ISSN 1870-4069

Search for Places on the Web for Wordnet Synonyms

Rafael Guzmán Cabrera

Universidad de Guanajuato,
Facultad de Ingeniería Mecánica, Eléctrica y Electrónica,
Mexico

guzmanc@salamanca.ugto.mx

Abstract. In this paper, we present a method that allows us, for a polysemic word given in English, to find collocations linked in a significant way with each of its senses. The synonymous that compose every sense are used as pattern of search in the Web with the purpose of creating a corpus by sense. Finally, we apply techniques of data mining that allow us to select the most relevant collocations or lexical patterns for every sense. We especially in two types of collocations especially: associations and lexical sequences. The obtained results show us that we can find collocations between words using the Web as linguistic corpus, as well as the feasibility of incorporation of the lexical pattern obtained in systems of word sense disambiguation that can be used in turn for example in machines of translation and of information retrieval.

Keywords: Text mining, text recognizing, WordNet synonyms.

1 Introduction

With the so-called information society, the amount of stored data increases daily, which leads to an increase in the difficulty of processing this information using traditional methods. To overcome this problem, a series of techniques and tools have emerged in recent years that facilitate advanced data processing, as well as in-depth automatic data analysis. One of these tools is data mining, whose key idea is that data contain more hidden information than meets the eye. Data mining can be defined as the non-trivial extraction of implicit, previously unknown, and potentially useful information from data (Frawley, 1992). Web mining, on the other hand, focuses on the use of data mining techniques to automatically discover and extract information from Web documents and services (Etzioni, 1996). Web mining can be classified into three main areas:

- Web usage mining: this method attempts to extract information (habits, user preferences, or content and relevance of documents) from the sessions and behavior of users and browsers, that is, it allows for the discovery of website access patterns.
- Structured Web mining: Focused on discovering a model from the topology of network links. This model can be useful for classifying or grouping documents.

pp. 5–14; rec. 2010-05-11; acc. 2010-08-01

- Web content mining: Allows you to find common information from web documents. It can be further classified as:
- Text mining: If the documents are in plain text.
- Hypertext mining: Whether documents contain links to other documents or to themselves.
- Markup mining: if the documents are structured, that is, with marks.
- Multimedia mining: if the documents contain images, audio or video.

This paper focuses on the latter, mining web content, using the web as a linguistic corpus for extracting collocations between words. A corpus is a collection of text in electronic format that can be processed by a computer for various purposes, such as linguistic research and language engineering (Leech, 1997).

Corpus linguistics involves the study of languages based on examples of their use. We used Google as our search engine, although recent research conducted to study the possibility of using the web to disambiguate nouns preceded by an adjective (Rosso, 2005) shows that the results do not depend much on the search engine used. Several investigations have been carried out using Web mining as a tool.

For example, (Mihalcea, 2004) presents the main lines of research regarding the exploitation of the Web as a linguistic resource in Word Sense Disambiguation (WSD) systems.

Furthermore, (Celina, 2003) has used the Web to enrich tagged corpora, which then facilitate the task of WSD. There are several reasons for using the Web as a linguistic resource, among them that it provides a means of quickly and easily accessing a wide variety of information stored in electronic format in different parts of the world. Furthermore, it is free and available with a mouse click.

The size of the Web in July 1999 was estimated at 56 million addresses, 125 million in January 2001, and 172 million in January 2003. This represents a massive growth of over 300% in just under five years. In 1999, 800 million indexed Web pages were available; if we estimate the average Web page size to be between 7 and 8 kilobytes of plain text, we have nearly 6 terabytes of available text in 1999 and approximately 30 terabytes in 2003.

With these figures, the Web is clearly an immense corpus, given the amount of information available to us. Furthermore, the Web is multilingual, since approximately: 71% of the pages are written in English, 6.8% in Japanese, 5.1% in German, 1.8% in French, 1.5% in Chinese, 1.1% in Spanish, 0.9% in Italian, and 0.7% in Swedish, the remaining 11.1% is spread across other languages and dialects (Kilgarriff, 2003). In Natural Language Processing (NLP) research, the use of corpora is important to extract language models, such as combinations of meaningful words that allow knowing which words are related to each other or which of them are from a certain domain.

