Automatic Validation of Terminology by Means of Formal Concept Analysis

Luis Felipe Melo Mora, Yannick Toussaint

Inria Nancy-Grand Est, BP 239 - 54506,
Villers-lès-Nancy, France
{luis-felipe.melo-mora, yannick.toussaint}@inria.fr

Abstract. Term extraction tools extract candidate terms and annotate their occurrences in the texts. However, not all these occurrences are terminological and, at present, this is still a very challenging issue to distinguish when a candidate term is really used with a terminological meaning. The validation of term annotations is presented as a bi-classification model that classifies each term occurrence as a terminological or non-terminological occurrence. A context-based hypothesis approach is applied to a training corpus: we assume that the words in the sentence which contains the studied occurrence can be used to build positive and negative hypotheses that are further used to classify undetermined examples. The method is applied and evaluated on a french corpus in the linguistic domain and we also mention some improvements suggested by a quantitative and qualitative evaluation.

1 Introduction

Terms in texts are important entities for any kind of document analysis: information retrieval, knowledge extraction or ontology building, etc. They are usually considered as linguistics entities that could be associated with meanings or concepts, their mirror at the ontological level [4]. However, there is no formal definition of what a term is, nor is there any reliable syntactic descriptions that could help term identification. Most of terms are noun phrases composed by a word or several words. Moreover, terms depend on the domain of speciality, and within a domain, terms are context-sensitive: a given string (a word or a set of words) may be a term in a given context with some meaning, a term in another context but with another meaning, or it could also be a non-term in a third context. Term extraction tools [17,18] extract candidate terms, i.e. groups of words that could be considered as terms. A candidate term fulfils linguistic (mainly syntactic schema) and/or statistic (based on occurrences) criteria. Once a candidate term is extracted by the tool, all its occurrences are annotated in the corpus. However, some of its occurrences correspond to a terminological use and some other correspond to a non-terminological use, *i.e.* these occurrences should be considered as words from general language. Thus, candidate terms and each of their occurrences should be manually validated by experts which makes it difficult for large-scale applications.

The paper presents how hypotheses built with formal concept analysis help to validate or invalidate candidate terms in texts of a specific domain. Some training data sets have been built on purpose but, as validation is time consuming, corpus are rather small and domain dependent. Thus, such a symbolic approach, based on itemset mining and classification, suits well the problem. In the longer term, linguists expect from this approach a better understanding of what term triggers are and how to find them.

The following examples with the candidate term `subject` remind us how ambiguous human language is: the same string may refer to different concepts. This is why term validation is so important for document indexing, automatic summarization, construction of ontologies and even for facilitating multilingual communication. The only help for meaning disambiguation is the context of the occurrences, *i.e.* the words that occurs with the term in the same sentence.

- (S_1) I **subject** him to a terrifying ordeal. <*Verb, non-term, general language*>.
- (S_2) This type of wound is highly **subject** to infection. <*Adjective, non-term, general language*>.
- (S_3) What is the **subject** in a sentence?<*Noun, term, linguistics*>.
- (S_4) Maths is not my best **subject**. <*Noun, term, pedagogy*>.
- (S_5)*A* moving picture of a train is more dramatic than a still picture of the same **subject**.<Noun, non-term, general language>.
- (S_6) The relation between the **subject** and predicate is identified by the use of: All, No, Some, ... <*Noun, term, logic*>.
- (S_7) The **subject of law** is a person (physical or juridical) who in law has the capacity to realize rights and juridical duties. <*Noun, complex term, law*>.

In the above examples, S_1, S_2, S_5 are contexts where the candidate `subject` is not a term, while S_3, S_4, S_6, S_7 are contexts where the candidate `subject` is a term in linguistics, in pedagogy, in logic or in law domains, respectively.

For each term candidate in a given domain, the goal is to validate or invalidate each of its occurrences. Each candidate term is studied separately and we propose a supervised learning method trained on a manually annotated corpus. For the learning phase, each occurrence of the candidate term is described by its textual context, *i.e.* the bag of the words of the sentence, and the occurrence is also tagged as "positive example" (belonging to the "T_+ class") if it is a terminological occurrence or as "negative example" ("T_- class") if it is not. Thus, from textual context our method extracts hypotheses, a notion that is formally introduced in the next section. Hypotheses are itemsets of words corresponding to the positive occurrences of a candidate term and, similarly, itemsets corresponding to negative occurrences of a candidate term. Then, during the test phase, a new occurrence of this candidate term in a new sentence is classified either as terminological occurrence or non-terminological occurrence according to the hypotheses that match the sentence.

The learning problem can be formulated in the paradigm of Formal Concept Analysis (FCA) [8], a formal method where ordered sets are classified in a lattice. FCA builds a bi-classification model from positive and negative examples.

