A novel framework for biomedical entity sense induction

Juan Antonio Lossio-Ventura, Clement Jonquet, Jiang Bian, Mathieu Roche, Maguelonne Teisseire

Noname manuscript No.
(will be inserted by the editor)

A novel framework for biomedical entity sense induction

Juan Antonio Lossio-Ventura · Jiang Bian · Clement Jonquet · Mathieu Roche · Maguelonne Teisseire.

Received: date / Accepted: date

Abstract

Background: Rapid advancements in biomedical research have accelerated the number of relevant electronic documents published online, ranging from scholarly articles to news, blogs, and user-generated social media content. Nevertheless, the vast amount of this information is poorly organized, making it difficult to navigate. Emerging technologies such as ontologies and knowledge bases (KBs) could help organize and track the information associated with biomedical research developments. A major challenge in the automatic construction of ontologies and KBs is the identification of words with its respective sense(s) from a free-text corpus. Word-sense induction (WSI) is a task to automatically induce the different senses of a target word in the

J.A. Lossio-Ventura
College of Medicine, University of Florida, U.S.A
E-mail: jlossioventura@ufl.edu

J. Bian
College of Medicine, University of Florida, U.S.A
E-mail: bianjiang@ufl.edu

C. Jonquet
University of Montpellier, LIRMM, CNRS - Montpellier, France
E-mail: jonquet@lirmm.fr

M. Roche
Cirad, TETIS - Montpellier, France
TETIS, Univ. Montpellier, APT, Cirad, Cnrs, Irstea, Montpellier, France
E-mail: mathieu.roche@cirad.fr

M. Teisseire
Irstea, TETIS - Montpellier, France
E-mail: maguelonne.teisseire@irstea.fr

This paper is a significant extension of our previous studies published in the proceedings of: 1) The 10th International Conference on Language Resources and Evaluation (LREC'2016) titled, "Automatic biomedical term polysemy detection"; and 2) The 19th International Conference on Extending Database Technology (EDBT'2016) titled, "A way to automatically enrich biomedical ontologies", with new scientific methodologies and results.

different contexts. In the last two decades, there have been several efforts on WSI. However, few methods are effective in biomedicine and life sciences.

Methods: We developed a framework for biomedical entity sense induction using a mixture of natural language processing, supervised, and unsupervised learning methods with promising results. It is composed of three main steps: 1) a polysemy detection method to determine if a biomedical entity has many possible meanings; 2) a clustering quality index-based approach to predict the number of senses for the biomedical entity; and 3) a method to induce the concept(s) (i.e., senses) of the biomedical entity in a given context.

Results: To evaluate our framework, we used the well-known MSH WSD polysemic dataset that contains 203 annotated ambiguous biomedical entities, where each entity is linked to 2 to 5 concepts. Our polysemy detection method obtained an *F-measure* of 98%. Second, our approach for predicting the number of senses achieved an *F-measure* of 93%. Finally, we induced the concepts of the biomedical entities based on a clustering algorithm and then extracted the keywords of reach cluster to represent the concept.

Conclusions: We have developed a framework for biomedical entity sense induction with promising results. Our study results can benefit a number of downstream applications, for example, help to resolve concept ambiguities when building Semantic Web KBs from biomedical text.

Keywords Word sense induction · Polysemy detection · Biomedicine · BioNLP · Clustering · Classification · Number of cluster prediction

1 Introduction

The World Wide Web is by far the most extensive information system available worldwide on the Internet, whose content has been growing exponentially every day with inputs from a large number of Internet users. Much of the information on the web is textual and contains rich information related to a wide range of domains. In particular, recent advances in biomedical research have accelerated the rate of health information being published on the web, ranging from scholar articles and news to blogs and user-generated social media content. The increasing capability and sophistication of biomedical tools and instruments have also contributed to the build-up of large volumes of biomedical data (e.g., the use of electronic health records for storing patient information).

Ontologies and knowledge bases (KBs) can help organize and track the information associated with biomedical research developments. Last few years have witnessed significant research efforts in automating ontology enrichment and KB construction leveraging the vast amount of electronic free-text data on the web. Nevertheless, one of the major challenges associated with automating KB and ontology constructions is the identification of words or phrases (entities) with their respectively sense(s), which has received attention only very recently [44,23,29].

Word-sense induction (WSI) is a task to automatically induce the different senses of a target word in a piece of text. The output of WSI is a sense inventory (a set of senses for the target word). Most existing WSI approaches are based on unsupervised learning algorithms with senses represented as clusters of tokens (e.g., words or phrases). There have been very few studies that use WSI in the context of information retrieval [74, 58].

In general, existing WSI approaches only consider sense induction for individual words, such as verbs, nouns, and adjectives [2, 53]. However, biomedical entities (or biomedical terms) are often composed of more than one word. Indeed, more than 80% of biomedical entities are composed of two or more words in the Unified Medical Language System (UMLS) metathesaurus[1].

Another issue with existing WSI methods is that they do not first check whether a target word is polysemic (i.e., ambiguous) or not. Thus, a significant amount of computing time is wasted on identifying the different senses for non-polysemic words. Reducing the runtime for WSI algorithms is crucial for real-world applications. A subsequent challenge is to determine the number of senses (i.e., the number of clusters) of an entity. Further, for a new entity (i.e., entities that do not exist in existing reference KBs or ontologies), there is no a priori knowledge about the candidate entities, which makes it more challenging to determine the exact number of clusters. Thus, the clustering algorithms for WSI often suffer from poor performance [25].

To address these challenges associated with applying WSI in biomedicine, we propose a novel framework for biomedical entity sense induction. Our framework is composed of three main steps: 1) a polysemy detection method to determine if a biomedical entity is ambiguous; 2) a clustering quality index-based approach to predict the number of senses for a biomedical entity; and 3) a method to induce the concept(s) (i.e., senses) of a biomedical entity.

The primary contributions of our work in comparison to our previous studies in [50, 52], are detailed below:

- We conducted a series of new evaluation experiments of our polysemy detection method (i.e., a supervised learning method based on 23 novel features we proposed in [50]) following the best practice in machine learning, which achieved an *F-measure* of 98%. We compared our polysemy detection approach with other similar methods (i.e., word classification tasks), which we adapted for the polysemy detection task.
- We presented in detail how we used a set of new clustering quality indexes and associated objective functions to predict the number of senses. Our method obtained an *F-measure* of 93%. We compared our method with 8 state-of-the-art clustering quality indexes that have been widely used for finding the number of clusters of a dataset. We also explored the 23 features [50] we used in polysemy detection for predicting the number of senses, which achieved an *F-measure* of 91%. Through these compre-

[1] UMLS is a large biomedical thesaurus that is organized by concept or meaning, and links similar names for the same concept from nearly 200 different ontologies and terminologies `http://www.nlm.nih.gov/research/umls`

hensive evaluation experiments, our clustering quality index-based method outperformed all other benchmark methods significantly.
- We presented our methods based on clustering and keyword extractions for concept induction to complete the proposed biomedical entity sense induction framework. We evaluated the concept induction methods against the gold-standard definitions of the senses in UMLS. Our concept induction method performed reasonably well in our evaluations.

We used a gold-standard dataset based on 203 ambiguous biomedical entities extracted from MEDLINE article abstracts, where each entity is annotated with one or more MeSH concepts[2] [37] and linked to the abstracts from which it was extracted. We have also experimented with a wide range of well-known supervised and unsupervised learning algorithms for the three steps in our framework.

