AUTOMATIC ACQUISITION OF LEXICAL SEMANTIC INFORMATION USING MEDIUM TO SMALL CORPORA

Mathias Rossignol, Pascale Sebillot***

* International Research Center MICA, Vietnam

** IRISA, France

mathias.rossignol@gmail.com; Pascale.Sebillot@irisa.fr

ABSTRACT

Since many speech and text processing techniques are portable with a limited amount of work from one language to another, the most daunting task for NLP and SP practitioners becomes to build the resources needing for those tools to operate, In particular, the constitution of "high-level" resources, such as advanced corpus annotations or linguistically motivated lexicons, can be extremely work-intensive. We present in this paper a system to assist the creation of semantic lexicons using small to medium-sized corpora, thanks to the combination of semantic class constitution and topic detection, and the development of specific statistical data analysis techniques for relatively small datasets. By reducing the amount of data needed for semi-automatic semantic lexicon acquisition, traditionally applied to 100 million-word corpus or more, we make this help for lexical resource acquisition applicable to the case of under-resourced languages.

Index Terms— Semantic classes, small corpora, statistical data analysis, topic detection

1. INTRODUCTION

Despite the fundamental differences that may exist between languages, the state of the art in Natural Language Processing (NLP) and Speech Processing (SP) is such that many techniques can be applied with minimal adaptations to a wide range of languages, thus moving the burden of NLP and SP practitioners to he task of building the resources needed for those tools to operate. In particular, the constitution of "high-level" resources, such as advanced corpus annotations or linguistically motivated lexicons, can be extremely work-intensive, and has a lot to gain from the development of corpus-based automatic or semi-automatic methods to assist it. Moreover, in the case of under-resourced languages, where the person building the resource might not be a native speaker, corpus-based methods provide a guide and safeguard founded on actual language use. The work we present in this context, tackles the task of gathering semantic information for lexicon building, and more precisely on the constitution of "semantic classes" bringing together words of similar meanings.

Unfortunately, most of the current research in that field focuses on the development of techniques able to take into account huge amounts of textual data: in [1], for example, Lin and Pantel present very good results for the constitution of semantic classes using 50 to 100 million words from archives of the *Wall Street Journal*. More recently, Adam Kilgarrif [2] mentions the very impressive results obtained by Google rsearchers using a 250 *billion*-word English corpus, and urges the NLP community to ccoperate for the development of a several billion-word academic corpus... Such figures seem to make semi-automatic lexicon acquisition totally un-attainable for under-resourced languages, despite the fact that it is probably for these languages that it is the most useful, since they do not have any of the hand-built lexical resources, such as WordNet [3] or FrameNet [4], that are available for English. It is however possible to develop techniques that take better advantage of limited amounts of data, and we present in this paper such an approach.

Most of the important amount of research dedicated to the automatic corpus-based acquisition of lexical semantic resources focuses on the task of building semantic classes, that is, bringing together words of similar meanings. Two main "families" of techniques can be

distinguished: the first, in the wake of Z. Harris's linguistic studies [5], gather words used in a similar way in a corpus, under the assumption that the observed functional equivalence implies semantic proximity [6, 7, 1]; the second exploits specific linguistic patterns, such as enumerations, that bring together in the text words of similar meanings [8, 9]. The work we present in this paper can be connected with the first family, but proposes an original solution to the problem of "lexical variability", or the phenomenon by which a same idea can be expressed in many different ways, using different words, thus greatly complicating word usage comparison. Some authors have chosen to circumvent that difficulty by restricting their research to the case of specialized corpora (technical manuals, surgery accounts, *etc.*), in which there often exists a "canonical" way of expressing key concepts, and linguistic elegance is of secondary importance. Moreover, word polysemy, an important source of noise when comparing word usage, is very limited in a specialized context. Works like [10, 11] have shown very interesting results on such texts, but are not directly applicable to non-specialized corpora. Others choose to employ textual corpora of such sizes that lexical variability is "naturally overcome" thanks to the amount of data available—thus fueling the abovementionned phenomenon of corpus size inflation.

To the extent of our knowledge, no existing system can effectively cope with non-specialized, small to medium-sized corpora, although such data is very frequent: archives of a monthly magazine, complete works of an author, institutional transcripts, or, in the case of languages of recent digitization, a few months of archives of a daily newspaper. Moreover, many systems depend on a syntactic analysis of the studied corpora, which further limits their applicability to those languages for which syntactic parsers exist.

