Mining the Web for Sense Discrimination Patterns

R. Guzmán-Cabrera[1,2], P. Rosso[1], M. Montes-y-Gómez[3] and J. M. Gómez-Soriano[1]

[1] *Departamento de Sistemas Informáticos y Computación*
Universidad Politécnica de Valencia, Spain
[2] *Facultad de Ingeniería Mecánica, Eléctrica y Electrónica*
Universidad de Guanajuato, Mexico.
[3] *Laboratorio de Tecnologías del Lenguaje*
Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico.

Emails: guzmanc@salamanca.ugto.mx; prosso@dsic.upv.es; jogomez@dsic.upv.es; mmontesg@inaoep.mx

ABSTRACT - ***In this paper we present a method for mining the Web in order to extract lexical patterns that help in discriminating the senses of a given polysemic word. These patters are defined as sets and sequences of words strongly related to each sense of the word. To discover the patterns, the method first determines the different senses of the word from a reference lexical database, and then it uses the set of synonyms from each sense as search patterns on the Web. The purpose is to create a corpus of usage cases per sense, downloading snippets via fast search engines. Finally, it applies a well-known association discovery data mining technique to select the most relevant lexical patterns for each word sense. The preliminary results indicate that making sense out of the Web is possible and the discovered patters should be of great benefit in tasks such as information retrieval and machine translation.***

1. INTRODUCTION

With the so-called information society every day the quantity of stored information multiplies, which involves an increase in the difficulty of processing this information with classic methods. To overcome this problem, in the last years a series of techniques have arisen. For instance, data mining that facilitate the prosecution and the analysis of information in an automatic way. The idea is based on the fact that the data contains more hidden information of that it is seen to simple sight. Therefore, data mining can be defined as the no trivial extraction of information implied, previously not acquaintance and potentially useful, from the data (Frawley, 1992). Web mining, on the other hand, focuses in the use of techniques of data mining to automatically discover and extract information from documents and services of the Web (Etzioni, 1996).

In Natural Language Processing (NLP) the use of corpora –huge collections of textual data– is important to extract language models: a list of word combinations that allows knowing the words frequently used together and the words belonging to a certain domain.

The use of the Web as corpus has great advantages for NLP tasks (Kilgarriff, 2003). Basically, the Web is an easy and fast way to access a great variety of stored information in electronic format in different parts of the world. There are different investigations that have been realized using Web as linguistic resource (Brill, 2001) (Solorio, 2004) (Grefenstette, 1999) (Bunescu, 2003) (Volk, 2001). For the case of Word Sense Disambiguation (WSD), (Mihalcea, 2004) uses the redundancy of the Web as source of knowledge for all kinds –supervised and unsupervised– of WSD systems. Whereas other authors, as (Celina, 2003), use the Web to enrich labeled corpora, which later facilitate the task of WSD. Nevertheless, the Web has several negative aspects. It is very heterogeneous and disorganized; also a lot of useless information exists. Furthermore it is not possible to be sure that everything found is correct, since nobody checks it. But thanks to the redundancy of the Web the correct information predominates.

In the present work we use the Web as a corpus and apply Web mining techniques to extract interesting relations between words. Basically, we focus in the extraction of lexical patters related to the different senses of a given polysemic word. These patterns are combinations of words that frequently co-occur and that correspond to a particular sense of the reference word. They are of two basic types: continuous strings of words, i.e., *sequences*, or sets of isolated words, i.e., *associations*. Both kinds of patterns are common in all languages, types of writing, and topical areas.

We expect the extracted patterns to enhance the performance of WSD systems, and also to contribute to other NLP applications such as: conceptual information retrieval (Montes, 2000), text classification (Kosala, 2000), and automatic translation (Smrz, 2001).

The following sections are organized as follows. Section 2 describes the general methodology for discovering the word sense discrimination patterns from the Web. Section 3 shows the results from the analysis of the polysemic word *peak*. Finally, the section 4 presents our conclusions and depicts the future work.

2. METHODOLOGY

The extraction of lexical disambiguation patterns from the Web considers three main steps:

1. The construction of a corpus of prospective usage examples for word sense.
2. The extraction of all restricted-size lexical patterns contained in the corpora.
3. The selection of the most relevant and discriminating patterns (associations and sequences of words) for each one of the word senses.

The following sections describe some details on these tasks.

2.1. Corpus construction

The purpose of this first step is to construct a corpus of usage examples for each sense of the given word. The word must be polysemic and must exist in a reference lexical database (e.g. WordNet for Englsih (Miller, 1995)). The lexical database allows determining the different senses attributable to the word and obtaining their corresponding set of synonyms. The set of synonyms are used as search patterns in the Web. In our case we are using Google as search engine, though recent investigations show that the results on word sense disambiguation do not depend much on the selected search engine (Rosso, 2005). The snippets returned by the search engine are downloaded and joined with the results of the rest of the synonyms of the same word sense, forming one corpus per sense. The figure 1 illustrates this process.