The rest of the paper is organized as follows. Section 2 discusses existing work related to WSI and its applications in the biomedical domain. We detail our biomedical entity sense induction framework in Section 3. We present and discuss our evaluation results in Section 4. We conclude the work and discuss future directions in Section 5.

2 Related work

2.1 Word-sense induction

WSI is one of the natural language processing (NLP) tasks that aims to automatically identify the different senses (i.e., meanings) of a word in a piece of text. WSI is closely related to word-sense disambiguation (WSD), which is the task of determining the correct sense of a word in a context [57]. Unsupervised WSD methods are considered as WSI techniques aimed at discovering senses automatically based on unlabeled corpora [57]. The output of a WSI algorithm is a set of senses (i.e., a sense inventory) for each target entity from the text corpora without any other knowledge resources (e.g., existing terminology services). WSI methods are mostly based on clustering algorithms, where each cluster represents a distinct sense of the word. Thus, a major problem of WSI is to determine the number of clusters (i.e., senses) within a given context, which is usually taken as a prior in most clustering algorithms.

Existing WSI approaches can be categorized in four types [57,83]: 1) *context clustering*, whose main idea is that the distributional profile of words in a corpus implicitly expresses their semantics [69,79,14,66,73,74,64,65,80,34,6,7], SenseClusters[3] [68]; 2) *word clustering*, which seeks to cluster words that are semantically similar and each cluster represents a specific sense [48,63,19,80,67,60,46]; 3) *co-occurrence graphs*, where the semantics of a word can be deduced by building and analyzing a word co-occurrence graph [58,85,81,82,1,

[2] `https://wsd.nlm.nih.gov/collaboration.shtml`

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

26, 3, 38, 42, 39]; and 4) *probabilistic clustering*, which can discover latent topic structures from the contexts of the words without involving feature engineering [15, 12, 88, 78, 45, 20, 77, 83].

Despite the number of WSI studies mentioned above, very few methods have been developed specially for the biomedical domain. In [61], a network of word co-occurrences was defined to induce both word senses and word contexts. Another work [27] presented an efficient graph-based algorithm to cluster words into groups for WSI. A comparison between graph-based approaches and topic modeling approaches was conducted to evaluate the state-of-the-art WSI methods in the clinical domain [18]. Another study has also been proposed to find semantic ambiguities [72] using agglomerative clustering methods on the context vectors of a particular target word in biomedical texts.

2.2 Polysemy detection

Polysemy detection seeks to detect whether an entity has more than one meaning (i.e., polysemic, true or false) based on the context of the entity. Term ambiguity detection is a task related to polysemy detection, as proposed in [10]: given a term and a corresponding topic domain, determining whether the term uniquely references a member of that topic domain. In [41], the authors proposed a rank-based distance measure to explore the vector-spatial properties of the ambiguous entity and to decide if a German preposition is polysemous or not. Polysemy detection is also related to other well-studied NLP topics such as named-entity disambiguation and WSD. However, named-entity disambiguation and WSD both assume that the number of senses for a target word is given. This assumption is inappropriate for enriching KBs, ontologies, and terminologies because the number of senses of a new candidate entity is not known.

In general, polysemy detection approaches (as well as WSI methods) discussed above characterize the text data into a vector of features, e.g., using the most well-known “bag-of-words” language model, and then use a learning algorithm (either supervised or unsupervised) to capture the polysemousness and the senses of a word.

2.3 Sense number prediction

In WSI, the number of senses (clusters) for ambiguous words is normally treated as an a priori knowledge; and most popular clustering methods require the number of clusters to be defined as an input parameter. Nevertheless, in clustering analysis, a major problem is to determine the most appropriate number of clusters, which can significantly affect the clustering results, often leading to poor performance [25]. Hence, learning the appropriate numbers of senses for ambiguous words is crucial for WSI tasks [56]. Klapaftis et al. [39] used the Hierarchical Dirichlet Processes [78] to predict the number of senses of a target word. More recently, Lau et al. [45] combined Hierarchical Dirichlet Processes with a non-parametric Bayesian method for the same purpose. Niu et al. [60] applied a cluster validation method to estimate the number of senses, where the number of clusters ranged from $k_{min} = 2$, to $k_{max} = 5$. However, all these approaches produced larger numbers of word senses (up to 89) on the gold standard SEMEVAL-2010 WSI dataset [45].

Several other strategies for estimating the optimal number of clusters have also been proposed [40,89,47,8,87]. In 1985, Milligan et al. conducted a very extensive comparative evaluation of 30 methods [55] for determining the number of clusters. In a more recent study [9], it was shown that *Calinski and Harabasz's* index was the most effective measure for determining the most appropriate number of clusters, followed by the *Duda and Hart's* method [35]. There is also a *R* package named *NbClust* which was developed for calculating a number of these measures to determinine the number of clusters discussed in [55].

Although many algorithms have been suggested to tackle the problem of determining the number of clusters, there does not appear to be a single method proven to be the most reliable, possibly due to the high complexity in real-world datasets. Thus, task-specific method for determining the number of clusters is always preferred.

3 A new framework for biomedical entity sense induction

Our framework is composed of three steps tackling two specific problems in automating the biomedical entity sense induction task, as depicted in Figure 1.

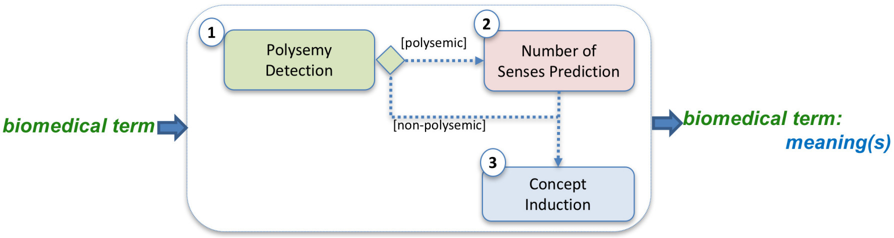

Fig. 1: The proposed framework for biomedical entity sense induction.

1. **Polysemy detection** is modeled as a binary classification task (i.e., *true*, whether a biomedical entity is associated with more than one sense; or *false*, otherwise in the specific context expressed in the source text). It is an important step as it narrows down the targets and reduces the number of options the downstream steps have to explore.

2. **Number of senses prediction** is to predict the number of concepts (k) associated with a biomedical entity, which is based on a set of clustering quality measures; and
3. **Concept induction** is to induce the concepts of a biomedical entity according to its context, based on the extracted keywords of the clusters.

Our dataset contained 203 ambiguous biomedical entities extracted from MEDLINE[4] abstracts as a part of the word sense disambiguation (WSD) test collections published by the National Library of Medicine [37]. Each biomedical entity in the dataset was annotated with one or more MeSH concepts and linked to the abstracts from which it was extracted. To construct the classification models for polysemy detection, we manually curated a dataset of 203 non-ambiguous biomedical entities as negative samples. The curation process of this dataset is described in the results section.

In the following sections, we will describe each step of the framework in detail.

3.1 Polysemy detection

In [50], we presented a set of statistical measures as features to characterize a piece of text. These features were extracted either directly from the text (i.e., direct features, e.g., the number of UMLS terms in the text) or from an undirected graph induced from the text based on co-occurrences (i.e., graph-based features, e.g., the unweighted degree of the target entity). A total of 23 features was proposed. A detailed description of these 23 features is presented in the supplemental material. We also used two terminology resources: UMLS (i.e., biomedical) and AGROVOC (i.e., agronomic)[5] to derive these features. These two dictionaries have a certain degree of overlapping concepts, which can be considered as polysemic entities that belong to both biomedical and agricultural domains. For instance, the entity *"cold"* can represent either a disease (i.e., the common cold) or the feeling of no warmth in UMLS, as well as the temperature of the weather in AGROVOC. Thus, we hypothesized that new entities (that did not appear in these two dictionaries) that co-occurred with existing polysemic entities were more likely to be polysemic as well.