However, the Web has several negative aspects, including the fact that the information found is very heterogeneous and disorganized, and there is a lot of junk information or information with tags that make it difficult to process. In addition, you can't be sure that everything you find is correct, since no one checks it. But thanks to the Web's redundancy, accurate information usually prevails.

The task of automatically finding semantic relationships between adjacent words has attracted the attention of many researchers in the field of natural NLP in recent decades. As a result of this research, important dictionaries of English collocations have been written, for example, the one created by (Benson, 1989).

There are also developed systems that allow analysis between collocations, such as the N-Gram Statistics Package. This system allows for the analysis of words in a corpus, such as the counting of frequent occurrences and various statistical measures that allow for an association between two or more words.

The possible applications of collocation detection and its relationship to the lexical meaning of the sentence are many and varied. Examples include translation from one language to another and the integration of lexical patterns into WSD systems.

In Section 2, we describe the method used to find meaningful collocations on the Web for a given word. In Section 3, we present the results of the experiments and some examples of the lexical patterns found by meaning. Finally, in Section 4, we present the conclusions of this work.

2 Methodology

In natural language, there are many combinations of words that frequently co-occur and correspond to a particular use of a word or a sense of a sentence. These combinations can be presented as an unbroken sequence of words, in this case simply called sequences, or the words in the combination may not occur contiguously in context, in this case called associations. Sequences and associations are two of the most important types of word collocations. The method used to obtain meaningful collocations consists of the following steps:

1. WordNet Senses. For a given polysemous word in English, we obtain its meanings from the WordNet lexical database, which is a lexical-conceptual database for English structured as a semantic network, so that access to lexical information is not restricted to merely alphabetical access. This is inspired by psycholinguistic theories about human lexical memory. WordNet stores information on words belonging to the syntactic categories of noun, verb, adjective, and adverb. The cost of having syntactic categories is a large amount of redundancy that conventional dictionaries lack.
2. Synsets. For each sense, we obtain a set of synonyms (synsets); Words in WordNet are organized into sets of synonyms, or synsets, each of which represents a different lexical concept. Each synset contains the list of synonymous words, as well as information on semantic relationships established with other words or synsets.
3. Snippets. Using each synonym (synset element) as a web search pattern and given a search engine (e.g., Google), we download snippets. When we perform a query on the web, the search engine provides us with a response, if the request is found, a set of web pages that match the request, as well as a brief summary of the content of each page so the user can choose the address that best matches the request. Each

set of addresses and summary provided by the search engine, stored in a plain text file, is a snippet. In our case, they represent commonly used expressions from WordNet synonyms for the given polysemous word.

4. Corpus by meaning. The snippets of all synonyms for a given sense are filtered and concatenated into a single file; that is, we create a WordNet corpus per sense. The corpus per sense is constructed by taking five context words from the right and five words from the left of the synonym.
5. Extraction of collocations. We found all significant collocations (sequences or associations) that occur in each corpus by meaning with the following criteria:

- Force. A sequence or association is relevant if it is frequent, that is, if it occurs more than a predetermined threshold or cutoff value; and is defined by:
- The threshold or cutoff frequency. It is defined as a value equal to the sum of the average frequency and the standard deviation and is the minimum frequency that words must have to pass this threshold. This ensures that only those occurrences that appear frequently in the contexts of WordNet's meaning are extracted, eliminating all words that may appear randomly.
- Local dispersion. To find the set of sequences and associations that are representative of each of the WordNet senses of a given word, it is desirable that, in addition to having passed the previous frequency-based filter, they be in the context of all the synonyms that make up the sense. This is precisely what the local dispersion measure does; it allows us to discard those words that are not in the context of all the synonyms that make up the WordNet sense.
- External dispersion. With this measure, what we seek is that the set of words that exceed the two previous measures is found only in the context of one of the senses, that is, all those words that appear in more than one sense are discarded.

