BRAC
BARCELONA RESEARCH ART CREATION

Hipatia Press
www.hipatiapress.com

Instructions for authors, subscriptions and further details:

http://brac.hipatiapress.com

Word Sense Discrimination Using Statistic Analysis of Texts

Rogelio Nazar[1]

1) Natural Language Processing Research Group, Dept. of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona.

Date of publication: June 3rd, 2013

To cite this article: Nazar, R. (2013). Word Sense Discrimination Using Statistic Analysis of Texts. *BRAC. Barcelona, Research, Art, Creation, 1*(1), 5-26. doi: 10.4471/brac.2013.01.

To link this article: http://dx.doi.org/10.4471/brac.2013.01

BRAC - Barcelona Research Art Creation. Vol. 1 No. 1, June 2013 pp. 5-26

Word Sense Discrimination Using Statistic Analysis of Texts

Rogelio Nazar
Universitat Pompeu Fabra, Barcelona

Abstract

For years, computer programs have been working to obtain information about certain entities such as persons, organizations or scientific concepts from the Web or from other sources. However, they have many challenges yet to overcome, for instance when texts refer to different entities that share the same name (e.g., a mouse can be an electronic device or a living creature). This article presents a method to solve this problem based on the frequency analysis of the words that are found in the vicinity of a target word. Each sense of the polysemous word or term will be represented as a different group of other vocabulary units that show a tendency to appear together with the target word in each of its different senses. The interest of the proposal is that it does not require previous knowledge about the language of the corpus or any other form of knowledge from the external world.

Keywords: computational linguistics, information extraction, word sense induction.

ISSN 2014-8992
DOI: 10.4471/brac.2013.01

BRAC - Barcelona Research Art Creation. Vol. 1 No. 1 June, 2013 pp. 5-26

Discriminación de Sentidos Basada en Análisis Estadístico de Textos

Rogelio Nazar
Universitat Pompeu Fabra, Barcelona

Resumen

Durante años han existido programas que de manera automática obtienen información acerca de entidades como personas, organizaciones o conceptos científicos a partir de repositorios de texto en formato digital tales como la Web u otras fuentes. Sin embargo, todavía existe una serie de dificultades que no se han podido resolver, por ejemplo cuando distintas entidades son designadas con un mismo nombre (como el ratón, que puede ser un dispositivo periférico en computación o bien un mamífero). El presente artículo propone un método para resolver este problema basado en el análisis de la frecuencia de las palabras que se encuentran en el contexto de aparición de la palabra ambigua.

Cada uno de los sentidos de una palabra polisémica se representan mediante los correspondientes grupos de otras unidades léxicas que muestran tendencia a aparecer en el contexto de esta palabra. El interés de esta propuesta reside en que no requiere ningún tipo de conocimiento externo al corpus, como conocimiento del mundo o de la lengua de los textos.

Palabras claves: extracción de información, inducción de sentidos, lingüística computacional.

ISSN 2014-8992
DOI: 10.4471/brac.2013.01

There currently exists a wide variety of computer programs[1] that are scanning the Web or vasts collections of scientific literature with the purpose of collecting information. These programs are designed to obtain data of different degrees of complexity from running text, such as, for instance, attributes of different entities like persons or organizations or information of a more technical nature, like drug-drug interaction or the possible relations between proteins and certain diseases, to name only a few possibilities. Of course, human beings are much more skilled than most computer programs in the task of reading and understanding a written document. However, the massive amount of text that is accumulating these days has reached such a point where it becomes difficult for a single individual or a group of researchers to retrieve all the relevant documents produced in their corresponding fields and assimilate the information in the traditional way, i.e., taking the time to read the documents one by one.

As a result of the massive grow experienced by the collections of technical and scientific literature, more and more researchers from different fields are using computers to search and extract information from electronic documents, offering a significant opportunity of application for the algorithms developed in computational linguistics, which is the general denomination of the field of research on semantic analysis of text by automatic means.

Information Retrieval (IR; cf. Manning et al., 2008) and Information Extraction (IE; cf. Grishman, 2012) are complex tasks that still have many challenges to be addressed by computational linguists. One of the most difficult problems is faced when different entities mentioned in the texts share the same name. This is the problem that is introduced in the present article, along with a proposal for a methodology to solve this kind of ambiguity by means of statistical analysis. More specifically, the article will show how the meaning of a word can be modeled using other words occurring in the vicinity of the ambiguous one. The article advocates for the use of statistical methods instead of rule based systems for a number of reasons that will become apparent later.