In the binary matrix associated to each candidate term, the objects are occurrences of the candidate term. Attributes are words coming from the different contexts and a positive/negative flag is introduced in accordance with to the manual annotation. Hypotheses [13,12] are generalised descriptions of positive or negative examples. These itemsets are non-redundant descriptions of either the positive class or the negative class. There is a high demand from linguists for such human-readable sets that could be considered as triggers and distinguish terminological occurrences from non-terminological ones. Moreover, hypotheses are applied to new (unannotated) occurrences of a candidate term to discover its terminological or non-terminological nature in new texts.

The paper is organized as follows. Section 2 provides a brief overview of the problem of validating term occurrences. Section 3 introduces Formal Concept Analysis and its application to learning problems. Then, Section 4 describes how positive and negative hypotheses can be applied to textual contexts of term occurrences in order to validate or invalidate them as a terms. In Section 5, we describe the dataset *i.e.* the corpus, the experiments and their results. Then, Section 6 concludes the paper.

2 Terminology Extraction

Eugen Wüster [19] emphasized on the role of terms, their link with concepts, and the importance of normalization of terms to avoid ambiguity, to ease indexing, thesaurus building or translation. He was *the* author who defined the general theory of terminology and worked within a standardization perspective. At that time, terminology was initially a prerogative of translators, with a rather normative approach.

However, in the 90's, the renewal of corpus linguistics with some new robust tools such as part of speech taggers or syntactic parsers showed that terms are not restricted to set phrases in a previously defined list but they are full linguistic entities whose form may vary in the texts (plural forms, modifiers, etc.). New software applications in information retrieval, summarizing, or ontology construction stimulate this new conception of terminology. Thus, there has been several initiatives for developing term extractors. Among them some are term locators: they locate in texts terms belonging to a controlled vocabulary [1,9]. Some others are working *ab-nihilo*, looking for candidate terms [17,18,2,7].

Thereby, term extraction consists in a set of computational techniques that allow to identify the linguistic realizations of domain-specific concepts known as terms. Frequently it is seen as an intermediate phase of Natural Language Processing, that bridges the gap with the knowledge level and enables different kind of reasoning. Few term extractors use only statistics on word occurrences and co-occurrences to propose term candidates. Most of them are now combining linguistic rules and statistical filters [17]. However, despite this configuration, there is still a lot of noise both in candidate term identification and in the distinction between their terminological and non-terminological occurrences.

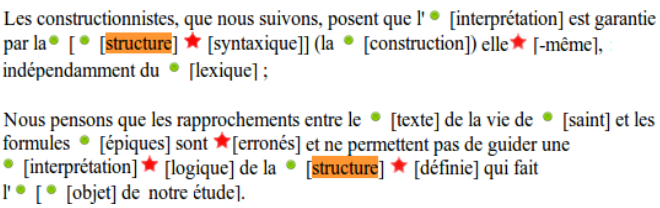

Fig. 1. Chunks of texts where candidate terms (simple and multi-words) are located with TTC Term Suite and represented by square brackets [ ]. The green dots indicate validated candidates (terms), whereas the red stars define candidate terms refused by the experts.

For instance, in Figure 1, extracted candidate terms are represented between square brackets. Some candidate terms include some others (nested brackets). Thus, `structure syntaxique` (*syntactic structure*) is proposed as a candidate term while `structure definie` (*defined structure*) is not. It should be noted that both elements have the same grammatical structure and in both sentences `structure` is also proposed as a candidate term. Occurrences which are marked by a green dot have been manually validated as terminological occurrences while non-terminological occurrences are marked by red stars.

In the next section, we introduce Formal Concept Analysis and its use for bi-classification of term candidate occurrences.

3 Formal Concept Analysis (FCA)-based Method

The main notions of Formal Concept Analysis (FCA) theory are introduced in this section. Afterwards, concept-based hypotheses (called also JSM-hypotheses from the John Stuart Mill method) are presented as a method for building a bi-classification model from positive and negative examples. Similarly, Jumping Emerging Patterns (JEPs) is an alternative formalism to identify the set of discriminating attributes which only occur in one class and are absent in the other.

3.1 Bases on FCA

FCA is a data analysis theory which builds conceptual structures defined by means of the attributes shared by objects. Formally, this theory is based on the triple $K = (G, M, I)$ called *formal context*, where G is a set of objects, M is a set of attributes and I is the the binary relation $I \subseteq G \times M$ between objects and attributes. Therefore, *(g, m)* $\in I$ means that g has the attribute m. For instance, some occurrences of the introductory examples with the candidate term `subject` are encoded in the formal context given by Table 1.

Two derivation operators are then defined:

$$A' := \{m \in M \mid \forall g \in A : gIm\} for A \subseteq G,$$
$$B' := \{g \in G \mid \forall m \in B : gIm\} for B \subseteq M.$$

Table 1. An example of a formal context where each row represents an occurrence of the candidate term `subject` with the words appearing in its textual context.