Based on these 23 features, we then experimented with a wide range of learning algorithms to determine whether an entity is polysemic.

3.2 Number of senses prediction

One of the most essential problems in WSI is to determine the number of senses k. Many algorithms have been proposed to identify k, but none was tailored for

[4] MEDLINE is a bibliographic database of life sciences and biomedical information.

[5] AGROVOC is a multilingual controlled vocabulary covering all areas of interest to the Food and Agriculture Organization of the United Nations: `http://aims.fao.org/agrovoc`

biomedical text [57,59,54,17]. Further, one limitation of these approaches is that they tend to predict a high number of senses, possibly due to the nature of the text they were targeting. In contrast, in the biomedical domain, according to UMLS version 2015 AA, polysemic terms were linked to only 2 to 5 senses. Thus, as we aim to identify possible senses for a new biomedical candidate term, we will limit the number of senses between 2 and 5, which was also used in [60].

Table 1 shows the descriptive statistics of polysemic entities in UMLS version 2015 AA for English, French, and Spanish. The English version of UMLS contained about $9,919,000$ distinct entities, $65,546$ of which were polysemic (i.e., ∼0.66%, roughly one out of every 200 UMLS entities was polysemic). Similarly, Table 2 shows the descriptive statistics of polysemic entities in MeSH[6] (Medical Subject Headings). The number of polysemic entities in the English version of MeSH was 179, which was ∼0.02% (i.e., roughly one out of every 5000 MeSH entities was polysemic). In short, there are more non-polysemic (monosemous) than polysemic entities in the biomedical domain for the three languages: English, French, and Spanish.

Table 1: Polysemic entities in UMLS.

# 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 2: Polysemic entities in MeSH.

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

To determine the number of senses, we executed a number of different clustering algorithms varying k (i.e., the number of clusters) from 2 to 5, then evaluated the quality indexes of the resulting clusters, and picked a k with the best clustering quality index[7]. We used the CLUTO[8] application – a flexible software program for clustering and analyzing the characteristics of the clusters, to conduct our experiments. We experimented with 5 different

[6] MeSH is the NLM controlled vocabulary used for indexing articles in PubMed

[7] This approach was an expansion of our previous study published as a poster in EDBT 2016 [52]

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

clustering algorithms (i.e., *rb, rbr, direct, agglo, graph,*) of three different types: partitional, agglomerative, and graph-partitioning.

Two common types of quality indexes [30], external and internal, are often used to evaluate the quality of the clustering results. External indexes use pre-labelled datasets with known cluster configurations; while, internal indexes are used to evaluate the goodness of a cluster configuration without any prior knowledge of the clusters. In this project, we proposed 5 new internal clustering quality indexes, as shown in Table 4, built upon the intra-cluster similarity (i.e., internal similarities, ISIM, of the objects within a cluster) and inter-cluster similarity (i.e., external similarities, ESIM, between clusters) measures offered by CLUTO. To find the optimal number of clusters, we also need an objective function to rank the quality of a clustering solution based on a quality measure [13]. We can obtain the optimal clustering results by optimizing (i.e. maximize/minimize) the objective function, which gives us an idea as to whether the obtained clusters are homogeneous. CLUTO has a set of built-in objective functions, as shown in Table 3. Nevertheless, each quality measure and each objective function have both strengths and weaknesses. Thus, we took an ensembing approach and examined the combinations of the different quality measures and objective functions.

Table 3: Objective functions for finding the number of clusters (k) in CLUTO.

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

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

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

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

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

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

Note that when optimizing the objective functions, we aim to maximize the internal similarity (*ISIM*), and minimize the external similarity (*ESIM*).

Table 4: New internal indexes for choosing the best k.

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

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

then we choose the maximal value of the *ISIM* average of all clusters:

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

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

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

then

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

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 $\mid S_i \mid$.

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

then we choose the maximal value as the clustering solution should have a high difference between *ISIM* and *ESIM*, showing that each cluster is compact and the clusters are well separated.

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

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 each cluster $\mid S_i \mid$, and the sum of *ESIM* multiplied by the number of objects in each cluster.

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

then, we choose the maximal average value, because the clustering solution should have 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) 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 of 10.

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

then, we choose the maximal value. In general, the value of the objective function is higher when the number of clusters is high, so we address this drawback via taking the logarithm of the number of clusters.

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

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

3.3 Concept induction

Following best practice for concept induction, we used the *rb* clustering algorithm that proved to perform well for text data, and used the predicted number of senses “k” from the previous step. Then we extracted the most relevant keywords of each cluster to represent the concept of the cluster. If a candidate entity was non-polysemic, then $k = 1$. Therefore, we did not need to apply a clustering algorithm, and directly extracted the most relevant keywords to represent the concept.

4 Results and discussion

We first discuss our datasets and then the results of our experiments for the proposed biomedical entity sense induction framework in this section. Each step of the framework was evaluated independently to show its effectiveness.

4.1 Datasets

4.1.1 Polysemic datatset

We extracted 203 ambiguous entities in English from the MSH WSD[9] [37] dataset, which contained 106 ambiguous abbreviations, 88 ambiguous terms, and 9 entities that were combinations of both ambiguous abbreviations and terms. Each ambiguous entity was linked to on average 180 titles/abstracts obtained from MEDLINE. The MSH WSD dataset was a well-known benchmark dataset in biomedical word sense disambiguation literature [36, 71, 62, 84].

Table 5 shows a few example entities in the dataset and their respective numbers of senses.

Table 5: Details of the polysemic dataset.

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

[9] https://wsd.nlm.nih.gov/collaboration.shtml

4.1.2 Non-polysemic dataset

We needed negative samples (i.e., non-ambiguous biomedical entities) to build the classifiers for polysemic detection. Therefore, we constructed a non-polysemic dataset using the MEDLINE MeSH terms in two steps: (i) selecting non-polysemic terms from MeSH, and (ii) extracting a set of titles and abstracts containing those terms from MEDLINE. We made this annotated dataset publicly available online at `http://simbig.org/NotPolysemicCorpus.zip`.

Table 6 summarizes our polysemic and non-polysemic datasets.

Table 6: Summary of the polysemic and non-polysemic datasets.

Description	
Number of Entities	406
Number of Ambiguous Entities	203
Number of Non-ambiguous Entities	203
Number of Tokens of the Context of Ambiguous Entities	7,597,337
Number 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

4.2 Experiments for polysemy detection

We proposed 23 novel features and conducted a preliminary analysis with these features for detecting polysemic biomedical entities in our previous work [50]. In this paper, we conducted a comprehensive evaluation and in-depth analysis of these 23 novel features. Following best practices in machine learning, we split the dataset into a *training set* (70%) and a *test set* (30%). The training set was used to build the model, while the remaining 30% of the dataset was a hold-out test set. Both types of biomedical terms, i.e., polysemic and non-polysemic terms, are important in the biomedical entity sense induction framework, since we seek to build/enrich dictionaries with new polysemic and non-polysemic biomedical entities. Therefore, it was important to evaluate the performance of our model over each class (polysemic as P and non-polysemic as NP). We reported the classification performance in terms of *F-measure (F)* for each class, on the hold-out test set, as well as the average of the two F-measure values. We experimented with a set of well-known supervised algorithms, implemented in the Weka[10] software with the default parameters for each algorithm.

