Detecting metaphors on domain-specific corpora

Stéphane Ferrari
LIMSI-CNRS, BP 133, F-91403 Orsay cedex, FRANCE
email: ferrari@limsi.fr

Abstract

In this paper, we propose a method for detecting metaphors on large domain-specific corpora, in order to provide NLU systems with probabilities for non-literal meanings. Our approach is based on the conjoint use of textual clues and heuristics concerning the frequencies of words. It comes in opposition to the classical approaches which used semantic information only for detecting and processing metaphors. The textual clues we found were firstly analyzed by hand on a small explanative corpus. Then they were generalized. They are now evaluated on a large electronic corpus. They are essentially lexical markers combined with syntactic regularities. We show how to theoretically improve metaphor detection by merging the information given by these textual clues with heuristics concerning frequencies of words in a domain-specific corpus. A prototype, STK, implementing this method for French, is currently under develoment. It integrates Brill's rule-based tagger. We describe the practical restrictions that constrain the method, as well as some encouraging experimental results.

1 Introduction

A metaphor can be considered as a commonly used meaning production process. Classical NLU (Natural Language Understanding) approaches to metaphor reveal at least difficulties in detecting the figure that bears the metaphor. As we show in a brief overview, in section 2, p. 2, the detection process, when existing, is here restricted to the failure of a semantic analysis. In opposition, other approaches, proposing multiple semantic analysis at the same time, led to complex word sense disambiguations. No proposition was made for detecting metaphors in texts before processing any semantic analysis. Therefore, no efficient method was proposed to avoid misunderstanding of ambiguous words when dealing with large corpora. Our aim is to provide NLU systems with information about where non-literal meanings are bound to happen, and with what probability, in order to improve a future semantic analysis.

Recent works proved the existence of textual clues related to specific linguistic phenomena, depending on the domain or the style, and especially concerning explanations. This motivated a first corpus analysis. We collected 26 explanatory texts, in French, of about 200 words each. They proved the existence of textual clues related to metaphors and analogies, exposed in section 3, p. 3. The textual clues we found are essentially lexical markers combined with syntactic regularities. They can be used for determining the words involved in the source and in the target of a possible metaphor. We expose how they can be generalized and we propose a semi-automatic method for evaluating their relevance as markers for metaphors, which results in a set of probabilities for the presence of metaphoric meanings.

Textual clues are obviously not sufficient for detecting all possible metaphors. Therefore, other tools must be added if we ought to obtain a robust detection. The conventional view of metaphors, previously exposed in section 2, which focusses on knowledge about common metaphors, may be extended. In section 4, p.5, we propose a method for automatically extracting lexica of commonly used terms from domain-specific corpora and thus determining in a text where and what metaphors are bound to happen : common ones or novel ones. We show how to exploit such an information, depending on the application.

Lastly, we propose, in section 5, p. 6, a hybrid method to help the detection, and, by this mean, the interpretation of metaphors. The general method is based on the previous results, combining the use of a syntactic parsing for retrieving the information provided by the textual clues, and heuristics concerning words frequencies in order to detect *possible* metaphors and to evaluate the probability of a *non-conventional* interpretation. STK, a prototype implementing the method, is currently under development. STK integrates the Brill's rule-based part-of-speech tagger, which has been trained on French language for this purpose. STK provides usual NLU tools with a new set of tags underlying the possibly metaphoric meanings, giving information about the probability for a non-conventional interpretation, as well as a list of domain-specific words for which a single conventional interpretation is bound to occur. In its current version, no syntactic parsing is provided. We show how this restriction practically constrains the method, and what experimental results can nonetheless be obtained.

A1.1 “attribute of the subject” (...)

A1.2 “attribute of the object”
A1.2.1
SSP $GN_1 \quad V_1 \quad GN_2 \quad (moins|plus|aussi) \quad Adj \quad (que|qu') \quad GN_3,$
A1.2.2
SSP $GN_1 \quad Ppobj_2 : (le|la|les|l') \quad V_1 \quad (moins|plus|aussi) \quad Adj \quad (que|qu') \quad GN_3$
where V_1 is a specific verb like *croire, juger, faire, estimer, rendre, trouver* for which French language has special syntactic rules.
target : $GN_2 \hookleftarrow Ppobj_2$
source : GN_3
tension : $\mathcal{V}^*(Adj) \cup \{p(GN_3)\}\backslash(\{p(GN_2)\} \cap \{p(GN_3)\})$

A1.3 “apposition” (...)