	I	a	this	subject	him	to	terrifying	ordeal	type	of	wound	what	is	the	in	sentence	highly	infection
***subject*(S_1)**	x			x	x	x	x	x										
***subject*(S_2)**			x	x		x			x	x	x		x				x	x
***subject*(S_3)**		x		x								x	x	x	x	x		

A *formal concept* is a pair *(A, B)*, satisfying $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. A is called the *extent* and B the *intent* of the (formal) concept. " is a closure operator which means that for any X, Y, $X'''' = X''$ (idempotent), $X'' \subseteq X$ (extensive), and $X \subseteq Y \longrightarrow X'' \subseteq Y''$ (monotone). Thus, the intent of a concept is the maximum set of attributes shared by all the objects of its extent. Moreover, an itemset $X \subseteq M$ is a *generator* of a formal concept *(A, B)*, if $X \subseteq B$ and $X' = A$. Likewise, a *minimal generator* for a concept is defined as a minimal subset of its intent which can similarly characterize the concept in question.

Formal concepts are organized into a complete *concept lattice* denoted by $\mathcal{L}$ following a partial ordering, called subsumption, ($\sqsubseteq$), defined as follows: $(A_1, B_1) \sqsubseteq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2$ (*or* $B_2 \subseteq B_1$).

3.2 Classification by FCA

A learning model from a concept lattice has been extensively studied through the notion of *concept-based hypothesis* [13,12]. This model is based on positive and negative examples of a target attribute. The idea laying beyond this model is to discover the attribute combinations which are shared by positive examples, but not by negative examples.

Let us consider the target attribute $w \notin M$, which may have one of the three values: positive, negative and undetermined. Thereby, the input data for learning is composed by sets of positive and negative examples. Positive examples are objects that are known to have the target attribute and negative examples are objects that are known not to have this attribute. The learning results are rules supposed to classify a third set of objects called the undetermined examples.

With regard to FCA theory, this classification method can be described by three sub-contexts : a positive context $K_+ = (G_+, M, I_+)$, a negative context $K_- = (G_-, M, I_-)$ and an undetermined context $K_\tau = (G_\tau, M, I_\tau)$. M is a set of attributes, w is the target attribute and $w \notin M$, G_+ is the set of positive examples whereas G_- is the set of negative examples. Alternatively, G_τ denotes the set of new examples to be classified. The *learning context* is denoted by $K_\pm = (G_+ \cup G_-, M \cup w, I_+ \cup I_- \cup G_+ \times \{w\})$. In addition, $K_c = (G_+ \cup G_- \cup G_\tau, M \cup w, I_+ \cup I_- \cup I_\tau \cup G_+ \times \{w\})$ is called the *classification context*.

For generalizing the G_+ subset and defining the cause of target attribute, we are interested in finding the sets of attributes that are shared by only positive ex-

amples. In the best case, the membership to G_+ supposes a particular attribute combination. However, in most cases it is necessary to find several attribute combinations called *positive hypotheses* to characterize only G_+ examples. Ideally, we would like to find enough positive hypotheses to cover all G_+ examples.

A positive hypothesis H_+ for w is defined as a non empty intent of K_+ which is not contained in the intent g' of any negative example $g \in G_-$. A *negative hypothesis* H_-, is defined accordingly.

Thereby, hypotheses can be used to classify an undetermined example $x \in G_r$. If the intent of x contains at least one positive hypothesis and no negative hypothesis, then, x is classified as a positive example. If the intent of x contains at least one negative hypothesis and no positive hypothesis, then it is a negative example. Otherwise, x remains unclassified.

In addition, we can restrict the number of useful hypotheses with regard to subsumption in the lattice. Formally, a positive hypothesis H_+ is a *minimal positive hypothesis* if there is no positive hypothesis H such that $H \subset H_+$. *Minimal negative hypothesis* is defined similarly. Hypotheses which are not minimal should not be considered for classification because they do not improve discrimination between positive and negative examples.

In a not-so-far context of itemset mining, the notion of Jumping Emergent Pattern (JEP) is very similar to concept-based hypothesis [5]. A JEP is an itemset that occurs only in objects of one class and not in objects of the other class. Clearly, a hypothesis is a JEP. On the other hand, a JEP is a generator of some hypotheses for this class. We can also define a *minimal JEP* as an itemset that does not contain any other JEP. Consequently, searching the minimal JEPs is equivalent to finding the minimal generators of the concept-based hypotheses for a class. Other important class of patterns that represent a contrast between the classes are exposed in [16]. For instance, an *emerging pattern(EP)* is an itemset whose frequency changes significantly from one data set (G_+ for example) to another (G_- respectively). Similarly, the *constrained emerging patterns (CEPs)* are defined as the minimal set of items which occur at most α times in one data set and at least β in the other. Unlike concept-based hypotheses or JEPs, EPs and CEPs are potentially more resistant to noise because they are less restrictive patterns [16].

