How to extract unit of measure in scientific documents?

Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie-Barthelemy, Mathieu Roche

How to Extract Unit of Measure in Scientific Documents?

Soumia Lilia Berrahou[1,2], Patrice Buche[1,2], Juliette Dibie-Barthelemy[3] and Mathieu Roche[1]
[1] *LIRMM – CNRS – Univ. Montpellier 2, 34095 Montpellier Cedex 5, France*
[2] *INRA – UMR IATE, INRIA – GraphIK 2, place Pierre Viala, 34060 Montpellier Cedex 2, France*
[3] *INRA – Mét@risk & AgroParisTech, 75231 Paris Cedex 5, France*
name@lirmm.fr, firstname.name@supagro.inra.fr, firstname.name@agroparistech.fr

Keywords: Information Retrieval, Unit of Measure Extraction, Ontological and Terminological Resource, Machine Learning.

Abstract: A large amount of quantitative data, related to experimental results, is reported in scientific documents in a free form of text. Each quantitative result is characterized by a numerical value often followed by a unit of measure. Extracting automatically quantitative data is a painstaking process because units suffer from different ways of writing within documents. In our paper, we propose to focus on the extraction and identification of the variant units, in order to enrich iteratively the terminological part of an Ontological and Terminological Resource (OTR) and in the end to allow the extraction of quantitative data. Focusing on unit extraction involves two main steps. Since we work on unstructured documents, units are completely drowned in textual information. In the first step, our method aims at handling the crucial time-consuming process of unit location using supervised learning methods. Once the units have been located in the text, the second step of our method consists in extracting and identifying candidate units in order to enrich the OTR. The extracted candidates are compared to units already known in the OTR using a new string distance measure to validate whether or not they are relevant variants. We have made concluding experiments on our two-step method on a set of more than 35000 sentences.

1 INTRODUCTION

Discovering and extracting information reported in published scientific documents is a crucial stake in several scientific domains in order to be able to reuse, manage, exploit and analyze the information they contain. Nevertheless, a large amount of data in any scientific field is still published in today's web in the form of unstructured text rather than as structured semantic information. Indeed, published scientific articles use natural language combined with domain-specific terminology that is extremely tedious to extract in the free form of text.

Linguistic methods for information extraction (IE), such as the ones proposed in GATE (Cunningham et al., 2002), are mostly rule-based techniques, which usually use a combination of gazetteer lists and pattern-matching rules in order to define if a candidate extracted term is valid or not. Those techniques are then included in an overall architecture providing various language processing tasks as building and annotating corpora. The system involves semantic analysis at the simplest level in the process with named entities. Other systems integrate directly formal ontologies in the natural language processing (NLP) environment, called ontology-based information extraction (OBIE), and incorporate semantic knowledge about terms in the context involving reasoning (Maynard et al., 2008; Wimalasuriya and Dou, 2010).

Our work deals with unit recognition and extraction, widely known as a challenging issue of IE as described in *Related Work* section below. Indeed, most of techniques used in classical IE systems, based on linguistic methods, cannot handle quantitative data extraction issues, especially units of measure because they rely on completely different syntactic definitions. New methods and rules of information extraction must be provided and take into account those particularities such as simple or complex units, prefix of multiple or submultiple of units, dimensions, considering the presence or absence of special characters, superscript numbers and so on. Currently, none of the existing NLP methods of the state-of-the-art consider efficiently the extraction of units of measure. However, it remains a paramount importance in many quantitative research fields such as physics, chemistry, biology and food science, where the process of quantitative data extraction in experimental results is

difficult to overcome.