Group A1: comparisons between nominal groups via comparatives.
Notation: GN and GV stand for nominal or verbal groups, Adj and Adv for adjectives and adverbs. $\{p(x)\}$ is used for the intuitive notion of “set of properties of x”, and $\mathcal{V}^*(x)$, inspired from a mathematical notation, represents the set of concepts close to x, excluding x. These last notations are currently unused in the implementation, which does not exploit the tension attribute. For future works on the semantic analysis, they may have to be redefined.

Figure 1: Example of textual clue representation : group A1

2 Overview of clasical NLU approaches to metaphors

The first set of methods for interpreting metaphors is based on the use of specific semantic analysis, reflecting the nature of the link between the source and the target of the metaphor. For instance, Gentner [8] proposes a method to solve analogies between the two structures representing the source and the target of a metaphor. This led to an implementation [3]. Dan Fass called this approach *the comparison view* [4]. Fass also distinguished an *interaction view*, focussing on the novelty created by the metaphor, as in [9], and a *selection restrictions violations view*. The latter is sometimes called the *anomaly view*.

In our point of view, all these methods consider metaphor as an anomaly. Indeed, the “metaphor-related” semantic analysis is processed only when the building of a *literal* meaning representation fails. No detection is proposed before such a semantic analysis. Metaphor is then considered only in opposition to the literal meaning, and the latter is supposed to be prevalent and easily computable. Martin showed that a metaphor can exist even when the building of a literal meaning representation is possible, as in example 1:

ex. 1 *Mc Enroe killed Connors*

Moreover, metaphors are not the only cause to the failure of literal interpretations : metonymies act the same, and when dealing with real corpora, numerous “anomalies” can happen too.

Fass approach [4] allows discriminating literal meanings, metonymies, metaphors and a kind of anomalies (the remainder). Multiple analysis can be processed at the same time, but no disambiguation process follows. Therefore, no detection is used at all. James Martin [12] proposes a method based on Lakoff theory on conceptual metaphors [10]. It uses knowledge about *common* metaphors, and is also called, by Fass, the *conventional metaphor view*. This knowledge is essentially a semantic one, represented by specific links in conceptual graphs [11]. It leads to a metaphor interpretation process using no previous detection. As Martin showed, in example 1, if the metaphor is ignored by the previous approaches, because no literal interpretation can be correctly computed, his model produces at least two meaning representations. Moreover, Martin proposes a method for disambiguating the possibly multiple meaning representations. Therefore, his model can be viewed as a complex tool for detecting metaphors: the disambiguation processed after semantic analysis is a kind of detection. However, such a detection depends on a large knowledge base in which metaphors must be exhaustively[1] represented [11]. We note also that no methodology was proposed for automatically creating

[1] Practically, only the *conceptual* metaphors, as introduced by Lakoff and Johnson [10], must be represented exhaustively. The lexical metaphors are then processed using the relations between concepts. However, an exhaustive list of conceptual metaphors also seems impracticable.

or extending the knowledge-base. Such a method is hardly computable in the scope of large corpora analysis.

In this paper, we propose two different approaches to help metaphor detection, one based on the existence of textual clues related to the figures that bear metaphors, in section 3, p. 3, the second based on an extension of the conventional view, focussing on domain-specific corpora, in section 4, p. 5. In section 5, p. 6, we present a prototype, named STK, implementing these two approaches in a restricted hybrid methodology. STK is specifically designed for corpus analysis.

3 Textual clues approach to metaphors

Previous works already revealed the existence of textual clues related to explanation [13]. They might be used for generating explanations in tutorial systems [2] [14] [7]. We shortly present here –3.1– the results of a corpus analysis we made in order to study the existence and the nature of the textual clues spefically related to metaphors. We propose a semi-automatic method –3.2– for evaluating the validity of the emerging clues, some of them issuing from empirical generalizations.

3.1 Corpus analysis

In a research project about "explanatory analogies and metaphors in teaching", involving psychologists and computer scientists of the LIMSI-CNRS laboratory (Orsay, France), we collected a corpus of 26 explanatory texts in French, of about 200 words each. The subjects were experts in computer science. They were asked to answer the question:

What is a computer, and what is it used for ?

They were also asked to consider their interlocutors as novices, knowing no computer science, and no technical terms. This was made with the purpose of collecting explanations, from experts to novices, in a specific domain about which we hoped to know enough for detecting the possible metaphors by hand. We also supposed the explanatory situation from an expert to a novice to be propitious to metaphors and analogies. This last hypothesis was authenticated. Indeed, the corpus analysis revealed metaphors as well as related regularities. The regularities relevant to metaphors fall in two main classes:

- conventional domain-specific metaphors, extended to different terms
- textual clues of different kind

In this section, we study the textual clues only. The other class of regularities partially motivated the approach exposed in section 4.