Our ambition is to address that limitation, and develop a generic methodology for the acquisition from medium corpora of semantic classes, as a first level of lexical data organization. The textual data we use for experimentation is composed of 14 years of archives of the French monthly newspaper *Le Monde Diplomatique*[1], a publication of great topical variety, geared towards analysis and reflexion, often featuring contributions by philosophers, sociologists, *etc*. It gathers approximately 11 million words, and has been morphosyntactically tagged and lemmatized. Its complex language and moderate size make it a good example of the "understudied" category of corpora we have mentioned. Although the experiments are carried on with French text for simplicity reasons, no language-specific method is employed, making the presented approach fully portable.

The following section presents our methodological choices and the structure of the system we present, whose operation is detailed in Sections 3 and 4. We expose sample results in Section 5 before concluding on the contribution, possible evolutions and potential applications of this work.

2. STRUCTURE OF OUR WORK

In order to reduce the problem of polysemy, by which a word may have different meanings in its various occurrences, thus making its usage data not representative of any meaning in particular, we choose to approximate text specialisation by topical homogeneity. Indeed, the knowledge of the topic of a sentence is often sufficient to raise any ambiguity about the meanings of its words: the meaning of *mouse* is self-evident if we know that it appears in a context dealing with computers. We therefore apply a topic characterization and detection method to the whole corpus to identify what topic(s) each of its paragraph deals with, and split it into topically homogenous subcorpora, with each of which is associated the corresponding lexical domain. The FAESTOS system, that we have developed for that task, is fully presented in [12]; we only quickly present it here in section 3.

We then split each lexical domain into semantic classes by bringing together words used in a similar way in the associated topical subcorpus. By first splitting the corpus into topical subcorpora and the vocabulary into domains, we let the studied words express distinct meanings in each of them, thus avoiding problems due to the irregular behaviour of polysemous words, and gathering what may be called *topicalized* lexical semantic information. Although that notion is akin to that of specialized lexical semantic information, it must be remembered that although topical filtering reduces word polysemy, the extracted subcorpora are still from all other points of view as linguistically complex as the complete corpus. That property, combined with their relatively small sizes, calls for the development of novel solutions to the problem of building semantic classes,

[1] This corresponds to $14 \times 12 = 168$ issues, less than six months of a daily newspaper.

that we present in section 4.

The two following sections are dedicated to the description of the developed method, beginning with a short overview of the domain constitution procedure.

3. BUILDING SUBCORPORA AND LEXICAL DOMAINS

A complete description of FAESTOS ("Fully Automatic Extraction of Sets of keywords for TOpic characterization and Spotting") can be found in [12], and we only describe here its purpose and capabilities.

Thanks to a series of statistical data analysis methods, FAESTOS builds a set of keyword classes, each of which characterizes a topic dealt with in the studied corpus and makes it possible to detect the occurrences of the topic by a simple keyword co-occurrence criterion. That is achieved totally automatically, without any foreknowledge of the topics dealt with in the corpus or external information of any kind.

On the *Le Monde Diplomatique* corpus, FAESTOS yields about 40 keyword classes of some 30 words each; for example, the keyword class for the topic we may call "information technology" contains words such as *computer*[2], *satellite*, *network*, *data*, *link*, *etc*.

Those classes let us split the initial corpus into topical subcorpora each containing all of the text segments of the complete corpus in which a given topic has been detected. With each of these subcorpora is associated the set of all words whose frequency in the subcorpus id at least double its average frequency in the whole corpus. This criterion lets us extract words prominently used—at least in one of their senses—to address a certain topic, a definition corresponding quite closely to that of lexical domains. We now detail the techniques we have developed to build semantic classes inside those domains.

4. BUILDING SEMANTIC CLASSES

This second stage of our "chain" of semantic lexical information acquisition methods addresses the problem of corpus-based constitution of semantic classes, to which much work has already been dedicated. More precisely, our technique belongs to the first "family" mentioned in the introduction, of works bringing together words used in a similar way. However, our situation of having to build those classes within domains leads to specific concerns, which we highlight in section 4.1. We then present the developed method in sections 4.2 and 4.3, and its results in section 5.

