this process by taking only two objective functions ***I1*** and ***I2*** into account, for the instance "Yellow Fever" (one instance

Objective function: **I2**, Algorithm: **Partitional (rb)**						
k	*Id of Cluster*	$\mid S_i \mid$	*ISIM*	*ESIM*	$a_{k,I2}$	$c_{k,I2}$
2	*Cluster-1*	110	0.058	0.025	0.053	**2.655**
	Cluster-2	73	0.048	0.025		
3	*Cluster-1*	43	0.087	0.029	0.070	2.374
	Cluster-2	67	0.074	0.030		
	Cluster-3	73	0.048	0.025		
4	*Cluster-1*	16	0.118	0.008	0.085	2.299
	Cluster-2	43	0.087	0.029		
	Cluster-3	67	0.074	0.030		
	Cluster-4	57	0.063	0.028		
5	*Cluster-1*	16	0.118	0.008	**0.094**	2.191
	Cluster-2	26	0.105	0.025		
	Cluster-3	43	0.087	0.029		
	Cluster-4	31	0.086	0.032		
	Cluster-5	67	0.074	0.030		
			$max(a_{k,I2})$		$k = 5$	
			$max(c_{k,I2})$			$k = 2$

Table 12.10: Choosing k according to $a_{k,I2}$ and $c_{k,I2}$ values.

in our data set). For instance, the second row shows that $max(a_{k,I1})$, with the clustering algorithms ***agglo*** and ***graph***, with ***I1*** as objective function, predicts **2** clusters for “Yellow Fever” (yellow cells).

Internal Indexes	*rb*	*rbr*	*direct*	*agglo*	*graph*
$max(a_{k,I1})$	5	5	4	5	2
$min(b_{k,I1})$	3	3	3	**2**	**2**
$max(c_{k,I1})$	3	3	2	2	2
$max(e_{k,I1})$	5	5	5	5	2
$max(f_{k,I1})$	2	2	2	2	2
$max(a_{k,I2})$	5	5	5	5	5
$min(b_{k,I2})$	4	4	4	2	2
$max(c_{k,I2})$	2	2	2	2	2
$max(e_{k,I2})$	5	5	5	5	5
$max(f_{k,I2})$	2	2	2	2	2

Table 12.11: k Prediction for *Yellow Fever*, with bag-of-words representation (2 classes)

To determine the performance of our new indexes, we carried out this process for our 203 ambiguous entities, while evaluating the accuracy of the cluster number prediction. Table 12.12 summarizes the accuracy for the determination of number of clusters in the entire data set, while taking only two objective functions ***I1*** and ***I2*** into account. Note that in several cases we achieve **93.10** % accuracy. It means that for 189 terms our methodology predicted the correct $\boldsymbol{k}$. Table 12.12, in the last row, shows that $\boldsymbol{max(f_{k,OF})}$ gives the best results for all of the clustering algorithms and for the ***I2*** objective function.

Internal Indexes	***rb***	***rbr***	***direct***	***agglo***	***graph***
$\boldsymbol{max(a_{k,I1})}$	6.40 %	5.42 %	6.90 %	1.97 %	91.63 %
$\boldsymbol{min(b_{k,I1})}$	36.45 %	38.92 %	34.98 %	92.12 %	**93.10 %**
$\boldsymbol{max(c_{k,I1})}$	32.02 %	30.54 %	31.53 %	42.86 %	**93.10 %**
$\boldsymbol{max(e_{k,I1})}$	0.99 %	1.48 %	1.48 %	8.87 %	**93.10 %**
$\boldsymbol{max(f_{k,I1})}$	92.12 %	92.12 %	**93.10 %**	**93.10 %**	92.61 %
$\boldsymbol{max(a_{k,I2})}$	0.99 %	0.99 %	0.49 %	1.97 %	91.63 %
$\boldsymbol{min(b_{k,I2})}$	81.77 %	84.73 %	86.21 %	92.12 %	**93.10 %**
$\boldsymbol{max(c_{k,I2})}$	88.67 %	87.68 %	91.63 %	42.86 %	**93.10 %**
$\boldsymbol{max(e_{k,I2})}$	3.45 %	2.96 %	4.93 %	8.87 %	**93.10 %**
$\boldsymbol{max(f_{k,I2})}$	**93.10 %**	**93.10 %**	**93.10 %**	**93.10 %**	**93.10 %**

Table 12.12: Accuracy (A) results for 203 Ambiguous Entities with bag-of-words.

12.2.2.2 Results With Graph Representation

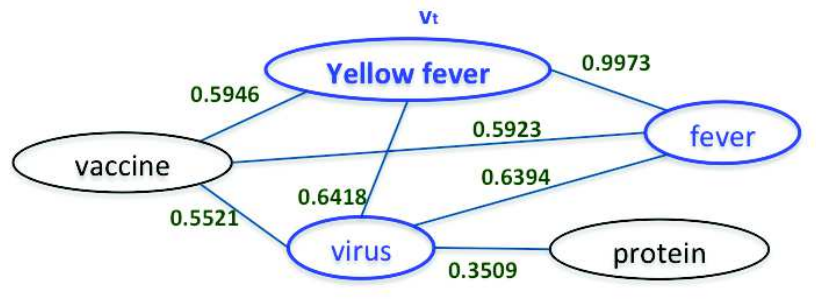
Experimental Protocol 3

Data set: Polysemic Data set:

Polysemic Term	**Sense Number**
Ca	4
Cold	3
Cortical	3
Yellow Fever	2
...	...

Algorithms: 5 clustering algorithms: *rb, rbr, direct, agglo, graph*

Input: Graph representation of our Data set, 1 000 vertex



Output: Prediction of the Number of Clusters (2,3,4 or 5)

Similar to the evaluation of bag-of-words representations, this process is performed for 5 objective functions, for 5 clustering algorithms, for the 5 new indexes, and for each instance of our data set (see Section 11.2.1). As previously described, our data set contains 203 ambiguous instances.

For this evaluation, we use the graph as defined in Section 11.1.1.2, with each graph containing 1 000 terms extracted with the BioTex application. Table 12.13 summarizes this process while taking only two objective functions ***I1*** and ***I2*** into account, for instance "Yellow Fever" (one instance in our data set). In our data set, the number of clusters (concepts) of the term Yellow Fever is 2. For instance, the first row of Table 12.13 shows that $\boldsymbol{max(a_{k,I1})}$, with the clustering algorithms ***rb, rbr, direct*** and ***graph***, with ***I1*** as objective function, predicts **2** clusters for "Yellow Fever" (yellow cells). In this same table, we observe that $\boldsymbol{max(f_{k,OF})}$ generally predicts the correct number of senses.

Similar to bag-of-words representations, we conducted this process for our 203 ambiguous entities in order to determine the performance of our indexes over the graph representation. Table 12.14 shows the accuracy for the prediction of the number of clusters in the entire data set. Note that in several cases we achieve **93.1** % accuracy, which means that our methodology has correctly predicted the number of senses of 189 entities out of 203. Consequently, with Table 12.13, $\boldsymbol{max(f_{k,OF})}$ gives the best accuracy results for all the clustering algorithms for both objective functions.

Internal Indexes	*rb*	*rbr*	*direct*	*agglo*	*graph*
$max(a_{k,I1})$	2	2	2	5	2
$min(b_{k,I1})$	2	2	2	2	2
$max(c_{k,I1})$	2	2	2	5	2
$max(e_{k,I1})$	5	5	5	5	2
$max(f_{k,I1})$	2	2	2	2	2
$max(a_{k,I2})$	5	5	4	5	2
$min(b_{k,I2})$	2	2	2	2	2
$max(c_{k,I2})$	2	2	2	5	2
$max(e_{k,I2})$	5	5	5	5	2
$max(f_{k,I2})$	2	2	2	2	2

Table 12.13: $\boldsymbol{k}$ Prediction for *Yellow Fever* with Graph representation (2 classes)

Internal Indexes	*rb*	*rbr*	*direct*	*agglo*	*graph*
$max(a_{k,I1})$	1.97 %	1.97 %	1.48 %	1.48 %	9.36 %
$min(b_{k,I1})$	77.83 %	77.83 %	75.86 %	**93.10 %**	64.04 %
$max(c_{k,I1})$	76.35 %	74.88 %	76.85 %	85.22 %	64.53 %
$min(c_{k,I1})$	8.37 %	7.88 %	7.39 %	0.49 %	21.67 %
$max(e_{k,I1})$	3.94 %	4.43 %	4.93 %	47.78 %	3.94 %
$max(f_{k,I1})$	**93.10 %**	**93.10 %**	**93.10 %**	**93.10 %**	**93.10 %**
$max(a_{k,I2})$	0.49 %	0.99 %	0.49 %	1.48 %	2.96 %
$min(b_{k,I2})$	82.27 %	82.76 %	86.21 %	**93.10 %**	80.3 %
$max(c_{k,I2})$	91.13 %	91.13 %	90.15 %	85.22 %	87.19 %
$min(c_{k,I2})$	0.99 %	0.99 %	0.99 %	0.49 %	1.48 %
$max(e_{k,I2})$	4.43 %	3.94 %	3.94 %	47.78 %	2.46 %
$max(f_{k,I2})$	**93.10 %**	**93.10 %**	**93.10 %**	**93.10 %**	**93.10 %**

Table 12.14: Accuracy results for 203 Ambiguous Entities with Graph representation.

As stated in the two previous sections, the new indexes evaluated with bag-of-words and graph representation obtained similar accuracy result values. From Tables 12.12, 12.14 shows that $\boldsymbol{max(f_{k,OF})}$ works better over the graph representation. These values confirm that the graphs are really useful, particularly with the methodologies proposed in this project.

12.2.2.3 Results With Meta-features

In this section, we report experiments carried out to evaluate the performance of the new meta-features for prediction of the number of clusters.We thus use direct and graph meta-features together. Algorithms cited in Section 11.1.2 are evaluated with a 10-cross-validation. The results are provided in terms of *Accuracy (A)*, *Precision (P)*, *Recall (R)*, and *F-Measure (F)* over the data set. On this occasion, the output is 4 classes (2,3,4,5).

Table 12.15 shows the results obtained on the previously defined data set with meta-features. Note that MB and M5P obtained the best results, with an *accuracy* of 0.936. This means that the supervised algorithms on our direct meta-features predicted the correct number of senses of 190 terms out of 203. We can prove that the meta-features are also useful for predicting the number of clusters.

Experimental Protocol 4

Data set: Polysemic Data set:

Polysemic Term	**Sense Number**
Ca	4
Cold	3
Cortical	3
Yellow Fever	2
...	...

Algorithms: 8 algorithms mentioned in Section 11.1.2: NB, AB, TD, SVM, MB, M5P, NN, MCC

Input: Meta-features extracted from our Data set (direct + graph-based)
Output: Prediction of the Number of Clusters (2,3,4 or 5)

	A	*P*	*R*	*F*
Naive Bayes (NB)	0.695	0.891	0.695	0.769
AdaBoost (AB)	0.921	0.866	0.921	0.893
Tree Decision (TD)	0.906	0.874	0.906	0.890
Support Vector Machine (SVM)	0.931	0.867	0.931	0.898
Meta Bagging (MB)	**0.936**	**0.930**	**0.936**	0.909
M5P Tree (M5P)	**0.936**	0.905	**0.936**	**0.911**
Multilayer Perceptron (NN)	0.897	0.884	0.897	0.890
MultiClassClassifier Logistic (MCC)	0.901	0.919	0.901	0.910

Table 12.15: Meta-features for Number of Sense Prediction.

12.2.2.4 Discussion

In the previous evaluations, bag-of-words and graph representations generally obtained similar accuracy values. For these two cases, the maximum value is 93.1%. In general, we can see in Tables 12.12, 12.14 that the best accuracy results are given by the $\boldsymbol{f_k}$ index (see Section 11.2.1.1). The $\boldsymbol{f_k}$ index is the division between the objective function and the logarithm of $\boldsymbol{k}$. As the objective function is stronger when more clusters are included for the solution, we divided by the logarithm of the number of clusters to try to overcome this drawback.

For meta-feature representation, MB (Meta Bagging model) was clearly one of the models which gave good results. The Meta Bagging model generates multiple versions of a predictor to build an aggregated predictor. This aggregated predictor is built from 10 regression trees, which is the main reasons for the accuracy gain in the results. As we mentioned before, meta-features assign a correct number of senses to 190 biomedical candidate terms out of 203, which represents 93.6% precision. This is very useful in the biomedical domain because it allows us to automatically determine the polysemy of candidate terms, including the associated number of senses (or concepts).

To our knowledge, there are no studies of Term Sense Induction focused on the biomedical domain. Therefore, a comparison with baselines measures is difficult. In contrast, several approaches are focused on general domains. The main studies use a particular data set to test their approaches. A well-known data set is that provided by the SemEval-2010 WSI shared task [Manandhar et al., 2010]. This data set contains 100 target words: 50 nouns and 50 verbs. The most relevant and known systems using this data set are, for instance: Baseline Random, Baseline MFS (Most Frequent Sense), Duluth-WSI [Pedersen, 2010], UoY [Korkontzelos and Manandhar, 2010], NMFlib [Van de Cruys and Apidianaki, 2011], NB [Choe and Charniak, 2013], RPCL [Huang et al., 2015], which obtain between 57.3% and 68.44% precision when predicting the number of senses for a word.
Another data set provided by SemEval-2007 task [Agirre and Soroa, 2007a], contains 65 target nouns and 35 target verbs. The most well-known approaches taking this data set for experiments on the sense number detection, such as Baseline MFS (Most Frequent Sense); UMND2 [Niu et al., 2007]; I2R [Niu et al., 2007]; 10w, 5w (BNC) [Brody and Lapata, 2009], HDP [Yao and Van Durme, 2011], HDP + position (tuned parameters) [Lau et al., 2012], obtain between 80.9% and 87.1% precision.

These two above-mentioned data sets differ from ours as they contain nouns and verbs and a fixed set of training and test instances are supplied for each target word, typically 1 to 3 sentences in length, each containing the target word. The approaches using these data sets should be adapted for the biomedical domain for our comparison, which represents a tough task. As mentioned previously, the sense number is the cluster number in the clustering task, and this is predicted by evaluating the cluster quality. Therefore, we expect to adapt our data set to compare the results of our proposed internal indexes with the R package*NbClust* (see Section 10.2.3), which implements several indexes to measure the clustering task quality.

Finally, next step is to evaluate the right position of new biomedical terms with the associated senses to be added in an existing biomedical ontology. This is described in the next section.

12.3 Semantic Linkage

From the previous section, we have new biomedical terms and their associated sense or senses (if polysemic). This information is very important for evaluating the position that a new biomedical could take in an already existing ontology to be added. This section thus describes the data set used and methodology results obtained on this data set. First we describe the data set, then the results.

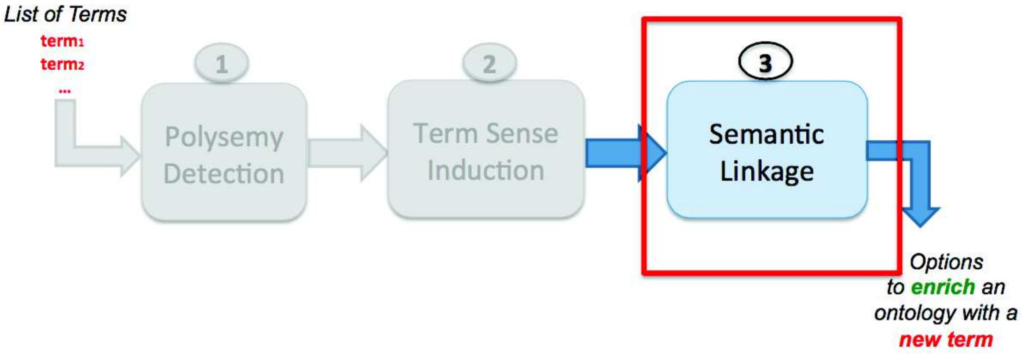


Figure 12.5: Semantic Linkage Evaluation.

12.3.1 Semantic LinkageData set

To properly evaluate our methodology, we need to know if this propose for a new biomedical terms a correct position in an ontology. So for our semantic linkage methodology experiments, we experiment with already existing terms (which will represent the new biomedical terms) from an already existing biomedical ontology. The aim is to determine if our methodology can propose the right position of these terms on the selected existing biomedical ontology. For this study, we selected MeSH ontology.