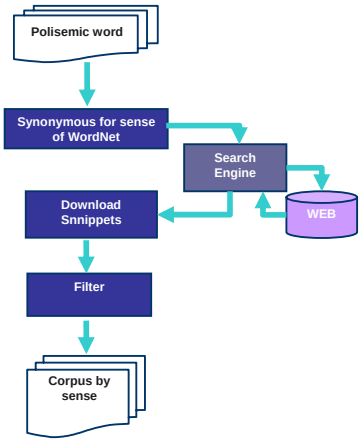

Figure 1. - Corpus construction

It is important to notice that the examples of the created corpora not necessarily correspond to the indicated word sense; they just contain some synonym of a sense of the given word. Further analysis is required to select the real examples and patterns for a specific word sense.

2.2 Pattern extraction

In this stage all the word associations and word sequences of given maximum size are extracted from the corpora.

In order to be interesting, a lexical association must be defined as a set of one or more words occurring at the neighborhood of a synonym of the word sense, while a lexical sequence must be a chain of one or more words linked to synonym of a specific word sense.

The extraction of the lexical associations is based on the use of traditional data mining techniques. In particular we adapted the well-known “a priori” algorithm for the association rule discovery (Agrawal, 1999).

The extraction of lexical sequences also considers several ideas from data mining. It is based on an iterative procedure that allows finding all maximal sequences (not included in any other sequence), of a maximum specified size. Details on the methods for association and sequences extraction are in (Guzmán, 2005).

2.3 Pattern selection

In order to determine the most relevant patterns per sense we applied the following criteria.

Strength measure

This measure is based on the frequency of occurrence of the pattern (sequence or association) in a sense corpus. It is defined as follows (Smadja, 1993).

$$S_P = \frac{f_P - f_\mu}{\sigma} \qquad (1)$$

where f_P indicates the frequency of the pattern P in a reference sense corpus, f_μ the average frequency of all patterns in this corpus, and σ their standard deviation. The score S_P is the strength of the pattern P.

Considering the patterns with strength greater than one, we assure the extraction of only those patterns highly related with a specific word sense, eliminating those appearing just by chance.

Dispersion levels

It is not sufficient to compute the strength of a pattern to select the most relevant ones. It is also necessary to consider the dispersion of the patterns among the whole set of synonyms of the each word sense, as well as their dispersion among the different senses of the word. Here are two basic assumptions:

Internal dispersion assumption: a pattern occurring within the near context (predefined window) of all (or the majority of) the synonyms of a word sense tends to be more relevant for that sense that a pattern happening with just some synonyms.

External dispersion assumption: a pattern happening in only one (or in a few) sense corpus tend to be more relevant for that sense that a pattern equally distributed in all senses.

The internal and external dispersions of a pattern indicates, in some degree, it qualitative condition. Used in combination with the strength measure –a quantitative characteristic of the pattern–, the dispersion measures ensure the selection of high quality patterns per sense.

3. RESULTS

In order to demonstrate our method we analyze the word *peak*. This word is extremely polysemic (it has 7 different senses) and has several synonyms per sense (5 synonyms per sense, in average).

Here is a brief description of the senses of the word peak. This description was taken from the WordNet database.

1. extremum, peak -- (the most extreme possible amount or value)
2. flower, prime, peak, heyday, bloom, blossom, efflorescence, flush -- (the period of greatest prosperity or productivity)
3. acme, height, elevation, peak, pinnacle, summit, superlative, top -- (the highest level or degree attainable)
4. peak, crown, crest, top, tip, summit -- (the top point of a mountain or hill)
5. point, tip, peak -- (a V shape)
6. vertex, peak, apex, acme -- (the highest point (of something))
7. bill, peak, eyeshade, visor, vizor -- (a brim that projects to the front to shade the eyes)

Table 1 shows some average data about the experiment. It is important to notice that even when we downloaded many usage examples for each word sense, they were insufficient. The average occurrence of a word was in most cases less than 5. As a consequence we could extract just a few relevant patterns for sense. The relevance criteria, specially the dispersion conditions, seemed to be very rigorous.

Table 1. - Statistics for the word peak.

Sense	1	2	3	4	5	6	7
Examples of use	7624	42066	49196	26895	11684	32772	22279
Different words	2279	7881	8359	5770	3127	6202	5516
Average	3.3	5.3	5.9	4.7	3.7	5.3	4.04
Standard deviation	4.9	10.9	10.1	9.1	6.6	8.2	7.75
Relevant associations	10	5	4	10	4	63	3
Relevant sequences	7	4	5	5	4	15	6

Table 2 presents a list of some lexical associations related to the different senses of the word peak.