Table 7 shows our experiment results. The Naive Bayes (NB) obtained the best results, with an average *F-measure* of 98.4%. Excluding ZeroR and OneR, the average *F-measures* of the other classifiers were between 94.25%

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

(i.e., MCC) and 98.4% (i.e., NB). These results show that our novel features performed well in the polysemy detection task.

Table 7: Evaluating classifiers with both direct and graph-based features.

	F_P	F_{NP}	$F_{Average}$
Zero Rule (ZeroR)	66.7 %	00.0 %	33.35 %
One Rule (OneR)	85.2 %	85.2 %	85.20 %
Naive Bayes (NB)	98.4 %	98.4 %	**98.40 %**
AdaBoost (AB)	96.0 %	95.8 %	95.90 %
Decision Tree (DT)	97.6 %	97.5 %	97.55 %
Support Vector Machine (SVM)	97.5 %	97.6 %	97.55 %
Meta Bagging (MB)	98.4 %	98.3 %	98.35 %
k-nearest neighbors (k-NN), $k = 1$	98.3 %	98.4 %	98.35 %
Multilayer Perceptron (NN)	95.8 %	96.0 %	95.90 %
MultiClassClassifier Logistic (MCC)	94.3 %	94.2 %	94.25 %

[*] Where F_P is the *F-measure* of the polysemic class and F_{NP} the *F-measure* of non-polysemic class.

4.2.1 Discussion

To the best of our knowledge, no previous studies has focused on polysemy detection (i.e., as a binary classification task) to determine whether a biomedical term is polysemic or not. However, a few studies have also used machine learning methods to classify words or phrases. For instance, in [4], Al-Mubaid et al. built a binary classification model to classify ambiguous entities into two semantic categories: genes or proteins. In a follow up study, the authors used similar methods for biomedical word-sense disambiguation [5]. In both studies, the authors built support vector machines (SVMs) and used neighborhood context words of the target word as features. In addition, they used a number feature selection methods such as chi-square (χ^2) [4], mutual information (MI) [5], and M2 [5]. Since these studies were framed as word classification tasks similar to our polysemy detection task, we adapted their methods to evaluate our polysemy detection approach.

We used the same processes presented in [4,5] for building the polysemy detection classifier. From the labeled training examples of the target biomedical entity, we built the feature vectors using their neighborhood context words. The top 20, 30, 50, 100, and 200 context words were selected using the feature selection methods: χ^2, MI, and M2. We then built SVMs[11] with the selected features. In our experiments, the best results were obtained with the top 200 context words. Table 8 summarizes the performance results of these SVMs in terms of their *F-measures*. The χ^2 feature selection method obtained the best results, with an average *F-measure* of 52.8% between polysemic and non-polysemic samples. These results show that using neighboring words as fea-

[11] The SVM algorithm uses a polynomial kernel.

tures with SVMs are suboptimal in classifying biomedical terms as polysemic or not.

Table 8: Results of using neighborhood context words as features with support vector machines for polysemy detection.

	F_P	F_{NP}	$F_{Average}$
χ^2	40.1 %	65.5 %	52.80 %
MI	01.7 %	66.0 %	33.85 %
$M2$	35.6 %	64.7 %	50.15 %

[*] Where F_P is the *F-measure* of the polysemic class and F_{NP} the *F-measure* of non-polysemic class.

As shown in Figure 2, our method achieved better performance than the baseline studies.

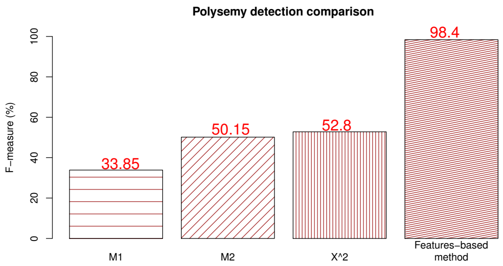

Fig. 2: Comparing our method with the baseline studies.

4.3 Predicting the number of senses

We evaluated our approach to induce the possible number of senses of an entity. Note that we only need to consider the entities that have been classified as polysemic. Thus, we only used the polysemic dataset.

Our evaluation experiments were conducted in three-fold: 1) applied clustering algorithms over the bag-of-words representation and evaluated the proposed new internal quality indexes of the clusters; 2) applied clustering algorithms over a graph representation of the titles/abstracts associated with each entity and evaluated the proposed new internal quality indexes of the clusters; and 3) evaluated the direct and graph-based features we used for polysemy detection above (see Section 3.1) with supervised algorithms to predict the number of senses.

For clustering tasks, we used five well-known clustering algorithms implemented in the CLUTO software, including *rb, rbr, direct, agglo, graph.*

4.3.1 Bag-of-words representation

We used the *cosine* similarity measure for our clustering experiments with the BoW representation. We tested various sizes of the feature vectors, and determined that the best results were obtained with $3,000$ BoW features. The BoW features were extracted with the BioTex application[12] [49,51].

Table 9 illustrates the process for determining the number of clusters (number of possible senses) for the entity *"yellow fever"* according to $a_{k,I2}$ and $c_{k,I2}$, where k is the number of clusters, a and c are two quality indexes, $I2$ is the objective function, and the clustering algorithm is *Partitional.* In our dataset, *"yellow fever"* was linked to 2 senses ("virus" and "vaccine"). Therefore, the correct number of clusters is 2. As shown in Table 9, we varied the number of clusters from 2 to 5, applied the clustering algorithm, and computed the quality indexes.

As shown in Table 9, according to the quality index c, the optimal number of clusters is $k = 2$; while according to the quality index a, the optimal number of clusters is $k = 5$.

Table 9: Choosing k according to $a_{k,I2}$ and $c_{k,I2}$ values.

Algorithm: **Partitional (rb)**						
k	***Cluster ID***	$\mid S_i \mid$	***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		
			$\boldsymbol{max(a_{k,I2})}$		$k = 5$	
			$\boldsymbol{max(c_{k,I2})}$			$k = 2$

* k is the number of clusters, $\mid S_i \mid$ is the number of entities in the cluster i, ISIM is the intra-cluster similarity, ESIM is the inter-cluster similarity, a and c are two quality indexes, and $I2$ is the objective function.

[12] BioTex is an application we previously built to automatically extract biomedical terms from free text: `http://tubo.lirmm.fr/biotex/`

Note that we considered 5 different objective functions for 5 different clustering algorithms with 5 different quality indexes, for each entity in our dataset. Table 10 summarizes the results for *"yellow fever"* considering two objective functions *I1* and *I2*.

Table 10: Predicting the number of senses (k) for *"yellow fever"*, with bag-of-words representation (The true number of senses is 2).

Internal Indexes	***rb***	***rbr***	***direct***	***agglo***	***graph***
$\boldsymbol{max(a_{k,I1})}$	5	5	4	5	2
$\boldsymbol{min(b_{k,I1})}$	3	3	3	**2**	**2**
$\boldsymbol{max(c_{k,I1})}$	3	3	2	2	2
$\boldsymbol{max(e_{k,I1})}$	5	5	5	5	2
$\boldsymbol{max(f_{k,I1})}$	2	2	2	2	2
$\boldsymbol{max(a_{k,I2})}$	5	5	5	5	5
$\boldsymbol{min(b_{k,I2})}$	4	4	4	2	2
$\boldsymbol{max(c_{k,I2})}$	2	2	2	2	2
$\boldsymbol{max(e_{k,I2})}$	5	5	5	5	5
$\boldsymbol{max(f_{k,I2})}$	2	2	2	2	2

[*] $\forall k = \{2, 3, 4, 5\}$; *rb, rbr, direct,* and *agglo* are the clustering algorithms; *a, b, c, e,* and *f* are the quality indexes.

