A Corpus-Based Approach for Metaphor Processing

Stéphane Ferrari

LIMSI-CNRS
BP 133
F-91403 Orsay cedex, FRANCE

phone : (+33) (1) 69.85.80.18
fax : (+33) (1) 69.85.80.88
email : ferrari@limsi.fr
URL : http://www.limsi.fr/Individu/ferrari

Introduction

Classical NLP (Natural Language Processing) approaches to metaphor used to focus on the semantic regularities related to metaphors. As we show in a brief overview, in section 1, p. 2, the detection process, when there is one, was mostly restricted to the failure of some semantic treatment methods. In opposition, some approaches, proposing multiple semantic treatments at the same time, led to complex word sense disambiguations. No proposition has been made to detect metaphors without semantic analysis. Therefore, no efficient method has ever been suggested to avoid misunderstanding of ambiguous words when dealing with large corpora. Our aim is to provide NLP systems with information about where metaphors are bound to happen, in order to improve a future semantic treatment. The same information may also be used for translation purposes, or in the scope of information retrieval, avoiding search about ambiguous words.

In previous works, we revealed the existence of textual clues related to metaphors, especially in explanatory discourse. But textual clues are not sufficient to provide a robust detection of metaphors. In this paper, we focus on how to combine the properties of such clues with other heuristics in a hybrid method in order to improve the detection and the interpretation of metaphors. Our approach is specifically designed for large, domain-specific corpora. The conventional view of metaphors, exposed in section 1, focusses on common metaphors, and may be easily extended. In section 2, p.3, we propose a method for automatically extracting lexicons of commonly used terms from domain-specific corpora and study their conventionality depending on their grammatical category. We show how some common words in a domain can be viewed as conventional metaphors.

In section 3, p. 5, we propose a method for chosing the most appropriate semantic treatment and for helping word sense disambiguation. A prototype implementing the method, STK, is currently under development. STK integrates Brill's rule-based part-of-speech tagger, which has been trained on the French language for this purpose. It is designed to provide usual NLP tools with a new set of tags underlying the possibly metaphoric meanings, giving information about the probability of a non-conventional interpretation, as well as a list of domain-specific words for which an unique conventional interpretation is bound to occur.

1 Overview of classical approaches to metaphors

In Natural Language Processing, approaches to metaphors consist in specific semantic analysis reflecting the nature of the link between the *source* and the *target* of the metaphor. For example, Gentner [6] proposes a resolution of analogies between the (conceptual) structures representing the source and the target of the metaphors, which led to an implementation [2]. This approach is called *the comparison view* by Dan Fass [3]. There also exists an *interaction view*, foccussing on the novelty created by the metaphor [7], and a *selection restrictions violation view*. This latter is sometimes called the *anomaly view*. In our point of view, all of these approaches consider metaphor as an anomaly. Indeed, the specific semantic treatment is processed only when the building of a literal meaning representation fails. No detection is processed before the semantic analysis. Metaphor is here opposed to literal meaning, and the latter is supposed prevalent.

Fass' approach [3] allows discriminating literal meanings, metonymies, metaphors and a kind of anomalies (the remainder). Multiple treatments can be processed at the same time, but disambiguating the possibly multiple meaning representations is not provided. James Martin [10] proposes a method based on Lakoff's theory of conceptual metaphors [8], using knowledge about common metaphors, a *conventional metaphor view*. This knowledge is essentially semantic knowledge [9]. It allows a metaphoric interpretation process without previous detection. As Martin showed, in the example

ex. 1 *Mc Enroe killed Connors*

the metaphor was ignored by the previous approaches, because a literal interpretation exists and can be correctly computed. The conventional view leads to two meaning representations : one where "McEnroe is a killer", another where "he is just a winner". Moreover, Martin proposes a method to disambiguate the possibly multiple meaning representations. This disambiguation can also viewed as a "post-semantic-analysis" detection. However, it depends on a large knowledge base in which metaphors must be exhaustively[1] represented, and the related semantic anal-

[1] Practically, only the *conceptual metaphors, as introduced by Lakoff and Johnson in [8], 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.*

ysis may be hardly computable on large corpora. We also note that no methodology was proposed for automatically creating the knowledge base, extracting possible conventional meanings.

In our point of view, a large corpus analysis requires more efficient methods than classical approaches would provide. In previous works, we revealed the existence of textual clues related to metaphors [5] [11]. In this paper, we propose an extension of the conventional view, detailed in section 2, p. 3. In section 3, p. 5, we present a hybrid method combining the use of textual clues and the previous extension of the conventional view for detecting and chosing the most appropriate semantic treatment for metaphors on large corpora, as well as a prototype partially implementing the method.