4.1. Problem specificities

Each of the studied words is likely to be used with a distinct meaning in each domain; therefore, when comparing word usages, we must only take into account the contexts where they are used with the meaning they have in the considered domain. This is most likely to be the case in the topical subcorpus corresponding to that domain; consequently, only the textual data contained in that subcorpus is used for word usage characterization. That decision raises serious problems because of the size of the topical subcorpora: on our 11 million word corpus, the most frequent topics give birth to subcorpora of some 500 000 words. Due to the importance of lexical variability, this amount of data is insufficient to efficiently compare word uses in a non-specialized (although topically homogenous) corpus.

In order to compensate for that lack of data, we define a 2-pass procedure. A first comparison of word usage is performed using the whole corpus, leading to the definition of an approximate "semantic distance" between words (section 4.2). That first information lets us "generalize" from usage data observed on topical subcorpora, thus allowing us to partially overcome the problems raised by lexical variability (section 4.3). In accordance with our goal of being able to exploit "ordinary" corpora, the developed methods do not make use of syntactic information.

4.2. Acquisition of a general "semantic distance"

Since it is quite "classical" in its principle, we only briefly describe here the procedure employed to compare words *via* their uses in the complete corpus. As in the subsequent stages of our work, our approach is a common one of statistical data analysis: defining a similarity measure between objects, whose values are stored in a similarity matrix then employed by a hierarchical classification method to build a classification tree.

Each word we wish to classify is characterized by the set of all full words[3] appearing within a window of

[2] For clarity and brevity, all the results and illustrations, initially obtained in French, are only presented in their translated English form.

[3] We use the term "full words" to designate nouns, verbs, adjectives and adverbs.

three words left and right of one of its occurrences. The similarity between two words is computed as a Jaccard measure between those sets of neighbours, followed by a normalisation procedure to center and reduce all lines and columns of the similarity matrix.

A classification tree is built by an average-link algorithm, and the "semantic distance" between words, henceforth noted d, is finally defined as an ultrametric between objects on the classification tree: $d(m_1, m_2)$ is defined as the $\log_2$ of the smallest class in the tree containing m_1 and m_2. Its values are normalized to 1, making it possible to use it simply as a "semantic proximity" $(1 - d)$.

That procedure is applied separately to all nouns, verbs, adjectives and adverbs appearing more than 100 times in the full corpus. We then have at our disposal four sets of semantic distances that, imperfect as they may be, constitute a very useful first element of semantic information we now use for the comparison of word usage within topical subcorpora.

4.3. Semantic similarity between words on a topical subcorpus

When comparing word uses on topical subcorpora, our goal is to fully exploit the scarce textual data available. From that point of view, the "bag of words" method we have used to compare word usage on the whole corpus is clearly not optimal: bringing together all the neighours of a word in a single set entails a considerable loss of information about the structure of the studied contexts, and what neighbours appear together.

In order to keep that last piece of information, we first compute a similarity between contexts considered individually and then use it to extrapolate a similarity between words. The methodological advantage of that approach is important, since it allows for a clear separation between the linguistic process of comparing individual text segments (contexts), and the statistical work necessary to generalize from the "context" level to the "word usage" level. Moreover, the extrapolation procedure we have developed (section 4.3.2) can be used with any similarity measure defined between contexts: the one we use and present here in section 4.3.1 is voluntarily simple and does not make use of any syntactic information, but many improvements are possible in this area.

In the remainder of this paper, we shall focus on the case of nouns, as the most commonly studied words in lexicography.

4.3.1. Individual context comparison

The similarity measure we define makes use of a simple representation of contexts: each of them is composed of the two sets of full words appearing in a four word window left and right, respectively, of the considered word occurrence.

The similarity distance d previously computed is used to "generalize" these characteristic word sets by treating them as "fuzzy sets" where each word considered in the calculation of d has an "intensity of belonging" equal to its greatest proximity with the words actually present in the set. A "cardinal of fuzzy intersection" cfi between two word sets E and E' is thus defined by:

$$\text{let } h\left(E, E'\right) = \sum_{w_1 \in E} \max_{w_2 \in E'} \left(1 - d\left(w_1, w_2\right)\right)$$

$$cfi\left(E, E'\right) = \frac{h\left(E, E'\right) + h\left(E', E\right)}{2}$$

The similarity between two contexts is simply defined as the sum of those "cardinals" between their respective left and right word sets, without any normalization: we choose to consider that each additional common word increases the semantic similarity, but that non common words do not necessarily increase dissimilarity. Therefore, two contexts of five words having three words in common have a stronger semantic similarity than two contexts of one word each having "all their words" in common.