Our work aims at leveraging knowledge in food science domain, which relies on an Ontological and Terminological Resource (OTR). The concepts related to units of measure in this OTR have been defined from the most extended ontology of units, Ontology of units of Measure and related Concepts (OM) (Rijgersberg et al., 2011), (Rijgersberg et al., 2013). OM respects the International System of Units (SI) (Thompson and Taylor, 2008), which organizes quantities and units of measure. The definitions of base and derived units of SI have been translated in a formal representation in OM. Our work focuses on the unit recognition and extraction issues in order to extract typographic variants, synonyms, abbreviations and new units of measure, which are then used to populate the OTR. This is a crucial step to be completed in order to build robust add-in semantic tagging and extracting tools concerning units of measure that could be integrated in OBIE or IE systems. These tools facilitate domain expert participation in the task of annotating textual corpus and data tables for gold standard corpus construction. Moreover, the OTR such populated is used as a basis of reasoning and to make multi-criteria analysis in decision-making systems (Hignette et al., 2008).

We propose a two-step approach in order to perform unit recognition and extraction, which focuses on the following contributions: (i) the first step of our work demonstrates how supervised learning methods can be successfully used to meet the challenge of reducing the time-consuming process of unit location in free form of text. (ii) the second step of our work deals with unit extraction and identification using a new string distance measure in order to overcome the scaling issue.

The paper is structured as follows. In section 2, we point out related work on unit recognition. In section 3, we describe our two-step approach, based on locating and extracting variant terms of units. In sections 4 and 5, each step is precisely detailed. In section 6 we deliver our results on real-world data and finally, discuss and conclude on our work.

2 RELATED WORK

Recent work (Touhami et al., 2011), (Jessop et al., 2011a) has revealed that most of extraction of experimental data fails because units suffer from a large variability of writing within scientific documents. Despite works to produce proper formal standard (Thompson and Taylor, 2008), (Rijgersberg et al., 2013) in order to exchange and process quantitative information, if a unit of measure is written differently in the text, the process of identification fails even in the smallest differences (e.g. *mol* written *mole*).

Other work in related fields as chemistry, focus not only on the identification and annotation of chemical entities (Jessop et al., 2011b) but also on the relationships linking these entities to each other (Hawizy et al., 2011) using ChemicalTagger. ChemicalTagger is the best performing state-of-the-art tool for text-mining in chemistry, a phrase-based semantic natural language processing tool for the language of chemical experiments. However, in (Jessop et al., 2011a), the observation is similar when the authors note that ChemicalTagger (Hawizy et al., 2011) fails in the process of recognizing chemical names as reagent because of the variability of writing units in the text. It perfectly matches the units of measure written in respect to the SI (Thompson and Taylor, 2008) but fails automatically when small changes occur in the text. The authors present this issue as a challenge to be met in future work.

In (Willems et al., 2012), in the same trend as we aim at working, the authors present a method for annotating and extracting quantities and units. Their approach relies on the use of semantic markups available in Latex files. Our approach must address the full-text unstructured documents.

In (Van Assem et al., 2010), the authors propose to overcome unit identification issue in data tables by using a string distance metric, called Jaro-Winkler-TFIDF (Cohen et al., 2003). We have used a new string distance measure, we adapted from Damerau-Levenshtein distance (Damerau, 1964), in order to tackle unit identification issue when units are drowned in textual information. Moreover, unlike our work, they do not need to consider unit location in order to overcome scalability issue since they work on structured data tables.

The current state-of-the-art methods of related domains is not suitable enough to handle efficiently the identification of variant terms of unit issue. Our approach aims at addressing this concern using a novel method to extract unit information in a two-step extraction process.

3 GLOBAL APPROACH

In this section, we present our two-step approach to extract units in order to enrich a domain OTR. We assume that the more we extract units from text, the more we will succeed in the experimental data identification process. Indeed, variant terms of units encountered in scientific documents, due to the variabil-

ity of writing, are seen as potential synonyms in the terminological part of the OTR. Once referenced in the OTR, they will increase the positive results in the identification and extraction process of experimental data. Thus, our work on unit identification task is seen as a sub-step of an overall process of experimental result identification and extraction but remains a crucial stake.