The textual clues are sometimes lexical markers only (e.g. a single word), sometimes more complex regularities involving both a lexical marker and a syntactic regularity. They have been categorized and described in regard of the underlying syntactic structures involved : comparison (ex. 2), identification (ex. 3), opposition (ex. 4), emphasis (ex. 3 too).

ex. 2 *comparison using a marker:* $\mathbf{Comme}_{LM}$ *le tracteur*$_S$ *est un outil qui permet de réaliser* $\mathbf{plus}_{LM}$ *rapidement ce* $\mathbf{que}_{LM}$ *l'homme faisait pour cultiver et traiter la terre, l'ordinateur*^{T} *est un outil qui permet de traiter des informations que l'homme [lui] donne [...].*

ex. 3 *identification through apposition & typographical emphasis: Un ordinateur [...] est généralement composé [...] d'une mémoire, et d'un processeur*^{T} $,_{LM}$ *le* "$_{LM}$ *cerveau*$_S$ "$_{LM}$ *de la machine.*

ex. 4 *opposition through a state verb: [...(drawing)] c'est-à-dire un cerveau*^{T} *qui* $\mathbf{ne}_{LM}$ $\mathbf{serait}_{LM}$ $\mathbf{pas}_{LM}$ *humain*$_S$ *car incapable d'effectuer une tâche pour laquelle il n'est pas conu.*

This led to three main groups involving an implicit comparison:

A through syntactic patterns only, (e.g. *lesser ... than* constructions)[2]

B using a lexical marker, (e.g. *to look like ..., such as ...*)[3]

C through an identification or an opposition process (e.g. *It is not a ..., but a real ...*)[3]

and a fourth one –group **D**– of lexical markers related to emphasis (e.g. *literally*).

For the three first groups, a generic description was made of the syntactic structure involved in the comparison. It makes it possible to detect, at the sentence level, which terms are involved in the source and in the target of a possible metaphor –respectively the comparative element, or the comparator itself, and the compared element (see example 5 below). When the comparator itself bears its own meaning (e.g. a comparative adjective or an adverb), the tension between the source and the target of the metaphor was, as far as possible, specified, in order to help a future semantic analysis.

ex. 5 *Yesterday, at home, after having already eaten more than ever, Peter*^{T} *threw*$^T{}_S$ *himself*^{T} *on his dessert*^{T} $\mathbf{like}_{LM}$ *a lion*$_S$.
where x^T *stands for possible target group,* x_S *possible source group and* $\mathbf{x}_{LM}$ *is the lexical marker.*

[2] The translation of a textual clue may not result in a textual clue. We propose here translations of French textual clues, their corresponding in English may be different, even they may not exist !

D.2 “adverb”
D2.1.1
SSP $Adv \in GV$, where Adv is an adverb like *littéralement, vraiment, réellement*.
target : GV
source : NIL
tension : NIL

Group D.2: emphasis through an adverb

Figure 2: Example of textual clue representation: group D

The resulting decriptions of textual clues are object-oriented ones (see figure1).

This object-oriented approach is used in order to give a generic description of a textual clue, here related to the metaphor. The attributes **source**, **target** and **tension** are specific to this kind of clue, whilst the **Surface Syntactic Pattern** (SSP) is inherited from the superclass “textual clue”. In our examples, a list of lexical markers may appear, which corresponds to multiple possible instanciations of the class, and does not appear as is in the code. The SSP is not used in our implementation, therefore we do not present here a formal syntactic description (see [6] for details about the object oriented representation, see section 5 for what remains of the whole information in the current implementation, and what is expected for future versions).

The clues involving emphasis (group D) are described the same way as previously, with a restricted SSP and some empty attributes (see figure 2).

In the explanatory corpus, 31 metaphors have been found coinjointly with a textual clue, distributed as shown in the following table:

group	A	B	C	D
number of metaphors	2	17	9	3

When a clue appeared, it was first described in terms of the SSP, and then different generalizations were made, in order to obtain a better covering. The generalization were made from:

- similar syntactic structures involving a same lexical marker, expressing the same meaning (see figure 1, where multiple syntactic patterns are proposed for a comparison through adjectives)
- words close to the lexical marker involved in the clue, and from the same grammatical category, resulting in a list of synonyms corresponding to multiple instanciations of the object. (ex: *true* → *real, literal*)
- different terms with the same root, resulting in multiple lexical markers and related syntactic patterns (ex: *literal* → *literally*)

The clues proposed are more numerous than the ones detected, and the ones issued from generalization can not be called clues unless we prove they are. This is why the whole clues are currently evaluated, within a protocole exposed in next subsection.