We hence created a data set containing the existing terms extracted from MeSH. Then we extracted a context for each term from PubMed. We explain this data set creation process in detail in the following paragraphs.

We collect MeSH terms that were added between 2009 and 2015, for instance the term *"aggregatibacter aphrophilus"*. As mentioned before, each term will represent a *"new biomedical candidate term"*. Then we retrieve the context of these terms using PubMed. As shown in Figure 11.9, the methodology needs biomedical candidate terms and its associated co-occurrence graph. Therefore we create a co-occurrence graph per term from the retrieved context. Figure 12.6 illustrates the principle of the process of collecting the data set for semantic linkage.
The steps for the creation of semantic linkage data set are:

- First, we select all the terms added to MeSH between 2009 and 2015.
- MeSH terms contain multiple signs, so we clean them by eliminating all terms containing (; , ? ! : { } [ ]).
- For this experimentation, we take 60 collected terms, because of the length of time necessary to retrieve its context and the context of their neighborhood. These 60 terms represent the *"new biomedical candidate terms"*, the input in our semantic linkage workflow (see Figure 11.9).
- We extract fewer than 100 titles/abstracts (a considerable amount for the experimentation) from PubMed for each of these 60 terms.
- We create 60 co-occurrence graphs of terms using the BioTex application to extract the terms. In this graph, terms represent vertices and co-occurrence values represent the edges. The graph is built in a similar way to that outlined in Section 5.3.1. So we ultimately take the 60 terms as input and their co-occurrence graphs (see Figure 12.6 and Figure 12.7).
- We take the graph neighborhood and their correspond context from PubMed. Table 12.16 shows the details of this data set.

Description	**data set**
Nb of MeSH Terms	60
Nb of Neighbors	8 263
Nb of Tokens of the Total Neighborhood	333 073 311
Mean of Tokens for each Graph Neighborhood	40 309

Table 12.16: Details of our Semantic Linkage Data set

Finally, we created a data set that could allow us to evaluate the results of our experiments. Then in next section we proceed to perform the experiments on this data set.

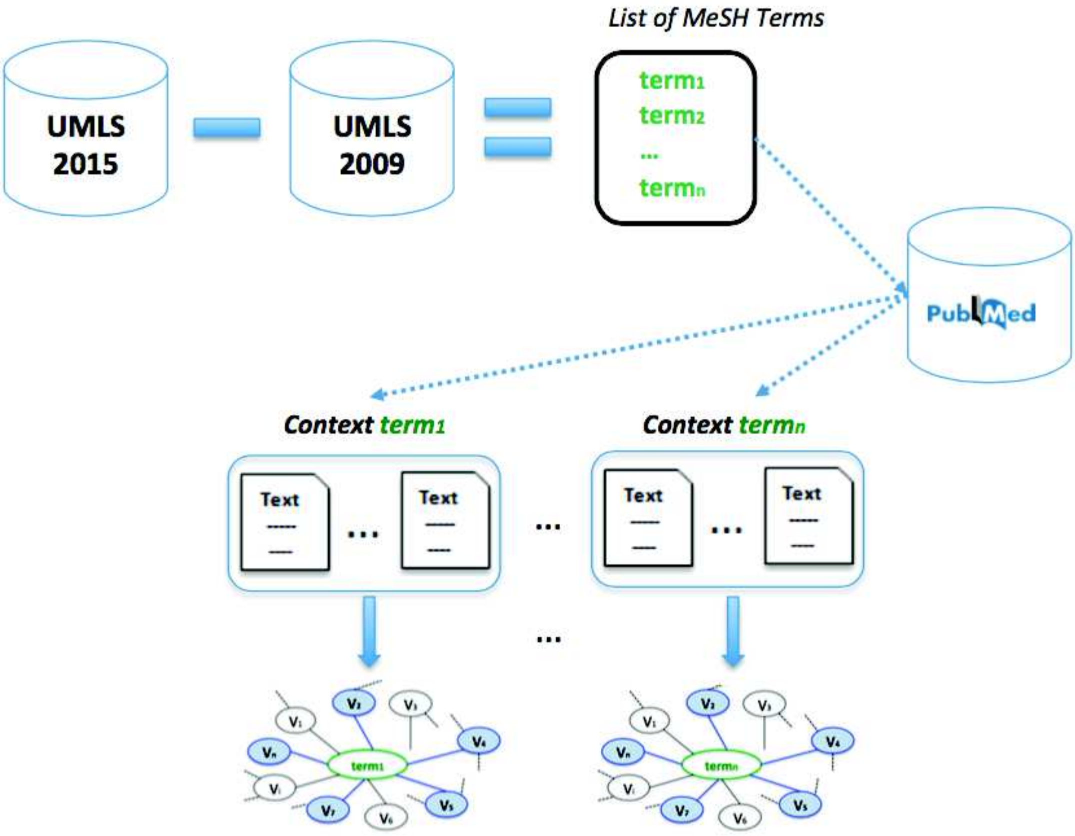


Figure 12.6: Semantic LinkageData set

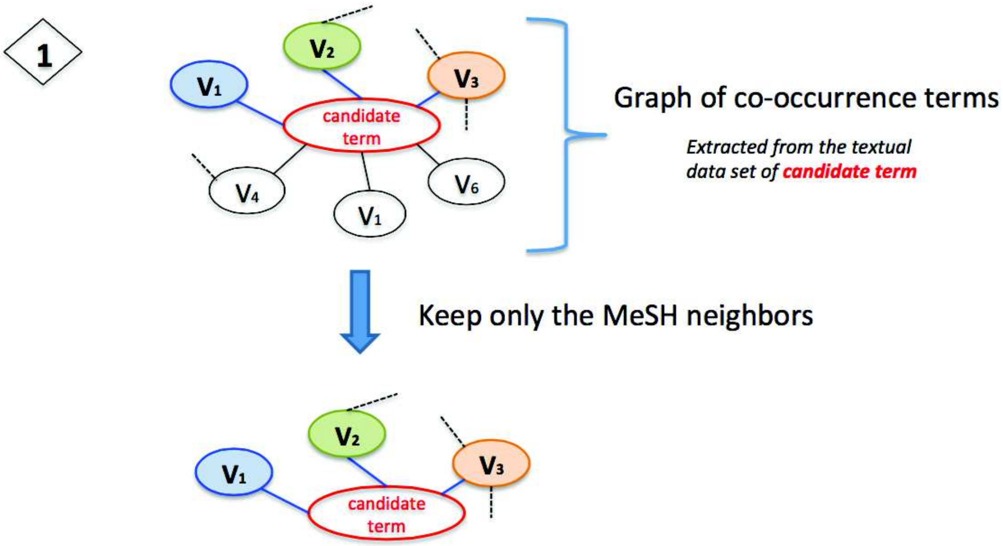
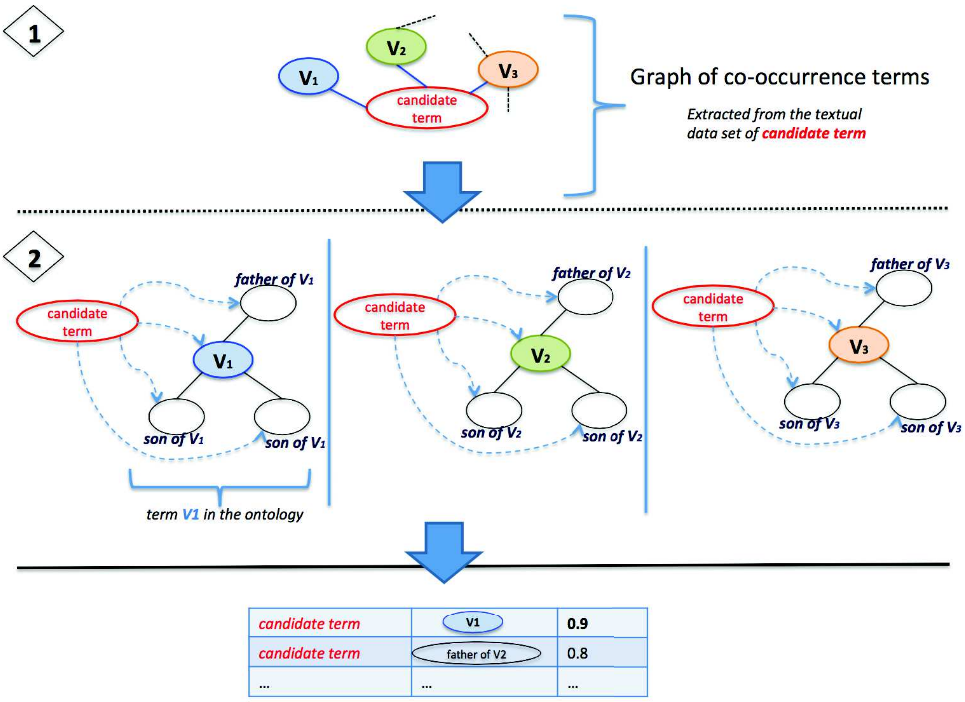

1
V1
V2
V3
V4
V6
V1
candidate term
Graph of co-occurrence terms
Extracted from the textual data set of candidate term
Keep only the MeSH neighbors
V1
V2
V3
candidate term

Figure 12.7: First Semantic Linkage step.

12.3.2 Semantic Linkage Results

In the previous section, we built the data set with already existing terms. That will allow us to evaluate the right position of these terms in a biomedical ontology. Therefore, in this section, we present the experimental results of our proposed methodology to locate a term on the MeSH ontology. To remind readers of the objective of the proposed methodology, we show the graphics of our methodology in Figure 12.8 (as shown in 11.3).



candidate term	V1	0.9
candidate term	father of V2	0.8
...	...	...

Figure 12.8: Semantic Linkage Workflow.

Among all the MeSH neighbors found on the co-occurence graph, we thus seek to evaluate where to locate the term. For each MeSH neighbor in the associated graph of the "candidate term", we thus select the closest terms to this neighbor in MeSH.

For this, we use a well-known pairwise measure. Pairwise measures are used to compare two semantically close terms. We used the measure proposed in [Rada et al., 1989] to evaluate the location of the candidate term. The authors defined the similarity of two terms as a function of the shortest path linking the two concepts. Therefore, the distance is defined as:

$$\boldsymbol{dist(u, v) = sp(u, v)}$$

Where $\boldsymbol{u}$ and $\boldsymbol{v}$ represent two concepts (or terms), and $\boldsymbol{sp(u, v)}$ is the shortest path linking these two concepts. This measure has shown to be the best among others [Dupuch, 2014], for the task of creating new clusters of pharmacovigilance terms. Then we use $\boldsymbol{sp(u,v) = \{1,2,3\}}$ for this evaluation. This principle is illustrated in Figure 12.9, where the term $\boldsymbol{V_3}$ hypothetically belongs to an ontology, and the $\boldsymbol{sp = 1}$ and $\boldsymbol{sp = 2}$ terms are identified.

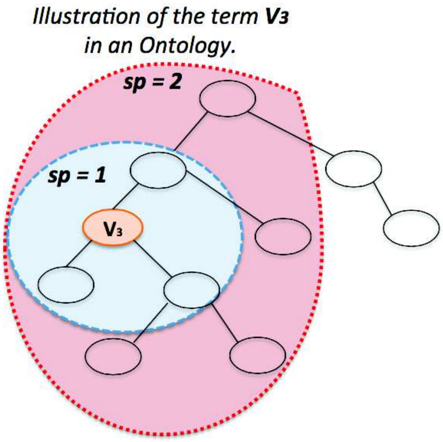


Figure 12.9: Shortest Path for the term $\boldsymbol{V_3}$.

For instance, we take the term ***"corneal injuries"*** added in MeSH between 2009 and 2015. In MeSH version 2015, this term is associated with the **C0339289** concept. Its synonyms are the terms associated with the same concept, such as: *corneal injury, corneal damage,* and *corneal trauma.* The fathers of ***"corneal injuries"*** are all terms associated with the concept **C0010034**, such as: *corneal diseases* and *eye injuries*. Figure 12.10 shows the term ***"corneal injuries"***, its synonyms and its fathers ($\boldsymbol{sp = 1}$) in the MeSH ontology version 2015.

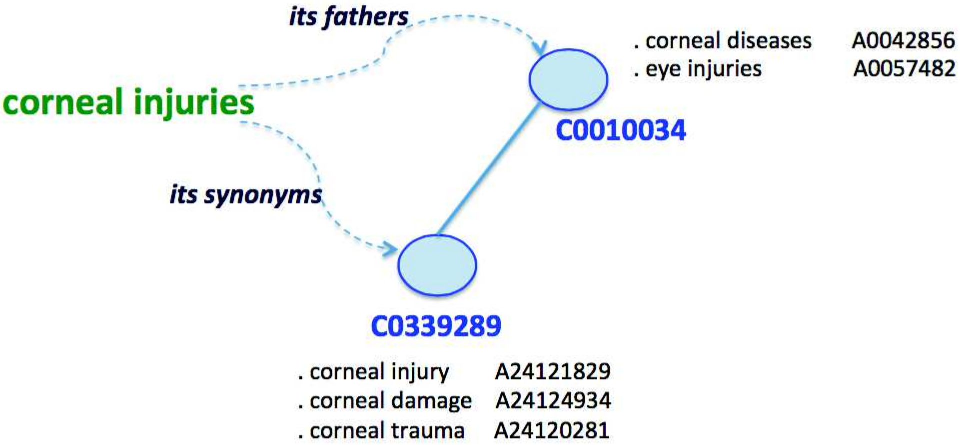

its fathers
corneal injuries
its synonyms
. corneal diseases A0042856
. eye injuries A0057482
C0010034
C0339289
. corneal injury A24121829
. corneal damage A24124934
. corneal trauma A24120281

Figure 12.10: Term *corneal injuries* in the MeSH ontology.

Then we apply our methodology to locate this term in MeSH. Table 12.17 shows the first 10 best propositions on where to locate this term. From these 10 propositions, we can see that 5 proposed terms have a direct connection (i.e. synonyms and fathers) with ***"corneal injuries"*** in MeSH version 2015 (yellow rows). More in detail, Table 12.18 shows the number of correct propositions varying the $\boldsymbol{k}$ first recommendations. For instance, in Table 12.18, the yellow cell shows that for the first $\boldsymbol{k = 2}$ propositions (i.e. best cosine value obtained), our methodology proposed 2 related terms that already exist in the MeSH ontology.

Term to be added	*Where*	*Cosine*
corneal injuries	***corneal injury***	0.4251
corneal injuries	***corneal damage***	0.4181
corneal injuries	*chemical burns*	0.4081
corneal injuries	***corneal diseases***	0.3696
corneal injuries	*corneal ulcer*	0.3689
corneal injuries	***eye injuries***	0.3681
corneal injuries	*amniotic membrane*	0.3639
corneal injuries	*re-epithelialization*	0.3588
corneal injuries	***corneal trauma***	0.3582
corneal injuries	*wound*	0.3472

Table 12.17: Proposition on where to add the term *corneal injuries*.

	Top 1	*Top 2*	*Top 5*	*Top 10*
corneal injuries	1	2	3	5

Table 12.18: Evaluation of propositions to add *corneal injuries*, with $\boldsymbol{sp = 1}$ (fathers, sons and synonyms).

Experimental Protocol 5

Data set: Contexts extracted from PubMed.
Algorithms: Cosine similarity between contexts.
Input: 60 MeSH Terms added between 2009 and 2015 and their co-occurrence graphs.
Output: 10 position propositions to add the terms in the MeSH ontology.

Table 12.19 shows an extract of 15 out the 60 terms added between 2009 and 2015. This table illustrates the evaluation of the propositions put forward for these 15 terms for $\boldsymbol{sp = \{1, 2, 3\}}$. Table 12.18 and Table 12.19 must be interpreted in a

similar manner. For instance, the yellow cell in Table 12.19 is interpreted from 5 propositions, with 2 existing in the MeSH ontology.
Table 12.20 shows the precision of the number of terms which have at least 1 correct proposition with our methodology for the *Top 1, Top 2, Top 5* and *Top 10*; taking the pairwise distance $\boldsymbol{sp = \{1, 2, 3\}}$ into account. For instance, the yellow cell shows that there is at least 1 correct proposition (i.e. existing in the MeSH ontology) for 36 of the 60 terms. (i.e. 40%).
As we have seen, this section introduced an approach to add a new biomedical candidate term in an existing ontology. The approach is based on the similarity between the context of new biomedical term and those existing in an ontology. This approach has generated relevant results on MeSH terms, i.e. finding the right position for the evaluated terms. The most well-known approaches are based on the type relation extraction task, counting with several baselines for different comparisons. This task is expected and represents a future research process for this work, in which comparisons with known baselines can be made.