To evaluate the performance of the combinations of different quality indexes, clustering algorithms and objective functions, we carried out the experiment for all of our 203 ambiguous entities. Table 11 summarizes the *F-measure* for determining the number of clusters on the entire dataset, while considering two objective functions *I1* and *I2*. Overall, Table 11 shows that $max(f_{k,OF})$ gave the best results for all of the clustering algorithms with the objective function *I2*.

Table 11: F-measures for predicting the number of senses with the 203 ambiguous entities using the bag-of-words representation.

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** %

[*] $\forall k = \{2, 3, 4, 5\}$; *rb, rbr, direct,* and *agglo* are the clustering algorithms; *a, b, c, e,* and *f* are the quality indexes.

4.3.2 Graph representation

Similar to the evaluation of bag-of-words representations, we evaluated the graph representations with the combinations of 5 objective functions, 5 clustering algorithms, and 5 quality indexes.

Graph construction: For each biomedical entity of interest, we built an undirected weighted graph (see Figure 3), where the vertices were biomedical entities, and the edges denoted co-occurrence relationships (i.e., thus undirected) between the biomedical entities. The weight of an edge represented the degree of association between two biomedical entities. Each entity graph contained the entity of interest and the top $1,000$ most important entities that co-occurred with the entity of interest. The co-occurring entities and their importance values were extracted with BioTex from the set of abstracts (A_t) that contain the entity of interest (t). We used the *Dice coefficient* (D), a measure to compute the degree of co-occurrence between two entities x and y in a set of texts (i.e., titles and abstracts) in the graph.

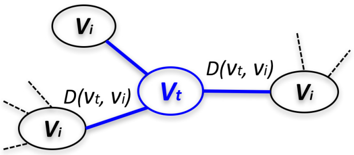

Fig. 3: An example of an entity graph, where t is the entity of interest.

In Figure 3, vertex v_t represents the entity of interest t, vertex v_i ($i = 1..n$) represents an entity i that co-occurred with t, the weight of the edge between v_t and v_i is the dice coefficient $D(v_t, v_i)$ between entity t and i (i.e., $weight(v_t, v_i) = D(v_t, v_i)$).

Table 12 summarizes the results for the entity *"yellow fever"* with two objective functions *I1* and *I2*. In our dataset, the number of clusters (concepts) of the term *"yellow fever"* is 2. We observed that $max(f_{k,I1})$ and $max(f_{k,I1})$ predicted the correct number of senses in general.

Similar to evaluating the bag-of-words representations, we conducted this experiment for all of our 203 ambiguous entities. Table 13 shows the *F-measure* results for the prediction of the number of clusters. As shown, $max(f_{k,OF})$ gives the best *F-measure* results for all the clustering algorithms for both objective functions.

4.3.3 Prediction with both the direct and graph-based features

Similar to our approach for polysemy detection (see Section 3.1), we used both the direct and graph-based features to form a multiclass supervised classification task to predict the number of senses (i.e., 4 classes: 2, 3, 4, 5). The

Table 12: Predicting the number of senses (k) for *"yellow fever"*, with graph representation (The true number of senses is 2).

Internal Indexes	***rb***	***rbr***	***direct***	***agglo***	***graph***
$\boldsymbol{max(a_{k,I1})}$	2	2	2	5	2
$\boldsymbol{min(b_{k,I1})}$	2	2	2	2	2
$\boldsymbol{max(c_{k,I1})}$	2	2	2	5	2
$\boldsymbol{max(e_{k,I1})}$	5	5	5	5	2
$\boldsymbol{max(f_{k,I1})}$	2	2	2	2	2
$\boldsymbol{max(a_{k,I2})}$	5	5	4	5	2
$\boldsymbol{min(b_{k,I2})}$	2	2	2	2	2
$\boldsymbol{max(c_{k,I2})}$	2	2	2	5	2
$\boldsymbol{max(e_{k,I2})}$	5	5	5	5	2
$\boldsymbol{max(f_{k,I2})}$	2	2	2	2	2

* $\forall k = \{2, 3, 4, 5\}$; *rb, rbr, direct,* and *agglo* are the clustering algorithms; *a, b, c, e,* and *f* are the quality indexes.

Table 13: F-measures for predicting the number of senses with the 203 ambiguous entities using the graph representation.

Internal Indexes	***rb***	***rbr***	***direct***	***agglo***	***graph***
$\boldsymbol{max(a_{k,I1})}$	1.97 %	1.97 %	1.48 %	1.48 %	9.36 %
$\boldsymbol{min(b_{k,I1})}$	77.83 %	77.83 %	75.86 %	**93.10** %	64.04 %
$\boldsymbol{max(c_{k,I1})}$	76.35 %	74.88 %	76.85 %	85.22 %	64.53 %
$\boldsymbol{min(c_{k,I1})}$	8.37 %	7.88 %	7.39 %	0.49 %	21.67 %
$\boldsymbol{max(e_{k,I1})}$	3.94 %	4.43 %	4.93 %	47.78 %	3.94 %
$\boldsymbol{max(f_{k,I1})}$	**93.10** %	**93.10** %	**93.10** %	**93.10** %	**93.10** %
$\boldsymbol{max(a_{k,I2})}$	0.49 %	0.99 %	0.49 %	1.48 %	2.96 %
$\boldsymbol{min(b_{k,I2})}$	82.27 %	82.76 %	86.21 %	**93.10** %	80.3 %
$\boldsymbol{max(c_{k,I2})}$	91.13 %	91.13 %	90.15 %	85.22 %	87.19 %
$\boldsymbol{min(c_{k,I2})}$	0.99 %	0.99 %	0.99 %	0.49 %	1.48 %
$\boldsymbol{max(e_{k,I2})}$	4.43 %	3.94 %	3.94 %	47.78 %	2.46 %
$\boldsymbol{max(f_{k,I2})}$	**93.10** %	**93.10** %	**93.10** %	**93.10** %	**93.10** %

* $\forall k = \{2, 3, 4, 5\}$; *rb, rbr, direct,* and *agglo* are the clustering algorithms; *a, b, c, e,* and *f* are the quality indexes.

results are provided in terms of *Accuracy (A), Precision (P), Recall (R)*, and *F-Measure (F).*

Table 14 summarizes the results on a hold-out independent test dataset. The MultiClassClassifier Logistic (MCC) obtained the best results, with an *F-measure* of 91.2%, followed by Meta Bagging (MB) with an *F-measure* of 90.9%. These results show that these features are useful for predicting the number of clusters.

4.3.4 Discussion

The bag-of-words and graph representations obtained similar *F-measure* values in predicting the number of senses. In both cases, the best *F-measure* is 93.1%. As shown in Tables 11 and 13, the best *F-measure* results are given by the f_k

Table 14: Number of senses prediction using both direct and graph-based features.

	$F_{average}$
Naive Bayes (NB)	76.9 %
AdaBoost (AB)	89.3 %
Tree Decision (TD)	89.0 %
Support Vector Machine (SVM)	89.8 %
Meta Bagging (MB)	90.9 %
k-nearest neighbors (k-NN), $k = 1$	88.6 %
Multilayer Perceptron (NN)	89.0 %
MultiClassClassifier Logistic (MCC)	**91.2 %**

index. The f_k index is the division between the value of the objective function and the logarithm of k.