Let us first clarify the terminology used to circumscribe the problem:

Word Sense Disambiguation (WSD) is defined as the operation by which an automaton assigns a determined sense to an ambiguous word in context from an inventory of senses available to the system. Word Sense Induction (WSI) or Discrimination, in contrast, is the operation of finding those senses from a sample of contexts of occurrence of a given word. Both operations are related, but they are often treated as independent problems in the literature (Ide & Véronis, 1998; Navigli, 2009).

The present article is focused on WSI only, and it discusses a methodology for its application to the results of a search on the Web or on other corpus. The paper reports promising results of experiments that were conducted to test if one can acquire the senses of acronyms in English and proper nouns in Spanish, although the same idea should also be useful for other languages and scenarios.

As an example of application of this algorithm, consider the case of a search engine like Google, which has not yet been able to find a proper solution to the problem of homographs in the search results. The problem of homographs is especially acute in the Google Alerts service, which provides an email alert when a new document appears on the Web containing a given query expression that has been set beforehand by the user. Of course, this user will only be interested in one of the possible senses or references of such query and not the noise produced by email alerts triggered by irrelevant documents.

There is increasing interest in this type of email alert services because of their relation to the field of “reputation management”, which is usually in the hands of public relations and communication professionals who are in turn creating a demand for IE technologies and motivating research in the extraction of information about persons and organizations (Artiles et al., 2010). In an increasingly digital world, the circulation of information about a given institution or individual can be of strategic importance, and we can expect to see a market of software solutions for the extraction of, for instance, the most frequent opinion of consumers about a certain brand or product. In order to fulfill this purpose, however, applications must be able to distinguish between different people that share the same name or different referents that are designated by the same acronym.

The present paper thus argues that clustering algorithms based on co-

occurrence graphs can be useful to resolve different kinds of ambiguity. It proposes an original algorithm that is based on graphs of lexical co-occurrence which takes an expression as input and retrieves contexts of occurrence from a corpus (a collection of documents from the Web or from any other source) and produces a graph based clustering of the contexts. This output, in turn, can serve as a representation (or a discrimination) of the different senses that the ambiguous word may have. The proposed model produces, for each input word, a representation of the syntagmatic associations of such word, i.e., those that have a significant frequency of co-occurrence with the analyzed word in the same sentence, paragraph, or in a context window of an arbitrary size. These are the units that are taken as disambiguation clues, disregarding all external sources of explicit linguistic or ontological knowledge.

The units found in the contexts are referred to as the vocabulary, which is represented as words and sequences of up to n orthographic words (called n-grams), with n=2 in the case of these experiments. This vocabulary is organized in a graph that represents the co-occurrence between the vocabulary units. Every node in the graph is a unit and the links between the nodes represent the associational strength between the words given by the number of times they appear together in the same contexts.

The graph is created from an input word (or term) A, thus the graph represents n-grams occurring with A with a significant frequency, set denoted as $\vec{A}$. The arcs of the graphs are not only between A and the nodes A_i, they are also created between nodes A_i and A_j if both tend to appear together in the contexts of A. The idea of applying these graphs to WSI is to *travel* through such graph to extract subgraphs and to treat them as separate clusters, each one representing a different sense of the analyzed word.

With a minimum number of contexts of occurrence of a given target word (experiments in this paper included 100 contexts per trial) representing different senses of such word, then the resulting graph of A will show different *hubs*, which are regions of densely interconnected nodes. Thanks to their singular geometric properties, co-occurrence graphs can be useful in the case of polysemous expressions. There are regions of the graph that attract a group of nodes related to one of the

senses, and this phenomenon can be used as a natural way to cluster contexts according to word senses. Some obvious examples are the cases of homographs such as the *computer mouse* and the *animal mouse*. In the graph for *mouse*, one region will be populated with computer-related terminology and the other with words related to the biological creature.