2 Domain-specific corpora approach

We have tried to characterize metaphors in relation with their frequencies of occurrence. Lakoff's theory on conventional metaphors [8], and James Martin's works on implementing this theory [10], motivated this approach. Indeed, the notion of *conventional metaphor* is closely related to the notion of commonly used metaphor. The main problem is that commonly used metaphors are not necessarily common independently of a topic or a semantic domain. Our hypothesis is that *conventionally metaphoric terms are bound to occur frequently within texts related to a specific domain*.

We first collected a small corpus of 26 explanatory texts of about 200 words each in order to study textual clues [5]. The texts were explanations from experts to novices telling what a computer is, and what it is used for. We noticed that the terms related to *intelligence* appeared frequently (stupid, intelligent, comparisons with human cognitive abilities), and were related to the common metaphor about “artificial intelligence” in computer science. In order to verify our hypothesis, a larger domain-specific electronic corpus has been used : about 500 articles (450 000 words) from the newspaper “Le Monde”, with the common topics “economics” and “stock market”. This electronic corpus has been tagged with grammatical categories, and the words counted.

Adverbs, verbs and adjectives do not 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, the nouns answer best 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 next section, p. 5, for details.

The most frequent verbs are very 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)).

The most frequent adverbs are not domain-specific ones at all: “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 (“proportionnaly” and “financially”, the first in the list we found at least related to the topic, come with a score equal to 1 !).

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.

The 40 most frequent nouns are listed in the table 1.

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	intérêt	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), the gender (Feminine or Masculine) and the number (Singular or Plural)

Table 1: 40 most frequent nouns

The most frequently used nouns are actually domain-specific ones. Practically, the nouns listed in table 1 correspond to two main topics, journalism and economics, defining exactly the domain of our corpus.

Most of the entries are polysemous words, with one specific meaning in the domain. *marché* (market), *groupe* (group), *actions* (actions), *valeurs* (values), *titres* (titles), …, correspond to highly polysemous French terms that are used for one meaning only in this corpus, which can be 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 [8]. They can be exceptionnaly extended using new related terms and then easily processed using Martin’s method [10]. In example 2, *réaliser un sans faute* (never losing in a competition) and *repasser sous la barre* (going under a threshold) are extensions of the conventional metaphor between the concept of “threshold” and the one of “physical bar” (like the “horizontal bar” in sport).

ex. 2 *En l'espace de quinze **jours**, la **Bourse** est devenue méconnaissable. Elle qui caracolait le 15 **décembre encore** sur des **sommets record** (l' **indice** CAC culminait à 414,3) a dû précipitamment en rabattre, pour repasser sous la **barre** des 400. Vilain pied de nez pour ce **marché** qui avait jusqu'ici réalisé un quasi sans faute, faisant la joie des fidèles du **palais Brongniart**.*

Martin's approach [10] can be enhanced by the results of such corpus analysis. Studying frequent nouns only seems sufficient to find the relations to add to the knowledge base [9] representing common metaphoric main relations. Part of the analysis must still be made by hand, but we consider this approach as a first step in automatizing the search for conventional metaphors.

The domain-specific corpus analysis shows the most frequent words are used with domain-specific meanings, sometimes conventional metaphors. This result allows us to consider only words with low frequencies as targets for novel (non-conventional) metaphoric meanings. We propose, in the following section, a consequent method for chosing the most appropriate treatment on domain-specific corpora.

3 Hybrid method and its implementation : STK

The textual clues related to metaphors (e.g. "like", "literally", comparisons, identifications, oppositions, emphasis...) are lexical markers combined with syntactic regularities [5]. They can be easily retrieved. When a clue is found in a sentence, a parsing of the sentence is sufficient to associate the information concerning the source and the target, when provided with the clue, as in example 3.

ex. 3 *Yesterday, at home, after having already eaten more than ever,* $\overline{\textit{Peter}}^T$ $\underline{\overline{\textit{threw}}}_S^T$ $\overline{\textit{himself}}^T$ $\overline{\textit{on his dessert}}^T$ ***like*** $\underline{\textit{a lion}}_S$.
where $\overline{x}^T$ *stands for possible target group (or part-of-sentence) and* $\underline{x}_S$ *possible source group and **like** is the lexical marker.*

Each clue has been evaluated on the large corpus. We shall consider the frequency of use of a clue in a metaphoric context, also called its value of relevance, as a probability for a metaphoric interpration of the source part-of-sentence.