Finally, we completed the second part of this thesis in order to extract concepts and add the new terms in the MeSH ontology. This part offered a first evaluation in order to enrich biomedical ontologies in a automatic manner via three steps: (i) Polysemy Detection, (ii) Term Sense Induction, and (iii) Semantic Linkage. This chapter outlines the experiments carried out on the previously mentioned steps, showing interesting results done automatically towards the main objective. This second part could be supplemented with a manual evaluation carried out by experts to identify the type of relation. In next chapter, we list the conclusions and perspectives of this second part.

	Distance = 1 ($\boldsymbol{sp} = \mathbf{1}$)			
	Top 1	*Top 2*	*Top 5*	*Top 10*
acanthocheilonema	0	0	0	1
ambulance diversion	0	1	1	1
apicoplasts	1	1	2	2
aurora kinase b	1	1	1	1
betacellulin	0	0	0	0
betamethasone valerate	0	0	1	2
British Virgin Islands	0	0	1	1
brown recluse spider	1	1	2	2
Buschke-Lowenstein tumor	1	2	3	3
central amygdaloid nucleus	1	1	2	2
cephalochordata	1	1	2	3
collagenous sprue	0	0	0	0
controlled before-after studies	0	0	0	0
corneal injuries	1	2	3	5
dexlansoprazole	0	0	0	0

	Distance = 2 ($\boldsymbol{sp} = \mathbf{2}$)			
	Top 1	*Top 2*	*Top 5*	*Top 10*
acanthocheilonema	0	0	3	5
ambulance diversion	0	1	3	5
apicoplasts	1	1	2	2
aurora kinase b	1	2	3	3
betacellulin	1	2	4	7
betamethasone valerate	1	1	2	3
British Virgin Islands	0	1	4	6
brown recluse spider	1	1	2	2
buschke-lowenstein tumor	1	2	3	3
central amygdaloid nucleus	1	1	2	2
cephalochordata	1	2	4	7
collagenous sprue	1	1	2	4
controlled before-after studies	1	1	2	2
corneal injuries	1	2	3	5
dexlansoprazole	0	0	0	0

	Distance = 3 ($\boldsymbol{sp} = \mathbf{3}$)			
	Top 1	*Top 2*	*Top 5*	*Top 10*
acanthocheilonema	1	2	5	9
ambulance diversion	0	1	3	5
apicoplasts	1	1	2	2
aurora kinase b	1	2	3	3
betacellulin	1	2	4	7
betamethasone valerate	1	1	2	3
British Virgin Islands	0	1	4	6
brown recluse spider	1	1	2	2
Buschke-Lowenstein tumor	1	2	3	3
central amygdaloid nucleus	1	1	2	2
cephalochordata	1	2	4	7
collagenous sprue	1	1	2	4
controlled before-after studies	1	1	2	2
corneal injuries	1	2	4	6
dexlansoprazole	1	1	1	1

Table 12.19: Evaluation of propositions put forward for the 15 terms for $\boldsymbol{sp} = \mathbf{\{1, 2, 3\}}$.

	Top 1	*Top 2*	*Top 5*	*Top 10*
$\boldsymbol{sp = 1}$	0.333	0.400	0.500	0.583
$\boldsymbol{sp = 2}$	0.417	0.533	0.617	0.733
$\boldsymbol{sp = 3}$	0.450	0.567	0.650	0.783

Table 12.20: Precision of the number of terms which have at least 1 correct proposition with our methodology for $\boldsymbol{sp = \{1, 2, 3\}}$.

Chapter 13

Discussion and Conclusions

In this chapter, we discuss and conclude the results obtained with the proposed methodology to enrich biomedical ontologies. The main objective of this second part is to introduce concepts and to find the correct position in an already existing biomedical ontology. As a reminder, our methodology for enriching biomedical ontologies has three main steps, as described in Figure 13.1. Each section discusses and concludes the methodology process, its performance, obtained for : (i) polysemy detection, (ii) term sense induction, and (iii) semantic linkage.

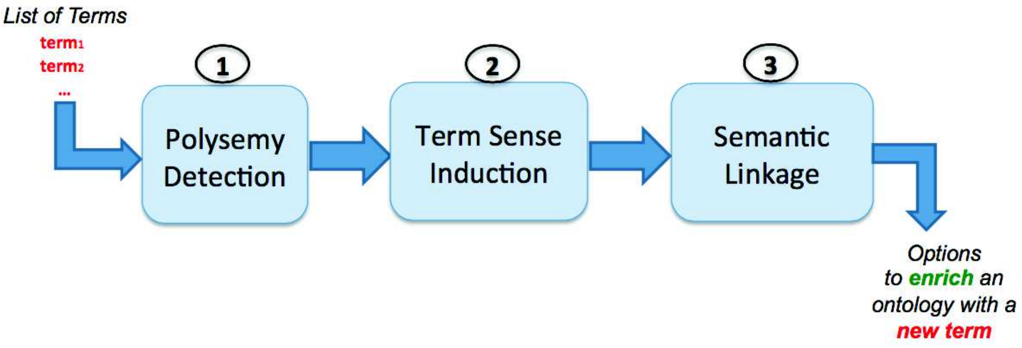


Figure 13.1: Workflow for Concept Extraction and Semantic Linkage.

13.1 Polysemy Detection

This section presents the conclusions of the polysemy detection methodology. For this, we present a novel approach focused on the biomedical domain to predict if a term is polysemic. The main contribution of this section is the definition of new meta-features, which are directly extracted from the text dataset and from an induced graph. Our novel approach is based on the extraction of new meta-features that better characterize our dataset. This allows a more efficient classification task (polysemy prediction). For this classification, we used the most well-known supervised algorithms over the whole meta-features.

Those meta-features are extracted in the following two ways. First, direct extraction from the dataset. This means that the characteristics more relevant and faster to obtain. Second, extraction from a graph, which is built according the dataset of each term. This allows us to take advantage of the graph properties to characterize the dataset.

First, we evaluated the direct meta-features, then the graph-based meta-features, and finally the performance when combining these two kinds of meta-features to obtain the best results. The results were calculated in terms of *Accuracy, Precision, Recall* and *F-Measure*. We obtained an accuracy (A) of between 95.3% and 97.8% when using the set of supervised algorithms on the combined meta-features.

Different strategies could be considered in the future, such as increasing the number of meta-features using other dictionaries like Wordnet associated with a general domain, or BabelNet, which contains general terms for several languages, including French and Spanish.

After the evaluation to determine if a term is polysemic, we applied the term sense induction methodology for all terms. The next section analyzes the term sense induction methodology.

13.2 Term Sense Induction

In this section, we describe the results of inducing sense for new terms. This represents the continuation of the Polysemy Detection process.

First, we presented a novel approach to predict the number of clusters (number of senses) for a new biomedical candidate term. The main contribution of this section is the definition of new internal indexes which, according to their values, can be used to predict the number of senses. This was also proved for a tweet clustering process.

We computed the values of these new internal indexes just for bag-of-words and graph representations. We obtained similar results, with 93.1% accuracy. For graph representation, we used the graph created in the previous sections for polysemy detection.

We also used the meta-features proposed in the previous section to predict the number of senses and, based on these, we could obtain the best results for this purpose. The meta-features for this task also allowed a good classification task. For the classification, we also used the most well-known supervised algorithms over the whole meta-features. The results for this task were calculated in terms of *Accuracy, Precision, Recall* and *F-Measure*. We observed maximum accuracy (A) of 93.6%.

Second, the selection of the clustering algorithm for textual data. This task is even easier than the first because we already know the number of parameter clusters. This is a well-known problem in the community that we solved specifically for biomedical term sense induction with the final objective of enriching ontologies.
With an induced sense (or concept), a term passed by an evaluation to be added to an ontology. That process is called semantic linkage. The next section presents the discussion an conclusions for this process.

13.3 Semantic Linkage

The aim of this process is to find the right position in an already established ontology for new biomedical terms associated with their senses.

We thus extracted the possible relations for a term. This relation was based only on the similarity context. We used the cosine similarity between the context of the new term and the context of an existing term in an ontology.

To conduct this step, we used a well-known biomedical ontology, called MeSH, i.e. the 2009 and 2015 version. We extracted 60 terms added between 2009 and 2015 in this ontology.

We computed the number of relations extracted for different pairwise distances $\boldsymbol{sp} = \{\mathbf{1}, \mathbf{2}, \mathbf{3}\}$. In major cases, the major number of relations proposition extracted happens when $\boldsymbol{sp} = \mathbf{1}$. We also evaluated this approach based on the top $\boldsymbol{k}$ first proposals.

This step could be used to extract the type of relation. This could be performed with the linguistic patterns, e.g. the verbs used between two terms, as detailed in the next chapter.

Chapter

14

GENERAL CONCLUSIONS

To conclude this PhD thesis, we first summarize the presented approaches, and end by describing the most important research prospects opened by them. This chapter presents an overview of our work, including the conclusions and perspectives of the two main parts: (i) Automatic biomedical term extraction, and (ii) Relation extraction and semantic linkage. Below, Section 14.1 presents a summary of our contributions, and then Section 14.2 presents the potential prospects and future directions of our research work.

14.1 Conclusions

As we mentioned in Chapter 1, NLP for biomedical ontology enrichment involves five challenges: (i) the text complexity and specialization to deal with; (ii) the complexity of extracting new terminology; (iii) the semantic concept of newly found terms, and the semantic relatedness; (iv) unifying several disciplines to set up a general workflow; and (v) to make this multidisciplinary aspect "user friendly" and "multilingual".

These challenges were divided in two groups: the first one representing the lexical complexity in the biomedical domain, and the second one representing the semantic complexity in the biomedical domain. Moreover, this thesis is divided in two main parts to address the previously mentioned challenges: (i) Automatic biomedical term extraction, and (ii) Concept extraction and semantic linkage. The following paragraphs summarize the contribution of each part, illustrating how they addressed the challenges.

Automatic Biomedical Term Extraction

This first part defined several measures for term extraction, while taking single-word and multi-word terms into account. These measures were classified as *ranking measures*, and *re-ranking measures*. These measures were based on linguistic, statistical, graph, and web information. We proposed new measures, as well as modifications of some baseline measures. The best ranking measure is *LIDF-value* (linguistic and statistic based), and the best re-ranking measures were *TeRGraph* (graph based), and *WAHI* (web based).

The use of biomedical linguistic features allows extraction of complex terminology associated with the biomedical domain. The use of statistical, graph, and web features improves the extraction of new terminology, which also addresses the second challenge of NLP, i.e. to build/enrich biomedical ontologies, due to their complex structure.

Our ranking measures and re-ranking measures were proposed to work on biological, life science, and medical texts, being those different. Indeed, they were applied on datasets such as GENIA, LabTestsOnline and Agricultural datasets, generating very good results. This addresses the first challenge associated with the specialized and complex biomedical text to be processed.

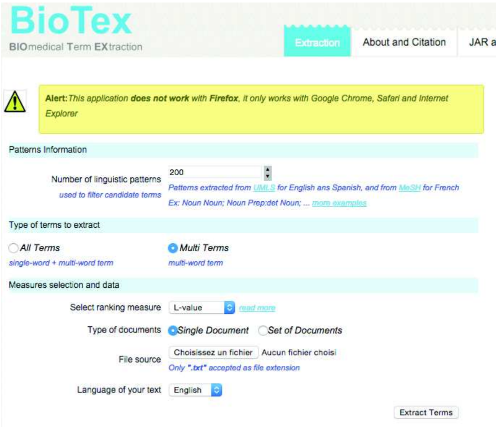
In addition, as previously mentioned, these measures were applied to different languages, i.e. English, French, and Spanish. Furthermore, we created an application called BioTex, which implements all of the ranking measures. Thus addressing the user-friendly and multilingual challenge (fifth challenge).

To fulfill the objectives of this part, we used linguistic, statistical, graph and web methods to propose a methodology, while also developing BioTex, which includes software development. For this, several disciplines were useful to address the fourth challenge.
As previously mentioned, we developed an application, called Biotex, which implements the proposed measures. We underline the main characteristics of this application in the next section.

BioTex

We presented the BioTex application for extracting biomedical terms. This involves the implementation of all the ranking measures previously defined. It is available for online testing and evaluation, but it can also be used in any program as a Java library (POS tagger not included). In contrast to other existing systems, our system allows us to analyze French and Spanish datasets, to manually validate extracted terms and to export the list of extracted terms.

BioTex is starting to be a valuable tool for the biomedical community. It has also generated interesting results for other domains. Indeed, it is currently used in other independent projects (i.e. use-cases). In addition, it is being used in a couple of test-beds within the SIFR project[1]. BioTex, as web application, presents two use cases: the first one, to extract terms (see Figure 14.1); and the second one, to validate the extracted terms (see Figure 14.2).



Alert: *This application* ***does not work*** *with* ***Firefox****, it only works with Google Chrome, Safari and Internet Explorer*

Patterns Information

Number of linguistic patterns *used to filter candidate terms*: 200
Patterns extracted from UMLS for English ans Spanish, and from MeSH for French
Ex: Noun Noun; Noun Prep:det Noun; ... more examples

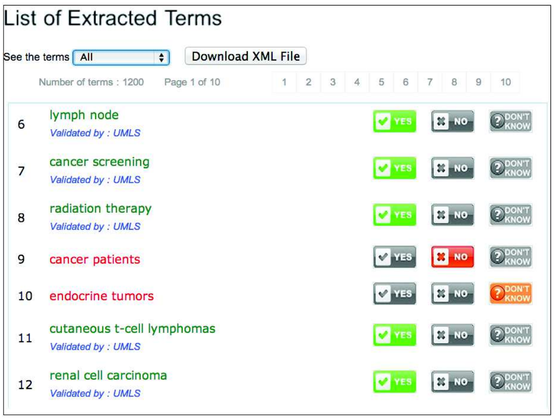
Type of terms to extract

○ *All Terms* — *single-word + multi-word term*
● *Multi Terms* — *multi-word term*

Measures selection and data

Select ranking measure: L-value *read more*
Type of documents: ● *Single Document* ○ *Set of Documents*
File source: Choisissez un fichier Aucun fichier choisi
Only ".txt" accepted as file extension
Language of your text: English

Figure 14.1: BioTex Term extraction interface.



List of Extracted Terms

See the terms: All

Number of terms : 1200 Page 1 of 10 1 2 3 4 5 6 7 8 9 10

6 lymph node
Validated by : UMLS

7 cancer screening
Validated by : UMLS

8 radiation therapy
Validated by : UMLS

9 cancer patients

10 endocrine tumors

11 cutaneous t-cell lymphomas
Validated by : UMLS

12 renal cell carcinoma
Validated by : UMLS

Figure 14.2: BioTex Term validation interface.

As previously mentioned, this part concerns extracting candidate terms. After this step, the terms have to be added to an ontology. For this, we proposed a methodology to determine the possible sense (s) of a term, and to figure out the position in an ontology. The next section highlights the main characteristics of our proposed approach.

Concept Extraction and Semantic Linkage

In this second part, we proposed a methodology to extract concepts and to add the new terms in the MeSH ontology. This second part offered a process workflow to enrich biomedical ontologies in an automatic manner, according to three steps: (i) Polysemy detection, (ii) Term sense induction, and (iii) Semantic linkage. These three steps were executed consecutively. The experiments carried out in all of the steps generated interesting results.

[1] http://www.lirmm.fr/sifr

For polysemy detection, we presented a novel approach to predict if a term is polysemic. This approach was based on the extraction of new meta-features. This allowed a more efficient classification task (polysemy prediction). We used the most well-known supervised algorithms for this classification. The meta-features were extracted in two ways. First, they were extracted directly from the dataset. Second, they were extracted from a term co-occurrence graph derived from the dataset.

For term sense induction, as cited in previous chapters, a major problem is to identify the right number of senses of a term. Therefore, we presented a novel approach to predict the number of clusters (number of senses) for a new biomedical candidate term. The main contribution is the definition of new internal indexes which, according their values, allow us to predict the number of senses. These internal indexes are based on the clustering task by using bag-of-words and graph-of-words approaches. These indexes have the same behavior for the evaluation of clustering of tweets, which proves that these indexes tend to find lower values for the number of senses.