The methodology described in these steps is applied to the extraction of lexical associations and uninterrupted sequences. We now present a description of each of these types of collocations.

Lexical associations.-Lexical associations are a set of words significantly linked to one of the WordNet senses of a polysemous word. To extract these, the synonyms that make up the WordNet sense in the corpus are identified. For each synonym found, the context words are entered into a table. Basically, for each context word found, the question "Does it exist in the table?" is asked. If the answer is no, it is included and its frequency value is initialized to one, while if the answer is yes, its frequency is increased by one. In this way, we go through the entire corpus. From the resulting table, we must select those context words that exceed the strength measure and the dispersion measures mentioned in point 5 of the methodology. It is worth mentioning that the words found in the table are not required to be adjacent to the synonym or between them; their position within the context varies within the defined window. In this way, we find those words that are significantly linked to the meaning.

Sequences.- To obtain uninterrupted word sequences, an automatic iterative process is performed that changes the window size, that is, the number of words taken to the left and right of the synonym, from one to five. For each window size, the process is as

Table 1. Number of snippets downloaded from the Web for the synonyms of Instance.

case	919
instance	924
example	983
illustration	987
representative	987

follows: the word or words are taken respecting their location relative to the synonym, that is, whether they are on the left or right. The result is a set of tables showing the context sequences to the left and right of the synonym for the different window sizes.

Statistics are obtained for each one. As with lexical associations, the resulting uninterrupted sequences are filtered, and only those that are significant in the corresponding WordNet sense are recovered. In this case, the words that make up the uninterrupted sequence are contiguous to each other, respecting their position within the context.

Sequences and associations are found using the redundancy of the Web as a corpus. This is intended to allow future incorporation of the results obtained into WSD systems, whose objective is to associate a word, given in a context, with a definition or meaning that distinguishes it from other meanings attributable to that word. Any NLP system requires a module with these characteristics. WSD is not an end in itself, but rather a necessary step for performing actions such as syntactic analysis or semantic interpretation in NLP tasks, as well as for the development of final applications such as information retrieval (Montes, 2000), text classification (Kosala, 2000), discourse analysis (Montes, 2002), and machine translation (Smrz, 2001), among others.

For example, a traditional information retrieval system will answer the question "What plants live in the desert?" with all documents containing the terms "plants" and "desert," regardless of their meaning. In some of these documents, the term "plant" would appear with the meaning of "living being," while in others, it would mean "industry." If the information retrieval system were able to distinguish the meanings of the query terms, it would return only the documents that use the meaning of "living being." To do this, the system must integrate a WSD module to disambiguate both the query terms and the terms in the indexed documents.

3 Results

We show the results obtained when applying the methodology to the word "instance." We chose this word because, in addition to being polysemous, it has two common synonyms among its meanings, which will allow us to observe the effect of the dispersion measures on the lexical patterns obtained. The WordNet meanings of "instance," as a noun, are:

1. case, instance, example -- (an occurrence of something)
2. example, illustration, instance, representative -- (an item of information that is representative of a type)

Table 2. Abstract of statistics for instance.

	Sense 1	Sense 2
Word: instance	Web	Web
Number of usage examples in the corpus	12684.0	15848.0
Number of distinct words	2831	3590
Average	4.5	4.4
Standard deviation	7.3	7.5
Cut-off frequency (mean + standard deviation)	11.8	11.96
Number of words that exceed measure 1	179	238
Number of words that exceed measure 2	87	67
Number of words that exceed measure 3	25	9

Table 3. Simple lexical associations for Instance

Sense 1		Sense 2
based	java	English
case	learning	free
code	multiple	government
data	net	library
date	number	Link
definition	org	members
documents	our	resources
example	process	section
examples	proposal	software
file	server	
index	use	
instance	will	
It		

Using synonyms for each sense as a Google search pattern, web snippets were downloaded containing examples of context of use, that is, commonly used expressions of the synonyms that make up the corresponding WordNet sense. The number of snippets downloaded per synonym is shown in Table 1.