In the case of proper nouns, homonymy can be resolved because the names of related people serve as hubs in the network of a given proper noun, separating the references to different persons because it is unlikely that homonyms will share the same friends and acquaintances. One would rather expect each person to know different people, which results in the creation of independent networks within the graph generated with the name of the homonyms. It is the same process as with words: every sense of a polysemous word has its own group of "word friends", according to a famous Firthian principle (Firth, 1957).

This paper is organized as follows: next section offers a brief comment of related work on the field of WSD and WSI; Section 3 outlines the proposed methodology while Section 4 discusses the results of two kinds of experiments designed to test the method. Finally, Section 5 draws some conclusions and lines for future research.

Related Work

Research in WSD has been carried out for many decades and it has long been regarded as one of the most difficult problems in Machine Translation (MT) since its early days (Weaver, 1949; Pierce et al, 1966). First attempts were based on hand-crafted disambiguation rules (Kelly & Stone, 1975). Lesk (1986), for instance, tried to disambiguate words using the definition of senses listed in machine readable dictionaries. In the same way, manually constructed lexical resources have been used in the attempt to disambiguate words, like Yarowsky (1992) with Roget's thesaurus or Voorhees (1993) with the WordNet ontology.

In contrast to rule based techniques, different statistical approaches have also been proposed. Gale et al. (1992) used parallel corpora (collections of documents with their translations to another language) to learn the different senses of a word, since ambiguous words are

translated to different words in the aligned sentences according to their different word senses. Thus, one sense of the English word *sentence* will be aligned with the French word *peine* and in another sense with the word *phrase*. Also in the statistical trend, Yarowsky (1995) later discovered that collocations are useful to disambiguate because there usually is only one sense per collocation.

In the present days, we are living a sort of critical moment in the history of the WSD field. After several decades of research and many competitions organized for comparing precision figures (such as the Senseval/Semeval competitions, Agirre & Soroa, 2007; Manandhar et al., 2010), signs of a loss of optimism have begun to appear, both in the theoretical and practical levels. The reliability of many common assumptions is now being scrutinized, such as if there really is such thing as an inventory of word senses (Kilgarriff, 1997). In the theoretic level, there is still a great amount of confusion around the topic of polysemy and ambiguity. At the same time, the levels of inter-annotator agreement –which measure how different persons agree upon the discrimination of senses of a word– are lower than one would expect (Véronis, 1998).

Probably as a consequence of the great complexity of the problem, WSD methods are often ignored in many real world applications (Ide & Wilks, 2007). In MT, for instance, disambiguation is done in an implicit manner, as a consequence of the fact that MT researchers in general use statistical models trained on parallel corpora to associate equivalent sequences of words in two languages, with the phrase context playing the role of a disambiguating factor. In IR, a debate exists on whether WSD modules can or not improve results, but no consensus has yet been reached (Sanderson, 1994; Voorhees, 1993; Véronis, 2004; Agirre & Edmonds, 2007).

The work on WSD has been criticized mainly from the lexicographic front for the rather naïve assumptions that underlie most approaches, mainly the already mentioned idea that words have a limited number of discrete senses that can be listed in a dictionary. Many authors have expressed that such an approach would only be useful for certain levels of word-sense distinctions such as homographs (Kilgarriff, 1997; Hanks, 2006; Jezek & Hanks, 2010). Regarding this debate, one has to raise the question of what is exactly the purpose of the research in WSD. As

pointed out by Ide & Wilks (2007), for the purpose of practical applications, it is just the homograph-type distinction that is really needed in NLP:

> We argue that there is rarely a need to make distinctions below the homograph-like level for understanding, human or automated; and in the unusual circumstance where it becomes necessary to explicitly throw one of the sub-senses away, we can expect there to be contextual clues that will enable both humans and machines to do so. (Ide & Wilks, 2007:66).

The application of unsupervised approaches of WSI could be the appropriate response to the objections raised by critics of the inventory-based WSD, because this approach is less rigid and would not need previous inventories. Many authors have been working in this direction since the early nineties, with the application of vector space models as used in IR to perform WSD oriented clustering (Schütze, 1992, 1998; Schütze & Pedersen, 1995; 1997; Purandare & Pedersen, 2004; among others). The interest on the vector based approach begun to decay, however, when Small World Graphs (Watts & Strogatz, 1999) were introduced in linguistics under the name of Co-occurrence Graphs (Ferrer-i-Cancho & Solé, 2001). According to a growing number of publications, different versions of the Co-occurrence Graph approach are the most promising solution for both WSI and WSD (Widdows & Dorow, 2002; Véronis, 2004; Biemann, 2006; Klapaftis & Manandhar, 2010, among many others).