3.2 Validity of the textual clues

Various observations can be made, and many questions debated, concerning the notion of textual clues related to metaphor:

- Not all the metaphors are introduced by a textual clue.
  (a clue is not necessary)
- A clue not always introduces a metaphor.
  (a clue is not sufficient)
- Lexical clues (including those involved in specific syntactic contex) seem better than syntactic ones only (group A: 2 vs others: 29).
- Some lexical markers were more frequently found than others: 15 uses of “comme”, “vrai” and “véritable”.
- Substituting the lexical marker with another similar is not always relevant (e.g. “vrai” (*true*) was found sometimes as a clue, sometimes not, but “réel” (*real*), issued from a generalization to close words, is never used this way).
- Most of the structures makes it possible to detect the source and the target of the metaphor (groups A, B and C). This corresponds to explicit comparisons and metaphors called *in praesentia*, in opposition to metaphors called *in extenso* for which the target is implicit [15]. On the other hand, the clues from group D are used with the *in extenso* ones, unless they conjointly appear with one from another group.

- As just said, clues of group D can be combined with the others.

The first two points are obvious. They mean that a detection using textual clues would not be robust, nor valid. It becomes necessary to combine the results of such a detection with other tools, or use them for probabilistic purpose only.

The next three points led us to a protocol of evaluation of the clues. The clues defined by syntax only can be evaluated as they are, but the other ones, which were characterized by a SSP (surface syntactic pattern), and a lexical marker, are too vague if considered for only one of these two attributes. What is a relevant clue seems to be indeed the combination of a SSP and a specific word –or group of words– which exactly maps to the object level in our specification. For instance, "littéralement" (*literally*) is frequently used within a metaphoric context (sometimes a common one). If we consider the grammatical category of the source (GN or GV), we note that it is more used for introducing a metaphor in a verbal group GV than in a nominal group GN.

Under such conditions, we shall then be able to distinguish the good instanciations issued from a generalization from the bad ones. Nonetheless, we assume that generalizing was here necessary, would it be only because the first corpus, analyzed by hand, was too short and thus restrained (forbade) the probability of finding an exhaustive list of clues, even if it proved, in spite of its insufficient size, similarities between certain clues. Note that in this particular case, "littéralement" was found after generalization.

Therefore, a second analysis is currently realized on a large electronic corpus (about 450,000 words) of articles extracted from the French newspaper "Le Monde", and focusing on the topic of economics[3], in order to evaluate the relevance of the supposed lexical clues. Each sentence containing a lexical clue is analyzed by hand. For instance, the French adverb "littéralement" (literally), not found in the first corpus, but issued from a generalization, was found 36 times, used 35 times in a metaphoric situation, in the large corpus. We propose to attach the result of such an evaluation to each specific clue in order to improve the word sense disambiguation. We call "value of relevance" of a clue its frequency of occurrence in a metaphoric situation. These "values of relevance" are currently determined semi-automatically on the electronic corpus, using STK tools –see section 5– and an analysis made by hand.

The two last points can be viewed as word-level information. The source and the target, as well as the tension in some cases, may be detected with syntactic parsing only. When a metaphor occurs with a textual clue providing such piece of information, a future semantic analysis can be improved by determining the target and restricting the scope of the interpretations. In the case of an automatic corpus analysis, it can be used to determine which terms are vague or ambiguous, rather than expecting an unknown part of the sentence to be metaphoric. In an information retrieval tool, it can be used for searching key-words on the whole text, but ignoring ambiguous parts of it. In the following example (6), the lexical marker **métaphore** has been found, for which no SSP is currently given. Therefore, the source, target and tension can not be determined. However, if searching for texts speaking about bombs and attemps in a large data base, using key-words like *explosion* or *detonator*, the whole sentences may be ignored, and the text in which they appear (practically an article from "Le Monde" speaking about stock market) will not be selected.

ex. 6 *il apparaît peu probable que l'on retrouve, du moins à court terme, la combinaison des "détonateurs*$_S$*" – pour reprendre la* $\mathbf{métaphore}_{LM}$ *d'un expert franais – à l'origine de l'explosion*$_S$ *du 19 octobre.[...] En fait, et pour poursuivre la* $\mathbf{métaphore}_{LM}$ *précédente, toute la question reste de savoir si l'explosif*$_S$ *à l'origine de l'effondrement*$_S$ *du 19 octobre est toujours présent.*

4 Domain-specific corpora approach

4.1 Hypothesis

Because textual clues are not sufficient to spot every metaphors in a text, we have tried to characterize metaphors in relation with their frequencies of occurrence. Lakoff's theory of conventional metaphors [10], and James Martin's works on implementing this theory [12], first motivated this approach. Indeed, the notion of *conventional metaphor* seems closely related to the notion of commonly used metaphor, if not defined this way.