When using supervised learning algorithms based on the direct and graph-based feature representations, MultiClassClassifier Logistic (MCC) and Meta Bagging (MB) were two of the best models. MB is a kind of ensemble learning algorithm that generates multiple versions of a predictor to build an aggregated predictor. MCC is also a meta-classifier that handles multi-class datasets with multiple binary classifiers. In short, the direct and graph-based features are effective in both determining the polysemy of candidate terms as well as predicting the associated number of senses (or concepts), which makes it easier to adopt in real-world implementations.

Most existing WSI studies were in the general domain (e.g., Duluth-WSI [65], UoY [42], NMFlib [80], NB [20], RPCL [33]) and have used the SemEval-2010 WSI shared task dataset [53] to test their approaches. The SemEval-2010 dataset contains 100 target words: 50 nouns and 50 verbs. A common step between the SemEval-2010 WSI task (task 14) [53] and our framework is the number of sense prediction. The dataset used in SemEval-2010 differs significantly from ours as: (1) it contains both single-word nouns and single-word verbs as target words, while ours contains single- and multi-word entities which are composed mainly of nouns and adjectives; (2) the target words are associated with a higher number of senses ($2 \leq k \leq 14$), while ours are between 2 and 5; and (3) texts supplied for each target word were extracted from different websites using Yahoo Search API, and directly from additional sources including Wall Street Journal, CNN, ABC and others; while our texts were extracted from PubMed. Nevertheless, we evaluated our number of sense prediction method over the SemEval-2010 dataset, which resulted in F-measures of 13.5% and 43% when predicting the number of senses for nouns and verbs, respectively. These poor results are mainly caused by the fact that our feature representations (i.e., the bag-of-words and graph-based representations) were based on knowledge from the biomedical domain (i.e., UMLS). In our feature representations, we used the LIDF-measure, which were derived based on the linguistic patterns of biomedical entities in UMLS. These results suggest that in these NLP tasks, using features tailored to the specific domain of the study will improve model performances significantly.

Moreover, a particular challenge related to WSI in the biomedical domain in comparison to the general domain is the different types of lexical ambiguity that exist and the unique characteristics of biomedical documents. In addition to ambiguous terms (words or phrases), abbreviations occur more frequently within biomedical documents [76] and they can have more than one possible expansion [76]. Moreover, the names of genes also contain lexical ambiguities, especially when standardized naming conventions are not always followed. More than one thousand gene terms overlap with generic English meanings [75,86]. Another challenge associated with the biomedical domain is the growing adoption of Electronic Health Record (EHR) and clinical documents that are manually created under a time constrain in which healthcare professionals often use shortened biomedical entity forms that are frequently ambiguous [11]. In addition, to the best of our knowledge, there is no biomedical WSI studies focusing on predicting the number of senses. Therefore, a direct baseline measure was not available. Nevertheless, as mentioned previously, the number of senses (clusters) was predicted by evaluating the clustering quality.

Thus, we used our dataset to our proposed internal clustering quality indexes with the indexes implemented in the *R* package *NbClust* including CH [16] (proved to be the most effective in [9]), DB [24], Duda [28] (proved to be the second most effective in [35]), KL [43], Pseudot2 [28], Sdbw [31], Sdindex [32], and Silhouette [70].

Another strong point of *NbClust* is that it allows to range the number of clusters between $k_{min} = 2$, to $k_{max} = 5$, as proposed in our method. Table 15 shows the results in terms of F-measure average of eight state-of-the-art indexes to predict the number of senses of polysemic biomedical terms. The CH index reached the highest performance with an F-measure of 69.9% followed by the Duda and Pseudot2 index with an F-measure of 64%, which proved the assertions done in [9,35].

Table 15: Results of predicting the number of senses using eight state-of-the-art clustering quality indexes.

	$F_{average}$
CH	69.9 %
DB	21.2 %
Duda	64.0 %
KL	40.9 %
Pseudot2	64.0 %
Sdbw	47.8 %
Sdindex	55.7 %
Silhouette	55.7 %

As shown in Figure 4, where the y-axis is the average F-measure (F), our proposed clustering quality index (f_x) outperformed other methods including our own feature-based method in predicting the number of senses on our dataset. Even though our feature-based method performed slightly worse than our clustering quality index-based method, it outperformed all other state-of-the-art clustering quality indexes. One possible reason is that our index and novel features were built using domain dictionaries such as UMLS, AGROVOC, which might provide more representative characteristics of our dataset.

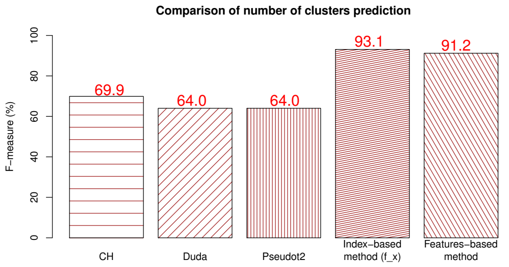

Fig. 4: Results of the comparison between three state-of-the-art indexes and our methods.

4.4 Concept induction

In this section we evaluate our method to induce the possible concept(s) of an entity. We considered both the polysemic and non-polysemic entities. The most difficult challenge is to identify the distinct senses of the entities that have been classified as polysemic. We evaluated the results of inducing the concept(s) of biomedical entities applying clustering algorithms over the bag-of-words representation. The number of senses predicted in the section above was used as an input of the clustering algorithms. For clustering tasks, we used the *rb* clustering algorithm, which gave the best results according to the CLUTO software. We used two objective functions: *I1* and *I2* (see Section 3.2), which are recommended in CLUTO. We then extracted the top ranked (5,10, and 20) keywords for each generated cluster with the *tf-idf, okapi-bm25,* and *LIDF-value* measures. Finally, we compared the keywords with the definitions of each entity in UMLS, and measured the overlapping of the keywords over the definition. The definitions of each entity were extracted from UMLS based on the CUIs provided in both datasets. Since for an entity, there are multiple UMLS definitions and clusters, there are also different combinations to match these UMLS definitions to clusters. For instance, “yellow fever” has two definitions and two clusters, there are two different combinations for its evaluation: “definition1-cluster1—definition2-cluster2” or “definition1-cluster2—definition2-cluster1”. In our evaluation, we automatically took the combination which maximized the overlapping rate between the keywords in clusters and the words in definitions. Table 16 summarizes the results of the overlapping of the keywords over the definitions. The *tf-idf* obtained the best results, with an average overlapping rate of 42%.

Table 16: The average overlapping rates ($KW@5, 10, 20$) of extracted keywords over UMLS entity definitions.

	KW@5	*KW@10*	*KW@20*
I1 *and* ***LIDF-value***	0.352	0.311	0.270
I1 *and* ***okapi***	0.380	0.326	0.259
I1 *and* ***tf-idf***	0.404	0.326	0.276
I2 *and* ***LIDF-value***	0.349	0.311	0.278
I2 *and* ***okapi***	0.399	0.324	0.270
I2 *and* ***tf-idf***	**0.420**	**0.355**	**0.280**

4.4.1 Discussion

The overlapping rates showed in Table 16 give meaningful information as a performance metric to evaluate concept induction results. However, overlapping rates do not give the full picture, as many entity definitions in UMLS do not contain the exact keywords, but contain words that have similar semantics to the extracted keywords.

Table 17: Samples of keywords extracted per cluster for the entity *"yellow fever"*.

CUI	**Definition**	**Keywords**
C0043395	An acute infectious disease primarily of the tropics, caused by a virus and transmitted to man by mosquitoes of the genera Aedes and Haemagogus. The severe form is characterized by fever, HEMOLYTIC JAUNDICE, and renal damage.	outbreak virus news outbreak news dengue disease epidemic mosquito fever fever virus
C0301508	Vaccine used to prevent YELLOW FEVER. It consists of a live attenuated 17D strain of the YELLOW FEVER VIRUS.	virus vaccination vaccine disease encephalitis fever vaccination yellow fever vaccination 17d vaccine response fever vaccine

Table 17 shows an example of keywords extracted per cluster for the entity *"yellow fever"*, where we can see that every cluster contains keywords strongly related to the concepts of the entity *"yellow fever"* (i.e., "vaccine" and "virus").