Before presenting our global approach, let us briefly present the unit concepts of the considered OTR. As shown in Figure 1, units are specifically organized around the concept *Unit_concept* and divided into 4 classes of units (units being considered as instances):

- *Singular_Unit* such as day or meter;
- *Unit_Exponentiation* such as Square_Meter;
- *Unit_Division_Or_Multiplication* such as Milliliter_Per_Square_Meter_Per_Day;
- *Unit_Multiple_Or_Submultiple* such as Micrometer.

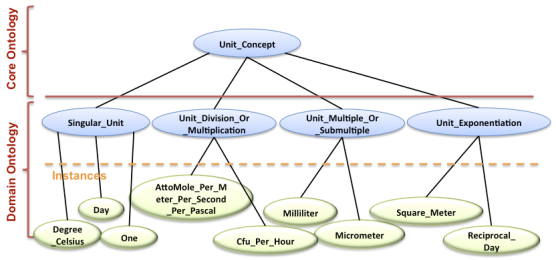

Figure 1: Excerpt of the OTR: Unit_concept.

Each unit has several possible labels, called terms, which are defined in the terminological part of the OTR and denote unit instances, called units in the following, defined in the conceptual part of the OTR.

In Figure 2, we give an example in order to clearly illustrate the main problem of unit recognition. The terms $amol.m^{-1}.s^{-1}.Pa^{-1}$ and *amol/m/s/Pa* referenced in the terminological part of the OTR denote the unit AttoMole_Per_Meter_Per_ Second_Per_Pascal, instance of the Unit_Division_Or_Multiplication concept. Those terms are not sufficient enough to recognize variant terms of units such as suggested in Figure 2. Our method aims at locating those variant terms and extracting them to enrich the OTR with new terms (e.g. amol x m^{-1} x s^{-1} x Pa^{-1} where a multiplication sign has been added).

The method must face several challenges: (i) the question of locating units within the text, do we need to skim all the text or can we predict the best information location without any assumption. The question is important to be asked because the major limit of classical methods of string distance metrics concerns

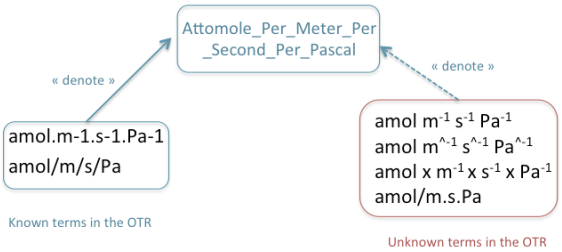

Figure 2: Relevant variant terms of units.

the scalability issue. Currently, these metrics can not be applied on a huge corpus. The first step is a major concern in order to reduce significantly the search space to the relevant sentences where we will be able to apply string distance metrics (ii) the question of extracting variant terms of units, once the best location has been identified, how do we extract variant terms of units from that part of text; and (iii) once extracted, the question of semantically identifying these variant terms of units and find which unit instances they can denote in the OTR. Our approach tackles these key points and intends at discovering more variant terms of units in order to enrich the terminological part of OTR as an iterative process.

The objective of the first step is to reduce significantly the size of search space and overcome the painstaking process of unit location. As a matter of fact, it is often assumed that experimental data are located in specific paragraphs of the text, e.g. *Experiments* or *Results and discussion* in (Hawizy et al., 2011). We aim at locating it without any of these assumptions. However, we are aware that going through all documents involves a time-consuming process. The method to reduce search space involves text mining and supervised machine learning approaches. Our aim in this first step is obtaining the most efficient model from machine learning algorithms that could be then applied in any corpus of a domain where extracting units is a major concern. In section 4, we give more details on different experiments that have been led with the purpose of achieving this best model.

The second step, detailed in section 5, works on textual fragments obtained from the learning model and focuses on terms of unit extraction. As soon as the optimal location for discovering variant terms of units is defined, we first highlight in this location the ones to be extracted. After extracting potential candidates, we compute a new string distance measure in order to distinguish which of the known units in the OTR can match these variant terms and so are potential synonyms to enrich the OTR.