The line of research that is explored in the present paper could be categorized in the same trend as the latter studies. The novel contribution, in comparison to related work, is that the proposed method is much more simple both in the conceptual and computational level and that it disregards all external sources of knowledge (both linguistic and ontological).

Methods

As briefly sketched in the introduction, for the experiments to take

place, this approach assumes that we have a determined polysemous word and a corpus where this word is instantiated. When the analyzed corpus is downloaded from the Web, a series of routines are needed to convert the format of the files (e.g., HTML, PDF, Word documents, etc.) to plain text documents. The conversion process is not strictly relevant to the design of the algorithm and is therefore not discussed in detail here.

Once the corpus is available in plain text format, each vocabulary unit of the contexts is represented as a node. A vocabulary unit, however, may have different surface realizations on the corpus due to inflectional derivation or other types of term variation. A proper solution to this problem would be to apply full text-handling, including tokenization, lemmatization, Part-of-Speech tagging and a series of operations some of which are of great complexity, such as the identification in text of named entities and technical terminology. For simplicity, and to allow for the application of the algorithm to different languages, the methodology disregards these sources of information and only performs a sort of "pseudo-lemmatization", using an orthographic similarity measure based on the Dice coefficient (1), which takes two strings (*i,j*) as arguments and returns a value between 0 and 1, representing how similar they are on the basis of how many sequences of two letters they have in common.

$$\text{Dice}(i,j) = 2\,|I \cap J| \,/\, (\,|I| + |J|) \qquad (1)$$

For instance, if we compare the French bigrams *détentions provisoires* and *détention provisoire*, the resulting similarity would be 0.9. It is important to mention, however, that this procedure is only justified for the purpose of evaluating this method without the benefits provided by services such as lemmatization and Part-of-Speech tagging. Of course, in a real world application, nothing would prevent the use of all available resources.

The next step is to reduce the vocabulary of the corpus in order to keep only the most informative units (i.e., to eliminate non-content words, words that are not conceptually related to the input word, or more precisely, words that could appear in any text, regardless of the topic). This is done by using a reference corpus of the analyzed

language. Reference corpora are expected to represent general vocabulary and should be a balanced collection of genres and registers. Instead of this, in the experiments the reference corpora consisted of collections of press articles of approximately two million tokens per language used, downloaded from the Wortschatz-Portal (Quasthoff et al., 2006).

Strictly speaking, it cannot be said that a corpus of press articles is representative of general vocabulary, but using the frequency of the analyzed words in this reference corpus, it is possible to filter out the least informative vocabulary units. This operation is undertaken with the help of association measures. In the case of these experiments, Pointwise Mutual Information (2) is used (Church & Hanks, 1990), calculating the score between the target word and each of the co-occurring *n*-grams. In this context, X would be the input word and Y each co-occurring word. P(X,Y) would be the relative frequency of co-occurrence in the analyzed corpus while P(X) and P(Y) represent their independent relative frequency in the reference corpus. Deleting nodes with an association score below a threshold produces an efficient reduction of the vocabulary.

$$\mathrm{PMI}(X,Y) = \log 2\ P(X,Y) / P(X)\ P(Y) \qquad (2)$$

In order to reduce computational effort, a further filtering of the vocabulary is performed by the construction of the co-occurrence graph. Arcs between nodes are created when the words in the nodes appear in the same context. Each arc is weighted according to Equation (2), where ($A_i\ A_j$) denotes the frequency of co-occurrence of nodes A_i and A_j while N is the total number of contexts.

$$R(A_i, A_j) = (A_i\ A_j) / N \qquad (3)$$

Once the graph for a given input word is constructed, the process of WSI consists of a clustering process of the graph by identifying and extracting hubs. Each extracted subgraph or cluster represents a sense of the analyzed word. If a given cluster shows a similarity over a given threshold (details of the overlapping function are given later), then the two clusters are merged into a single cluster and this process continues

until no more clusters are created.