We also performed a manual evaluation of the induced concepts to validate if the extracted keywords conveyed the right semantics of each biomedical entity (i.e., the reference definitions are concepts extracted from UMLS). The evaluation was conducted on the polysemic dataset for 174 entities. 73.33% of the induced concepts can adequately represent their UMLS definitions. We also reviewed the remaining 26.67% entities whose induced concepts were considered of low quality to represent their UMLS definitions. We made several observations of these low-quality clusters (induced concepts). First, in our framework, we extracted multiple clusters for each polysemic entity, where each cluster was supposed to represent a distinct sense. Often in these low-quality induced concepts, we observed that only part of the clusters could convey the UMLS definitions of the corresponding biomedical entity, while other clusters did not. However, the clusters that did not correspond to any UMLS definitions themselves were still cohesive internally. One possibility is that these clusters might represent new meanings (i.e., senses) of the biomedical entity that were not captured in UMLS. Second, some of the UMLS definitions associated with a single entity were semantically close. Table 18 illustrates an example of the automatically detected low-quality results induced for the entity *"HIV"*, where the two distinct concepts (i.e., C0019693 and C0019682) defined in the UMLS for *"HIV"* have very similar meanings. Third, the keywords we extracted were single words (rather than phrases), which might not be able to convey the semantic information of the definition. Fourth, our evaluation can also have been affected by the quality of the terminologies in the UMLS. UMLS merged hundreds of different terminology resources, and it contains many different types of errors, including semantic [21,90] and lexical errors [90], and numerous redundancy[22]. For instance, a quality assurance study of UMLS in 2009 found errors in 81% of the concepts studied [90].

5 Conclusions

In this paper, we present a framework for biomedical entity sense induction with three steps: 1) to predict whether a biomedical term is polysemic, 2) to induce the number of senses for a biomedical entity, and 3) to induce the concepts (senses) through ranked keyword extraction. The first two steps are novel, which existing WSI frameworks have often neglected.

Through extensive evaluations, we have shown that the novel features allowed a more effective polysemy classification task. We also presented a novel approach to predict the number of senses (clusters) for candidate biomedical entities. Our contribution is the internal clustering quality indexes, which are then used to predict the number of senses. We also experimented with the same direct and graph-based features used for polysemy detection to predict the number of senses, which have also shown promising results. Using the predicted number of senses as a priori, the clustering task for concept induction is then straightforward. From the clusters, we can easily extract keywords to represent the different senses (concepts) of the biomedical entity.

Table 18: Keywords extracted of the induced concepts for the entity *"HIV"*.

CUI	Definition	Keywords
C0019693	Includes the spectrum of human immunodeficiency virus infections that range from asymptomatic seropositivity, thru AIDS-related complex (ARC), to acquired immunodeficiency syndrome (AIDS).	virus peptide replication vaccine activity infection antibody immunodeficiency host fusion
C0019682	Human immunodeficiency virus. A non-taxonomic and historical term referring to any of two species, specifically HIV-1 and/or HIV-2. Prior to 1986, this was called human T-lymphotropic virus type III/lymphadenopathy-associated virus (HTLV-III/LAV). From 1986-1990, it was an official species called HIV. Since 1991, HIV was no longer considered an official species name; the two species were designated HIV-1 and HIV-2.	treatment prevention testing transmission research infection study youth risk group

As future work, different strategies could be considered to improve the proposed framework, such as introducing more features using other dictionaries like WordNet and BabelNet.

Acknowledgements This work was supported in part by the French National Research Agency (ANR) grant ANR-12-JS02-01001, the University of Montpellier, the French National Center for Scientific Research (CNRS), the FINCyT program, Peru, the National Science Foundation (NSF) award #1734134, and the National Insitutes of Health (NIH) grant UL1TR001427.

References

1. E. Agirre and P. Edmonds. *Word Sense Disambiguation: Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2007.
2. E. Agirre and A. Soroa. 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, 2007. Association for Computational Linguistics.
3. E. Agirre and A. Soroa. 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, 2007. Association for Computational Linguistics.
4. H. Al-Mubaid and P. Chen. Biomedical term disambiguation: An application to gene-protein name disambiguation. In *Proceedings of the Third International Conference on Information Technology: New Generations*, ITNG '06, pages 606–612, Washington, DC, USA, 2006. IEEE Computer Society.
5. H. Al-Mubaid and S. Gungu. A learning-based approach for biomedical word sense disambiguation. *The Scientific World Journal*, 2012, 2012.

6. L. Albano, D. Beneventano, and S. Bergamaschi. 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, 2014.
7. L. Albano, D. Beneventano, and S. Bergamaschi. 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, 2015. International World Wide Web Conferences Steering Committee, ACM.
8. A. N. Albatineh and M. Niewiadomska-Bugaj. Mcs: A method for finding the number of clusters. *Journal of classification*, 28(2):184–209, 2011.
9. M. J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1):32–46, 2001.
10. T. Baldwin, Y. Li, B. Alexe, and I. R. Stanoi. Automatic term ambiguity detection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 804–809, Sofia, Bulgaria, 2013.
11. W. Blair and B. Smith. Nursing documentation: frameworks and barriers. *Contemporary nurse*, 41(2):160–168, 2012.
12. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022, 2003.
13. J. G. Booth, G. Casella, and J. P. Hobert. Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):119–139, 2008.
14. S. Bordag. Word sense induction: Triplet-based clustering and automatic evaluation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'06, pages 137–144, Trento, Italy, April 2006. Association for Computational Linguistics.
15. S. Brody and M. Lapata. 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, 2009.
16. T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.
17. J. Camacho-Collados, M. T. Pilehvar, and R. Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64, 2016.
18. R. Chasin, A. Rumshisky, O. Uzuner, and P. Szolovits. 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, 2014.
19. P. Chen, W. Ding, C. Bowes, and D. Brown. 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, 2009. Association for Computational Linguistics.
20. D. K. Choe and E. Charniak. 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, October 2013. Association for Computational Linguistics.
21. J. J. Cimino. Auditing the unified medical language system with semantic methods. *Journal of the American Medical Informatics Association*, 5(1):41–51, 1998.
22. J. J. Cimino. Battling scylla and charybdis: the search for redundancy and ambiguity in the 2001 umls metathesaurus. In *Proceedings of the AMIA Symposium*, page 120. American Medical Informatics Association, 2001.
23. P. Cook, J. H. Lau, D. McCarthy, and T. Baldwin. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.
24. D. L. Davies and D. W. Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227, 1979.