Meta-features used for the polysemy detection task were also efficient to predict the number of senses, which generated the best results for this purpose. For the classification, we also used the most well-known supervised algorithms for all the meta-features.
Then the sense or concept was induced by selecting the most representative features after having applied a clustering algorithm for textual data, taking the previous computed number of clusters as input.

Finally, the aim of the the semantic linkage task was to find the relevant position (or location) of a term in an already established ontology. Note that at this step we have a candidate term associated with its sense. Therefore, we proposed to extract several possible relations for a term. This relation is based on the similarity context. We used the cosine similarity between the context of the new term and the context of an existing term in an ontology. To perform a quantitative assessment (right position) of our methodology, we used MeSH terms, i.e. those added between 2009 and 2015. This approach gave relevant results on MeSH terms, i.e. finding the right position for the evaluated terms.

These three steps together provided the semantic sense of a new term. It also offered the semantic relatedness of a new term with a term already existing in an ontology. Hence, they addressed the third challenge, providing semantic (sense and linkage) for a new term. As we have seen, several disciplines, such as word sense induction, polysemy detection, machine learning, clustering and meta-learning are involved to meet the main objective.

To address the fifth challenge, the methodology containing the three steps has to be implemented and also evaluated for French and Spanish. This methodology did not provide the type of relation between terms (e.g. hyperonymy, synonymy, etc.). Hence, for biomedical ontology enrichment, we could propose expert intervention at the end of the entire workflow.

In this thesis, we identified several future biomedical ontology prospects, which are described in next section.

14.2 Perspectives

All of the proposed approaches in this thesis open new perspectives and future research directions. These form the basis for future work or perspectives required for the main objective. The following paragraphs summarize the identified prospects of each part.

Automatic Biomedical Term Extraction

Four features, such as linguistic, statistical, graph, and web features were used for automatic biomedical term extraction. It could also be interesting to use a fifth feature, i.e. the knowledge feature. The main idea is to create a measure to favour candidate terms consisting of nested terms existing in a terminology. The measure will be focused on multi-word term extraction, e.g. the extracted term "pancreatic cancer center". Its nested terms belonging to UMLS terminology are: *cancer, pancreatic cancer*. According to these terms, the knowledge based measure will favour this term.

The relation value between terms could be used for the graph-based measure. Another idea is to use other graph ranking computations, e.g. PageRank, tailored for automatic term extraction.

Moreover, future work will involve using the web to extract more terms than those already extracted. Moreover, another search engine could be used, such as Exalead[2], which offers terms related to that used in the query. This will generate semantic information for candidate terms to improve the term extraction. The three above-mentioned prospects are currently being evaluated in collaboration with the NaCTeM team, at the University of Manchester.

Finally, we could potentially modify our measures in order to standardize the possible variants, looking towards for a preferred term for those variants.

[2] `https://www.exalead.com/search/`

BioTex

Concerning our application, a very interesting strategy could be to add the context visualization of each candidate term. For instance, to show the most relevant sentences containing the candidate term. This might enhance the validation of candidate terms. This could also allow manual recognition of whether a candidate term is ambiguous (or polysemic).

In addition, a possible future work will be to present a graph of term co-occurrences, showing the relatedness between the candidate terms, while determinging the right position for new terms in a biomedical ontology. Note that the candidate terms are terms present in a terminology (green terms), as well as new candidate terms (red terms).

Concept Extraction and Semantic Linkage

The most important direction of future work of this thesis concerns the "relation extraction" process. This process seeks to identify the type of relationship between two terms. For instance, in the biological domain, gene-gene interactions and protein-disease interactions are types of relation, as well as relations between cancer-related genes, drugs and cell lines.

Actually, the extraction of relation type between terms in natural language text is a crucial step towards automatic biomedical ontology enrichment/construction. Relation extraction approaches can be categorized as [Bach and Badaskar, 2007]: (i) unsupervised relation discovery, and (ii) supervised classification. Hence, we might propose both types of approaches. The annotated dataset GENIA could be used to test the supervised approaches. However, Big Data currently does not contain annotated datasets. Therefore, there is a need to further develop unsupervised approaches.

Following the distributional approach, the syntactic information of specific verbs could be combined, as done in [Nguyen et al., 2015]. Indeed, we consider that the relevant context is not the set of co-occurrences but rather the set of elements which are in a syntactic relationship with the target term. So the main idea is to find the verbs used most between two terms. Normalize verbs with lemmatization, and propose the most representative ones. At the end of the workflow, the expert will decide on the most relevant type of relation, thus substantially decreasing the need for expert intervention.

One more strategy could be considered in the future, such as increasing the number of meta-features using other dictionaries like Wordnet, which is associated with a general domain. However this dictionary is only available for English, so it would thus be more interesting to use BabelNet, which contains general terms for several languages, including French and Spanish.

Finally, we think it will be very useful to develop an application to unify the issues covered in the first and the second part of this thesis.

List of Figures

1.1 Processes of the SIFR Lyfecicle. 5

5.1 Workflow Methodology for Biomedical Term Extraction. 34
5.2 Section to Extract Patterns. 35
5.3 Section to Extract Candidate Terms. 39
5.4 Section Describing the Ranking Measures. 40
5.5 Merging lists. 43
5.6 Section Describing the Re-ranking Measures. 46
5.7 Importance of a term in a domain 47
5.8 TeRGrpah's value for $\boldsymbol{k = \{0.5; 1; 1.5; 2\}}$ 49
5.9 *TeRGraph*'s value for ***chloramphenicol acetyltransferase reporter*** 49

6.1 resultats with R 61
6.2 Precision comparison with *LIDF-value* and baseline measures 63
6.3 Precision comparison of 2-gram terms 65
6.4 Precision comparison of 3-gram terms 65
6.5 Precision comparison of 4+ gram terms 65
6.6 Precision comparison of *LIDF-value* and *TeRGraph* 70

8.1 BioTex Term extraction interface. 79
8.2 BioTex Term validation interface. 80

9.1 Workflow for Concept Extraction and Semantic Linkage. 90
9.2 Final Proposed Workflow for Concept Extraction and Semantic Linkage. 92

10.1 Workflow for Concept Extraction and Semantic Linkage. 93

11.1 Workflow for Concept Extraction and Semantic Linkage. 121
11.2 Solving the Polysemy Detection issue 122
11.3 Workflow Methodology for Polysemy Prediction. 123
11.4 Graph created for the $\boldsymbol{t}$ term. 126
11.5 Subgraph created for the term t = **Yellow Fever** 129
11.6 Solving the Term Sense Induction. 132
11.7 Term Sense Induction Workflow. 132
11.8 Semantic Linkage of a Candidate Term. 140
11.9 Semantic Linkage Workflow. 140
11.10First Semantic Linkage step. 141
11.11Second Semantic Linkage step. 142

12.1 Data sets and Results of the Proposed Methodology. 143
12.2 Evaluation of Polysemy Detection 144
12.3 Decision Tree obtained from the Polysemic and Non-polysemic Data set. 149
12.4 Evaluation of Term Sense Induction 150
12.5 Semantic Linkage Evaluation. 159
12.6 Semantic LinkageData set 161
12.7 First Semantic Linkage step. 161
12.8 Semantic Linkage Workflow. 162
12.9 Shortest Path for the term $\boldsymbol{V_3}$. 163
12.10Term *corneal injuries* in the MeSH ontology. 163

13.1 Workflow for Concept Extraction and Semantic Linkage. 169

14.1 BioTex Term extraction interface. 175
14.2 BioTex Term validation interface. 175

List of Tables

4.1 Summarization of Methodologies for Automatic Term Extraction. . . 32

5.1 Example of with complex linguistic pattern (where *NN* is a noun, *IN* a preposition or subordinating conjunction, *JJ* an adjective, and *CD* a cardinal number) 37
5.2 Example of pattern construction (where *NN* is a noun, *IN* a preposition or subordinating conjunction, *JJ* an adjective, and *CD* a cardinal number) 38
5.3 Calculation of $\boldsymbol{w(A)}$ 41

6.1 Details of LabTestsOnline data set. 56
6.2 Details of Available Resources for Validation. 57
6.3 Biomedical Term Extraction for English 58
6.4 Biomedical Term Extraction for French 59
6.5 Biomedical Term Extraction for Spanish 60
6.6 *p-value* between *LIDF* and *TF-IDF*$_{\boldsymbol{S}}$. 61
6.7 Precision comparison of *LIDF-value* with baseline measures 62
6.8 Precision comparison of 2-gram terms, 3-gram terms, and 4+ gram terms 64
6.9 Comparison of top-20 ranked 2-gram terms (irrelevant terms are italicized and marked with *). 66
6.10 Comparison of the top-10 ranked 3-gram terms (irrelevant terms are italicized and marked with *). 66

6.11 Comparison of the top-10 ranked 4+ gram terms (irrelevant terms are italicized and marked with *). 67
6.12 Precision performance of *TeRGraph* when varying $\boldsymbol{\delta}$ (*threshold* parameter for Dice) 68
6.13 Precision comparison of *WAHI with YAHOO* and word association measures 69
6.14 Precision comparison of *WAHI with BING* and word association measures 69
6.15 Time execution comparison between Bing and Yahoo in seconds. . . 70
6.16 Precision comparison of *LIDF-value* and *TeRGraph* 71
6.17 Precision comparison *LIDF-value*, *TeRGraph*, and *WAHI* 72

7.1 Details of Cirad corpus. 74
7.2 Precision comparison of Term Extraction with Agronomic and Biomedical Patterns 75

8.1 Results of validation in percentage. 81
8.2 Alignment of Biomedical Terms and Patient Terms. 82
8.3 Brief comparison of biomedical terminology extraction applications. . 84

11.1 An Extract of the Titles/Abstracts dataset for the term **Yellow Fever** 128
11.2 Direct Meta-features for t = **Yellow Fever** 129
11.3 Graph-based Meta-features for t = **Yellow Fever** 130
11.4 Details of Polysemic Terms in UMLS. 133
11.5 Details of Polysemic Terms in MeSH. 133

12.1 Principle of Non-polysemic Data set Creation. 145
12.2 Extract of Polysemic and Non-polysemic Data set. 146
12.3 Details of our Polysemic/Non-polysemic Data set 146
12.4 Direct Meta-features 148
12.5 Graph-based Meta-features 148
12.6 Combining two kinds of meta-features 149
12.7 Confusion Matrix based on the Polysemic and Non-polysemic Data set. 150
12.8 Sense Number per Term. 151
12.9 Details of our Polysemic Data set 151
12.10 Choosing $\boldsymbol{k}$ according to $\boldsymbol{a_{k,I2}}$ and $\boldsymbol{c_{k,I2}}$ values. 153
12.11 $\boldsymbol{k}$ Prediction for *Yellow Fever*, with bag-of-words representation (2 classes) 153
12.12 Accuracy (A) results for 203 Ambiguous Entities with bag-of-words. . 154
12.13 $\boldsymbol{k}$ Prediction for *Yellow Fever* with Graph representation (2 classes) . 156
12.14 Accuracy results for 203 Ambiguous Entities with Graph representation. 156
12.15 Meta-features for Number of Sense Prediction. 157
12.16 Details of our Semantic Linkage Data set 160
12.17 Proposition on where to add the term *corneal injuries*. 164

12.18 Evaluation of propositions to add *corneal injuries*, with $\boldsymbol{sp = 1}$ (fathers, sons and synonyms). 164
12.19 Evaluation of propositions put forward for the 15 terms for $\boldsymbol{sp = \{1, 2, 3\}}$. 166
12.20 Precision of the number of terms which have at least 1 correct proposition with our methodology for $\boldsymbol{sp = \{1, 2, 3\}}$. 167

Bibliography

[Abacha and Zweigenbaum, 2011] Abacha, A. B. and Zweigenbaum, P. (2011). Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of Biomedical Semantics*, 2. *2 citations in pages 85 and 114*

[Adam et al., 2013] Adam, C., Fabre, C., and Muller, P. (2013). Évaluer et améliorer une ressource distributionnelle : protocole d'annotation de liens sémantiques en contexte. *Traitement Automatique des Langues*, vol. 54(n° 1):pp. 71–97. *Cited in page 115*

[Aggarwal and Zhai, 2012] Aggarwal, C. C. and Zhai, C. (2012). A survey of text clustering algorithms. In *Mining Text Data*, pages 77–128. Springer. *Cited in page 139*

[Agirre et al., 2014] Agirre, E., de Lacalle, O. L., and Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84. *2 citations in pages 105 and 106*

[Agirre and Edmonds, 2007] Agirre, E. and Edmonds, P. (2007). *Word Sense Disambiguation: Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition. *Cited in page 102*

[Agirre et al., 2006] Agirre, E., Martínez, D., de Lacalle, O. L., and Soroa, A. (2006). Two graph-based algorithms for state-of-the-art wsd. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 585–593, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 106*

[Agirre and Soroa, 2007a] Agirre, E. and Soroa, A. (2007a). Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 7–12, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 158*

[Agirre and Soroa, 2007b] Agirre, E. and Soroa, A. (2007b). Ubc-as: A graph based unsupervised system for induction and classification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 346–349, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 102*

[Agirre and Soroa, 2009] Agirre, E. and Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics. *2 citations in pages 91 and 106*

[Agirre et al., 2010] Agirre, E., Soroa, A., and Stevenson, M. (2010). Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896. *2 citations in pages 91 and 107*

[Ahmad et al., 1999] Ahmad, K., Gillam, L., and Tostevin, L. (1999). University of surrey participation in TREC-8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *TREC*. *2 citations in pages 22 and 32*

[Aimé, 2015] Aimé, X. (2015). Eléments de réflexion sur l'utilisation de corpus pour la construction d'ontologies. In *Actes de 25es journées francophones d'Ingénierie des Connaissances*, IC'2015. *Cited in page 81*

[Aimé et al., 2012] Aimé, X., Dhombres, F., Fürst, F., Kuntz-Cosperec, P., Trichet, F., and Charlet, J. (2012). Rare diseases knowledge management: the contribution of proximity measurements in OntoOrpha and OMIM. In *The 24th European Medical Informatics Conference*, EMIC'2012, pages 88–92, Pise, Italy. *Cited in page 2*

[Albano et al., 2014] Albano, L., Beneventano, D., and Bergamaschi, S. (2014). Word sense induction with multilingual features representation. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 2, pages 343–349. IEEE. *Cited in page 99*

[Albano et al., 2015] Albano, L., Beneventano, D., and Bergamaschi, S. (2015). Multilingual word sense induction to improve web search result clustering. In *Proceedings of the 24th International Conference on World Wide Web Companion*, WWW '15, pages 835–839, New York, NY, USA. International World Wide Web Conferences Steering Committee, ACM. *Cited in page 99*

[Albatineh and Niewiadomska-Bugaj, 2011] Albatineh, A. N. and Niewiadomska-Bugaj, M. (2011). Mcs: A method for finding the number of clusters. *Journal of classification*, 28(2):184–209. *Cited in page 108*

[an tSaoir, 2014] an tSaoir, R. M. (2014). Using spreading activation to evaluate and improve ontologies. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING'14, pages 2237–2248. *Cited in page 106*

[Anderson, 2001] Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1):32–46. *Cited in page 107*

[Anguiano, 2013] Anguiano, E. H. (2013). *Efficient large-context dependency parsing and correction with distributional lexical resources.* PhD thesis, Université Paris-Diderot-Paris VII. *Cited in page 115*

[Arsevska et al., 2014] Arsevska, E., Roche, M., Lancelot, R., Hendrikx, P., and Dufour, B. (2014). Exploiting textual source information for epidemiosurveillance. *Metadata and Semantics Research*, 359. *Cited in page 83*

[Assadi, 1997] Assadi, H. (1997). Knowledge acquisition from texts: Using an automatic clustering method based on noun-modifier relationship. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 504–506. Association for Computational Linguistics. *Cited in page 29*

[Aubin and Hamon, 2006] Aubin, S. and Hamon, T. (2006). Improving term extraction with terminological resources. In *Proceedings of the 5th International Conference Natural Language Processing*, FinTAL'06, pages 380–387. Springer, Turku, Finland. *Cited in page 25*

[Aussenac-Gilles et al., 2000] Aussenac-Gilles, N., Biebow, B., and Szulman, S. (2000). Revisiting ontology design: A methodology based on corpus analysis. In *Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*, EKAW'00, pages 172–188, London, UK, UK. Springer-Verlag. *Cited in page 10*