Table 1 shows the pseudo code for the WSI algorithm. Perhaps the most suitable way to explain this algorithm is by means of an example. Consider, for instance, the Spanish word *ratón* (mouse) in a corpus consisting of 50 contexts of occurrence of *ratón* in a Computer Science corpus and 50 contexts from a Genomics corpus.

Table 1

Pseudo-code for the proposed WSI algorithms

Word Sense Induction Algorithm

```
For each word j co-occurring with input word t {
  Initialize counter;
  Initialize rank;
  Create cluster s with documents containing j ;
  Next if first cluster;
  For each previously created cluster p {
       o = overlap ( p, s );
       if ( o > k ) {
            rank(p) = o;
           counter++;
        }
   }
   if (counter > 1 ) {
     destroy cluster s;
   } else if ( counter == 1 ) {
     collapse ( max(rank), s );
  }
}

Do pairwise comparison of clusters;
```

The process of word sense induction starts in the first loop with the word *icono* (icon) as the value of the variable *j*, because it is the word with the most significant frequency of co-occurrence with the target word. Since it is the first word, the loop ends here, creating a first cluster with pointers to all the contexts of *ratón* that also contain the word *icono*. The next value of *j* is the word *teclado* (keyboard), another word frequently found in the contexts of *ratón* when the word is used to designate the computer device. In this case, since there already is a cluster created, the new cluster of contexts *teclado* + *ratón* is compared to determine the number of contexts they share using the overlap function (4).

$$\text{overlap}\ (i,j) = |I \cap J| / \min(|I|, |J|) \qquad (4)$$

The result of the comparison between the two clusters of *teclado* and icono yields a coincidence of 33% of the contexts (i.e., in 33% of the cases, the word *ratón* occurs both with *teclado* and *icono*. Since the arbitrary threshold *k* of the overlap was set precisely to 33%, these clusters are considered different and therefore they are not merged and, as a consequence, a new cluster for the contexts of the word *teclado* is thus created. This is a somewhat unfortunate decision, since both terms are indeed related. However, as we will see in a moment, the algorithm is robust enough to recover later from this "mistake": when the clusters grow in number of members, then there is sufficient overlapping to allow for their assimilation into a single, larger cluster.

For the moment, let us continue with the simulation of the process. The next word to be the value of *j* is *transgénico* (transgenic). Again, the same process unfolds: the cluster of contexts of *transgénico* is compared with each of the two clusters recently created and the result in both cases is that the overlap is zero. Consequently, the contexts of *transgénico* are treated as a new cluster. The following word is *seleccionar* (to select). Again, the cluster of contexts of this word is compared with the three clusters previously created. This time, the comparison shows an overlap of 20% with the first cluster, the one for *icono*, but since it is again below the threshold *k*, a new cluster is created. In the case of the following cycle, with the word *puntero* (pointer), the overlap with the first cluster is 75%, thus the new cluster

is collapsed with the first one.

The process goes on like this until there are no more co-occurring words or no more contexts of occurrence. When this is the case, the algorithm performs a pairwise comparison of each of the created clusters, with the purpose of collapsing those clusters that, again, surpass the overlap threshold (as in the already mentioned case of *teclado* and *icono*). The mechanism is the same as in the comparison just described: a cluster *s* will be collapsed to another cluster *p* if the overlap between *s* and *p* is greater than *k* and if there is no other cluster apart from *p* that *s* has such an overlap with. For illustration, figures 1 and 2 depict fragments of the two clusters or sub-graphs representing each of the two uses of the word *ratón*.

Figure 1. One of the regions of the graph for the Spanish word *ratón* meaning *computer mouse*.

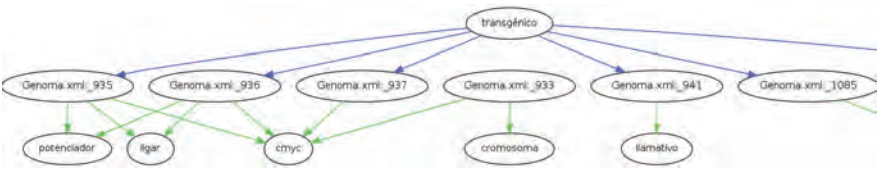

Figure 2. Another region of the graph of *ratón*, now representing the *animal* sense.

