AUTOMATIC ACQUISITION OF MEANING ELEMENTS FOR THE CREATION OF SEMANTIC LEXICONS

Ronan Pichon, Pascale Sébillot
IRISA,
Campus de Beaulieu, 35042 Rennes cedex, France
tel: 33 2 99 84 74 50 / 73 17, fax: 33 2 99 84 71 71
email: rpichon@irisa.fr, sebillot@irisa.fr

ABSTRACT

This paper presents, in a unified way, two new trends in natural language processing, that is a new kind of lexicons that are cornerstones of a lot of current natural language applications which tackle the problem of meaning, and different corpus-based lexical knowledge acquisition studies that have emerged with the big amounts of electronic texts available on the nets. More precisely, this paper is an overview of corpus-based methods that could assist a lexicographer in order to build semantic lexicons. The presentation is directed by the lexicographer's tasks that these studies can facilitate, i.e. term acquisition in order to circumscribe the vocabulary of the lexicon of a domain, structuring of a lexicon with linked semantic classes, adjustment to a language, and determination of the predicates of a domain. We end up by a discussion about the integration of these methods into a fully automatic construction of semantic lexicons.

1. INTRODUCTION

A lot of natural language applications which tackle interpretation or understanding problems (translation, man-machine interfaces, intelligent information retrieval, text summarization, etc.) need lexicons. The traditional lexicons, which were lists of words associated with their own morpho-syntactical or semantic information, have evolved toward much more structured ones, named *semantic lexicons*, in which the meaning of a word is defined by its relations with the meanings of the other words. Such lexicons are much more complete, because they permit to express the different aspects of a word and its links with the other words of the vocabulary of the application [1]. Particularly they enable to express how the interpretation associated with a given utterance of a word is influenced by the presence of another word in its environment.

However building such a lexicon for every new application and/or domain is a very heavy and expensive task. Yet, the fact that lots of electronic texts are now available has led to the spreading of numerous works in the domain of corpus-based knowledge acquisition, especially the acquisition of information about the elements of the vocabulary associated with a domain of knowledge. Even if these works do not all aim at building semantic lexicons, or even parts of them, they can be relevant in order to help a lexicographer in this task.

Therefore this paper has two aims: first, it *presents two new trends* in natural language processing, that is the *use of a new kind of rich lexicons* in current natural language applications, and *the corpus-based lexical knowledge acquisition*, a domain whose development is closely connected with the emergence of big amounts of electronic texts through nets (Internet, for example). Moreover, it *unifies the presentation* of these two trends, by showing how they can be linked. Actually we propose an overview of the different corpus-based studies, in which we try to bring to light the various aspects of the task of a lexicographer that they could assist. Therefore the second and main interest of this paper is to propose an organization of an active research domain which mixes natural language and statistical techniques, and to show how some elements of these various works that have been developed for quite different purposes can all be taken into account in order to facilitate the production of a semantic lexicon.

We first briefly define more precisely what a semantic lexicon is. We then present the overview of the works about corpus-based approaches for lexical acquisition. Our conclusion is a discussion of the integration of these researches into a fully automatic construction of semantic lexicons, which points out some general ideas for such a work.

2. SEMANTIC LEXICONS

Two aspects are traditionally studied in semantics: the *signifier*, that is the (linguistic) sign, often named lexeme, and the *signified*, that is the meaning, often connected with the notion of concept. In lexical semantics, defining a given lexeme firstly consists in associating it with concepts of the domain of an application, but also in linking it to other words of this field.

A semantic lexicon is therefore a structure which displays semantic relations between lexemes and concepts, but also between lexemes, and between concepts. A lexical entry of the lexicon may correspond to a couple (lexeme, concept).

The first step in building such a lexicon consists in determining which lexemes are associated with which concepts (of course, if both lexemes and concepts of the domain are already known). If several lexemes are linked to the same concept, they form a semantic class of lexemes.

Within this class, it is possible to find (ontological) relations that enable to distinguish the various elements from each other. For example, chair and armchair are both "objects on which someone can sit" (concept), but may be distinguished by the fact that only one of them "has arms". After that, a second step consists in determining relations between lexical entries, like synonymy, or hyperonymy (is-a link) for example.