3.3 Relevant Hypotheses

A concept-based hypothesis generalizes a class of positive or a class of negative examples. Each hypothesis is a closed itemset, ı.e. the intent of a concept. Nevertheless, because of the noise in the data, these hypotheses are not all relevant. *Stability* [14] is a measure that qualifies the tendency of a concept (and its intent) to persist when some objects are randomly removed from its extent. Thereby, stability measures how much a concept depends on each particular object of its extent. As a consequence, a stable concept will be independent of data noise.

Thus, it may happen that some minimal hypotheses have a low stability value. In that case, the intent of subsumed concepts $H_1, ..., H_n \supset H$, which are hypotheses but non-minimal, may have a higher stability value. These hypotheses

are more restrictive when applied to undetermined object classification and the precision of the overall system could be improved. However, such a strategy may reduce the coverage of the positive (resp. negative) examples by the set of hypotheses, with a possible deterioration of the system recall.

Other measures have been proposed in [11] to recognise relevant concepts in noisy data. Among these measures are the support, concept probability or separation index which can be useful in different kinds of contexts. However this comparison concludes that stability is the most effective and the less independent of the type of the context. Stability is the only measure we kept to evaluate hypotheses relevance in our study.

Accordingly, we adopted a FCA classification model to identify the patterns that represent the largest shared textual contexts from the occurrences of a term on a specific domain. In the following section, we present practical aspects and some other considerations for our method.

4 Term Validation as a bi-classification Problem

In order to minimize the human intervention and to improve the terminology validation scalability, Formal Concept Analysis (FCA) can be used for learning the hypotheses from positive and negative examples. As shown in [15], the textual context is the key for validating terminological occurrences. So, we assume that the textual context around each candidate occurrence gives us relevant information on its class. For a given domain and a given candidate term, we thus focus on the differentiation between a terminological use (T_+) and a non-terminological use (T_-). If the candidate term is multi-words, words are joined together. To build the context of a candidate term occurrence, its (textual) context, i.e. the sentence, is represented as a set of words $S_i = \{W_1, W_2, W_3, ...W_n\}$. Table 2 illustrates how occurrences of the candidate term `subject` in the linguistic domain is encoded as a formal context.

Table 2. Part of the `subject` formal context where each occurrence is defined by its textual context. The target attributes T_+ and T_- show the terminological nature of the occurrence in the linguistic domain.

	I	a	this	subject	him	to	terrifying	ordeal	type	of	wound	what	is	the	in	sentence	highly	infection	T_	T+
***subject*(S_1)**	x			x	x	x	x	x											x	
***subject*(S_2)**			x	x		x			x	x	x		x				x	x	x	
***subject*(S_3)**		x		x								x	x	x	x	x				x

The lattice is built according to the formal context and then, positive and negative hypotheses are extracted. The preliminary results show that noise in data significantly reduces the quality of the results and increases drastically the

size of the lattice. The next section is dedicated to noise reduction in the original data.

4.1 Reducing noise in the learning process

In order to reduce the noise in data, we assume that some words in the textual context are more relevant than others. Such words should show an intrinsic semantic. Likewise, function words which semantics depends on the words they govern and the words they are governed by loose their semantics when placed within an unordered bag of words. Therefore, these function words are removed. Similarly, as we have a rather small number of examples for each term candidate, words are lemmatized to tackle the different forms of a word and reduce dispersion. A lemma is the canonical form shared by a set of words expressing the same meaning. For example, walk is the lemma of walking, walks and walked.

After several experimentations, the most relevant configuration to reduce the formal context for a candidate term is the following:

- The set of objects G: Each occurrence of the studied candidate;
- The set of attributes M: Lemmas of content words(nouns, verbs, adjectives and adverbs) for each textual context (*i.e.* the sentence) where the candidate term occurs;
- The binary relation I: It sets which lemma co-occurs with which candidate term occurrence.
- The target attributes (T_+ and T_-): Corresponding to the manual annotation in the corpus.

An example of such formal context is shown in the Table 3.

Table 3. An excerpt of the formal context with lemmas of content words for the candidate term `subject` and its class (T_+ or T_-).

	subject(cT)	terrify	ordeal	type	wound	be	sentence	highly	infection	T_	T+
***subject*(S_1)**	x	x	x							x	
***subject*(S_2)**	x			x	x	x		x	x	x	
***subject*(S_3)**	x					x	x				x

5 Experiments

The experiments and evaluations of our method aim at demonstrating the quality and interest of extracted hypotheses as well as helping linguists in defining

new features, *i.e.* new annotations, to improve term validation. In any corpus, there exist candidate terms whose occurrences are almost always terminological, some other candidate terms whose occurrences are mostly non-terminological and some other with a rather balanced distribution between the two classes as shown by the column *category* of Table 4. To ease reading, tables presented in this section are translated from French.

Table 4. Selected candidates for evaluation and values observed in the whole corpus.