[Aussenac-Gilles and Bourigault, 2000] Aussenac-Gilles, N. and Bourigault, D. (2000). The th (ic) 2 initiative: Corpus-based thesaurus construction for indexing www documents. In *Proceedings of the EKAW conference*, page 3. *2 citations in pages 20 and 32*

[Aussenac-Gilles et al., 2008] Aussenac-Gilles, N., Despres, S., and Szulman, S. (2008). The terminae method and platform for ontology engineering from texts. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge*, pages 199–223, Amsterdam, The Netherlands, The Netherlands. IOS Press. *Cited in page 10*

[Bach and Badaskar, 2007] Bach, N. and Badaskar, S. (2007). A review of relation extraction. *3 citations in pages 91, 113, and 178*

[Baker and Hubert, 1975] Baker, F. B. and Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349):31–38. *Cited in page 109*

[Baldwin et al., 2013] Baldwin, T., Li, Y., Alexe, B., and Stanoi, I. R. (2013). Automatic term ambiguity detection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 804–809, Sofia, Bulgaria. *Cited in page 94*

[Ball and Hall, 1965] Ball, G. H. and Hall, D. J. (1965). Isodata, a novel method of data analysis and pattern classification. Technical report, DTIC Document. *Cited in page 109*

[Banerjee et al., 2014] Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2014). Gossip: Identifying central individuals in a social network. Technical report, National Bureau of Economic Research. *Cited in page 27*

[Baneyx et al., 2007] Baneyx, A., Charlet, J., and Jaulent, M.-C. (2007). Building an ontology of pulmonary diseases with natural language processing tools using textual corpora. *International Journal of Medical Informatics*, 76(2):208–215. *Cited in page 114*

[Banko and Etzioni, 2008] Banko, M. and Etzioni, O. (2008). The tradeoffs between open and traditional relation extraction. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, ACL'08, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 114*

[Barrón-Cedeño et al., 2009] Barrón-Cedeño, A., Sierra, G., Drouin, P., and Ananiadou, S. (2009). An improved automatic term recognition method for spanish. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'09, pages 125–136. Springer. *2 citations in pages 23 and 41*

[Bensusan et al., 2000] Bensusan, H., Giraud-Carrier, C., and Kennedy, C. (2000). A higher-order approach to meta-learning. Technical report, Bristol, UK, UK. *Cited in page 112*

[Bhatt et al., 2012] Bhatt, N., Thakkar, A., and Ganatra, A. (2012). A. a survey & current research challenges in meta learning approaches based on dataset characteristics. *International Journal of Soft Computing and Engineering*, 10(2):234–247. *Cited in page 110*

[Bhatt et al., 2013] Bhatt, N., Thakkar, A., Ganatra, A., and Bhatt, N. (2013). Ranking of classifiers based on dataset characteristics using active meta learning. *International Journal of Computer Applications*, 69(20):31–36. Full text available. *Cited in page 110*

[Blanco and Lioma, 2012] Blanco, R. and Lioma, C. (2012). Graph-based term weighting for information retrieval. *Information Retrieval*, 15(1):54–92. *2 citations in pages 27 and 32*

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022. *2 citations in pages 94 and 104*

[Blomqvist, 2009] Blomqvist, E. (2009). Semi-automatic ontology construction based on patterns. *PhD Thesis*. *Cited in page 2*

[Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270. *Cited in page 106*

[Bodenreider, 2008] Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of medical informatics*, page 67. *Cited in page 2*

[Boldi and Vigna, 2014] Boldi, P. and Vigna, S. (2014). Axioms for centrality. *Internet Mathematics*, 10(3-4):222–262. *Cited in page 27*

[Bollegala et al., 2010] Bollegala, D. T., Matsuo, Y., and Ishizuka, M. (2010). Relational duality: Unsupervised extraction of semantic relations between entities on the web. In *Proceedings of the 19th International Conference on World Wide Web*, WWW'10, pages 151–160, New York, NY, USA. ACM. *Cited in page 114*

[Booth et al., 2008] Booth, J. G., Casella, G., and Hobert, J. P. (2008). Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):119–139. *Cited in page 134*

[Bordag, 2004] Bordag, S. (2004). Word sense induction: Triplet-based clustering and automatic evaluation. In *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics*, EACL '04. *Cited in page 98*

[Borgatti, 2005] Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, 27(1):55–71. *Cited in page 27*

[Borgatti et al., 2009] Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. (2009). Network analysis in the social sciences. *science*, 323(5916):892–895. *Cited in page 27*

[Bourigault, 1993] Bourigault, D. (1993). Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *TAL. Traitement automatique des langues*, 34(2):105–117. *2 citations in pages 20 and 32*

[Bourigault and Fabre, 2000] Bourigault, D. and Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, 25:131–151. *3 citations in pages 23, 31, and 32*

[Bourigault et al., 2005] Bourigault, D., Fabre, C., Frérot, C., Jacques, M.-P., and Ozdowska, S. (2005). Syntex, analyseur syntaxique de corpus. In *TALN'2005*, pages 17–20. *3 citations in pages 23, 31, and 32*

[Bourigault and Jacquemin, 1999] Bourigault, D. and Jacquemin, C. (1999). Term extraction + term clustering: An integrated platform for computer-aided terminology. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, pages 15–22. *2 citations in pages 20 and 32*

[Bowker and Pearson, 2002] Bowker, L. and Pearson, J. (2002). *Working with specialized language: a practical guide to using corpora*. Routledge. *Cited in page 24*

[Boyd-Graber and Blei, 2007] Boyd-Graber, J. and Blei, D. (2007). Putop: Turning predominant senses into a topic model for word sense disambiguation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 277–281, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 105*

[Boyd-Graber et al., 2007] Boyd-Graber, J. L., Blei, D. M., and Zhu, X. (2007). A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 1024–1033. *Cited in page 105*

[Brazdil et al., 2008] Brazdil, P., Giraud-Carrier, C., Soares, C., and Vilalta, R. (2008). *Metalearning: Applications to Data Mining*. Springer Publishing Company, Incorporated, 1 edition. *Cited in page 112*

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117. *Cited in page 102*

[Brody and Lapata, 2009] Brody, S. and Lapata, M. (2009). Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–111. Association for Computational Linguistics. *2 citations in pages 104 and 158*

[Budanitsky and Hirst, 2006a] Budanitsky, A. and Hirst, G. (2006a). Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, 32(1):13–47. *Cited in page 115*

[Budanitsky and Hirst, 2006b] Budanitsky, A. and Hirst, G. (2006b). Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, 32(1):13–47. *Cited in page 115*

[Buitelaar et al., 2004] Buitelaar, P., Olejnik, D., and Sintek, M. (2004). A protégé plug-in for ontology extraction from text based on linguistic analysis. In *Proceeding of the Semantic Web: Research and Applications, First European Semantic Web Symposium*, ESWS'04, pages 31–44, Heraklion, Crete, Greece. *Cited in page 11*

[Bunescu and Mooney, 2005] Bunescu, R. C. and Mooney, R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT-EMNLP'05, pages 724–731, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 114*

[Cai et al., 2007] Cai, J., Lee, W. S., and Teh, Y. W. (2007). Improving word sense disambiguation using topic features. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '07, pages 1015–1023. *Cited in page 105*

[Caliński and Harabasz, 1974] Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27. *Cited in page 109*

[Casellas, 2011] Casellas, N. (2011). *Legal ontology engineering: methodologies, modelling trends, and the ontology of professional judicial knowledge*, volume 3. Springer Science & Business Media. *Cited in page 2*

[Castiello et al., 2005] Castiello, C., Castellano, G., and Fanelli, A. M. (2005). Meta-data: Characterization of input features for meta-learning. In *Modeling Decisions for Artificial Intelligence*, pages 457–468. Springer. *Cited in page 112*

[Chan et al., 2007] Chan, Y. S., Ng, H. T., and Zhong, Z. (2007). Nus-pt: exploiting parallel texts for word sense disambiguation in the english all-words tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 253–256, Prague, Czech Republic. Association for Computational Linguistics. *Cited in page 106*

[Charlet et al., 2006] Charlet, J., Bachimont, B., and Jaulent, M.-C. (2006). Building medical ontologies by terminology extraction from texts: an experiment for the intensive care units. *Computers in biology and medicine*, 36(7):857–870. *Cited in page 114*

[Chasin et al., 2014] Chasin, R., Rumshisky, A., Uzuner, O., and Szolovits, P. (2014). Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods. *Journal of the American Medical Informatics Association*, 21(5):842–849. *Cited in page 104*

[Chaudhari et al., 2011] Chaudhari, D. L., Damani, O. P., and Laxman, S. (2011). Lexical co-occurrence, statistical significance, and word association. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'11, pages 1058–1068, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 28*

[Chen et al., 2005] Chen, J., Ji, D., Tan, C. L., and Niu, Z. (2005). Unsupervised feature selection for relation extraction. In *Proceedings of the International Joint Conference on Natural Language Processing*, IJCNLP'05. *Cited in page 113*

[Chen et al., 2009] Chen, P., Ding, W., Bowes, C., and Brown, D. (2009). A fully unsupervised word sense disambiguation method using dependency knowledge. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 100*

[Choe and Charniak, 2013] Choe, D. K. and Charniak, E. (2013). Naive bayes word sense induction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 1433–1437, Seattle, Washington, USA. Association for Computational Linguistics. *2 citations in pages 104 and 158*

[Chuang et al., 2012] Chuang, J., Manning, C. D., and Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI'12, pages 74–77, New York, NY, USA. ACM. *Cited in page 85*

[Church and Hanks, 1990] Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29. *2 citations in pages 21 and 32*

[Cilibrasi and Vitanyi, 2007] Cilibrasi, R. L. and Vitanyi, P. M. (2007). The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383. *3 citations in pages 28, 32, and 50*

[Cimiano and Völker, 2005] Cimiano, P. and Völker, J. (2005). Text2onto: A framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th International Conference on Natural Language Processing and Information Systems*, NLDB'05, pages 227–238, Berlin, Heidelberg. Springer-Verlag. *2 citations in pages 2 and 10*

[Clauset et al., 2007] Clauset, A., Moore, C., and Newman, M. E. (2007). Structural inference of hierarchies in networks. In *Statistical network analysis: models, issues, and new directions*, pages 1–13. Springer. *Cited in page 103*

[Claveau and Kijak, 2015] Claveau, V. and Kijak, E. (2015). Thésaurus distributionnels pour la recherche d'information et vice-versa. In *CORIA 2015 - Conférence en Recherche d'Infomations et Applications - 12th French Information Retrieval Conference, Paris, France, March 18-20, 2015.*, pages 405–420. *Cited in page 115*

[Claveau et al., 2014] Claveau, V., Kijak, E., and Ferret, O. (2014). Improving distributional thesauri by exploring the graph of neighbors. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 709–720, Dublin, Ireland. Dublin City University and Association for Computational Linguistics. *Cited in page 115*

[Clouet and Daille, 2014] Clouet, E. and Daille, B. (2014). Compound terms and their multi-word variants: Case of german and russian languages. In *Computational Linguistics and Intelligent Text Processing*, pages 68–78. Springer. *Cited in page 22*

[Conrado et al., 2013] Conrado, M. S., Pardo, T. A., and Rezende, S. O. (2013). Exploration of a rich feature set for automatic term extraction. In *Advances in Artificial Intelligence and Its Applications*, volume 8265 of *Lecture Notes in Computer Science*, pages 342–354. Springer Berlin Heidelberg. *3 citations in pages 16, 22, and 32*

[Corcho et al., 2005] Corcho, O., Fernández-López, M., Gómez-Pérez, A., and López-Cima, A. (2005). Law and the semantic web. chapter Building Legal Ontologies with METHONTOLOGY and WebODE, pages 142–157. Springer-Verlag, Berlin, Heidelberg. *Cited in page 9*

[Crossley et al., 2010] Crossley, S., Salsbury, T., and McNamara, D. (2010). The development of polysemy and frequency use in english second language speakers. *Language Learning*, 60(3):573–605. *2 citations in pages 90 and 122*

[Cunningham et al., 2002] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). Gate: an architecture for development of robust hlt applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL'02, pages 168–175. Association for Computational Linguistics. *Cited in page 10*

[Curran and Moens, 2002] Curran, J. R. and Moens, M. (2002). Scaling context space. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 231–238, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 115*

[Dagan and Church, 1997] Dagan, I. and Church, K. (1997). Termight: Coordinating humans and machines in bilingual terminology acquisition. *Machine Translation*, 12(1-2):89–107. *2 citations in pages 23 and 32*

[Daille, 1994] Daille, B. (1994). *APPROCHE MIXTE POUR L'EXTRACTION DE TERMINOLOGIE: STATISTIQUE LEXICALE ET FILTRES LINGUISTIQUES.* *5 citations in pages 21, 23, 30, 32, and 36*

[Daille, 1996] Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. *The balancing act: Combining symbolic and statistical approaches to language*, 1(1):49–66. *4 citations in pages 23, 30, 32, and 36*

[Daille, 1998] Daille, B. (1998). An evaluation of statistical scores for word association. In *Proceedings of the Tbilisi Symposium on Logic, Language and Computation: Selected Papers, CSLI Publications*, pages 177–188. *Cited in page 23*

[Daille et al., 1994] Daille, B., Gaussier, E., and Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, COLING'94, pages 515–521, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 36*

[Daille and Morin, 2005] Daille, B. and Morin, E. (2005). French-english terminology extraction from comparable corpora. In *Proceedings of the 2nd International Joint Conference Natural Language Processing*, IJCNLP'05, pages 707–718. Springer. *Cited in page 24*

[Darmoni et al., 2009] Darmoni, S. J., Pereira, S., Sakji, S., Merabti, T., Prieur, É., Joubert, M., and Thirion, B. (2009). Multiple terminologies in a health portal: automatic indexing and information retrieval. In *Artificial Intelligence in Medicine*, pages 255–259. Springer. *Cited in page 85*

[David and Plante, 1990] David, S. and Plante, P. (1990). De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *Intelligence artificielle et sciences cognitives au Québec*, 3(3):140–154. *2 citations in pages 20 and 32*

[Davies and Bouldin, 1979] Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227. *Cited in page 109*

[Deborah et al., 2011] Deborah, L. J., Baskaran, R., and Kannan, A. (2011). Ontology construction using computational linguistics for e-learning. In *Visual Informatics: Sustaining Research and Innovations*, pages 50–63. Springer. *Cited in page 2*

[Decadt et al., 2004] Decadt, B., , Decadt, B., Hoste, V., Daelemans, W., and Bosch, A. V. D. (2004). Gambl: Genetic algorithm optimization of memory-based wsd. In *Proceedings of the 3rd Interna- tional Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 108–112, Barcelona, Spain. *Cited in page 106*

[Dehkordi et al., 2009] Dehkordi, M. Y., Boostani, R., and Tahmasebi, M. (2009). A novel hybrid structure for clustering. In *Advances in Computer Science and Engineering*, pages 888–891. Springer. *2 citations in pages 90 and 107*

[Déjean and Gaussier, 2002] Déjean, H. and Gaussier, E. (2002). *Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables.* *Cited in page 24*

[Deléger et al., 2009] Deléger, L., Merkel, M., and Zweigenbaum, P. (2009). Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4):692–701. *Cited in page 24*

[Di Marco and Navigli, 2011] Di Marco, A. and Navigli, R. (2011). Clustering web search results with maximum spanning trees. In *Proceedings of the 12th International Conference on Artificial Intelligence Around Man and Beyond*, AI*IA'11, pages 201–212, Berlin, Heidelberg. Springer-Verlag. *Cited in page 103*

[Di Marco and Navigli, 2013] Di Marco, A. and Navigli, R. (2013). Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754. *Cited in page 103*

[Dixit et al., 2012] Dixit, P., Sethi, S., Sharma, A., and Dixit, A. (2012). Design of an automatic ontology construction mechanism using semantic analysis of the documents. In *Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on*, pages 611–616. IEEE. *Cited in page 2*

[Dobrov and Loukachevitch, 2011] Dobrov, B. and Loukachevitch, N. (2011). Multiple evidence for term extraction in broad domains. In *Proceeding of Recent Advances in Natural Language Processing*, RANLP'11, pages 710–715, Hissar, Bulgaria. *Cited in page 16*