4 LOCATING UNITS

In the first step of the approach, our objective is to focus on locating relevant data concerning units of measure. In the work of (Hawizy et al., 2011), the authors use a Regex-tagger based on regular expressions in order to capture the sentences where quantitative data, chemical entities and units appear. However, when testing the tagger in (Jessop et al., 2011a), the process of recognition fails because of typographic variations of units in the text. This observation has motivated our choices towards supervised learning methods instead of using regular expressions. Hence, our process consists of using supervised classifier to determine whether a part of text is classified as containing "units" or "non-units". In order to obtain the best model to be applied on the corpus, we work on several training sets, each one corresponds to a specific textual context. Thus, as soon as the process recognizes a unit annotation in the sentence of the text, we observe the context around it. We consider different case studies to build these training sets for supervised learning methods:

- the first case study takes into account only the sentence where at least one unit appears and is recognized according to the units defined in the OTR;
- the second and third case study take respectively into account one sentence before and one sentence after the one where at least one unit appears;
- the fourth and fifth case study take respectively into account two sentences before and two sentences after the one where at least one unit appears.

The following is a description of the transformation process tasks and explains our choices in order to obtain the model.

4.1 Preparing Data

Data preparation involves text processing and text transformation tasks. Text processing sub-step consists of:

- text segmentation in order to generate a set of sentences, which deals with our issues, e.g. not going to new sentence when units with point separators are encountered;
- text cleanup, which removes punctuation and special characters from text, except those involved in units;
- text tokenization, which splits up a string of characters into a set of tokens;
- text reducing, which prunes away tokens containing junk words according to a list of stop words;
- text tagging, which annotates the units in the text.

Text tagging relies on comparing the terms denoting the referenced units in the OTR with tokens of each sentence. As soon as the process recognizes a unit or a part of unit, this one is tagged. Our main purpose is obtaining a tagged data used to test our different case studies under learning algorithms.

4.2 Transforming Data

As soon as we have obtained our training data, we can proceed with text transformation, which consists of representing a text according to the words it contains and their occurrences in the sentences. The representative words (features) form the *bag-of-words* according to a relevant threshold we can pre-select (e.g. words appearing more than just once), and we compute their occurrences in each sentence according to three word weighting measures:

- Term Frequency (TF), which considers that the word is more important if it appears several times in sentences;
- Term Frequency-Inverse Document Frequency (TF.IDF (Hiemstra, 2000)), which considers that the word is more important if it appears in less sentences;
- Okapi BM25 (Jones et al., 2000), which takes also into consideration the sentences length where the word appears to define its relevancy.

Sentences become vectors that constitute our training corpus in order to define, with supervised learning methods, whether we are in a "unit" or "non-unit" context.

4.3 Learning Model

Our five case studies performed under those three weight-based measures are run under several learning algorithms. Our aim is to obtain exhaustive experiments and results in order to conclude about the best model in our issue. Various learning algorithms have been tested in our experiments:

- Naive Bayes classifier and the Discriminative Multinominal Naive Bayes (DMNB) classifier;
- J48 decision tree classifier ;
- Sequential Minimal Optimization (SMO), which is a Support Vector Machines classifier (SVM).

The motivation of our choices relies on comparing the behavior of those widely known learning algorithms on a corpus containing many quantitative data.

Naive Bayes (John and Langley, 1995) is competitive for computational efficiency. Decision tree (Kohavi and Quinlan, 2002) classifiers are known to obtain good classification results but are less competitive in speed of execution. SMO (Platt, 1999) is a discriminative classifier known to efficiently behave on text classification and, DMNB is a new text classification algorithm (Su et al., 2008) which performs competitively with discriminative classifiers in terms of accuracy, without losing the advantages of computational efficiency of Naive Bayes classifiers.

The conclusion on the best model is taken after paying a particular attention to several key points:

- Training time and best classification balance;
- Precision, Recall and F-measure results, we focus on the recall, which means we want to maximize the fraction of relevant sentences that are retrieved, without losing too much precision in the results ;
- Confusion matrix, which compares the results of the classifier under test with trusted external judgements.