The similarity between contexts we have defined, henceforth noted s, is used as the basis for noun usage comparison, as we now describe.

4.3.2. From context similarity to word similarity

From the definition of semantic classes as "groups of interchangeable words", we can infer that two nouns w_1 and w_2 are semantically similar if, for each context where w_1 is observed, we can find at least one similar context featuring w_2. The similarity S between w_1 and w_2 should therefore be computed by considering, for each context c_{1i} $(1 \leq i \leq n_1)$ of m_1, the one that is most similar to it amongst the contexts of m_2 (c_{2j}, $1 \leq j \leq n_2$). Hence the following formula for S:

$$S(w_1, w_2) = \frac{1}{n_1}\left[\sum_{i=1}^{n_1} \max_{1 \le j \le n_2} (s(c_{1i}, c_{2j}))\right]$$

The proposed measure is asymetrical, which calls for an averaging of reciprocal values, but more importantly, it is very sensitive to variations in numbers of occurrences of the compared nouns. To overcome that difficulty, we use a method inspired by statistical random sampling techniques such as Monte Carlo [13].

With each noun w we wish to study is associated a population of "sample-words" characterized by a selection of contexts randomly drawn from the contexts of w, all sample-words being defined by the same number of contexts. When comparing two nouns, the similarities between all possible pairs of sample-words associated with them are computed using the expression given for S above, this time without any difficulty due to the variations in numbers of occurrences. The similarity between the compared nouns is then obtained by averaging all those values, thus resulting in a naturally normalized similarity measure.

As when working on the complete corpus, the semantic similarities between words are exploited by a classical average-link hierarchical classification method to build a classification tree of words, which we now present and explore.

5. RESULTS

Figure 1 presents an excerpt of the classification tree obtained on the 300 most frequent nouns of the domain "information technology" (appearing more than 20 times in the corresponding subcorpus).

As can be seen, the classification tree proposes many relevant word groupings; a particularly interesting result is the proximity between *highway* and *infrastructure*, showing that *highway* is clearly understood here with the meaning it takes in the expression "information highway", now rather deprecated but used in our corpus to designate the Internet infrastructure. Similarly, *line* is classified in a way consistent with its "phone line" meaning, and *operator* is considered as a "telecommunication operator". The topical restriction we have imposed on word usage does seem to play its role and allow the sense of the word relevant to the studied domain to be prominent in the considered data.

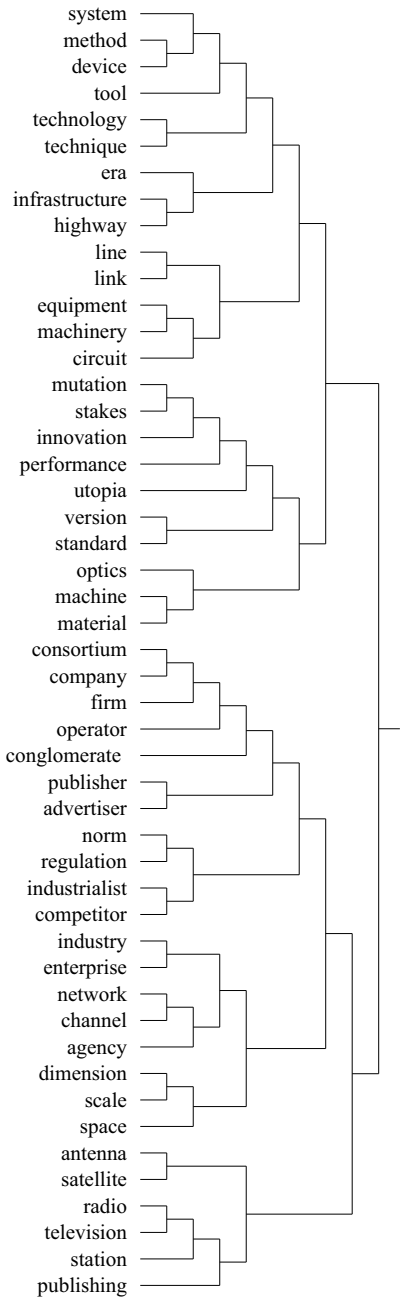

Fig. 1. Excerpt from the classification tree obtained on the 300 most frequent nouns of the domain "information technology"