[Doing-Harris et al., 2015] Doing-Harris, K., Livnat, Y., and Meystre, S. (2015). Automated concept and relationship extraction for the semi-automated ontology management (seam) system. *Journal of biomedical semantics*, 6(1):15. *Cited in page 114*

[Dongen, 2000] Dongen, S. (2000). Performance criteria for graph clustering and markov cluster experiments. *Cited in page 102*

[Dorow and Widdows, 2003] Dorow, B. and Widdows, D. (2003). Discovering corpus-specific word senses. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 79–82, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 102*

[Drouin, 2003] Drouin, P. (2003). Acquisition des termes simples fondée sur les pivots lexicaux spécialisés. pages 183–186. *2 citations in pages 20 and 32*

[D'Souza and Ng, 2015] D'Souza, J. and Ng, V. (2015). Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 297–302, Beijing, China. Association for Computational Linguistics. *2 citations in pages 116 and 117*

[Duan et al., 2009] Duan, W., Song, M., and Yates, A. (2009). Fast max-margin clustering for unsupervised word sense disambiguation in biomedical texts. *BMC Bioinformatics*, 10(Suppl 3). *Cited in page 103*

[Duch et al., 2011] Duch, W., Maszczyk, T., and Grochowski, M. (2011). Optimal support features for meta-learning. In *Meta-Learning in Computational Intelligence*, pages 317–358. Springer. *Cited in page 110*

[Duda et al., 1973] Duda, R. O., Hart, P. E., et al. (1973). *Pattern classification and scene analysis*, volume 3. Wiley New York. *Cited in page 109*

[Dunn, 1974] Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104. *Cited in page 109*

[Dupuch, 2014] Dupuch, M. (2014). *Détection de termes sémantiquement proches. Clustering non supervisé basé sur les relations sémantiques et le degré d'apparenté sémantique. (PhD Thesis Dissertation).* Université Pierre et Marie Curie - Paris6. *Cited in page 163*

[Dupuch et al., 2011] Dupuch, M., Périnet, A., Hamon, T., and Grabar, N. (2011). Utilisation de méthodes de structuration de terminologies pour la création de groupements de termes de pharmacovigilance. In *9th International Conference on Terminology and Artificial Intelligence*, page 3. *Cited in page 29*

[El-Rab et al., 2013] El-Rab, W. G., Zaiane, O. R., and El-Hajj, M. (2013). Biomedical text disambiguation using umls. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 943–947. ACM. *2 citations in pages 89 and 107*

[Embley and Liddle, 2013] Embley, D. W. and Liddle, S. W. (2013). Big data—conceptual modeling to the rescue. In *Conceptual Modeling*, ER'13, pages 1–8. LNCS, Springer. *Cited in page 1*

[Enguehard and Pantera, 1995] Enguehard, C. and Pantera, L. (1995). Automatic natural acquisition of a terminology*. *Journal of quantitative linguistics*, 2(1):27–32. *2 citations in pages 21 and 32*

[Enguehard et al., 1993] Enguehard, C., Trigano, P., and Malvache, P. (1993). ANA: automatic natural acquisition. *IJPRAI*, 7(2):353–375. *2 citations in pages 21 and 32*

[Faraj et al., 1996] Faraj, N., Godin, R., Missaoui, R., David, S., and Plante, P. (1996). Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique de texte. *Canadian Journal of Information and Library Science/Revue l'information et de bibliothéconomie*, 21(1):1–21. *2 citations in pages 20 and 32*

[Faure and Nedellec, 1999] Faure, D. and Nedellec, C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium. In *Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management*, EKAW'99, pages 329–334, London, UK, UK. Springer-Verlag. *Cited in page 10*

[Faure et al., 1998] Faure, D., Nédellec, C., and Rouveirol, C. (1998). Acquisition of semantic knowledge using machine learning methods: The system" asium". In *Universite Paris Sud*. Citeseer. *Cited in page 10*

[Faure and Nédellec, 1998] Faure, D. and Nédellec, C. (1998). A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *In LREC workshop on*, pages 5–12. *3 citations in pages 10, 29, and 100*

[Fernández-López et al., 1997] Fernández-López, M., Gómez-Pérez, A., and Juristo, N. (1997). Methontology: from ontological art towards ontological engineering. *Cited in page 9*

[Ferret, 2010] Ferret, O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. In *LREC*, volume 10, pages 3338–3343. *Cited in page 116*

[Ferret, 2012] Ferret, O. (2012). Combining bootstrapping and feature selection for improving a distributional thesaurus. In *ECAI*, pages 336–341. *Cited in page 116*

[Ferret, 2013] Ferret, O. (2013). Identifying bad semantic neighbors for improving distributional thesauri. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 561–571, Sofia, Bulgaria. Association for Computational Linguistics. *2 citations in pages 115 and 116*

[Ferret, 2015] Ferret, O. (2015). Early and late combinations of criteria for reranking distributional thesauri. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 470–476, Beijing, China. Association for Computational Linguistics. *2 citations in pages 115 and 116*

[Fischl and Scharl, 2014] Fischl, D. and Scharl, A. (2014). Metadata enriched visualization of keywords in context. In *Proceedings of the 2014 ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, EICS'14, pages 193–196, New York, NY, USA. ACM. *Cited in page 85*

[Fortuna et al., 2006] Fortuna, B., Grobelnik, M., and Mladenic, D. (2006). Semi-automatic data-driven ontology construction system. In *Proceedings of the 9th International Multi-Conference Information Society*, IS'06, pages 223–226, Ljubljana, Slovenia. *Cited in page 10*

[Fortuna et al., 2007] Fortuna, B., Grobelnik, M., and Mladenic, D. (2007). Ontogen: Semi-automatic ontology editor. In *Proceedings of the 2007 Conference on Human Interface: Part II*, pages 309–318, Berlin, Heidelberg. Springer-Verlag. *Cited in page 10*

[Frantzi et al., 2000] Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130. *7 citations in pages 16, 23, 30, 32, 36, 40, and 44*

[Freeman, 1979] Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239. *Cited in page 27*

[Friedman and Rubin, 1967] Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320):1159–1178. *Cited in page 109*

[Gaizauskas et al., 2000] Gaizauskas, R., Demetriou, G., and Humphreys, K. (2000). Term recognition and classification in biological science journal articles. In *Proceeding of the Computional Terminology for Medical and Biological Applications Workshop of the 2 nd International Conference on NLP*, pages 37–44. *Cited in page 19*

[Gale et al., 1992] Gale, W. A., Church, K. W., and Yarowsky, D. (1992). Work on statistical methods for word sense disambiguation. In *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, volume 54. *Cited in page 98*

[Gherasim, 2013] Gherasim, T. (2013). *Detection of quality problems in ontologies constructed automatically from texts.* Theses, Université de Nantes. *3 citations in pages 3, 5, and 9*

[Ghiasvand and Kate, 2014] Ghiasvand, O. and Kate, R. (2014). Uwm: Disorder mention extraction from clinical text using crfs and normalization using learned edit distance patterns. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 828–832, Dublin, Ireland. Association for Computational Linguistics and Dublin City University. *Cited in page 117*

[Giraud-Carrier, 2008] Giraud-Carrier, C. (2008). Metalearning-a tutorial. In *Proceedings of the 7th international conference on machine learning and applications.* *Cited in page 110*

[Golik et al., 2013] Golik, W., Bossy, R., Ratkovic, Z., and Nédellec, C. (2013). Improving term extraction with linguistic analysis in the biomedical domain. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics, Special Issue of the journal Research in Computing Science*, CICLing'13, pages 24–30. *4 citations in pages 20, 25, 31, and 32*

[Golub and Van Loan, 1989] Golub, G. H. and Van Loan, C. F. (1989). Matrix computations. *Cited in page 96*

[Gómez-Pérez et al., 2007] Gómez-Pérez, A., Fernández-López, M., and Corcho, O. (2007). *Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. (Advanced Information and Knowledge Processing).* Springer-Verlag New York, Inc., Secaucus, NJ, USA. *Cited in page 9*

[Gordon, 1999] Gordon, A. D. (1999). Classification, (chapman & hall/crc monographs on statistics & applied probability). *Cited in page 108*

[Grefenstette, 1994] Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, Norwell, MA, USA. *Cited in page 115*

[Grozea, 2004] Grozea, C. (2004). Finding optimal parameter settings for high performance word sense disambiguation. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 125–128, Barcelona, Spain. *Cited in page 106*

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220. *Cited in page 1*

[Habert et al., 1996] Habert, B., Naulleau, E., and Nazarenko, A. (1996). Symbolic word clustering for medium-size corpora. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages

490–495, Stroudsburg, PA, USA. Association for Computational Linguistics. *2 citations in pages 29 and 48*

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18. *Cited in page 131*

[Halliday and Hasan, 1976] Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in english*. Longman, London. *Cited in page 115*

[Hamon et al., 2014] Hamon, T., Engström, C., and Silvestrov, S. (2014). Term ranking adaptation to the domain: Genetic algorithm-based optimisation of the c-value. In *Proceedings of the 9th International Conference on Natural Language Processing*, PolTAL'2014 - LNAI, pages 71–83. Springer, Warsaw, Poland. *Cited in page 23*

[Harispe et al., 2014] Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2014). The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, 30(5):740–742. *Cited in page 28*

[Harris, 1954] Harris, Z. S. (1954). Distributional structure. *Word*. *3 citations in pages 91, 96, and 113*

[Hartigan, 1975] Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc. *Cited in page 109*

[Hasegawa et al., 2004] Hasegawa, T., Sekine, S., and Grishman, R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL'04, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 113*

[Henriksson et al., 2014] Henriksson, A., Moen, H., Skeppstedt, M., Daudaravicius, V., and Duneld, M. (2014). Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5(6). *Cited in page 114*

[Heylen et al., 2008] Heylen, K., Peirsman, Y., Geeraerts, D., and Speelman, D. (2008). Modelling word similarity: an evaluation of automatic synonymy extraction algorithms. *Proceedings of the sixth international language resources and evaluation (LREC'08)*, pages 3243–3249. *Cited in page 115*

[Hliaoutakis et al., 2009] Hliaoutakis, A., Zervanou, K., and Petrakis, E. G. (2009). The amtex approach in the medical document indexing and retrieval application. *Data & Knowledge Engineering*, 68(3):380–392. *Cited in page 23*

[Hoste et al., 2002] Hoste, V., Hendrickx, I., Daelemans, W., and van den Bosch, A. (2002). Parameter optimization for machine-learning of word sense disambiguation. *Natural Language Engineering*, 8(04):311–325. *Cited in page 106*

[Huang et al., 2015] Huang, Y., Shi, X., Su, J., Chen, Y., and Huang, G. (2015). Unsupervised word sense induction using rival penalized competitive learning. *Engineering Applications of Artificial Intelligence*, 41(C):166–174. *2 citations in pages 105 and 158*

[Hubert and Levin, 1976] Hubert, L. J. and Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological bulletin*, 83(6):1072. *Cited in page 109*

[Hullman et al., 2015] Hullman, J., Diakopoulos, N., Momeni, E., and Adar, E. (2015). Content, context, and critique: Commenting on a data visualization blog. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, CSCW'15, pages 1170–1175, New York, NY, USA. ACM. *Cited in page 85*

[Ide and Erjavec, 2001] Ide, N. and Erjavec, T. (2001). Automatic sense tagging using parallel corpora. In *Natural Language Pacific Rim Symposium (artificial intelligence)*, NLPRS '01. *Cited in page 99*

[Jacquemin, 1996] Jacquemin, C. (1996). A symbolic and surgical acquisition of terms through variation. In *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438, London, UK, UK. Springer-Verlag. *2 citations in pages 20 and 32*

[Jacquemin, 1999] Jacquemin, C. (1999). Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 341–348, Stroudsburg, PA, USA. Association for Computational Linguistics. *2 citations in pages 20 and 32*

[Javed et al., 2008] Javed, O., Shafique, K., Rasheed, Z., and Shah, M. (2008). Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162. *Cited in page 107*

[Jayapandian et al., 2014] Jayapandian, C., Chen, C.-H., Dabir, A., Lhatoo, S., Zhang, G.-Q., and Sahoo, S. S. (2014). Domain ontology as conceptual model for big data management: application in biomedical informatics. In *Conceptual Modeling*, pages 144–157. Springer. *Cited in page 2*

[Ji et al., 2007] Ji, L., Sum, M., Lu, Q., Li, W., and Chen, Y. (2007). Chinese terminology extraction using window-based contextual information. In *Proceedings* *of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'07, pages 62–74, Berlin, Heidelberg. Springer-Verlag. *Cited in page 23*

[Jimeno Yepes and Aronson, 2012] Jimeno Yepes, A. and Aronson, A. R. (2012). Knowledge-based and knowledge-lean methods combined in unsupervised word sense disambiguation. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 733–736. ACM. *Cited in page 106*

[Jimeno-Yepes et al., 2011] Jimeno-Yepes, A. J., McInnes, B. T., and Aronson, A. R. (2011). Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12(1):223. *2 citations in pages 144 and 150*

[Jing et al., 2006] Jing, L., Ng, M. K., Yang, X., and Huang, J. Z. (2006). A text clustering system based on k-means type subspace clustering and ontology. *International Journal of Intelligent Technology*, 1(2):91–103. *Cited in page 139*

[John and Langley, 1995] John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc. *Cited in page 131*

[Jonquet et al., 2009] Jonquet, C., Shah, N. H., Youn, C. H., Callendar, C., Storey, M.-A., and Musen, M. A. (2009). Ncbo annotator: Semantic annotation of biomedical data. In *8th International Semantic Web Conference, Poster and Demonstration Session*, ISWC'09, Washington DC, USA. *Cited in page 77*

[Kageura and Umino, 1996] Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289. *Cited in page 16*

[Kambhatla, 2004] Kambhatla, N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, ACLdemo'04, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 114*

[Kinnunen et al., 2011] Kinnunen, T., Sidoroff, I., Tuononen, M., and Fränti, P. (2011). Comparison of clustering methods: A case study of text-independent speaker modeling. *Pattern Recognition Letters*, 32(13):1604–1617. *Cited in page 139*

[Klapaftis and Manandhar, 2008] Klapaftis, I. P. and Manandhar, S. (2008). Word sense induction using graphs of collocations. In *Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, ECAI '08, pages 298–302, Amsterdam, The Netherlands, The Netherlands. IOS Press. *Cited in page 102*

[Klapaftis and Manandhar, 2010] Klapaftis, I. P. and Manandhar, S. (2010). Word sense induction & disambiguation using hierarchical random graphs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 745–755, Stroudsburg, PA, USA. Association for Computational Linguistics. *2 citations in pages 103 and 107*

[Kolesnikov et al., 2015] Kolesnikov, A., Trichina, E., and Kauranne, T. (2015). Estimating the number of clusters in a numerical data set via quantization error modeling. *Pattern Recognition*, 48(3):941–952. *Cited in page 108*

[Kontonatsios et al., 2014a] Kontonatsios, G., Korkontzelos, I., Tsujii, J., and Ananiadou, S. (2014a). Combining string and context similarity for bilingual term alignment from comparable corpora. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP'14, pages 1701–1712, Doha, Qatar. Association for Computational Linguistics. *Cited in page 24*

[Kontonatsios et al., 2014b] Kontonatsios, G., Mihăilă, C., Korkontzelos, I., Thompson, P., and Ananiadou, S. (2014b). A hybrid approach to compiling bilingual dictionaries of medical terms from parallel corpora. In *Statistical Language and Speech Processing*, pages 57–69. Springer. *Cited in page 24*

[Köper and im Walde, 2014] Köper, M. and im Walde, S. S. (2014). A rank-based distance measure to detect polysemy and to determine salient vector-space features for german prepositions. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC'14, pages 4459–4466, Reykjavik, Iceland. European Language Resources Association (ELRA). *Cited in page 94*

[Köpf and Iglezakis, 2002] Köpf, C. and Iglezakis, I. (2002). Combination of task description strategies and case base properties for meta-learning. In *Proceedings of the 2nd international workshop on integration and collaboration aspects of data mining, decision support and meta-learning*, pages 65–76. *Cited in page 112*

[Korkontzelos and Manandhar, 2010] Korkontzelos, I. and Manandhar, S. (2010). Uoy: Graphs of unambiguous vertices for word sense induction and disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 355–358, Stroudsburg, PA, USA. Association for Computational Linguistics. *2 citations in pages 102 and 158*