As we want to estimate how accurately the model of each classifier will perform in practice, we use a 10-fold cross-validation: the original sample is randomly partitioned into 10 equal size subsamples. One subsample is used as the validation data for testing the model while the other subsamples are used as training data. This process is repeated 10 times with each of the subsample used once as the validation data. The averaged result produces the estimation of the model. Using cross-validation is a crucial task in order to avoid "overfitting" effect of the model.

5 EXTRACTING UNITS

In the second step of the approach, we work on textual fragments obtained from the first step. We have significantly reduced the search space and need now to detect, extract and identify variant terms of units or new units. It involves working on textual fragments as follows:

- detecting terms of units inside textual fragments: the tags used in the annotation process of the first step are our entry point;
- extracting terms (or fragment of terms of units) from the entry point: both sides of the tag are scanned and the process stops when it encounters a common word of the dictionary;
- identifying variant terms of units or new units using the new string distance measure.

The identifying process relies on a new string distance measure, adapted from Damerau-Levenshtein (DL) distance (Damerau, 1964). The DL distance between two strings is defined as the minimum number of edits needed to transform one string into another, with the edit operations being insertion, deletion, or substitution of a single character. The distance can then be normalized by using (Maedche and Staab, 2002) approach:

$$SM_{DL}(u1,u2) = max[0; \frac{\min(|u1|,|u2|) - DL(u1,u2)}{min(|u1|,|u2|)}]$$
$$\in [0;1]$$

The similarity measure is computed and the higher it is, the closer the variant terms of units is to the units of the OTR.

Example 1. *Let us consider the similarity between amol/(m.s.Pa) and amol/m.sec.Pa according to the classical DL distance. 2 characters are removed ("(" and ")"), 2 characters are inserted ("e" and "c"). The D_c (the distance between those units) is* 4 *and the similarity distance normalized according to D_c is then:*

$$SM_{D_c}(amol/(m.s.Pa), amol/m.sec.Pa) = max[0; \frac{13-4}{13}]$$
$$SM_{D_c} = 0.69$$

We adapted this distance by considering each block of characters forming the terms of units rather than the simple characters. Thus, in our previous example, the process considers each block as *amol*, *m*, *s*, *Pa* and compares each of them with each block of characters from the units referenced in the OTR. The edit distance adapted to the block becomes D_b = 1 since only the block *sec* varies and

$$SM_{D_b}(amol/(m.s.Pa), amol/m.sec.Pa) = max[0; \frac{4-1}{4}]$$
$$SM_{D_b} = 0.75$$

The new string distance measure is more accurate and more relevant to address units of measure similarity.

6 EXPERIMENTS AND RESULTS

6.1 Data Description

Experiments have been led on real-world dataset composed of 115 scientific documents in food science domain, which represents a set of more than 35 000 sentences. The positive corpus, where at least one unit appears, represents more than 5 000 sentences. We

use a list of 211 terms denoting units of the OTR in order to conduct unit identification. During experiments, we can set the number of instances that will constitute our training data. The results delivered in our paper rely on a training set of 2 000 instances randomly chosen from the corpus, and size balanced between the positive and negative instances. According to the different case studies, the *bag-of-words* varies from 3 000 to 4 800 features.

6.2 Learning the Best Location

6.2.1 Learning under Various Supervised Algorithms

We have tested our five case studies. Only the 3 more relevant cases are reported in Table 1. Each case is represented by a symbol in order to make a readable table:

- XP1: only the sentence where at least one unit appears;
- XP2: two sentences after the one where at least one unit appears;
- XP3: two sentences before the one where at least one unit appears.

Table 1 restitutes the results with all weight-based measures included and according to the case studies. This first table helps us to determine which case is the most relevant context.

Table 1: Results of "Unit" instances: Precision (P), Recall (R), F-measure (F) are given for each case study. Best recall in bold considering P and F.