WordNet [2], (manually) developed at Princeton university, is an example of such a lexicon, which is currently often used in natural language applications. In this lexical database, concepts are represented by *synsets*, which are sets of words that may be exchanged within a given linguistic context, typically a sentence. A word may belong to several synsets. Concerning the concepts, the stress is laid on the lexical relations between them, much more than on what they can represent.

The cost of the creation of such a lexicon for a new application can be largely reduced if some parts of the tasks that it implies are automatically generated. We present now an overview of the numerous studies in *lexical acquisition*, that is the domain of automatic acquisition of lexical information from corpora. These works extract information that could be integrated into a lexicon. This presentation is directed by the lexicographer's tasks that these studies could facilitate. It is not exhaustive because this field is extremely productive, but we try to point out some families within this set of studies in order to get a good general idea of these researches, even if some works that we have chosen to classify into a single family for clarity reasons also possess some aspects which could justify their belongings to another family.

3. LEXICAL ACQUISITION

The first step in building a lexicon is to get the lexemes that must be incorporated in it; therefore we begin our presentation by works dedicated to the extraction of the terminology of a domain. We then summarize various researches which concern the main part of the creation, that is the structuring of the lexicon, by establishing sets of "similar" words which are words that are "similarly" used within the corpus of texts of a domain of knowledge; this regrouping may be used to determine the concepts of the domain; moreover some of these studies also establish lexical relations between the different sets that they have discovered. We thirdly present works which try to update a given lexicon for a special use, by getting, eliminating or adapting lexical entries. We finally conclude with researches which aim at finding predicates (verbs, roughly speaking) of a domain and their argument structures, that is the elements that their uses imply (subject/agent and direct object/theme for a transitive verb for example).

For all the families of works we give an idea of the techniques that are used to get the different results.

3.1 Extraction of Terminology

These works are not exactly situated within the field of lexical acquisition because they aim at constituting the nomenclature of the terms of domains from corpora. But they can be a very interesting starting point for the elaboration of a semantic lexicon.

Two kinds of techniques are used to isolate the terms, whether they are simple (unique nouns (N) for example) or complex (N N compounds for example).

Some works use *syntactic methods* for this task, that is they determine the terms in the texts by recognizing special syntactic categories of elements like N N or Adjective N for example, or conversely by detecting parts of the texts that cannot be terms and can be used as frontiers between them. This last technique is used by Bourigault [3, 4] in his extractor LEXTER.

Other works use *statistical methods*. For example, a common method to isolate complex terms consists in studying the collocations, that is the words that have a higher probability to be encountered together than just by chance. This "together" notion may correspond to the juxtaposition of the two elements or their common presence within a window of *n* words. One criterion often used to determine the collocations is the *mutual information* defined by Fano [5].

Finally, these *two methods* can be used *together*. For example, Smadja [6] first collects terms through a mutual information criterion and then only accepts those that have particular syntactic forms, whereas Daille [7, 8] first filters terms through automata and then only keeps the most frequent ones.

3.2 Structuring of the Lexicon

We have chosen to organize these different works according to their aims, that is *creating groups of words* - these groups may correspond to words linked to a word, or within a semantic class -, *adapting a classification* - a given semantic classification which is insufficient is used to determine a new one -, *constituting an ontology* - the words of a specialized language are structured in order to represent the relations that link their signifieds in the terminology of the considered domain -, *characterizing linguistic relations* - semantic classes that are linked within a given relation may be found.

Groups of words: The word "group" is vague because it covers different kinds of works. It both corresponds to studies which form sets of words that can be linked to each other by a lexical relation, or within a semantic class. For example, Grefenstette [9] presents a method to automatically generate a specialized thesaurus through a corpus analysis. For each most frequent nouns of his corpus, he determines the following information (relations): the nouns that are encountered in similar contexts within the corpus; the verbs of which the studied noun is typically the

subject or object; the compound nouns where this noun appears, and those which appear in similar contexts; the words that typically appear in the same texts than the considered noun. The technique which is used to determine the associations is quite simple: each noun gets an attribute vector which consists of a list of the adjectives with which it is associated, the verbs for which it appears as a subject or object, the nouns that are used is the same prepositional groups, etc. Similarity between vectors is calculated according to the Jacquard coefficient. For example, two nouns are linked if they both appear within the ten most strongly associated nouns with the other one. Quite similar techniques can also be used to get good semantic classes. For example, Agarwal [10] calculates for each noun a few attributes like the verbs for which it is the most frequently the subject or object, the preposition often used with it, etc. Then a classification is made that groups terms that share similar vectors of attributes. A class is submitted to an expert for validation purpose only if there exists in WordNet a synset which contains all its elements.