[Kozakov et al., 2007] Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, N., and Confino, T. (2007). Glossary extraction and knowledge in large organisations via semantic web technologies. In *Proceedings of the 6th International Semantic Web Conference and he 2nd Asian Semantic Web Conference (Semantic Web Challenge Track)*, ISWC-ASWC'07. Springer. *2 citations in pages 22 and 32*

[Krauthammer and Nenadic, 2004] Krauthammer, M. and Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526. *2 citations in pages 20 and 32*

[Krzanowski and Lai, 1988] Krzanowski, W. J. and Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pages 23–34. *Cited in page 109*

[Kuo et al., 2007] Kuo, Y.-T., Lonie, A., Sonenberg, L., and Paizis, K. (2007). Domain ontology driven data mining: A medical case study. In *Proceedings of the 2007 International Workshop on Domain Driven Data Mining*, DDDM'07, pages 11–17, New York, NY, USA. ACM. *Cited in page 114*

[Lau et al., 2013] Lau, J. H., Cook, P., and Baldwin, T. (2013). unimelb: Topic modelling-based word sense induction for web snippet clustering. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 217–221. *Cited in page 103*

[Lau et al., 2012] Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 591–601, Stroudsburg, PA, USA. Association for Computational Linguistics. *4 citations in pages 90, 104, 107, and 158*

[Leaman et al., 2013] Leaman, R., Doğan, R. I., and Lu, Z. (2013). Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, page btt474. *Cited in page 117*

[Leaman et al., 2011] Leaman, R., Khare, R., and Lu, Z. (2011). Ncbi at 2013 share/clef ehealth shared task: disorder normalization in clinical notes with dnorm. *Radiology*, 42(21.1):1–941. *Cited in page 117*

[Lee and Ng, 2002] Lee, Y. K. and Ng, H. T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 41–48, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 101*

[Lemke et al., 2013] Lemke, C., Budka, M., and Gabrys, B. (2013). Metalearning: a survey of trends and technologies. *Artificial Intelligence Review*, 44(1):117–130. *Cited in page 110*

[Lemke et al., 2009] Lemke, C., Riedel, S., and Gabrys, B. (2009). Dynamic combination of forecasts generated by diversification procedures applied to forecasting of airline cancellations. In *Proceedings of the 6th European conference on principles and practice of knowledge discovery in database*, Helsinki, Finland. *Cited in page 112*

[Liang et al., 2012] Liang, J., Zhao, X., Li, D., Cao, F., and Dang, C. (2012). Determining the number of clusters using information entropy for mixed data. *Pattern Recognition*, 45(6):2251–2265. *Cited in page 108*

[Lin, 1998a] Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 2 of *ACL-COLING '98*, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 99*

[Lin, 1998b] Lin, D. (1998b). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 115*

[Lin and Pantel, 2001] Lin, D. and Pantel, P. (2001). Dirt - sbt - discovery of inference rules from text. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'01, pages 323–328, New York, NY, USA. ACM. *Cited in page 114*

[Liu et al., 2011] Liu, K., Chapman, W., Savova, G., Chute, C., Sioutos, N., Crowley, R. S., et al. (2011). Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents. *Methods of information in medicine*, 50(5):397. *Cited in page 114*

[Loginova Clouet, 2014] Loginova Clouet, E. (2014). *Processing of Compound Terms: Segmentation, Translation and Variation*. Theses, Université de Nantes. *Cited in page 22*

[Lossio-Ventura et al., 2012] Lossio-Ventura, J. A., Hacid, H., Ansiaux, A., and Maag, M. L. (2012). Conversations reconstruction in the social web. In *Proceedings of the 21st International Conference on World Wide Web*, WWW'12 Companion, pages 573–574, New York, NY, USA. ACM. *Cited in page 16*

[Lossio-Ventura et al., 2014a] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014a). Biomedical terminology extraction: A new combination of statistical and web mining approaches. In *Proceedings of the Journées internationales d'Analyse statistique des Données Textuelles*, JADT'14, Paris, France. *Cited in page 78*

[Lossio-Ventura et al., 2014b] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014b). BIOTEX: A system for biomedical terminology extraction, ranking, and validation. In *Proceedings of the 13th International Semantic Web Conference, Posters & Demonstrations Track*, ISWC'14, pages 157–160. *Cited in page 77*

[Lossio-Ventura et al., 2014c] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014c). Integration of linguistic and web information to improve biomedical terminology extraction. In ACM, editor, *Proceedings of the 18th International Database Engineering & Applications Symposium*, IDEAS'14, pages 265–269, Porto, Portugal. ACM. *Cited in page 46*

[Lossio-Ventura et al., 2014d] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014d). Yet another ranking function for automatic multiword term extraction. In *Proceedings of the 9th International Conference on Natural Language Processing*, number 8686 in PolTAL'2014 - LNAI, pages 52–64, Warsaw, Poland. Springer. *3 citations in pages 16, 44, and 78*

[Lossio-Ventura et al., 2015] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2015). Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, 19(1):59–99. *Cited in page 32*

[Lossio-Ventura et al., 2016a] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2016a). Automatic biomedical term polysemy detection. In *Proceedings of the 10th International Language Resources and Evaluation Conference*, LREC'2016, page *(to appear)*. *Cited in page 123*

[Lossio-Ventura et al., 2016b] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2016b). Communication overload management through social interactions clustering. In *Proceedings of the 31st ACM/SIGAPP Symposium on Applied Computing*, SAC'2016, page *(to appear)*, New York, NY, USA. ACM. *Cited in page 139*

[Lossio-Ventura et al., 2016c] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2016c). A way to automatically enrich biomedical ontologies. In *Proceedings of the 19th International Conference on Extending Database Technology*, EDBT'2016, page *(to appear)*, New York, NY, USA. ACM. *Cited in page 135*

[Lundqvist et al., 2011] Lundqvist, K. Ø., Baker, K., and Williams, S. (2011). Ontology supported competency system. *International Journal of Knowledge and Learning*, 7(3-4):197–219. *Cited in page 2*

[Lv and Zhai, 2011a] Lv, Y. and Zhai, C. (2011a). Adaptive term frequency normalization for BM25. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM'11, pages 1985–1988, New York, NY, USA. ACM. *Cited in page 40*

[Lv and Zhai, 2011b] Lv, Y. and Zhai, C. (2011b). When documents are very long, BM25 fails! In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'11, pages 1103–1104, New York, NY, USA. ACM. *Cited in page 40*

[Madden, 2012] Madden, S. (2012). From databases to big data. *IEEE Internet Computing*, 16(3):4–6. *Cited in page 1*

[Maedche and Staab, 2000] Maedche, A. and Staab, S. (2000). Discovering conceptual relations from text. In *Proceedings of the 14th European Conference on Artificial Intelligence*, volume 321 of *ECAI'00*, page 27. *Cited in page 10*

[Maedche and Volz, 2001] Maedche, A. and Volz, R. (2001). The ontology extraction & maintenance framework text-to-onto. In *Proceedings of the Workshop on Integrating Data Mining and Knowledge Management, in the 2001 IEEE International Conference on Data Mining*, Workshop-ICDM'01, pages 1–12. *Cited in page 10*

[Manandhar et al., 2010] Manandhar, S., Klapaftis, I. P., Dligach, D., and Pradhan, S. S. (2010). Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 63–68, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 158*

[Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press. *Cited in page 100*

[Marelli et al., 2014] Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *SemEval-2014*. *Cited in page 113*

[Marriott, 1971] Marriott, F. (1971). Practical problems in a method of cluster analysis. *Biometrics*, pages 501–514. *Cited in page 109*

[Matsuo and Ishizuka, 2004] Matsuo, Y. and Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169. *2 citations in pages 21 and 32*

[Maudsley, 1979] Maudsley, D. (1979). *A Theory of Meta-learning and Principles of Facilitation : an Organismic Perspective*. Thesis (Ed.D.)–University of Toronto. *Cited in page 110*

[Maynard et al., 2009] Maynard, D., Funk, A., and Peters, W. (2009). SPRAT: a tool for automatic semantic pattern-based ontology population. In *International Conference for Digital Libraries and the Semantic Web*, Trento, Italy. Citeseer. *2 citations in pages 2 and 10*

[McInnes, 2008] McInnes, B. T. (2008). An unsupervised vector approach to biomedical term disambiguation: integrating umls and medline. In *Proceedings* *of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop*, pages 49–54. Association for Computational Linguistics. *Cited in page 106*

[McInnes and Pedersen, 2013] McInnes, B. T. and Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of biomedical informatics*, 46(6):1116–1124. *Cited in page 106*

[McInnes et al., 2011] McInnes, B. T., Pedersen, T., Liu, Y., Melton, G. B., and Pakhomov, S. V. (2011). Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. In *AMIA Annual Symposium Proceedings*, volume 2011, page 895. American Medical Informatics Association. *Cited in page 106*

[McInnes and Stevenson, 2014] McInnes, B. T. and Stevenson, M. (2014). Determining the difficulty of word sense disambiguation. *Journal of biomedical informatics*, 47:83–90. *Cited in page 106*

[Medelyan et al., 2009] Medelyan, O., Frank, E., and Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP'09, pages 1318–1327, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 25*

[Medelyan and Witten, 2006] Medelyan, O. and Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '06, pages 296–297, New York, NY, USA. ACM. *Cited in page 25*

[Mihalcea and Faruque, 2004] Mihalcea, R. and Faruque, E. (2004). Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of the 3rd Interna- tional Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 155–158, Barcelona, Spain. *Cited in page 106*

[Milligan and Cooper, 1985] Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179. *2 citations in pages 107 and 109*

[Min et al., 2012] Min, B., Shi, S., Grishman, R., and Lin, C.-Y. (2012). Ensemble semantics for large-scale unsupervised relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1027–1037, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 115*

[Mirkin, 2011] Mirkin, B. (2011). Choosing the number of clusters. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):252–260. *2 citations in pages 90 and 107*

[Mondary, 2011] Mondary, T. (2011). *Construction d'ontologies à partir de textes. L'apport de l'analyse de concepts formels.* Theses, Université Paris-Nord - Paris XIII. Equipe RCLN. *Cited in page 2*

[Moon et al., 2015] Moon, S., McInnes, B., and Melton, G. B. (2015). Challenges and practical approaches with word sense disambiguation of acronyms and abbreviations in the clinical domain. *Healthcare informatics research*, 21(1):35–42. *Cited in page 107*

[Morin and Prochasson, 2011] Morin, E. and Prochasson, E. (2011). Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 27–34. Association for Computational Linguistics. *Cited in page 24*

[Morris and Hirst, 2004] Morris, J. and Hirst, G. (2004). Non-classical lexical semantic relations. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, CLS '04, pages 46–51, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 115*

[Murdoch and Detsky, 2013] Murdoch, T. B. and Detsky, A. S. (2013). The inevitable application of big data to health care. *Journal of the American Medical Association, JAMA*, 309(13):1351–1352. *Cited in page 15*

[Nakagawa and Mori, 2002] Nakagawa, H. and Mori, T. (2002). A simple but powerful automatic term extraction method. In *COLING-02 on COMPUTERM 2002: Second International Workshop on Computational Terminology - Volume 14*, COMPUTERM '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 23*

[Nasiruddin, 2013] Nasiruddin, M. (2013). A state of the art of word sense induction: a way towards word sense disambiguation for under-resourced languages. *TALN-RECITAL 2013*, pages 192–205. *Cited in page 96*

[Navigli, 2009] Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10. *Cited in page 105*

[Navigli, 2012] Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. In *SOFSEM 2012: Theory and practice of computer science*, pages 115–129. Springer. *2 citations in pages 91 and 96*

[Navigli and Crisafulli, 2010] Navigli, R. and Crisafulli, G. (2010). Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 116–126, Stroudsburg, PA, USA. Association for Computational Linguistics. *2 citations in pages 101 and 103*

[Navigli and Lapata, 2010] Navigli, R. and Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(4):678–692. *Cited in page 106*

[Navigli et al., 2003] Navigli, R., Velardi, P., and Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, 18(1):22–31. *Cited in page 10*

[Névéol et al., 2014] Névéol, A., Grosjean, J., Darmoni, S. J., and Zweigenbaum, P. (2014). Language resources for french in the biomedical domain. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC'14. Association for Computational Linguistics. *2 citations in pages 4 and 16*

[Newman et al., 2012] Newman, D., Koilada, N., Lau, J. H., and Baldwin, T. (2012). Bayesian text segmentation for index term identification and keyphrase extraction. In *Proceedings of 24th International Conference on Computational Linguistics*, COLING'12, pages 2077–2092, Mumbai, India. *3 citations in pages 16, 22, and 32*

[Nguyen et al., 2015] Nguyen, N. T., Miwa, M., Tsuruoka, Y., Chikayama, T., and Tojo, S. (2015). Wide-coverage relation extraction from medline using deep syntax. *BMC bioinformatics*, 16(1):107. *Cited in page 178*

[Niu et al., 2007] Niu, Z.-Y., Ji, D.-H., and Tan, C.-L. (2007). I2r: Three systems for word sense discrimination, chinese word sense disambiguation, and english word sense disambiguation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 177–182, Stroudsburg, PA, USA. Association for Computational Linguistics. *3 citations in pages 101, 107, and 158*

[Noh et al., 2010] Noh, T.-G., Park, S.-B., and Lee, S.-J. (2010). Unsupervised word sense disambiguation in biomedical texts with co-occurrence network and graph kernel. In *Proceedings of the ACM Fourth International Workshop on Data and Text Mining in Biomedical Informatics*, DTMBIO '10, pages 61–64, New York, NY, USA. ACM. *Cited in page 103*

[Noh et al., 2009] Noh, T.-G., Park, S.-B., Yoon, H.-G., Lee, S.-J., and Park, S.-Y. (2009). An automatic translation of tags for multimedia contents using folksonomy networks. In *Proceedings of the 32Nd International ACM SIGIR Conference on*

Research and Development in Information Retrieval, SIGIR'09, pages 492–499, New York, NY, USA. ACM. *Cited in page 27*

[Noy et al., 2009] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N. B., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., and Musen., M. A. (2009). Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37:170–173. *2 citations in pages 16 and 85*

[Nzali et al., 2015] Nzali, M. D.-T., Bringay, S., Lavergne, C., Opitz, T., Azé, J., and Mollevi, C. (2015). Construction d'un vocabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux. In *Actes de 25es journées francophones d'Ingénierie des Connaissances*, IC'2015. *Cited in page 82*

[Opsahl et al., 2010] Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251. *Cited in page 27*

[Padró et al., 2014] Padró, M., Idiart, M., Villavicencio, A., and Ramisch, C. (2014). Nothing like good old frequency: Studying context filters for distributional thesauri. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 419–424, Doha, Qatar. Association for Computational Linguistics. *Cited in page 115*

[Page et al., 1999] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. *Cited in page 27*

[Pantel et al., 2009] Pantel, P., Crestan, E., Borkovsky, A., Popescu, A.-M., and Vyas, V. (2009). Web-scale distributional similarity and entity set expansion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'09, pages 938–947, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 28*

[Pantel and Lin, 2002] Pantel, P. and Lin, D. (2002). Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM. *Cited in page 99*

[Pedersen, 2007] Pedersen, T. (2007). Umnd2: Senseclusters applied to the sense induction task of senseval-4. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 394–397, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 98*

[Pedersen, 2010] Pedersen, T. (2010). Duluth-wsi: Senseclusters applied to the sense induction task of semeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 363–366, Stroudsburg, PA, USA. Association for Computational Linguistics. *2 citations in pages 98 and 158*

[Pedersen and Bruce, 1997] Pedersen, T. and Bruce, R. (1997). Distinguishing word senses in untagged text. In *Second Conference on Empirical Methods in Natural Language Processing*, EMNLP '97, pages 197–207. *Cited in page 97*

[Peng et al., 2002] Peng, Y., Flach, P. A., Soares, C., and Brazdil, P. (2002). Improved dataset characterisation for meta-learning. In *Discovery Science*, pages 141–152. Springer. *3 citations in pages 110, 111, and 123*

[Pérez et al., 2008] Pérez, A. G., de Figueroa Baonza, M. C. S., and Villazón, B. (2008). Neon methodology for building ontology networks: Ontology specification. *Methodology*, pages 1–18. *Cited in page 9*

[Pfahringer et al., 2000] Pfahringer, B., Bensusan, H., and Giraud-Carrier, C. (2000). Tell me who can learn you and i can tell you who you are: Landmarking various learning algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML2000, pages 743–750. *Cited in page 113*