Candidate	Frequency	Positive Examples	Terminological Degree	Category
Adjective	216	207	95.83%	highly terminological
Lexical relation	55	52	94.54%	highly terminological
Collocation	109	90	82.56%	highly terminological
Sentence	311	238	76.52%	enough terminological
Speaker	233	178	76.39%	enough terminological
Corpus	688	510	74.12%	enough terminological
Language	926	549	59.28%	ambiguous
Statement	289	164	56.74%	ambiguous
Context	302	147	48.67%	ambiguous
Text	568	266	46.83%	ambiguous
Speech	534	248	46.44%	ambiguous
Form	462	122	26.40%	slightly terminological
Relation	676	171	25.29%	slightly terminological
Expression	197	48	24.36%	slightly terminological
Semantic	413	80	19.37%	very slightly terminological
Lexical	477	84	17.61%	very slightly terminological
Model	250	13	5.20%	very slightly terminological

5.1 Dataset

The training corpus is composed of 60 free ScienceText documents in french from the linguistics domain. This corpus has been automatically enriched with different annotations: tokenization, sentence splitting and part-of-speech tagging (PoS) performed by the TreeTagger. For normalization issues, an XML-based format has been defined and applied to the documents. Subsequently, TTC Term Suite (Terminology Extraction, Translation Tools and Comparable Corpora project) extracted 5,038 different candidate terms and 69,007 occurrences of them. Finally, each occurrence of candidate terms was manually validated thanks to a dedicated annotation interface[1].

[1] Smarties: The annotation interface by stickers (https://apps.atilf.fr/smarties/, last visit 01.04.15)

Two annotators evaluated each occurrence of a candidate term considering different linguistics aspects: syntagmatics considerations, membership to a scientific lexicon, membership to a linguistic lexicon and terminological nature. For each of these aspects, experts assign a class (positive or negative) to each occurrence as shown in [10]. To perform cross-validation, this corpus has been split into several parts for training and then, for classification of undetermined examples (testing).

In order to achieve a reliable evaluation of experiments, we selected a list of candidate terms which occur frequently and that belong to different categories as show in Table 4: `adjectif` (*adjective*), `relation lexical` (*lexical relation*), `collocation`, `phrase` (*sentence*), `locuteur` (*speaker*), `corpus`, `langue` (*language*), `énnoncé` (*statement*), `contexte` (*context*), `texte` (*text*), `discours` (*speech*), `forme` (*form*), `relation`, `expression`, `sémantique` (*semantic*), `lexical`, `modèle` (*model*). We also introduce a measure of the *terminological degree* of each candidate term (named *ambiguity rate* in [3]). This measure gives the ratio between the number of positive examples (terminological occurrences) of the candidate term with regards to all of its occurrences.

5.2 Implementation

For each experiment, a formal context is generated candidate per candidate. Attributes are the lemmas of content words (verbs, adverbs, nouns and adjectives) that co-occur with the candidate term in the same sentence.

Afterwards, concept-based hypotheses are extracted by means of Formal Concept Analysis to build a set of positive hypotheses (for terminological occurrences) and a set of negative hypotheses (for non-terminological occurrences).

To extract hypotheses, we developed a pipeline within the GATE Natural Language Engineering platform GATE [6]. This pipeline uses several plugins that deal with the specific XML-based format to represent a formal context and extract hypotheses. During the evaluation phase, these hypotheses are matched with sentences in the testing dataset in order to classify undetermined examples.

Table 5. Classification summary of candidate occurrences.

Word	**Frequency**	**Positive Examples**	**Terminolog. Degree**	**Words used only in T_+**	**Hypotheses Generated in T_+**	**Proportion of Positive Hypotheses**	**Shared Words**	**Negative Examples**	**Words used only in T_-**	**Hypotheses Generated in T_-**
Adjective	216	207	95.83%	966	301	97,41%	64	9	59	8
Corpus	688	510	74.12%	1035	1347	81,93%	713	178	535	297
Text	568	266	46.83%	735	913	52,32%	772	302	792	832
Relation	676	171	25.29%	159	183	11,48%	629	505	1427	1410
Semantic	413	80	19.37%	272	108	8,88%	560	333	1258	1107

Table 5 presents a summary of the results obtained on the whole corpus for some candidate terms selected among the different categories. We observed that certain words are shared by textual contexts of both positive and negative classes (`Shared Words`). The more ambiguous or frequent the candidate is, the bigger is the shared set. We also remark that the proportion of positive hypotheses with regards to the global number of hypotheses (positive and negative) is quite similar to the ratio of positive examples with regards to the whole set of examples (*i.e.* the `Terminological Degree`).

5.3 Results

This section presents our experimental results. Evaluation aims at measuring how good are hypotheses for classification of undetermined examples. We used a k-fold cross-validation over our annotated ScienceText corpus (partitioned in 8 folds with a length per fold of 7 texts). Thus, for each experiment, annotations of candidate terms in 7 texts were removed and texts were used for testing; the rest was used for training.