Table 2.- Associations

Sense	Words
1	{global, conditions, large, educations}
2	{most, series, time, sites}
3	{America, great, die}
4	{sports, college, university, science}
5	{jobs, magazine, district}
6	{performance, standard, class}
7	{accessories, bill}

4. CONCLUSIONS

In this paper we present a method for extracting sense discrimination patterns from the Web. The method allows finding lexical associations and sequences for each sense of a given polysemic word. Our first experiments showed the potential of the Web as linguistic corpus. Our principal contribution is the search in the Web for (quantitative and qualitative) relevant patterns for each word sense. The preliminary results presented in this work are for English nouns, nevertheless, this methodology can be applied to other syntactic categories, as well as to other languages, providing that in those languages a lexical database exists.

At present, the dispersion conditions are very rigid, since the patterns must be in the context of all the synonyms that compose the sense and just in that sense, causing that many relevant patterns stay out of the analysis. For this reason it is desirable to implement a weighting scheme of dispersion that allows considering a pattern even if it does not appear in the context of all the synonyms of the sense or in just one sense.

Acknowledgments

The work was partially supported by the R2D2 (CICYT TIC2003-07158-C04-03), ICT EU-India (ALA/95/23/2003/077-054), Conacyt - Mexico (J43990-Y) and PROMEP (UGTO-121).

REFERENCES

Agrawal, R., and Srikant, R., 1994, Fast algorithms for mining association rules, VLDB-94.

Brill, E., Lin, J., Banko, M., Dumais, S., and Ng, A., 2001, Data-intensive Question Answering. Proceedings of the Tenth Text REtrieval Conference TREC-2001.

Bunescu, R., 2003, Associative Anaphora Resolution: A Web-Based Approach. Proceedings of the EACL-2003, Workshop on the Computational Treatment of Anaphora, Budapest, Hungary.

Celina, S., Gonzalo, J., and Verdejo, F., 2003, Automatic association of web directories with word senses, *Computacional Linguistics*, Volume 29, Number 3, pp.485-502.

Etzioni, O., 1996, The World Wide Web: Quagmire or Gold Mine?, *Communications of the ACM*, Vol.39, No.11, pp. 65-68.

Frawley, W., and Piatesky-Shapiro, G., 1992, Knowledge Discovery in Databases: An Overview, *AI Magazine*, pp. 213-228.

Grefenstette, G. 1999, The World Wide Web as a resource for example-based Machine Translation Tasks. Proceedings of Aslib Conference on Translating and the Computer. London.

Guzmán-Cabrera, R., Montes -y-Gómez, M., and Rosso, P.., 2005, Búsqueda de Colocaciones en la Web Para Sinónimos de WordNet. *Acta Universitaria*. Accepted.

Kilgarriff, A., and Greffenstette, G., 2003, Introduction to the Special Issue on Web as Corpus, *Computational Linguistics*, 29(3), pp.1-15.

Kosala, R., and Blockeel, H., 2000, Web Mining Research: a survey, *SIG KDD Explorations*, Vol. 2, pp. 1-15.

Mihalcea, R., 2004, Making Sense Out of the Web, Workshop on Lexical Resources and the Web for Word Sense Disambiguation, IBERAMIA, Mexico.

Miller, A., 1995, Wordnet: A lexical Database for English, *Comunications of the ACM*, 38 (11), pp 39-41.

Montes-y-Gómez, M., López-López, A. and Gelbukh, A., 2000, Information Retrieval with Conceptual Graph Matching, Proceedings of the 11th International Conference on Database and Expert Systems Applications DEXA-2000, Springer-Verlag.

Rosso, P., Montes, M., Buscaldi, D., Pancardo, A., and Villaseñor, A., 2005, Two Web-based Approaches for Noun Sense Disambiguation. In: Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2005, Springer Verlag, LNCS (3406), Mexico D.F., Mexico, pp. 261-273.

Solorio, T., Pérez, M., Montes, M., Villaseñor, L., and López, A., 2004, A Language Independent Method for Question Classification. Proceedings of the 20th Int. Conf. on Computational Linguistics (COLING-04). Geneva, Switzerland.

Smadja, F., 1993, Retrieving collocations from text: Xtract, *Computational Linguistics*, 7(4),pp. 143–177.

Smrz, P., 2001, Finding Semantically Related Words in Large Corpora, FIMU Report Series:Masaryk University.

Volk, M., 2001, Exploiting the WWW as a Corpus to Resolve PP Attachment Ambiguities. Proceedings of Corpus Linguistics, Lancaster.