For example, in the explanatory corpus we analyzed by hand, we noticed that terms related to *intelligence* appeared frequently, even if not coinjointly with terms related to *articial*. This could be seen as extensions of the conventional metaphor about *artificial intelligence*. We shall assume that *conventionaly metaphoric terms are bound to occur frequently with a same metaphoric meaning in texts related to a specific domain*.

4.2 Analysis

In order to verify our hypothesis, a larger domain-specific electronic corpus was collected: about 500 articles (450 000 words) from the newspaper "Le Monde", with the common topics "economics" and "stock market"[4]. Words

[3] This electronic corpus, on CD-ROM, allows extraction of articles related to a subject, an author, a date, pages, and so on.

have been tagged with their grammatical categories[4], then counted. Adverbs and verbs don't seem to be used metaphorically in a common way (we had a look at the most frequent ones only, for no automatic analysis can be made to determine if there is a metaphoric use or not). Practically, only the nouns and adjectives answer our hypothesis.

Note: in the following generic results, the score in parenthesis correspond to the number of occurrences of the most frequently used form related to the terms given, and not to the number of occurrences of the multiple forms. See following section, p. 6, for details.

The most frequent verbs are quite frequent ones (like "to do"(330), "to have to" (240), "to be able to" (190), ...), specialized ones appear with lower rates ("to change" (56), "to record" (50)). Having a look at the whole list of verbs used in our corpus and their context of occurrence, we note that some are used metaphorically, but no relevant relation with their frequency of use was found. More over, we observe that they sometimes correspond to extensions of common metaphors that are better reflected in a nominal form. For instance "tomber" (*to fall*) is used less than 50 times in a verbal form, when the name "chute" (*a fall*) itself, which is very close in meaning, is used 140 times, always in the singular form.

The most frequent adverbs are not domain-specific ones: "more" (1505), "still" (523), "very" (435), "well" (435), "few" (404), "then" (385), …Having a look at the whole list, no adverb seems specific to economics ("proportionaly" and "financially", the first in the list we found at least related to the topic, come with a score equal to 1 or 2 !).

The most frequent adjectives are domain-specific ones, like "financial", *boursier*, the French adjective meaning "related to the Stock Market", "French", "general", "commercial". However, we notice they are not metaphoric ones, neither polysemous in the domain.

The 40 most frequent nouns are listed in table 1.

Comments on table 1

The noun category answers best our hypothesis. The most frequently used nouns are domain-specific ones. One can notice the terms related to dates, mostly used in journalistic text. Practically, the vocabulary listed in the table corresponds to two topics: journalism and economics. These topics define exactly the domain of our corpus.

The entry *page* corresponds to an indication added to each article (its page number). The entry *vendredi* (friday), corresponds to an important date for people working in a Stock Market place or in trading rooms. Friday is somewhat a domain-specific word: a person working on the "forex" (foreign exchange market) explained us that, for instance, on fridays, each week, every participant square their positions, and once a month, the american unemployment data fall. Some also speak about a "week-end effect": statistically, the important events such as devaluations always fall before the week-end (note also the terms *fin* (end) and *semaine* (week) which co-occured to mean week-end). This year, in March, we could hear journalists naming "vendredi $noir_S$" ($black_S$ friday) the one concerning the unexpected results of the american unemployment data.

Concerning metaphors, entries like *marché* (market), *groupe*

(group), *actions* (actions), *valeurs* (values), *titres* (titles), …, correspond to highly polysemous terms that are used for one specific meaning in this corpus, which can be a viewed as a common metaphor for some of them (e.g. "market" and "change"). Indeed, there are conceptual metaphors underlying these domain-specific meanings, as the ones proposed by Lakoff and Johnson in [10], that can be exceptionaly extended[5] using new related terms and then easily processed with Martin's method [12]. We noticed that the conventional metaphor of the domain could be exhaustively detected using the lexicon of common nouns only. Thereby, Martin's approach can be enhanced by the results of such a corpus analysis, providing a set of new relations to add to the knowledge base (described in [11]) representing common metaphoric main relations. We will not detail here how to find back the *roots* of the conceptual metaphors, what can be made by hand only at the present time.

In the aim of this paper, the domain-specific corpus analysis shows the most frequent words are monosemic in the corpus, used with domain-specific meanings, which allows us to consider only words with low frequencies for non-conventional metaphoric meanings. The remaining problem is how to assign a value over which words will be frequent and under which they will be rare. In the corpus we analyzed, by hand, the threshold was in low frequencies. We plan to test different statistical distributions laws (e.g. χ^2) in order to see to what extent such threshold can be determined automatically.