Table 6 shows average values over the different runs. The `Ex2Cla` value is the number of undetermined examples to classify. Generated hypotheses is the number of hypotheses extracted from a training set. Accordingly, projected hypotheses is the number of hypotheses that matched undetermined examples. Positive (resp. neg.) unclassified examples are undetermined examples (know as being positive (resp. neg.) in the corpus) that do not contain any positive or negative hypothesis and thus, they have not been classified.

As could be expected, the amount of positive hypotheses is greater than the negative hypotheses if the candidate tends to have a terminological nature. Conversely, the number of negative hypotheses is greater than the positive hypotheses if the candidate tends to be not terminological. However, the ambiguous candidates contain a similar amount of positives and negatives hypotheses. The cause of this behaviour is related to the number of positive and negative occurrences (frequency) of each candidate by category.

The number of hypotheses (projected) used to classify examples is greater than the number of undetermined examples but the proportion between these two values varies a lot. Candidates at the top or at the bottom of the table have good results with a low number of unclassified examples. However, `corpus`, which is frequent, enough terminological, and with a very high number of positive hypotheses has a high number of unclassified positive examples. Thus, a high number of training examples does not always seem to garantee a better result. Candidate terms which are ambiguous are, of course, the most difficult to classify. Here again, one candidate term, `language`, seems apart : it is very frequent, generated lot of (+/-) hypotheses, but the number of unclassified examples (positive or negative) is high.

Table 7 gives the average of some performance measures (precision, recall and F-measure) over the 8 runs. In general, if a class (positive or negative) has a high number of training occurences, then this class gets a better precision and

Table 6. Average of kept hypotheses and unnamed examples in k-fold cross-validation ($k = 8$).

Candidate	Cat.	Freq.	Ex2Cla	Hypotheses				Unclassified	
				Generated (+)	Projected (+)	Generated (-)	Projected (-)	Pos. Examp.	Neg. Examp.
Adjective	highly term.	216	27	263	46	7	0	1.375	0.125
Lexical relation	highly term.	55	4.71	58	3	1	0	3.14	0
Collocation	highly term.	109	13.62	132	10	16	0	3.5	0.875
Sentence	enough term.	311	38.87	394	94	77	14	4	1.875
Speaker	enough term.	233	29.12	397	96	64	10	2.875	0.25
Corpus	enough term.	688	86	1126	268	249	61	25.5	5.25
Language	ambig.	926	115,75	1201	418	617	231	21.5	13.125
Statement	ambig.	289	36.12	309	51	146	18	5.5	2.125
Context	ambig.	302	37.75	261	76	326	74	4.5	4.5
Text	ambig.	568	71	757	194	706	145	7.625	5.75
Speech	ambig.	534	66.75	396	80	535	83	5.75	6.125
Form	slightly term.	462	57.75	134	33	955	333	1.125	4.875
Relation	slightly term.	676	142,25	160	16	1172	244	1.5	9.5
Expression	slightly term.	197	24.62	44	5	275	67	1.25	3.0
Semantic	very slightly term.	413	51.62	92	20	915	288	0.5	5.5
Lexical	very slightly term.	477	59.62	77	12	950	312	0.375	2.625
Model	very slightly term.	250	31.25	8	0	384	80	0	1

recall. On the opposite, the coverage of the training examples by hypotheses does not seem to impact precision and recal.

5.4 Qualitative Analysis

The second goal of this study is to help linguists to better understand what are the mechanisms that take part to the decision on the terminological status of an occurrence. The ideal process would be when validating occurrences of candidate terms is independent of the term candidate or, even better, when it is independent of the domain. To reach such a goal, we should identify new features that should be added to the initial annotation set. We still are far from reaching the goal but the qualitative analysis already helps us in interacting with linguists.

We carried out a qualitative analysis of patterns. We give here the way patterns are analysed looking abitrarily at positive and negative patterns for the candidate term `argument`. `argument` has 92 occurrences in the corpus and 66.30% of them are positives (classify between the "enough terminological" and

Table 7. K-fold cross-validation over the collected ScienceText corpus ($k = 8$).