	Dec. Tree J48			**Naive Bayes**			**DMNB**			**SMO**		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
XP1	0.99	0.87	0.93	0.83	0.93	0.88	0.95	0.96	**0.95**	0.99	0.99	**0.99**
XP2	0.99	0.92	0.96	0.95	0.77	0.85	0.93	0.96	**0.95**	0.99	0.97	**0.99**
XP3	0.99	0.92	0.95	0.77	0.98	0.86	0.94	0.96	**0.95**	0.99	0.97	**0.98**

Table 2: Results of "Unit" instances: Precision (P), Recall (R), F-measure (F) are given for each weight-based measures and boolean matrix. Best results in bold.

Algo.	**Dec. Tree J48**			**Naive Bayes**			**DMNB**			**SMO**		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Boolean	0.99	0.87	0.93	0.83	0.93	0.88	**0.95**	**0.96**	**0.95**	0.99	0.99	0.99
TF	0.99	0.86	0.92	0.69	0.85	0.76	**0.95**	**0.96**	**0.95**	0.84	0.90	0.87
TF.IDF	0.99	0.86	0.92	0.69	0.85	0.76	**0.95**	**0.96**	**0.95**	0.84	0.90	0.87
Okapi	0.99	0.86	0.92	0.69	0.86	0.76	**0.95**	**0.96**	**0.95**	0.77	0.88	0.82

We take into consideration several parameters in order to conclude about the most relevant context: recall, overall classification, F-measure, stability of the various results in the weight-based measures that are presented in section 6.2.2 and confusion matrix results. We are particularly interested in recall since our aim is obtaining the most relevant instances that are retrieved considering "unit" class but, without losing too much precision in the results, which is described by the F-measure. Firstly, we can say that Naive Bayes returns F-measure rates ranging from 0.85 to 0.88. Decision tree (i.e. J48) returns better rates from 0.93 to 0.96. DMNB and SMO return better values (0.95 to 0.99).

Secondly, we can observe that a larger context (i.e. composed of two sentences – XP2 and XP3) does not improve the results. We can conclude that considering the smallest context based on one sentence (i.e. XP1) is enough for unit extraction. This allows us to significantly reduce the search space while being in an optimal context of discovery.

6.2.2 Learning under Various Weight-based Measure

Table 2 restitutes the results on the XP1 context, previously underlined, according to the three weight-based measures and the boolean matrix. This second table allows us to conclude whether the learning algorithm is constant on the weight-based measures.

As the previous analysis, we take into consideration recall, overall classification and F-measure. Firstly, we can say that, all weight-based measures included, Naive Bayes returns rates declining from 86.9% (boolean matrix) to 73.4% (okapi measure), decision tree J48 overall classification falls a bit around $92-93\%$. SMO loses almost 20% with a rate declining from 99.6% to 80.9%. DMNB stays constant with 95.5% regardless of weight-based measures. The DMNB model stays constant on precision, recall and F-measure and achieves a very good recall at 0.96.

6.2.3 Discussion

In this first step, we have studied how supervised learning methods can be successfully used in order to optimize the process of unit location drowned in unstructured scientific documents. For this, we have compared several algorithms and weight-based measures before concluding that the best context for discovering variant terms of units and even new units (cf. next sub-section) is the sentence where at least one unit appears. This result is constantly confirmed regardless weight-based measures under DMNB algorithm with a recall at 0.96 and a precision at 0.95. Thus, we have successfully overcome the time-consuming process of location issue.

Table 4: Identifying variant terms of units.