The results of our analysis were made after automatic tagging of the corpus. We briefly present in the following section the tool we have developped to test our hypothesis, as well as how we combine the two approaches.

5 Hybrid method and its implementation: STK

5.1 Hybrid method

The previously detailled approaches were theoretically based on the use of a syntactic parser and a lemmatizer.

[4] see section 5 for details about the tagging.

[5] see example 8, p.7

score	1575	861	659	605	589	589	577	573	571	552
noun	marché	francs	semaine	hausse	groupe	page	mois	actions	milliards	valeurs
tag	NMS	NMP	NFS	NFS	NMS	NFS	NMS	NFP	NMP	NFP
score	547	503	502	481	459	440	436	434	403	396
noun	titres	baisse	marchés	année	octobre	cours	millions	vendredi	fin	société
tag	NMP	NFS	NMP	NFS	NMS	NMS	NMP	NMS	NFS	NFS
score	392	362	357	351	349	344	340	338	335	332
noun	terme	place	fois	entreprises	taux	lundi	jours	sociétés	change	début
tag	NMS	NFS	NFS	NFP	NMP	NMS	NMP	NFP	NMS	NMS
score	314	304	303	301	298	297	297	290	277	275
noun	jeudi	cours	président	indice	janvier	intért	investisseurs	niveau	prix	effet
tag	NMS	NMP	NMS	NMS	NMS	NMS	NMP	NMS	NMS	NMS

This table corresponds to the 40 most frequent nouns in the corpus of articles extracted from the newspaper “Le Monde”, related to “economics” and “stock market”. The score indicates the total number of occurrences, the tag corresponds to the category (Noun), gender (Feminine or Masculine) and number (Singular or Plural)

Table 1: 40 most frequent nouns

Parsing is required to extract which parts-of-sentence are supposed to be metaphoric when finding a textual clue. Lemmatizing is required to find back the terminological roots of the words and complete the words occurrences count.

The textual clues, including syntactic descriptions, can be easily retrieved, using the lexical markers and their categories. When a clue is found in a sentence, a parsing of the sentence is sufficient to associate the information concerning the source, the target and the tension, when provided with the clue. Thereby, we consider the value of relevance of the clue as a probability for a metaphoric interpration of the source part-of-sentence. This value is theoretically lesser or equal to 1. When n clues appear for a same source, like in example 7, we propose to combine the values $\{v_i\}_{[1..n]}$ according to the following recursive law:

$$\begin{aligned} V^1 &= v_1, \qquad \text{if} \quad n = 1, \\ V^n &= V^{n-1} + (1 - V^{n-1})v_n, \qquad \text{if} \quad n > 1, \end{aligned}$$

which is obviously independent of the order of the v_i and verifies $\forall n, V^n \leq 1$. This allows us to consider two clues as more relevant than one alone, and so on... Practically, this almost never occurred in the examples we analyzed ! This is why we do not detail this theory further.

ex. 7 *La bourse s’est* **littéralement**$_{LM}$ *effondrée*$_S$, **comme**$_{LM}$ *souflée par l’explosion d’hier*$_S$.

The domain-specific lexicon we analyzed did not show that, under a determined threshold, the words with lowest scores would be more easily used in a metaphoric way than the other ones under the threshold. Therefore, we propose a more “binary” law, just considering words with high scores as conventional, and the others as possibly metaphoric, with no probability provided. As said before, the threshold is currently determined by hand. We note its value may depend on the application. Indeed, if searching for texts related to a domain in a large data base, using keywords, a low threshold may be usefull not to avoid rarely used technical terms. On the opposite, if wanting to help a semantic treatment process, then the highest threshold, the best, in order to spot every metaphors.

In the example 8, bold terms are marked monosemic, using a high threshold. The metaphoric expressions *caracoller*, *repasser sous la barre*, *sans faute* can be considered as specific to other domains. Therefore, the method may be improved considering terms from other domains, if possible, rather than terms rare in this domain only. This will be part of our future works.

ex. 8 *En l’espace de quinze* **jours**, *la* **Bourse** *est devenue méconnaissable. Elle qui caracolait*$_S$ *le 15* **décembre encore** *sur des* **sommets record** *(l’* **indice** *CAC culminait à 414,3) a dû précipitamment en rabattre*$_S$, *pour repasser sous la* **barre**^{T} *des 400. Vilain pied de nez pour ce* **marché** *qui avait jusqu’ici réalisé un quasi sans faute*$_S$, *faisant la joie des fidèles*$_S$ *du* **palais Brongniart**.