Adaptation of a classification: Works here try to adapt a given semantic classification to a specialized domain. One way is to discover more precise sub-classes within a class, or to combine existing classes to form a new classification. For example, Velardi et al. [11] use a hierarchical classification of general concepts of a domain and a set of relations between two general classes (for eg. *apply(action, animate entity)*); they then search complex terms within their corpus which are instances of a general relation (for eg. *apply(rearing, cow)*), and generalize the semantic classes of the constituents of the terms to obtain an intermediate representation (for eg. *apply(rearing, animal)*). This method permits to determine couples of semantic classes linked by a relation within complex terms of a domain.

Creation of an ontology: Some studies try to build a representation of the concepts of a domain of knowledge by bringing to light sets of words which represent similar concepts, or by detecting predicative links between elements of such a set. For example, Zweigenbaum et al. [12] first extract nominal terms from a corpus and then group those which share syntactic contexts. They build a graph whose nodes are the words, and arcs the syntactic contexts that the nodes share. Only the arcs that correspond to a certain number of contexts are kept; the sets that are obtained, and their relations, form an ontology of the domain.

Characterization of linguistic relations: These researches aim at characterizing semantic classes which are linked in a given linguistic relation. For example, Resnik [13] calculates a criterion, named *selectional association*, to determine, for a given predicate (verb) and a given argument (complement), the semantic class of the argument in a sentence. This criterion is based on the comparison between the probability for an element of the class *c* to be an argument of a predicate *p* and the probability for an element of the class *c* to be an argument of any predicate. It enables for example to determine when the word *baseball* is used as an object (the ball), a game or a diversion.

3.3 Update of the Lexicon

Updating a lexicon consists in adding, removing or modifying lexical entries, for example in order to adapt it to the vocabulary of a domain.

Some works focus on the determination of a preferential meaning for a word in a domain among all its possible meanings. For example, Resnik [14] tries to achieve this goal by determining, among the different synsets of WordNet in which two same nouns appear, the synset that is the best representative of their common points in a domain. He uses a measure of the information borne by each synset, which increases when one gets down in the is-a taxonomy of WordNet. This technique may be extended to groups of nouns, and may be used to facilitate the adaptation of WordNet to a specialized domain, for example by only keeping the synsets that are proved relevant.

Pustejovsky et al. [15] develop a means to incorporate a new unknown noun in a taxonomy of a lexicon. To discover its place, they suppose that its mother node is a noun that is the head constituent (rightmost, roughly speaking) in N N compounds in which the new noun is the modifier (leftmost constituent). This leads to a first list of candidates. Moreover, two nouns can be in a is-a relation if they appear as arguments of the same verbs. Therefore a determination of the mutual information between the new noun and verbs of which it is a direct object, and of the mutual information between the mother candidates and the same verbs permits to finally estimate the place of the unknown noun in the hierarchy.

3.4 Determination of Predicates

The meaning of a syntactically complex structure is generally expressed by a predicate, which is frequently associated with a verb. This predicate possesses an argument structure, that is a list of arguments with which a thematic role (agent, theme, etc.) is associated, and which must verify some semantic constraints. This argument structure dictates the role of each argument in the relation borne by the predicate. The studies about predicates try to automatically discover possible argument structures associated with given predicates from their uses in sentences. Some of them focus on the identification of the syntactic forms of these argument structures, which are named subcategorization frames; others try to find the semantic information linked to the arguments, or their thematic roles. For example, from a set of possible

subcategorization frames and a syntactic parser, Briscoe et al. [16] get the subset of possible frames for a given predicate. The only ones that are finally kept are those that successfully undergo a statistical filtering based on the binomial law.