Variant term	Reference	SMDc	SMDb
10e10 (cm3.m-1.sec-1.Pa-1)	10e10.cm3.m-1.sec-1.Pa-1	0.87	1
10e-14(cm3/m.s.Pa)	10e-14.cm3/(m.s.Pa)	0.89	1
10e-16cm3.cm/cm.cm2.s.Pa	(10e-16cm3.cm)/(cm2.s.Pa)	0.76	0.8
10e18 (mol.m/Pa.sec.m2)	10e18.mol.m/(Pa.sec.m2)	0.87	1
amol.m-1.s-1.Pa-1	amol.s-1.m-1.Pa-1	0.88	0.75
amol/m.s.Pa	amol/(m.s.Pa)	0.84	1
amol/m.sec.Pa	amol/(m.s.Pa)	0.69	0.75
cm3.um/m2.d.kPa	cm3.μm/(m2.d.kPa)	0.77	0.8

Table 3: Discovering new terms of units.

Class	Nb	Relevancy	Terms of units
Singular_Unit	9	55.5%	*mol, s, m, g, d*, etc.
Unit_Exponentiation	7	85.7%	m^{-2}, m^{-1}, d^{-1}, cm^3, s^{-1}, m^2, etc.
Unit_Division_or_Multi.	39	43.6%	*mol.L^{-1}*, *mol/L, mL/kg/h*, etc.
Unit_Multiple_or_Submulti.	4	50.0%	*amol, mL*, etc.
Total	59	50.8%	All previous units

Besides that, we have performed some extra experiments in order to know what kind of features are involved in the learning methods. The question asked is: are the terms of units referenced in the OTR the best features used in learning step? To answer this question, we have removed from the *bag-of-words* "unit" features and we have relaunched the learning algorithm DMNB in the same context. The recall and the precision drop slightly but remains quite honorable with respectively 0.81 and 0.76. This result shows that other features are involved in the learning process. We have then conducted a feature classification according to their different weight-based measures. The top k classification with k=10 shows that there are not only units but features like nouns, verbs, or adverbs describing a context where unit may appear (e.g. *water, temperature, RH for relative humidity, thickness, modified, solutions, samples, concentration, min, permeability, respectively*). We can say that units and other types of terms are closely related in the overall expression of an experimental data.

6.3 Extracting Variant Terms of Units and Discovering New Ones

From the results obtained in step 1, we conclude that the best location to work in, is the sentence where at least one unit appears. We apply the process according to different thresholds of the new similarity measure, adapted to the block and compare those experiments to the results obtained with the classical string distance measure as described in section 5.

In our experiments, we assume that a similarity measure between 0.2 and 0.4 is favourable for **discovering new terms of units**. Indeed, these values represent the case where at least one part of the blocks, a subset of units is recognized using the new string distance measure. Let us consider the following derived unit $g.m^{-1}.d^{-1}.Pa^{-1}$. This unit is referenced in the OTR and used in the process of recognition. In our approach, we consider 4 blocks g, m^{-1}, d^{-1} and Pa^{-1}. In the process of recognition, the base unit g, found in the text and not referenced in the OTR yet, is recognized compared to the derived unit with a similarity measure equals 0.2. The g block is identified. It is then relevant to propose this new base unit to be introduced in the OTR. Considering this assumption, we obtained 59 new possible units for the four classes of the *Unit_concept*, organized as described in Table 3.

The similarity measure between 0.5 and 1 is favourable for **identifying variant terms of units**. Indeed, the more the similarity measure is close to 1, the more terms of units are the same. We obtained 17 variant terms of units, we have 8 relevant variant terms that can be introduced in the OTR. The relevant variant terms of units are described in Table 4.

In this second step, we have successfully detected and extracted either relevant variant terms of units or new units. We can potentially enrich the OTR of 38 terms of units or 18% of enrichment (the actual OTR references 211 units). We have adapted a new string distance measure in order to compare variant and new terms of units to the ones referenced in the OTR. We compare blocks of units rather than characters, which proves to be more efficient in our issue. We have noticed that this process is quite relevant to identify new terms of units but is under our expectations concerning the identification of variant terms of units.