[Pinto et al., 2007] Pinto, D., Rosso, P., and Jimenez-Salazar, H. (2007). Upv-si: Word sense induction using self term expansion. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 430–433. Association for Computational Linguistics. *Cited in page 100*

[Pinto et al., 2004] Pinto, H. S., Staab, S., and Tempich, C. (2004). Diligent: Towards a fine-grained methodology for distributed, loosely-controlled and evolving. In *Proceedings of the 16th European Conference on Artificial Intelligence*, volume 110 of *ECAI'2004*, page 393. *Cited in page 9*

[Platt, 1999] Platt, J. C. (1999). Advances in kernel methods. chapter Fast training of support vector machines using sequential minimal optimization, pages 185–208. MIT Press, Cambridge, MA, USA. *Cited in page 131*

[Polajnar and Clark, 2014] Polajnar, T. and Clark, S. (2014). Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 230–238, Gothenburg, Sweden. Association for Computational Linguistics. *Cited in page 115*

[Ponzetto and Navigli, 2010] Ponzetto, S. P. and Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1522–1531. Association for Computational Linguistics. *Cited in page 106*

[Purandare and Pedersen, 2004a] Purandare, A. and Pedersen, T. (2004a). Senseclusters: finding clusters that represent word senses. In *Demonstration Papers at HLT-NAACL 2004*, pages 26–29. Association for Computational Linguistics. *Cited in page 98*

[Purandare and Pedersen, 2004b] Purandare, A. and Pedersen, T. (2004b). Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning*, volume 72 of *CoNLL*. *Cited in page 97*

[Qian et al., 2008] Qian, L., Zhou, G., Kong, F., Zhu, Q., and Qian, P. (2008). Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, COLING'08, pages 697–704. Association for Computational Linguistics. *Cited in page 114*

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. *Cited in page 131*

[Qureshi et al., 2012] Qureshi, M. A., O'Riordan, C., and Pasi, G. (2012). Short-text domain specific key terms/phrases extraction using an n-gram model with wikipedia. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM'12, pages 2515–2518, New York, NY, USA. ACM. *Cited in page 16*

[Rada et al., 1989] Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(1):17–30. *Cited in page 162*

[Rao et al., 2013] Rao, D., McNamee, P., and Dredze, M. (2013). Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer. *Cited in page 116*

[Ratkowsky and Lance, 1978] Ratkowsky, D. and Lance, G. (1978). A criterion for determining the number of groups in a classification. *Australian Computer Journal*, 10(3):115–117. *Cited in page 109*

[Ravichandran and Hovy, 2002] Ravichandran, D. and Hovy, E. (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL'02, pages 41–47, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 113*

[Rebholz-Schuhmann et al., 2008] Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., and Jimeno, A. (2008). Text processing through web services: calling whatizit. *Bioinformatics*, 24(2):296–298. *Cited in page 26*

[Reif et al., 2012a] Reif, M., Shafait, F., and Dengel, A. (2012a). Dataset generation for meta-learning. *35th German Conference on Artificial Intelligence.* *Cited in page 112*

[Reif et al., 2012b] Reif, M., Shafait, F., and Dengel, A. (2012b). Meta2-features: Providing meta-learners more information. *35th German Conference on Artificial Intelligence*. Cited in page 112

[Reif et al., 2014] Reif, M., Shafait, F., Goldstein, M., Breuel, T., and Dengel, A. (2014). Automatic classifier selection for non-experts. *Pattern Analysis and Applications*, 17(1):83–96. Cited in page 113

[Richard et al., 2015] Richard, M., Aimé, X., Krebs, M.-O., and Charlet, J. (2015). Enrich classifications in psychiatry with textual data: an ontology for psychiatry including social concepts. *Studies in health technology and informatics*. Cited in page 81

[Robertson et al., 1999] Robertson, S. E., Walker, S., and Beaulieu, M. (1999). Okapi at TREC-7: automatic ad hoc, filtering, vlc and interactive track. *IN*, 21:253–264. 3 citations in pages 21, 32, and 42

[Roche and Fortuno, 2014] Roche, M. and Fortuno, S. (2014). La fouille de textes au service de la documentation. *Arabesques*, 76:13–14. Cited in page 83

[Roche et al., 2015] Roche, M., Fortuno, S., Lossio-Ventura, J. A., Akli, A., Belkebir, S., Lounis, T., and Toure, S. (2015). Extraction automatique des mots-clés à partir de publications scientifiques pour l'indexation et l'ouverture des données en agronomie. In *Cahiers Agricultures*. Cited in page 83

[Roche et al., 2004] Roche, M., Heitz, T., Matte-Tailliez, O., and Kodratoff, Y. (2004). EXIT: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. In *7th International Conference on Statistical Analysis of Textual Data*, volume 2 of *JADT'04*, pages 946–956, Louvain-la-Neuve, France. Cited in page 25

[Roche and Prince, 2010] Roche, M. and Prince, V. (2010). A web-mining approach to disambiguate biomedical acronym expansions. *Informatica*, 34(2). 2 citations in pages 28 and 32

[Roche and Teissire, 2015] Roche, M. and Teissire, M. (2015). Traitement automatique des données hétérogènes liées à l'aménagement des territoires. In *Proceedings of Association de Science Régionale de Langue Française*. Cited in page 83

[Rose et al., 2010] Rose, S., Engel, D., Cramer, N., and Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Theory and Applications*, pages 1–20. 2 citations in pages 21 and 32

[Rousseau and Vazirgiannis, 2013] Rousseau, F. and Vazirgiannis, M. (2013). Graph-of-word and tw-idf: New approach to ad hoc ir. In *Proceedings of the 22Nd ACM International Conference on Information and Knowl-*

edge Management, CIKM'13, pages 59–68, New York, NY, USA. ACM. *4 citations in pages 27, 30, 31, and 32*

[Rousseau and Vazirgiannis, 2015] Rousseau, F. and Vazirgiannis, M. (2015). Main core retention on graph-of-words for single-document keyword extraction. In *Advances in Information Retrieval*, pages 382–393. Springer. *2 citations in pages 27 and 48*

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65. *Cited in page 109*

[Rubin et al., 2008] Rubin, D. L., Shah, N. H., and Noy, N. F. (2008). Biomedical ontologies: a functional perspective. *Briefings in bioinformatics*, 9(1):75–90. *Cited in page 15*

[Rychlý and Kilgarriff, 2007] Rychlý, P. and Kilgarriff, A. (2007). An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 41–44, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 115*

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523. *4 citations in pages 21, 32, 42, and 99*

[Sánchez and Moreno, 2008] Sánchez, D. and Moreno, A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. *Data & Knowledge Engineering*, 64(3):600–623. *Cited in page 2*

[Sarle, 1983] Sarle, W. S. (1983). *Cubic clustering criterion*. SAS Institute. *Cited in page 109*

[Savova and Pedersen, 2005] Savova, G. and Pedersen, T. (2005). Resolving ambiguities in biomedical text with unsupervised clustering approaches. *University of Minnesota Supercomputing Institute Research Report*. *2 citations in pages 97 and 98*

[Savova et al., 2008] Savova, G. K., Coden, A. R., Sominsky, I. L., Johnson, R., Ogren, P. V., de Groen, P. C., and Chute, C. G. (2008). Word sense disambiguation across two domains: biomedical literature and clinical notes. *Journal of biomedical informatics*, 41(6):1088–1100. *Cited in page 107*

[Schuemie et al., 2005] Schuemie, M. J., Kors, J. A., and Mons, B. (2005). Word sense disambiguation in the biomedical domain: an overview. *Journal of Computational Biology*, 12(5):554–565. *Cited in page 106*

[Schutze, 1992] Schutze, H. (1992). Dimensions of meaning. In *Proceedings of Supercomputing'92*, pages 787–796. IEEE. *3 citations in pages 96, 97, and 98*

[Schütze, 1998] Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123. *Cited in page 97*

[Sclano and Velardi, 2007] Sclano, F. and Velardi, P. (2007). Termextractor: a web application to learn the shared terminology of emergent web communities. In *Enterprise Interoperability II*, pages 287–290. Springer. *Cited in page 25*

[Scott and Symons, 1971] Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, pages 387–397. *Cited in page 109*

[Sebag, 2002] Sebag, M. (2002). Genetic programming applied to model identification. In *Principles of nonparametric learning*, pages 271–335. Springer. *2 citations in pages 21 and 32*

[Séguéla and Aussenac-Gilles, 1999] Séguéla, P. and Aussenac-Gilles, N. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Conférence ingénierie des connaissances*, pages 79–88. *Cited in page 29*

[Sharoff, 2011] Sharoff, S. (2011). In the garden and in the jungle. In *Genres on the Web*, pages 149–166. Springer. *Cited in page 28*

[Shen et al., 2015] Shen, W., Wang, J., and Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):443–460. *Cited in page 116*

[Singhal et al., 1996] Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'96, pages 21–29, New York, NY, USA. ACM. *Cited in page 40*

[Smadja, 1993] Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computation Linguistic*, 19(1):143–177. *2 citations in pages 22 and 32*

[Sparck Jones, 1986] Sparck Jones, K. (1986). *Synonymy and Semantic Classification.* Edinburgh University Press, Edinburgh, Scotland, Scotland. *Cited in page 115*

[Spasic et al., 2013] Spasic, I., Greenwood, M., Preece, A., Francis, N., and Elwyn, G. (2013). FlexiTerm: a flexible term recognition method. *Biomedical Semantics*, 4(27). *Cited in page 25*

[Staab et al., 2001] Staab, S., Studer, R., Schnurr, H.-P., and Sure, Y. (2001). Knowledge processes and ontologies. *IEEE Intelligent Systems*, 16(1):26–34. *Cited in page 9*

[Stevenson et al., 2008] Stevenson, M., Guo, Y., Gaizauskas, R., and Martinez, D. (2008). Disambiguation of biomedical text using diverse sources of information. *BMC bioinformatics*, 9(Suppl 11):S7. *2 citations in pages 89 and 106*

[Stoykova and Petkova, 2012] Stoykova, V. and Petkova, E. (2012). Automatic extraction of mathematical terms for precalculus. *Procedia Technology Journal*, 1:464–468. *Cited in page 16*

[Suchanek et al., 2006] Suchanek, F. M., Ifrim, G., and Weikum, G. (2006). Combining linguistic and statistical analysis to extract relations from web documents. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'06, pages 712–717, New York, NY, USA. ACM. *Cited in page 114*

[Sy et al., 2012] Sy, M.-F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., and Ranwez, V. (2012). User centered and ontology based information retrieval system for life sciences. *BMC bioinformatics*, 13(Suppl 1):S4. *Cited in page 2*

[Tamura et al., 2012] Tamura, A., Watanabe, T., and Sumita, E. (2012). Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'12, pages 24–36, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 24*

[Tang et al., 2014] Tang, G., Xia, Y., Sun, J., Zhang, M., and Zheng, T. F. (2014). Statistical word sense aware topic models. *Soft Computing*, 19:1–15. *Cited in page 104*

[Teh et al., 2006] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581. *2 citations in pages 104 and 107*

[Tian and Lo, 2015] Tian, Y. and Lo, D. (2015). A comparative study on the effectiveness of part-of-speech tagging techniques on bug reports. In *Proceedings of the 22nd International IEEE Conference on Software Analysis, Evolution, and Reengineering*, SANER'15, pages 570–574, Montreal, Canada. IEEE. *Cited in page 35*

[Tibshirani et al., 2001] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423. *Cited in page 109*

[Turenne and Barbier, 2004] Turenne, N. and Barbier, M. (2004). Beluga: un outil pour l'analyse dynamique des connaissances de la littérature scientifique d'un domaine-première application au cas des maladies à prions. In *EGC*, pages 423–428. *2 citations in pages 20 and 32*

[Turney, 2001] Turney, P. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl. *2 citations in pages 28 and 32*

[Udani et al., 2005] Udani, G., Dave, S., Davis, A., and Sibley, T. (2005). Noun sense induction using web search results. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 657–658, New York, NY, USA. ACM. *Cited in page 97*

[Van de Cruys and Apidianaki, 2011] Van de Cruys, T. and Apidianaki, M. (2011). Latent semantic word sense induction and disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1476–1485. Association for Computational Linguistics. *3 citations in pages 98, 100, and 158*

[Van Dongen, 2001] Van Dongen, S. M. (2001). Graph clustering by flow simulation. *Cited in page 101*

[Van Eck et al., 2010] Van Eck, N. J., Waltman, L., Noyons, E. C., and Buter, R. K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics*, 82(3):581–596. *Cited in page 20*

[Vanschoren, 2010] Vanschoren, J. (2010). Understanding machine learning performance with experiment databases. *lirias. kuleuven. be, no. May.* *Cited in page 113*

[Velardi et al., 2007] Velardi, P., Cucchiarelli, A., and Petit, M. (2007). A taxonomy learning method and its application to characterize a scientific web community. *Knowledge and Data Engineering, IEEE Transactions on*, 19(2):180–191. *2 citations in pages 2 and 10*

[Véronis, 2004] Véronis, J. (2004). Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252. *Cited in page 101*

[Vilalta and Drissi, 2009] Vilalta, R. and Drissi, Y. (2009). A characterization of difficult problems in classification. In *Proceedings of the 6th European conference on principles and practice of knowledge discovery in databases*, pages 85–91. IEEE. *Cited in page 112*

[Wang et al., 2015a] Wang, C., Cao, L., and Zhou, B. (2015a). Medical synonym extraction with concept space models. *arXiv preprint.* *Cited in page 114*

[Wang et al., 2015b] Wang, J., Bansal, M., Gimpel, K., Ziebart, B., and Yu, C. (2015b). A sense-topic model for word sense induction with unsupervised data enrichment. *Transactions of the Association for Computational Linguistics*, 3:59–71. *2 citations in pages 96 and 104*

[Weeds, 2003] Weeds, J. E. (2003). *Measures and applications of lexical distributional similarity*. PhD thesis, University of Sussex. *Cited in page 115*

[Widdows and Dorow, 2002] Widdows, D. and Dorow, B. (2002). A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th International Conference on Computational Linguistics*, volume 1 of *COLING '02*, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 101*

[Yan, 2005] Yan, M. (2005). Methods of determining the number of clusters in a data set and a new clustering criterion. *Cited in page 108*

[Yang et al., 2009] Yang, Y., Zhao, T., Lu, Q., Zheng, D., and Yu, H. (2009). Chinese term extraction using different types of relevance. In *Proceedings of the International Joint Conference on Natural Language Processing*, ACL-IJCNLP'09, pages 213–216, Suntec, Singapore. Association for Computational Linguistics. *Cited in page 16*

[Yao et al., 2011] Yao, L., Haghighi, A., Riedel, S., and McCallum, A. (2011). Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'11, pages 1456–1466, Stroudsburg, PA, USA. Association for Computational Linguistics. *Cited in page 114*

[Yao and Van Durme, 2011] Yao, X. and Van Durme, B. (2011). Nonparametric bayesian word sense induction. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, TextGraphs-6, pages 10–14, Stroudsburg, PA, USA. Association for Computational Linguistics. *2 citations in pages 104 and 158*

[Yu et al., 2014] Yu, H., Liu, Z., and Wang, G. (2014). An automatic method to determine the number of clusters using decision-theoretic rough set. *International Journal of Approximate Reasoning*, 55(1):101–115. *Cited in page 108*

[Zadeh and Goel, 2013] Zadeh, R. B. and Goel, A. (2013). Dimension independent similarity computation. *Journal of Machine Learning Research*, 14(1):1605–1626. *Cited in page 28*

[Zesch and Gurevych, 2010] Zesch, T. and Gurevych, I. (2010). Wisdom of crowds versus wisdom of linguists – measuring the semantic relatedness of words. *Nat. Lang. Eng.*, 16(1):25–59. *Cited in page 115*

[Zhang et al., 2010] Zhang, X., Song, Y., and Fang, A. (2010). Term recognition using conditional random fields. In *International Conference on Natural Language Processing and Knowledge Engineering*, NLP-KE'10, pages 1–6. IEEE. *2 citations in pages 22 and 32*

[Zhang et al., 2008] Zhang, Z., Iria, J., Brewster, C., and Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC'08, Marrakech, Morocco. *3 citations in pages 16, 23, and 25*

[Zhong and Ng, 2010] Zhong, Z. and Ng, H. T. (2010). It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83. Association for Computational Linguistics. *Cited in page 106*