25. M. Y. Dehkordi, R. Boostani, and M. Tahmasebi. A novel hybrid structure for clustering. In *Advances in Computer Science and Engineering*, pages 888–891. Springer, 2009.
26. B. Dorow and D. Widdows. 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, 2003. Association for Computational Linguistics.
27. W. Duan, M. Song, and A. Yates. Fast max-margin clustering for unsupervised word sense disambiguation in biomedical texts. *BMC Bioinformatics*, 10(Suppl 3), 2009.
28. R. O. Duda, P. E. Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.
29. L. Frermann and M. Lapata. A bayesian model of diachronic meaning change. *TACL*, 4:31–45, 2016.
30. A. D. Gordon. Classification, (chapman & hall/crc monographs on statistics & applied probability). 1999.
31. M. Halkidi and M. Vazirgiannis. Clustering validity assessment: Finding the optimal partitioning of a data set. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 187–194, Washington, DC, USA, 2001. IEEE Computer Society.
32. M. Halkidi, M. Vazirgiannis, and Y. Batistakis. Quality scheme assessment in the clustering process. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '00, pages 265–276, London, UK, UK, 2000. Springer-Verlag.
33. Y. Huang, X. Shi, J. Su, Y. Chen, and G. Huang. Unsupervised word sense induction using rival penalized competitive learning. *Engineering Applications of Artificial Intelligence*, 41(C):166–174, May 2015.
34. N. Ide and T. Erjavec. Automatic sense tagging using parallel corpora. In *Natural Language Pacific Rim Symposium (artificial intelligence)*, NLPRS '01, 2001.
35. O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, 2008.
36. A. Jimeno-Yepes. Higher order features and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation. *CoRR*, abs/1604.02506, 2016.
37. A. J. Jimeno-Yepes, B. T. McInnes, and A. R. Aronson. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12(1):223, 2011.
38. I. P. Klapaftis and S. Manandhar. 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, 2008. IOS Press.
39. I. P. Klapaftis and S. Manandhar. 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, 2010. Association for Computational Linguistics.
40. A. Kolesnikov, E. Trichina, and T. Kauranne. Estimating the number of clusters in a numerical data set via quantization error modeling. *Pattern Recognition*, 48(3):941–952, 2015.
41. M. Köper and S. S. im Walde. 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, May 2014. European Language Resources Association (ELRA).
42. I. Korkontzelos and S. Manandhar. 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, 2010. Association for Computational Linguistics.
43. W. J. Krzanowski and Y. Lai. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pages 23–34, 1988.

44. J. H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics, 2012.
45. J. H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. 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, 2012. Association for Computational Linguistics.
46. Y. K. Lee and H. T. Ng. 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, 2002. Association for Computational Linguistics.
47. J. Liang, X. Zhao, D. Li, F. Cao, and C. Dang. Determining the number of clusters using information entropy for mixed data. *Pattern Recognition*, 45(6):2251–2265, 2012.
48. D. Lin. 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, 1998. Association for Computational Linguistics.
49. J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire. 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, 2014.
50. J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire. Automatic biomedical term polysemy detection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC'2016, pages 1684–1688, Paris, France, may 2016. European Language Resources Association (ELRA).
51. J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, 19(1-2):59–99, 2016.
52. J. A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire. A way to automatically enrich biomedical ontologies. In *Proceedings of the 19th International Conference on Extending Database Technology*, EDBT'2016, pages 676–677, New York, NY, USA, 2016. ACM.
53. S. Manandhar, I. P. Klapaftis, D. Dligach, and S. S. Pradhan. 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, 2010. Association for Computational Linguistics.
54. D. McCarthy, M. Apidianaki, and K. Erk. Word sense clustering and clusterability. *Comput. Linguist.*, 42(2):245–275, June 2016.
55. G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.
56. B. Mirkin. Choosing the number of clusters. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):252–260, 2011.
57. R. Navigli. A quick tour of word sense disambiguation, induction and related approaches. In *SOFSEM 2012: Theory and practice of computer science*, pages 115–129. Springer, 2012.
58. R. Navigli and G. Crisafulli. 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, 2010. Association for Computational Linguistics.
59. R. Navigli and D. Vannella. Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. volume 2, pages 167–174, 2013.
60. Z.-Y. Niu, D.-H. Ji, and C.-L. Tan. 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, 2007. Association for Computational Linguistics.
61. T.-G. Noh, S.-B. Park, and S.-J. Lee. 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, 2010. ACM.

62. S. V. Pakhomov, G. Finley, R. McEwan, Y. Wang, and G. B. Melton. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, 32(23):3635–3644, 2016.
63. P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM, 2002.
64. T. Pedersen. 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, 2007. Association for Computational Linguistics.
65. T. Pedersen. 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, July 2010. Association for Computational Linguistics.
66. T. Pedersen and R. Bruce. Distinguishing word senses in untagged text. In *Second Conference on Empirical Methods in Natural Language Processing*, EMNLP '97, pages 197–207, 1997.
67. D. Pinto, P. Rosso, and H. Jimenez-Salazar. 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, 2007.
68. A. Purandare and T. Pedersen. Senseclusters: finding clusters that represent word senses. In *Demonstration Papers at HLT-NAACL 2004*, pages 26–29. Association for Computational Linguistics, 2004.
69. A. Purandare and T. Pedersen. 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*, 2004.
70. P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
71. A. K. M. Sabbir, A. Jimeno-Yepes, and R. Kavuluru. Knowledge-based biomedical word sense disambiguation with neural concept embeddings and distant supervision. *CoRR*, abs/1610.08557, 2016.
72. G. Savova and T. Pedersen. Resolving ambiguities in biomedical text with unsupervised clustering approaches. *University of Minnesota Supercomputing Institute Research Report*, 2005.
73. H. Schutze. Dimensions of meaning. In *Proceedings of Supercomputing'92*, pages 787–796. IEEE, 1992.
74. H. Schütze. Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123, Mar. 1998.
75. A. K. Sehgal, P. Srinivasan, and O. Bodenreider. Gene terms and english words: An ambiguous mix. In *Proc. of the ACM SIGIR Workshop on Search and Discovery for Bioinformatics, Sheffield, UK*. Citeseer, 2004.
76. M. Stevenson and Y. Guo. Disambiguation in the biomedical domain: the role of ambiguity type. *Journal of biomedical informatics*, 43(6):972–981, 2010.
77. G. Tang, Y. Xia, J. Sun, M. Zhang, and T. F. Zheng. Statistical word sense aware topic models. *Soft Computing*, 19:1–15, 2014.
78. Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
79. G. Udani, S. Dave, A. Davis, and T. Sibley. 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, 2005. ACM.
80. T. Van de Cruys and M. Apidianaki. 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, 2011.
81. S. van Dongen. *Graph clustering by flow simulation.* PhD thesis, University of Utrecht, May 2000.
82. J. Véronis. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252, 2004.

83. J. Wang, M. Bansal, K. Gimpel, B. Ziebart, and C. Yu. A sense-topic model for word sense induction with unsupervised data enrichment. *Transactions of the Association for Computational Linguistics*, 3:59–71, 2015.
84. Y. Wang, K. Zheng, H. Xu, and Q. Mei. Clinical word sense disambiguation with interactive search and classification. In *AMIA Annual Symposium Proceedings*, volume 2016, page 2062. American Medical Informatics Association, 2016.
85. D. Widdows and B. Dorow. 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, 2002. Association for Computational Linguistics.
86. H. Xu, M. Markatou, R. Dimova, H. Liu, and C. Friedman. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. *BMC bioinformatics*, 7(1):334, 2006.
87. M. Yan. *Methods of determining the number of clusters in a data set and a new clustering criterion*. PhD thesis, Virginia Polytechnic Institute and State University, 2005.
88. X. Yao and B. Van Durme. 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, 2011. Association for Computational Linguistics.
89. H. Yu, Z. Liu, and G. Wang. An automatic method to determine the number of clusters using decision-theoretic rough set. *International Journal of Approximate Reasoning*, 55(1):101–115, 2014.
90. X. Zhu, J.-W. Fan, D. M. Baorto, C. Weng, and J. J. Cimino. A review of auditing methods applied to the content of controlled biomedical terminologies. *Journal of biomedical informatics*, 42(3):413–425, 2009.