Experimental results

Analysis of acronyms

The first experiment was undertaken with a set of acronyms in English, in order to cluster documents downloaded from the Web according to the different referents of these symbols. The case of the acronyms is interesting because their disambiguation with a previously available

inventory of senses (or references, in this case) is virtually impossible. The number of entities that can be referred to by these symbols in large digital corpora such as the Web would make it very difficult to keep those inventories updated. Thus, the ideal solution would be a WSI technique such as the present proposal, because it would take nothing for granted apart from a given input word and a corpus where this word is instantiated.

This first run of experiments was conducted using English acronyms of an extension of three and four letters, randomly sampled from Wikipedia (e.g. ASG, NCO, PCR, etc.). For each acronym, the algorithm downloaded 100 documents from the Web and generated clusters as subgraphs. In total, 25 experiments like these were undertaken.

Table 2
Results of the experiment in the case of the acronym AASC

Clusters	Docs	Errors	Omissions	Precision (%)	Recall (%)
Asian American Studies	9	2	0	66	100
American Association of State Climatologists	4	1	0	75	100
Arizona Association of Student Councils	3	0	0	100	100

After a thorough examination of the results, it became apparent that each generated cluster represents a different entity referred to by the acronym. Table 2 is an example of the clusters generated for the acronym *AASC*. There we can see that the algorithm correctly identified three distinct clusters: the *Asian American Studies Center, the American Association of State Climatologists* and the *Arizona Association of*

Student Councils. There were also two other referents, however, that the algorithm could not discriminate. These are two groups of three documents each, referring to the *Asian American Students Center* and the *Academic Assembly Steering Committee.* The rest of the documents, up to 100, is a long tail of single references to different uses of the expression, a very stable pattern in acronyms that is naturally not seen in general vocabulary units because, normally, general vocabulary units do not have so many senses.

Table 3
Examples of the overall results of the experiment with English acronyms

Acronym	Detected Uses	Undetected Uses
AASC	3	2
APCS	5	1
ASG	2	0
BVM	4	0
CKD	2	0
DDO	3	0
ETN	1	1
FYI	2	0
IED	4	0
JUB	6	1
KPS	7	0
KSP	5	1
LEP	4	1
...	...	...
Total	93	11

Table 3 reports more details on the number of detected vs. non-detected senses for some of the 25 trials. There we can see that, in general, the number of detected senses in each case is far larger than those that went undetected. After 25 experiments, the algorithm was able to detect 93 senses (89% of the total).

Analysis of homonyms

The second experiment is a case of disambiguation of homonyms in proper nouns. In this case, the analyzed unit refers to two different persons who share the same uncommon name. What makes this case interesting from a WSI perspective is that, by pure chance (since they are not related), both have approximately the same age and were born and raised in the same city. Because of their professions, both have raised a certain public profile, being present on the Web and the media (names were deleted to preserve anonymity).

This case makes any attempt of automatic disambiguation difficult because these two persons share an important number of vocabulary items in the texts that refer to them, due to their common origin and circumstances. The idea, as in the previous experiments, is to take a collection of documents referring to both persons (the first 100 documents served by Google with their name used as a query expression) and to separate this collection in coherent clusters of documents according to each person.

Despite the complexity of the problem, the results obtained using the same algorithm as in the previous experiment were very promising. Table 4 presents the distribution of contexts per cluster. Each context of occurrence is labeled as *A* or *B*, depending on the subject, along with a context identification number. Ideally, we would have expected only two clusters (one per person) instead of six. However, the subdivision is meaningful, as it will be seen shortly, and the internal consistency of the clusters is very high. As table 4 shows, there is no single cluster that mixes documents referring to both persons. There is only one context in cluster 1 which is labeled as N, that is, neither *A* or *B*, and is just an irrelevant document –not related with any of the two persons– that was returned by the search engine.