Therefore, the hybrid method we propose for detecting metaphors on domain-specific corpora can be decomposed as follows:

step 1 tagging the corpus with grammatical categories, in order to avoid lexical ambiguities

step 2 extracting common nouns and studying on this restricted lexicon the possible conventional metaphors, marking other nouns and related adjectives, verbs and adverbs as possibly metaphoric

step 3 searching for parts-of-sentence co-occurring with textual clues related to metaphors, marking them as possibly metaphoric, with a probability corresponding to the "value of relevance" of the clue

A prototype implementing this method, STK, is currently under development. STK includes tokenization, tagging and lexicon extraction. The search for metaphors is partially implemented : without syntactic parsing, parts-of-sentence described in the syntactic regularities of the textual clues can not be retrieved.

The tagger assigns grammatical labels to words. These labels provide information about the category (word, verb, adjective, adverb, ...), the gender and the number, and so on (see table 1). The search of the lexical markers involved in the textual clues is made on the couples (word, tag) rather than words only, avoiding possible lexical ambiguities. The whole sentences are marked, rather than well determined parts of them. Note that the couples (word, tag) are also used for the lexicon extraction (see entries "marché" and "marchés" in table 1). The current results tend to prove that domain-specific words can be used differently depending on the number, as for the entry *barre*. Therefore, lemmatizing may not be relevant, because more information are provided with the couples (word, tag). This result was unexpected.

The tagger we use is Brill's tagger, described in [1]. It has been trained on extracts of the electronic corpus dedicated to economics. The tagged by hand extracts are two articles, a very short size. A large amount of tagged-by-hand data has now been provided, allowing a new training on heterogeneous data (currently realized). The lexicon we use derives from BDLEX[2]. 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, where the first label corresponds 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. A tokenization process has been developped, specially designed to the corpus used and the entry Brill's tagger needs (one sentence per line). The tagging part, including segmentation and tagging itself, is evaluated as is in GRACE[3] session for French taggers.

In its current version, STK computes multiple files, depending on the options choosen : tokenized file, tagged file, extracted ordered lexicons, and a "metagged" file which is the text with the tags concerning possible metaphors. STK in its whole, and the method for extracting conventional metaphors, are evaluated under a coordinated research action of the AUPELF-UREF[4] on automatic extraction of termi-

[2] a lexicon 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

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

[4] Francophone Agency for Education and Research

nology. In example 2, words in bold type are domain-related ones marked using STK, and table 1 is the head of a lexicon extracted by STK.

STK is also used for evaluating the textual clues previously found on a smaller corpus. A "value of relevance" is assigned to each clue, corresponding to its frequency of use in metaphoric context. For instance, the lexical marker "literally" has been found within 35 metaphors out of 36 uses. Therefore, the clues provide a probability for metaphopric interpretations. Considering the classical approaches to metaphors, our analysis leads to the following methodology for chosing the appropriate semantic treatment at the word level :

A corresponds to Martin's treatment, including interpretation of conventional metaphors found in the domain

B corresponds to a specific treatment dedicated to novel metaphors

step 4 nouns of high frequencies for which no clue is associated are processed with treatment **A** only, tagged *conventional*

step 5 other nouns and related verbs, adjectives and adverbs, as well as the ones found with a clue, are processed with **A** and **B**, tagged *possibly metaphoric*

Conclusion, future works

We have shown that the classical approaches to metaphors are hardly computable on large corpora, and a robust detection process has to be designed. We proposed to extend the conventional view of metaphors in order to be able to deal with domain-specific corpora. Studying the most frequent words in a domain, we note the nouns provide information on the conventional metaphors of the domain. This result led to a methodology for spotting possible metaphors. Previous results concerning textual clues related to metaphors can be combined with this approach.

A restricted method for detecting metaphors and extracting lexicon have been designed and implemented in STK. No syntactic parsing is currently used, therefore, when a textual clue is found, the whole sentence is marked, and the methodology proposed is constrained by this restriction. We currently study how to spot the parts-of-sentence relevant to the different textual clues in order to improve the current results of our prototype. We also plan to deal with SGML format, in order to design a reusable tool. Depending on the results of the evaluations, STK may be distributed.

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

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

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

[5] S. Ferrari. Traitement automatique des métaphores : une approche par marquage textuel. In *Actes du Deuxième Colloque Jeunes Chercheurs en Sciences Cognitives*. Presqu'île de Giens, France, 5, 6 et 7 juin 1996, 1996. à paraître.

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

[7] J.R. Hobbs. Metaphor and abduction. In Fass et al. [4], pages 52–61.

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

[9] J.H. Martin. MetaBank : a Knowledge-Base of Metaphoric Language Conventions. In Fass et al. [4], pages 74–82.

[10] J.H. Martin. Computer Understanding of Conventional Metaphoric Language. *Cognitive Science*, 16:233–270, 1992.

[11] V. Prince and S. Ferrari. A textual clues approach for generating metaphors as explanations by an intelligent tutoring system. In *Proceedings of TALC96, Conference on Teaching And Language Corpora*. Lancaster University, U.K., 9-12th August, 1996. to be published.