4. CONCLUSIONS

Many natural language applications do need rich lexicons, and among the possible lexicons, semantic ones have proved to be particularly attractive [1, 2]. In this paper, we have shown that the different works that are currently developed in the field of automatic acquisition of lexical information, even if they do not aim at building lexicons, may really be relevant to facilitate the task of a lexicographer. In addition to the presentation of two new trends in natural language processing, and especially to the overview of a quite new domain, the interest of this text concerns the presentation of totally separated works within an homogeneous framework.

The integration of these researches into an automatic construction of semantic lexicons for applications can now be discussed. The question is: is it really possible to get a completely automatic version of a semantic lexicon for a new natural language application in a new domain, if we suppose that we can get enough electronic texts of this domain on the net? We have already seen that some parts of the whole can be obtained. The first step in building a lexicon is to circumscribe the vocabulary of the domain that is studied; this task can be achieved by terminology extractors. Moreover a corpus-based work can permit to bring to light lexical and semantic relations that will determine the structure of the lexicon, that is the relations between the lexical entries. One last task concerns the discovery of the information which enable to express the meanings that are attached to these lexical entries; studies on the determination of the predicates and on the way they express relations between the syntactically complex entities of the language can be helpful. However, the automatic construction of semantic lexicons is still a future prospect, and still has to prove its feasibility, and its interest by a precise comparison between a simple use of the techniques to assist a lexicographer and the global costs (in terms of data preparation, etc.) of a fully automatic version.

5. REFERENCES

1. Pustejovsky, J. (1995), *The Generative Lexicon*, MIT Press.
2. Miller, G. and Beckwith, R., and Fellbaum, C. and Gross, D. and Miller, K. (1990), "Five papers on WordNet", Technical report, Cognitive Science Laboratory, Princeton University.
3. Bourigault, D. (1992), "Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases", *Proc. Coling-92*, Nantes, Aug. 1992.
4. Bourigault, D. (1994), *Acquisition de terminologie*, PhD thesis, EHESS.
5. Fano, R. (1961), *Transmission of Information: A Statistical Theory of Communications*, MIT Press, Cambridge, MA.
6. Smadja, F. (1993), "Retrieving collocations from text: Xtract". *Computational Linguistics*, Vol.19(1), pp. 143-177.
7. Daille, B. (1994), *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques*, PhD thesis, Paris VII University.
8. Daille, B. (1994), "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology", In: *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, Workshop at the 32nd Annual Meeting of the ACL, Las Cruces, New Mexico, Jul. 1994.
9. Grefenstette, G. (1993), "Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques", MLTT 01, Rank Xerox Research Center, Grenoble.
10. Agarwal, R. (1995), *Semantic Feature Extraction from Technical Texts with Limited Human Intervention*, PhD thesis, Mississippi State University.
11. Velardi, P. and Fasolo, M and Pazienza, M.T. (1991), "How to Encode Semantic Knowledge: a Method for Meaning Representation and Computer-Aided Acquisition". *Computational Linguistics*, Vol. 17(2).
12. Zweigenbaum, P. and Bouaud, J. and Habert, B. and Nazarenko, A. (1997), "Coopération apprentissage en corpus et connaissance du domaine pour la construction d'ontologies", *Proc. Journées scientifiques et techniques*, Avignon, Apr. 1997.
13. Resnik, P. (1993), *Selection and Information: A Class-Based Approach to Lexical Relationships*, PhD thesis, University of Pennsylvania.
14. Resnik, P. (1995), "Disambiguating Noun Groupings with Respect to WordNet Senses", *Proc. 3rd Workshop on Very Large Corpora*, Cambridge, Jun. 1995.
15. Pustejovsky, J. and Anick, P. and Bergler, S. (1993), "Lexical Semantic Techniques for Corpus Analysis". *Computational Linguistics*, Vol. 19(2).
16. Briscoe, T. and Carroll, J. (1997), "Automatic Extraction of Subcategorisation from Corpora", *Proc. 5th ACL conference on Applied Natural Language Processing*, Washington, Apr. 1997.