As all the metaphoric terms with scores over the threshold are related to conventional metaphors previously analyzed and coded, the only possible novel metaphors can not occur over the threshold. This is part of our main hypothesis for domain-specific corpora. Even most conventional terms can theoretically be used for novel metaphors. In order to avoid such figure, we can here combine the textual clues with the domain-related lexicon, adding a possibly not conventional meaning to the frequent terms found coinjointly with a clue. No relevant example was found of such a combination in the corpus we analyzed, and we could not evaluate this theory.

Lastly, note that the threshold is currently determined by hand in regard of the list of nouns only. In the corpus we work on, it appears that conventional metaphors are best reflected in this category. This interesting result may depend on the corpus, and the threshold comming from the nouns may be wrong for the other categories in another domain. Here again, future works on different corpora should answer this question.

Controlling semantic analysis

Depending on the results of a search for textual clues and the frequencies of words, we propose a generic methodology for controlling deeper semantic analysis. For plain words (nouns, verbs, adjectives and adverbs), after syntactic parsing:

- if a clue is found, mark the source items as possibly metaphoric, with the "value of relevance" of the clue for probability, process a specific analysis using the target and tension attributes if present
- if the frequency of the word is under the threshold for conventional meanings, with no clue, just mark as possibly metaphoric with no other information, process a specific analysis
- if the frequency of the word is over the threshold, with no clue, consider as conventional

A generic semantic analysis should be processed in each case, including the model proposed by Martin [12] for processing conventional metaphors and their extensions.

5.2 Prototype: STK

Parsing and lemmatizing tools designed for corpora were not available for French language. As a matter of fact, some tools exist but are limited by the lexicon they use, and thus can not be applied for a large corpus analysis. We supplied this lack using the results of a part-of-speech tagger, and practically restricting the method.

Textual clues search

The part-of-speech tagger assigns grammatical labels to words. These labels provide restricted information depending on the grammatical category:

- nouns, participles, pronouns, determiners → gender and number
- verbs, pronouns → person and number
- others (adverbs, punctuations, numbers, ...)

The search for the words relevant to the textual clues is made on the couples *(word, tag)* rather than words only, avoiding possible lexical ambiguities. But the search of the parts-of-sentence is not yet performed. The whole sentences are marked, rather than well determined parts of them. However, we note that novel metaphors appear on are rare words only even if coinjointly with a textual clue, which should allow us to mark parts of sentences only, but this part has not yet been developed.

Words count

The couples *(word, tag)* are also used for the counting step. This results in scores far different from what expected. However, the results concerning the domain-related words seem correct (see table 1). By correct we mean that the most frequently appearing couples actually correspond to words frequently used. Some constricting problems remain, like distinctions between the plural and singular forms, (see entries "marché" and "marchés" in the appendix). However, the current results tend to prove that domain-specific words can be used with specific number, as for the entry *barre* in table 1, or the noun "chute", as exposed in subsection 4.2. Therefore, lemmatizing may not be relevant. This was an unespected result our incremental approach revealed.

STK

The methods have been implemented in STK, a prototype we plan to use in different context [7]. We shortly expose here its general design and its possible applications.

The tagger used is Brill's tagger, described in [1]. It has been trained on extracts of the electronic corpus dedicated to economics. The extracts tagged by hand are two articles, a very short size. A large amount of tagged-by-hand data has now been provided, for a new training on heterogeneous data (currently performed). The lexicon we use derives from BDLEX[6]. No information about the use of the lexically ambiguous forms were given, and Brill's tagger needs partially ordered lists of labels for each entry, the first label corresponding to the most frequently used grammatical form. Some heuristics were used to order the different labels, but there are still corrections to do by hand. For instance, "son" sometimes stands for *his* or *her*, sometimes for *a sound*. Considering the noun as the most frequent form significantly corrupts the results of the tagging...

A specific tokenizing process has been developped, specially designed for the corpus used and the kind of entry Brill's tagger requires (one sentence per line). Usually, the search for the sentence unit can be improved by a partial syntactic parsing. In STK, the tagger used constraints this search, and specific processes for punctuation, insertions and lists were added to the tokenizing main process (see [7] for details).

The tagging part of STK, including tokenizing and tagging itself, is evaluated as is in a GRACE[7] session of evaluation for taggers for French corpora. This project should be accomplished in the present year.

[6] developped in 1993 by Guy Perennou Martine de Calmes and Isabelle Ferrane at the IRIT-CERFIA laboratory under the BDLEX action of the GRECO CNRS, French National Center for Scientific Research

[7] Grammars and Resources for Corpora Analysis and their Evaluation

STK in its whole has been developed in C++, and was designed as a common UNIX command, with on-line options and manual. In its current version, it computes multiple files depending on the options choosen: tokenized data, tagged data, ordered lists of occurrences, and what we call the "metagged" data, corresponding to the text with tags concerning metaphors and words frequencies, for future analysis.