The reason why there are more than two clusters is that each one is devoted to a particular aspect of the life of these individuals. For

Table 4

Clusters of documents generated from a corpus of two homonym subjects (A and B)

Cluster 1	Cluster 2	Cluster 3
Subject_A_34	Subject_A_11	Subject_B_13
Subject_A_42	Subject_A_19	Subject_B_16
Subject_A_46	Subject_A_27	Subject_B_21
Subject_A_5	Subject_A_32	Subject_B_22
Subject_A_71	Subject_A_46	Subject_B_23
Subject_A_73	Subject_A_55	Subject_B_48
Subject_N_53	Subject_A_57	Subject_B_58
	Subject_A_69	
	Subject_A_71	
	Subject_A_72	
	Subject_A_75	
Cluster 4	**Cluster 5**	**Cluster 6**
Subject_B_13	Subject_B_15	Subject_A_1
Subject_B_14	Subject_B_64	Subject_A_24
Subject_B_29	Subject_B_67	Subject_A_62
Subject_B_39	Subject_B_70	Subject_A_65
Subject_B_44		Subject_A_8
Subject_B_52		

instance, one of the clusters is related to the professional life of one of the subjects, while others are related to his social life, with many of the nodes representing the names of friends or places where he usually goes. Similarly, in the case of the other man, it can be seen that the division in clusters reflect the different domains in which he has worked, with many of the nodes representing the names of places, colleagues and friends.

Another aspect that should be mentioned about these results is that, from 100 documents downloaded from the web, only a limited proportion (40) was assigned to a cluster. This, however, is not really a problem in this case because the task was to infer senses (referents in this case) and not to perform an exhaustive classification of the downloaded documents. This could be done indeed, assigning each remaining document to the most likely cluster based on shared

vocabulary but, again, this attempt was not undertaken because it was not the purpose of the experiment (it would be, indeed, a kind of WSD operation).

Conclusions

This paper has presented a co-occurrence graph based approach for WSI which does not demand great conceptual or computational complexity and disregards external knowledge such as lemmatization, Part-of-Speech tagging as well as dictionaries, ontologies or other semantic resources. In the practical level, the result is a fast and flexible algorithm that can be adapted to different languages and domains. Moreover, the paper can be of theoretical interest as well, as it may model or at least offer clues on how humans use contextual information to disambiguate words or to acquire word-senses, as a complement to other types of psycholinguistic evidence.

Despite the complexity of the problem, the results obtained in general in both experiments are very promising. The results presented here are not those of a ready-to-use tool, and much work is still needed to refine and further evaluate these results. However, there is no doubt that this algorithm, as it is in its present state, could already be an important improvement for the email alert services of most mainstream search engines, as well as for other multiple uses such as the improvement of Web searches in general and the automatic creation of glossaries, among other possibilities.

At the moment, it has only been tested on a few European languages, and much more experiments have to be undertaken before claiming the algorithm is language independent. The evidence already gathered, however, suggests that co-occurrence patterns are a property of language in general.

Co-occurrence graphs have proven to be useful tools for the linguistic analysis of polysemy, but the number of possible practical applications

in lexicography and related domains are very diverse. One of them, which is a line of future work for this research, is to apply co-occurrence graphs to the study of semantic neology, i.e., to detect novel senses in known words, which can be seen as a special case of polysemy. Other directions of future work would be to conduct more experimental research comparing the performance of this method with language-specific approaches and, moreover, to try to seek ways to integrate different solutions to the problem (both statistical and knowledge-based).

Notes

[1] This paper is partially based on a chapter of the author's Ph.D. Thesis.

References

Agirre, E. & Edmonds, P. (2007). *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht, Springer.

Agirre, E. & Soroa, A. (2007). SemEval 2007: Evaluating Word Sense Induction and Discrimination Systems. *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 7-12). Prague.

Artiles, J. & Borthwick, A., Gonzalo, J., Sekine, S., Amigó, E. (2010). WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Task. *Proceedings of Conference on Multilingual and Multimodal Information Access Evaluation* (CLEF).

Biemann, C. (2006). Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. *Workshop on TextGraphs at HLT-NAACL* (pp. 73-80).

Church, K. & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics 16*(1), 22-29.

Ferrer-i-Cancho, R. & Solé, R. (2001). The Small World of Human Language. *Proceedings of the Royal Society of London 268*, 2261-2265.

Firth, J. (1957). *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Gale, W., Church, K. & Yarowsky, D. (1992). *A Method for Disambiguating Word Senses in a Large Corpus*. Technical report, AT&T Bell Laboratories.

Grishman, R. (2012). *Information Extraction: Capabilities and Challenges*. [Lecture Notes]. Retrieved from http://cs.nyu.edu/grishman/tarragona.pdf

Hanks, P. (2006). The Organization of the Lexicon: Semantic Types and Lexical Sets. *Proceedings of XII Euralex* (pp. 1165-1168), Prague.