Candidate	Cat.	Freq.	Positive Examples covered by H_+	Precision in T_+	Recall in T_+	F1 in T_+	Negative Examples covered by H_-	Precision in T_-	Recall in T_-	F1 in T_-
Adjective	highly term.	216	99.57%	0.8229	0.8452	0.8339	100%	0.0	0.0	0.0
Lexical relation	highly term.	55	100%	0.2689	0.1820	0.2170	100%	0.0	0.0	0.0
Collocation	highly term.	109	96.11%	0.4864	0.4194	0.4504	95,39%	0.0	0.0	0.0
Sentence	enough term.	311	97.05%	0.8811	0.8828	0.8819	90.41%	0.1968	0.6	0.2963
Speaker	enough term.	233	98.52%	0.8436	0.7729	0.8067	95.22%	0.4375	0.3779	0.4055
Corpus	enough term.	688	78.11%	0.7087	0.3955	0.5076	84.05%	0.6626	0.4227	0.5161
Language	ambig.	926	82.60%	0.7879	0.590	0.6747	84.15%	0.7739	0.5256	0.6260
Statement	ambig.	289	87.19%	0.6572	0.5683	0.6095	85.30%	0.5177	0.3134	0.3904
Context	ambig.	302	98.21%	0.7649	0.5936	0.6684	98.30%	0.6897	0.5655	0.6214
Text	ambig.	568	97.03%	0.5675	0.4355	0.4928	96.77%	0.5353	0.4819	0.5071
Speech	ambig.	534	88.70%	0.7523	0.5108	0.6084	91.43%	0.4846	0.5220	0.5026
Form	slightly term.	462	94.97%	0.75	0.1124	0.1955	98.19%	0.8828	0.7252	0.7962
Relation	slightly term.	676	74.92%	0.25	0.0535	0.0881	91.70%	0.8337	0.8735	0.8531
Expression	slightly term.	197	94.53%	0.3125	0.0631	0.1049	97.06%	0.7490	0.8437	0.7935
Semantic	very slightly term.	413	90.93%	0.4583	0.1107	0.1783	97.86%	0.7735	0.8761	0.8216
Lexical	very slightly term.	477	89.58%	0.0	0.0	0.0	97.10%	0.8791	0.9408	0.9089
Model	very slightly term.	250	100%	0.0	0.0	0.0	100%	0.8780	0.9660	0.9199

the "ambiguous" category), our method generates 48 positives and 40 negatives hypotheses. Tables 8 and 9 show positive hypotheses (resp. negative) ranked following support and stability.

The most stable positive hypotheses include, in addition to the candidate term itself, a "high" terminological term in linguistics `sdrt` (which stands for Segmented Discourse Representation Theory) and the meaningless verb `be`. There is no doubt that `sdrt`, a linguistic theory which study relation between arguments in a discourse, is a very good trigger for positive occurences. Afterwards, some other high terminological terms `syntactic` or `verbal` also contribute to a positive validation. Others hypotheses with a lower support, but not less important, are related to Rhetorical Structure Theory (*rst*) representing the distinction between `nucleus` and satellite `arguments`.

However, we should notice that it is quite easy to find counter-examples. Considering the hypothesis `[sdrt, be, argument]` and the sentence *"An argument in favor of SDRT is also that ..."* (which is not in the initial corpus), the `ar-`

Table 8. Set of the most representative positive hypotheses for the `argument` candidate term.

Support	Stability	Hypotheses in T_+	Hypotheses in T_+ *-english-*
7	0.7968	[sdrt, être, argument]	[sdrt, be, argument]
9	0.7792	[argument, plus]	[argument, more]
6	0.73437	[être, argument, aussi]	[be, argument, also]
6	0.7187	[argument, verbal]	[argument, verbal]
...	...	...	...
5	0.6562	[être, argument, indique]	[be, argument, denote]
4	0.5	[argument, syntaxique]	[argument, syntactic]
...	...	...	...
6	0.3281	[être, argument, rst]	[be, argument, rst]
8	0.25	[argument, nucleus]	[argument, nucleus]

`gument` candidate term will be wrongly classified as positive. Similarly, the third positive hypothesis in the table, `[be, argument, also]`, could fit the negative example *"An argument in favor of SDRT is also that . . . "* and, of course, it could also fit the positive example like *" This relationship, which raises issues concerning the linear order of its arguments is studied in (Redeker and Egg, 2006) and also in (Hunter et al., 2006) . . . . "* . The two last examples show that some additional information could probably produce better hypotheses: preserving order in the sentence (working with sequences instead of bag of words), using syntactic role (subject, object . . .), syntactic dependencies between the studied occurrence and some other words, or keeping information about the type of determiner it is linked with, like definite (ex: *the*) or indefinite (*a*). . .

Table 9. Set of the most representative negative hypotheses for the *argument* candidate term.

Support	Stability	Hypotheses in T_-	Hypotheses in T_- *-english-*
3	0.5	[argument, prendre]	[argument, assume]
1	0.5	[trancher, pas, ne, argument, permettre, décisif, position, avoir]	[settle, not, argument, allow, decisive, position, have]
...	...	...	...
4	0.375	[argument, hypothèse]	[argument, hypothesis]
4	0.3125	[dire, argument]	[say, argument]
...	...	...	...
2	0.25	[trouver, même, argument]	[find, same, argument]

By contrast, the sets of negative hypotheses showed in Table 9 showed another usage of the `argument` candidate. Mainly, `argument` refers to authors trying

to convince the reader about an idea, an hypothesis or a theory by using an evidence. Consequently, the large diversity of situations leads to hypotheses which include meaningless words like `dire` (*to say*), `prendre` (*to assume*), `trouver` (*to find*).