These snippets formed two corpora, one for each sense. The corpus for sense 1 was formed by concatenating 2,826 snippets, while the corpus for sense 2 was formed with 3,881 snippets; this number is higher because it includes one more synonym. Table 2 shows a summary of the results obtained. In the corpus for sense 1, 12,684 examples of common context use of the synonyms that comprise it were found, while in the corpus for sense 2, 15,848 examples of context use were found.

A total of 2,831 different context words were found in the 12,684 usage examples for sense 1. Of these, 179 exceed the threshold frequency (frequency greater than the average frequency plus the standard deviation). The words that, in addition to having exceeded the threshold frequency, are found in the context of all synonyms for sense 1

Table 5.- Sequences to the left of instance

Instance-1 Sequences	
Customers	Bottle
Home	Us party
Yam	The bottle
Party	Studies customers
Resources	To the
This	The us party
To	In the bottle
Tools	

Instance-2 Sequences	
Design	Page layout
art	Visual arts
Bouchard	Of the mouth
Medical	Proactive core component
The	Multimedia design at
Fanny Bouchard	Object web proactive core component
Graphic design	Illustration of the mouth

are 87. In the end, we have 34 words that could help us disambiguate the word instance (25 for sense 1 and 9 for sense 2); these words are shown in Table 3.

In simple lexical associations, the significant words encountered do not necessarily have to appear contiguously. To illustrate the use of significant words associated with instance, we present some commonly used sentences. To reinforce the idea of using these words in lexical disambiguation systems, we separate the examples by meaning. The first meaning of instance relates to "the occurrence of something," while the second sense refers to "an item of information that is representative of a type."

Sense 1:

…another instance of the same process already running on the current machine….

...Enforcing a rule that only one instance of process is running is an interesting task….

Sense 2:

...ACTIVITY in this instance involves the use of government facilities and equipment for …

…. Another difference between instance members and class members is that class ….

Since the examples presented are extracts from sentences, coupled with the fact that we are talking about ambiguous words, it may be difficult to clearly and concisely distinguish between one meaning and the other. In this case, it would be worthwhile to increase the number of immediate context words surrounding the ambiguous word. However, this task is not easy, even manually. It should be noted that in the Senseval-2 competition (a competition that compares the performance of different WSD systems in different languages), there was only 75% agreement between human annotators for English.

Table 6.- Sequences to the right of instance.

Sequences Instance-1	
Design	Code
Edu	Western reserve
Index	Studies case
Law	Studies catalog
Studies	Studies in
Study	Western reserve university
Western	Studies catalogs resources

Sequences Instance-2	
And	A formal example with
Clients	Of the mouth
In	Livres d enfants children
For	Employees post a jobs
Is	Of the mouth illustration
Of the	A formal example with a
Of a	Government by john Stuart
In the	Of the secretary general
Livres d enfants	Livres d enfants children book´s
Employees post a	Employees post a jobs and
Of judicial activism	Government by john Stuart mill
And fine art	Of the mouth illustration of
Of the secretary	
Government by john	
Of next at	

For uninterrupted word sequences around the instance senses, the results obtained for different window (V) values are shown in Table 4, as well as the number of different and significant sequences. Negative values in the window column represent the number of words taken to the left of the synonym.

As the window size increases, the number of sequences decreases. This is because the sequence (of one, two, or more words) must be part of the context of all synonyms and also have the same order of appearance. The correlation between the number of sequences and significant sequences is 0.94. This value tells us that as we increase the number of sequences, the number of significant sequences will increase, which is to be expected since it presents a nearly normal distribution. For a sequence that has passed the strength and dispersion measures, the higher the frequency, the more significant it will be (Guzmán, 2005a).

Unbroken sequences typically begin or end with a polysemous word, instance in our case. For this reason, we separated the results obtained, differentiating the unbroken sequences not only by direction but also by their left or right location. Table 5 shows the unbroken sequences on the left, which are found in commonly used expressions such as customers instance or graphic design instance.

The significant sequences to the right of instance are shown in Table 6. We will find these sequences in commonly used expressions, such as instance design or instance studies case.