Ide, N. & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, *24*(1), 1-40.

Ide, N. & Wilks, Y. (2007). Making Sense About Sense. In Agirre & Edmonds (eds.) *Word Sense Disambiguation, Algorithms and Applications*, (pp. 47-73). Springer.

Jezek, E. & Hanks, P. (2010). What Lexical Sets Tell us about Conceptual Categories. *Lexis 4*, 7-22.

Kelly, E. & Stone, P. (1975). *Computer Recognition of English Word Senses*. North-Holland, Amsterdam.

Kilgarriff, A. (1997). I don't Believe in Word Senses. *Computers and the Humanities 31*(2), 91-113.

Klapaftis, I., Manandhar, S. (2010). Taxonomy Learning Using Word Sense Induction. *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Los Angeles, USA, ACL.

Lesk, M. (1986). Automatic Sense Disambiguation: How to Tell a Pine Cone from an Ice Cream Cone. *Proceedings of the 1986 SIGDOC Conference*.

Manandhar, S., Klapaftis, I., Dligach, D. & Pradhan, S. (2010) SemEval-2010 Task 14: Word Sense Induction & Disambiguation. *Fifth International Workshop on Semantic Evaluation*, (pp. 63- 68) Uppsala, Sweden.

Manning, C., Raghavan, P. & Schütze, H. (2008) *Introduction to Information Retrieval*. Cambridge University Press.

Navigli, R. (2009). Word Sense Disambiguation: a Survey. *ACM Computing Surveys, 41*(2), 1-69. ACM Press.

Pierce, J., Carroll, J., Hamp, E., Hays, D., Hockett, C., Oettinger, A. & Perlis, A. (1966). *Language and machines computers in translation and linguistics: A Report*. Automatic Language Processing Advisory Committee. Division of Behavioral Sciences. National Academy of Sciences. National Research Council. Washington, D. C.

Purandare, A. & Pedersen, T. (2004). Word sense discrimination by clustering contexts in vector and similarity spaces. *Conference on Computational Natural Language Learning* (pp. 41-48). Boston, USA.

Quasthoff, U., Richter, M. & Biemann, C. (2006). Corpus portal for search in monolingual corpora. *Proceedings of the LREC 2006* (pp- 1799-1802) Genoa, Italy.

Sanderson, M. (1994). Word sense disambiguation and information retrieval. *Proceedings of ACM SIGIR Conference* (pp. 142-151).

Schütze, H. (1992). Dimensions of Meaning. *Proceedings of Supercomputing* (pp. 787-796).

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics, 24*(1), 97-123.

Schütze, H. & Pedersen, J. (1995). Information Retrieval Based on Word Senses. *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval* (pp. 161-175).

Schütze, H. & Pedersen, J. (1997). A Co-occurrence-based Thesaurus and Two Applications to Information Retrieval. *Information Processing and Management,* 33(3), 307- 318.

Véronis, J. (1998). A study of polysemy judgments and inter-annotator agreement. *Senseval Workshop*, Herstmonceux Castle, UK.

Véronis, J. (2004). HyperLex: Lexical Cartography for Information Retrieval. *Computer Speech & Language, 18*(3), 223-252.

Voorhees, E. (1993). Using WordNet to Disambiguate Word Senses for Text Retrieval. *Proceedings of the 16th ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 171-180).

Watts, D. & Strogatz, S. (1998). Collective Dynamics of 'Small- World' Networks. *Nature, 393*, 440-442.

Weaver, W. (1949). Translation. Reprinted in *Readings in Machine Translation*, Nirenburg et al. (eds.), MIT Press, 2003.

Widdows, D. & Dorow, B. (2002). A Graph Model for Unsupervised Lexical Acquisition. *Proceedings of the 19th International Conference on Computational Linguistics* (pp. 1093-1099). Taipei.

Yarowsky, D. (1992). Word-sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora. *Proceedings of Coling-92* (pp. 454-460). Nantes.

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (pp. 189-196). Stroudsburg, PA, USA.

Rogelio Nazar Natural Language Processing Research Group
Dept. of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona, Spain.

Contact Address: Roc Boronat, 138. 08018 Barcelona. Spain.
Email: rogelio.nazar@upf.edu.