We could not find a criterion to automatically extract semantic classes from the classification tree, and it is therefore up to the user to perform that selection manually. On average, the classification trees obtained thanks to the presented method give birth to semantic classes gathering about 60 % of the studied nouns.

6. CONCLUSION

We have presented a generic methodology for the acquisition of lexical semantic information from non-specialized, medium sized textual corpora that, thanks to a "topicalisation" of lexical semantic information acquisition, bridges a gap between works on small, specialized corpora and very large, non-specialized ones.

The method we have defined to compare word usage by first defining a similarity between individual contexts and then exploiting it to compare words opens many possibilities for the definition of more sophisticated context comparison techniques. Indeed, it makes it possible to specify arbitrary criteria taking into account whatever particular knowledge of a given corpus or language may be available.

Finally, the two-pass procedure we have developed for the constitution of semantic classes can easily be adapted to replace the "semantic distance" acquired on the whole corpus by external, *a priori* generalist semantic information, thus making it possible to extract topicalized lexical semantic information from small, isolated topically homogenous corpora, and not only subcorpora of a larger non-specialized corpus. Since the "semantic distance" does not need to be highly accurate, that seems to be an especially promising direction for under-resourced languages, for example for the structuration of task-oriented lexicons.

7. REFERENCES

[1] D. Lin and P. Pantel, "Induction of Semantic Classes from Natural Language Text," in *7th Int. Conference on Knowledge Discovery and Data Mining (SIGKDD 01)*, San Francisco, CA, USA, 2001.

[2] A. Kilgarriff, "Googleology is Bad Science," *Computional Linguistics*, vol. 33, no. 1, pp. 147–151, 2007.

[3] C. Fellbaum, Ed., *WordNet, an Electronic Lexical Database*, MIT Press, Cambridge, MA, EU, 1998.

[4] C. F. Baker, C. J. Fillmore, and B. Cronin, "The Structure of the Framenet Database," *Int. Journal of Lexicography*, vol. 16, no. 3, pp. 281–296, 2003.

[5] Z. Harris, *Mathematical Structures of Language*, John Wiley & Sons, New York, NJ, USA, 1968.

[6] D. Hindle, "Noun Classification from Predicate-Argument Structures," in *28st Annual Meeting of the Association for Computational Linguistics (ACL 90)*, Pittsburgh, PA, USA, 1990.

[7] G. Grefenstette, "Automatic Thesaurus Generation from Raw Text Using Knowledge-Poor Techniques," in *Making Sense of Words, 9th Annual Conference of the UW Centre for the New* Oxford English Dictionary *and Text Research*, Oxford, UK, 1993.

[8] M. A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," in *14th International Conference on Computational Linguistics (COLING 92)*, Nantes, France, 1992.

[9] E. Riloff and J. Shepherd, "A Corpus-based Bootstrapping Algorithm for Semi-Automated Semantic Lexicon Construction," *Natural Language Engineering*, vol. 5, no. 2, pp. 147–156, 1999.

[10] B. Habert, E. Naulleau, and A. Nazarenko, "Symbolic Word Clustering for Medium-Size Corpora," in *16th Int. Conference on Computational Linguistics (COLING 96)*, Copenhagen, Denmark, 1996.

[11] V. Claveau and M.-C. L'Homme, "Discovering Specific Semantic Relationships between Nouns and Verbs in a Specialized French Corpus," in *3rd Workshop on Computational Terminology (CompuTerm'04)*, Geneva, Switzerland, 2004.

[12] (anonymized), "Combining Statistical Data Analysis Techniques to Extract Topical Keyword Classes from Corpora," *IDA (Intelligent Data Analysis)*, vol. 9, no. 1, pp. 105–127, 2005.

[13] B. Efron and R. Tibshirani, "Statistical Analysis in the Computer Age," *Science*, vol. 253, pp. 390–395, 1991.