6 Conclusion and perspectives

In this paper, we describe a method for validating occurrences of candidate terms using Context-based Hypotheses. It starts with a corpus on a specific domain, where each occurrence of candidate terms has been manually annotated as terminological or non-terminological occurrence. We built a formal context for hypotheses extraction. Each positive hypothesis represents a textual context where the candidate is used as a term. Similarly, a negative hypothesis describes the textual context where the candidate is used as a non-terminological entity.

Some plugins have been developed to run under the GATE the open source solution for text processing. In that way, some higher-level linguistic annotations could be used to improve the process. Among them we could mention syntactic trees, dependencies or the use of linguistic resources such as a transdisciplinary lexicon. As mentioned in Section 5.4, we have several options to improve annotations and better discriminate positive and negative occurrences defining hypotheses which are not only based on words (lexical level) but also on more elaborated linguistic features.

We would like to thank the ANR Agency for supporting this work which is part of the Termith project (ANR-12-CORD-0029-05) and we would like to thank all the linguists involved in the project for the work they did on the corpus.

References

1. Aronson, A., Lang, F.M.: An overview of metamap: historical perspective and recent advances. JAMIA 17(3), 229–236 (2010)
2. Aubin, S., Hamon, T.: Improving term extraction with terminological resources. In: Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (eds.) Advances in Natural Language Processing, pp. 380–387. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2006)
3. Boumedyen, M., Camacho, J., Jacquey, E., Kister, L.: Annotation sémantique et validation terminologique en texte intégral en shs. In: Actes de la 21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN'2014). Marseille, France (2014)
4. Bourigault, D., Jacquemin, C., L'Homme, M.: Searching for and identifyng conceptual relationships via a corpus-based approach to a terminological knowledge base(ctkb): Method and results. In: Condamines, A., Rebeyrolle, J. (eds.) Recent Advances in Computational Terminology, chap. 6. Natural language processing, J. Benjamins Publishing Company (2001)
5. Buzmakov, A., Kuznetsov, S., Napoli, A.: A new approach to classification by means of jumping emerging patterns. In: in "FCA4AI: International Workshop "What can FCA do for Artificial Intelligence?" - ECAI 2012 (2012)

6. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M., Saggion, H., Petrak, J., Li, Y., Peters, W.: Text Processing with GATE (Version 6) (2011), http://tinyurl.com/gatebook
7. David, S., Plante, P.: Termino version 1.0. Rapport du Centre dAnalyse de Textes par Ordinateur. Université du Québec à Montréal (1990)
8. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st edn. (1997)
9. Jacquemin, C.: Fastr : A unification-based front-end to automatic indexing. In: Funck-Brentano, J.L., Seitz, F. (eds.) RIAO. pp. 34–48. CID (1994)
10. Kister, L., Jacquey, E.: Relations syntaxiques entre lexiques terminologique et transdisciplinaire : analyse en texte intégral. In: In Actes du Congrès Mondial de Linguistique Franaise. pp. 909–919. Lyon, France (2012)
11. Klimushkin, M., Obiedkov, S., Roth, C.: Approaches to the selection of relevant concepts in the case of noisy data. In: Kwuida, L., Sertkaya, B. (eds.) Formal Concept Analysis, Lecture Notes in Computer Science, vol. 5986, pp. 255–266. Springer Berlin Heidelberg (2010)
12. Kuznetsov, S.: Machine learning on the basis of formal concept analysis. Autom. Remote Control 62(10), 1543–1564 (2001)
13. Kuznetsov, S.: Complexity of learning in concept lattices from positive and negative examples. Discrete Applied Mathematics 142(13), 111 – 125 (2004)
14. Kuznetsov, S.: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence 49(1-4), 101–115 (2007)
15. Maynard, D., Ananiadou, S.: Term extraction using a similarity-based approach. In: In Recent Advances in Computational Terminology. John Benjamins. pp. 261–278 (1999)
16. Ramamohanarao, K., Bailey, J.: Discovery of emerging patterns and their use in classification. In: Gedeon, T., Fung, L. (eds.) AI 2003: Advances in Artificial Intelligence, Lecture Notes in Computer Science, vol. 2903, pp. 1–11. Springer Berlin Heidelberg (2003)
17. Rocheteau, J., Daille, B.: Ttc termsuite: A uima application for multilingual terminology extraction from comparable corpora. In: Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP). Chiang Mai, Thailand (2011)
18. Sclano, F., Velardi, P.: Termextractor: a web application to learn the shared terminology of emergent web communities. In: Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007) (2007)
19. Wüster E., E.d.M.E.: La théorie générale de la terminologie un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et les sciences des objets. In: Actes du colloque international de terminologie (Québec, Manoir du lac Delage, 5-8 octobre 1975) (1976)