In our process of identifying blocks, we currently ignore all separators between those blocks, which are very specific in the case of units (e.g. /, ., (,), x). Those characters have semantic meanings that need to be added in the concern of extraction and identification, in order to improve the results of the similarity measure. We intend to carry out this work given these promising results. Another future work is completing the process with a new step in order to identify what

kind of terms of units are discovered (e.g. a temperature unit, a relative humidity unit, a permeability unit) and allow an automatic integration of variant or new units directly in the OTR.

7 CONCLUSIONS AND FUTURE WORK

The paper presents our overall two-step approach in order to identify variant terms of units or new units in unstructured scientific documents. We have successfully demonstrated how supervised learning methods can help reduce the search space of terms of units drowned in textual information. We have achieved almost 86% of reducing. For this purpose, we have used many algorithms and weight-based measures to prove the relevancy of our approach. We have also presented our first results to detect and extract variant terms and new units, using a new string distance measure adapted to our unit identification issue. As presented in the paper, this second step needs to be deepened. Our approach, presented in this paper, addresses the entire issue: location, extraction and identification of units. Further work on semantical identification of variant terms and new units should be addressed in order to automatically populate the OTR.

REFERENCES

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proc. of the Workshop on Information Integration on the Web (IIWeb-03)*, volume 47.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA*.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176.

Hawizy, L., Jessop, D., Adams, N., and Murray-Rust, P. (2011). ChemicalTagger: a tool for semantic text-mining in chemistry. *Journal of cheminformatics*, 3(1):17.

Hiemstra, D. (2000). A probabilistic justification for using tf x idf term weighting in information retrieval. *Int. J. on Digital Libraries*, 3(2):131–139.

Hignette, G., Buche, P., Couvert, O., Dibie-Barthélemy, J., Doussot, D., Haemmerlé, O., Mettler, E., and Soler, L. (2008). Semantic annotation of Web data applied to risk in food. *International Journal of Food Microbiology*, 128(1):174–180.

Jessop, D. M., Adams, S. E., and Murray-Rust, P. (2011a). Mining chemical information from open patents. *Journal of cheminformatics*, 3(1):40.

Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L., and Murray-Rust, P. (2011b). OSCAR4: a flexible architecture for chemical text-mining. *Journal of cheminformatics*, 3(1):1–12.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proc. of the conf. on Uncertainty in artificial intelligence*, pages 338–345.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments - part 1. *Inf. Process. Manage.*, 36(6):779–808.

Kohavi, R. and Quinlan, J. R. (2002). Data mining tasks and methods: Classification: decision-tree discovery. In *Handbook of data mining and knowledge discovery*, pages 267–276. Oxford University Press, Inc.

Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, volume 2473 of *LNCS*, pages 251–263. Springer.

Maynard, D., Li, Y., and Peters, W. (2008). Nlp techniques for term extraction and ontology population. In *Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, page 107127.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods*, pages 185–208. MIT Press.

Rijgersberg, H., van Assem, M., and Top, J. (2013). Ontology of units of measure and related concepts. *Semantic Web*.

Rijgersberg, H., Wigham, M., and Top, J. (2011). How semantics can improve engineering processes: A case of units of measure and quantities. *Advanced Engineering Informatics*, 25(2):276–287.

Su, J., Zhang, H., Ling, C. X., and Matwin, S. (2008). Discriminative parameter learning for bayesian networks. In *Proc. of the int. conf. on Machine learning*, pages 1016–1023.

Thompson, A. and Taylor, B. N. (2008). Guide for the use of the international system of units (SI).

Touhami, R., Buche, P., Dibie-Barthélemy, J., and Ibanescu, L. (2011). An ontological and terminological resource for n-ary relation annotation in web data tables. *On the Move to Meaningful Internet Systems: OTM 2011*, pages 662–679.

Van Assem, M., Rijgersberg, H., Wigham, M., and Top, J. (2010). Converting and annotating quantitative data tables. *The Semantic Web–ISWC 2010*, pages 16–31.

Willems, D. J., Rijgersberg, H., and Top, J. (2012). Identifying and extracting quantitative data in annotated text. *SWAIE*.

Wimalasuriya, D. C. and Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306323.