If the method we exposed introduces STK as a tool for word sense disambiguation, we see other possible applications. The terminology extraction is one of them, already used for automatic extraction of specialized lexica [7]. STK in its whole is evaluated under a coordinated research action of the AUPELF[8] on automatic extraction of terminology.

STK is also used for the evaluation of the textual clues previously proposed. The tokenizing process allows us to easilly extract sentences, and the interface with the corpus includes the extraction of whole articles. By this mean, STK can be used for a semi-automatic analysis of textual clues, independently of what kind of style or domain these clues are related to.

Lastly, STK "metagged" file contains sentence-level and word-level information providing probabilities for non-conventional metaphoric meanings (involving the "value of relevance" of the clues).

Conclusion

A method for detecting metaphors on domain-specific corpora has been proposed in order to complement previous works in NLU. This method partly comes in opposition to the semantic approaches to metaphor. It is essentially based on the use of textual clues, easy to spot, and frequencies of words occurrences, easy to determine on corpora.

A corpus analysis determined a set of clues which were generalized and are currently evaluated. Their use has been proposed to help word sense disambiguation, spotting possible metaphors and providing probabilities for metaphoric interpretations. At the same time, we note that domain-specific corpora makes it possible for us to semi-automatically detect the conventional metaphors of its domain, and by this mean, improves the search for all novel metaphors without any semantic analysis.

The method has been partially implemented in a prototype, STK, integrating Brill's tagger trained on French language, and currently under evaluation. We plan to design a generic tool for detecting non-literal meanings on corpora. Our aim is to provide helps for ambiguous words analysis for different other tools: interpretation, semantic analysis and information retrieval.

[8] Association des Universités Partiellement ou Entièrement En Langue Française

References

[1] E. Brill, 'A simple rule-based part of speech tagger.', in *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, (1992). ACL.

[2] M.P. Daniel, L. Nicaud, V. Prince, and M.P. Pery-Woodley, 'Apport du style Linguistique à la Modélisation Cognitive de l'Élève', *Lecture Notes in Computer Sciences*, **608**, 252–260, (1992). Proceedings of the International Conference on Intelligent Tutoring Systems (ITS-92), Montréal.

[3] B. Falkenhainer, K.D. Forbus, and D. Gentner, 'The Structure-Mapping Engine : Algorithm and Examples', *Artificial Intelligence*, **41**, 1–63, (November 1989).

[4] D. Fass, 'met : A Method for Discriminating Metonymy and Metaphor by Computer', *Computational Linguistics*, **17**(1), 49–90, (1991).

[5] D. Fass, E. Hinkelman, and J. Martin, eds. *Proceedings of the IJCAI Workshop on Computational Approaches to Non-Literal Language*, Sydney, Australia, 1991. 1991.

[6] S. Ferrari, 'Using textual clues to improve metaphor processing', in *Proceedings of the Student Sessions at the 34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, CA, USA. June 23-28 1996, (1996). to be published.

[7] S. Ferrari and V. Prince, 'Création et extension automatiques de dictionnaires terminologiques multilingues spécialisés à partir de corpus monolingues', in *Actes de la Conférence internationale sur le traitement automatique des langues et ses applications industrielles TAL+AI 96*. Moncton, New-Brunswick, Canada. 4-6 juin 1996, (1996). à paraître.

[8] D. Gentner, *Analogical Inference and Analogical Access*, chapter 3, 63–88, Pitman Publishing, London, Morgan Kaufmann Publishers, Inc., Los Altos, California, 1988.

[9] J.R. Hobbs, 'Metaphor and abduction', In Fass et al. [5], pp. 52–61.

[10] G. Lakoff and M. Johnson, *Metaphors we live by*, University of Chicago Press, Chicago, U.S.A., 1980.

[11] J.H. Martin, 'MetaBank : a Knowledge-Base of Metaphoric Language Conventions', In Fass et al. [5], pp. 74–82.

[12] J.H. Martin, 'Computer Understanding of Conventional Metaphoric Language', *Cognitive Science*, **16**, 233–270, (1992).

[13] M.P. Pery-Woodley, *Textual Clues for user modeling in an intelligent tutoring system*, Master's thesis, University of Manchester, England, U.K., 1990.

[14] V. Prince, 'Indices linguistiques pour la construction d'un modèle automatique d'analyse et de production des explications', in *Actes de l'atelier de recherche GENE du PRC IA*, ed., Sup.Telecom, pp. 141–153. 15-16 décembre, (1994).

[15] P. Ricoeur, *La métaphore vive*, Seuil, Paris, France, 1975.