In our work, we have not ignored stop words, or empty words, such as prepositions, determiners, etc., which appear in almost every sentence. However, some of these words play an important role in assigning meaning to a sentence (Guzmán, 2005b). For example, in the case of "for," we found it significantly associated with sense 2 of "instance" and is used in commonly used expressions such as "for instance"; this sequence has a single meaning in WordNet (its meaning refers to an example).

4 Conclusions

The initial experiments conducted demonstrate the potential of the Web as a linguistic corpus. Furthermore, they demonstrate the feasibility of incorporating the extracted lexical patterns into disambiguation systems. The methodology can be applied to other morpho-syntactic categories, as well as to other languages, provided that a lexical database exists in these languages, such as WordNet for English, that allows us to determine the meanings attributable to a polysemous word. Furthermore, it can be applied to finite corpora. Although the number of significant collocations per meaning is generally encouraging, we must increase the size of the corpus to have more examples of contexts of use for each of the synonyms that make up the meaning in WordNet and thus obtain better lexical patterns

References

1. Benson, M.: The BBI combinatorial dictionary of English. John Benjamin Publ., Amsterdam, Philadelphia (1989)
2. Benson, M.: Collocations and General Purpose dictionaries. International Journal of Lexicography, 3, pp. 23–35 (1990)
3. Buscaldi, D., Rosso, P., Masulli, F.: The Upv-unige-CIAOSENSO WSD System. In: Proceedings of Senseval-3, pp. 77–82 (2004)
4. Celina, S., Gonzalo, J., Verdejo, F.: Automatic Association of Web Directories with Word Senses. Computational Linguistics, 29(3), pp 485–502 (2003)
5. Choueka, Y.; Klein, S.T., Neuwitz, E.: Automatic retrieval of frequent idiomatic and collocational expressions in large corpus. Association for Literary and Linguistic Computing Journal, 4(1):34–38 (1983)

6. Etzioni, O.: The World Wide Web Quagmire or Gold Mine? Communications of the ACM, 39(11), pp. 65–68 (1996)
7. Fellbaum, Ch.: WordNet as Electronic Lexical Database. MIT Press (1998)
8. Frawley, W., Piatesky-Shapiro, G.: Knowledge Discovery in Databases: An Overview, AI Magazine, pp. 213–228 (1992)
9. Guzmán-Cabrera, R., Rosso, P., Montes-y-Gomez, M., Gomez-Soriano, J. M.: Mining the Web for word sense discrimination, In: Information and communication technologies international symposium, Tetouan, Morocco (2005)
10. Guzmán-Cabrera, R., Montes-y-Gomez, M., Rosso, P.: Searching the Web for word sense collocations. In: IADIS international conference, Algarve, Portugal (2006)
11. Kilgarriff, A., Greffenstette, G.: Introduction to the Special Issue on Web as Corpus, Computational Linguistics 29(3), pp. 1–15 (2003)
12. Kosala, R., Blockeel, H.: Web mining research: a survey. GIS KDD Explorations, pp. 1–15 (2000)
13. Leech Garcide, G., McEnery, T.: Corpus annotation. Linguistic Information from Computer Text Corpora, Grammatical Tagging. chap 2, pp. 19–33 (1997)
14. Mihalcea, R.: Making Sense out of the Web. In: Workshop on Lexical Resources and the Web for Word Sense Disambiguation, IBERAMIA, Mexico (2004)
15. Montes-y-Gómez, M., López-López A., Gelbukh, A.: Information Retrieval with Conceptual Graph Matching. In: 11th International Conference on Database and Expert Systems Applications DEXA 2000, Springer-Verlag (2000)
16. Montes y Gómez, M.: Text Mining using Similarity between Semantic Structures. Doctoral Thesis, Computing Research Center (CIC), National Polytechnic Institute (IPN), Mexico (2002)
17. Resnick, P. S.: Selection and Information: A Class-Based Approach to Lexical Relationship. Doctoral Thesis, University of Pennsylvania (1993)