A Representation Model based on Wordnet Meanings for Internet Search

Andrés Carretero
Department of Information Systems and Technologies, University of Castilla La Mancha, 13071 Ciudad Real
Andres.Carretero@alu.uclm.es

Francisco P. Romero
Department of Information Systems and Technologies, University of Castilla La Mancha, 13071 Ciudad Real
FranciscoP.Romero@uclm.es

Jose A. Olivas
Department of Information Systems and Technologies, University of Castilla La Mancha, 13071 Ciudad Real
JoseA.Olivas@uclm.es

Abstract

The majority of search engines obviate the different meanings of the words used in the searching process. Introducing meanings in this process could improve the results, but the use of wrong meanings could produce worse results. To deal with this problem, in this work is presented a document representation model based on meanings. In this model is developed an automatic word sense disambiguation process. Index terms are associated with their corresponding meanings in WordNet. The model proposed is implemented in an experimental search engine called Bianca which has been tested using standard measures. The achieved results are compared with a classic word searcher.

Keywords: Representation Model, WordNet, Information Retrieval, Word Sense Disambiguation.

1. Introduction

Users find many problems using Internet search engines. They leave them frustrated when a search engine retrieves a large amount of irrelevant information. They do not know how to improve the results. There are contents unreachable by search engines and actually users do not understand how the queries are processed by the search engine [6]. Generally, search engines process documents eliminating stop-words, and make a stemming process, but finally each document will be represented by means of a bag of words.

In this work, the search engine Bianca replaces the literal comparison between words by the meanings comparison. This technique would improve the results of a usual query thrown in a search engine. It is a very important aspect because is not the same asking about how to take care of a tree than making the same question about an oak. An oak is a tree, and both of them are plants. The majority of search engines do not take into account these kinds of problems [15].

Bianca works with WordNet [14]. WordNet is a lexical Database whose design is inspired by psycholinguistic theories about human lexical memory and where the words are organized in synonyms groups called synset (synonym set). Wordnet is a thesaurus [1] but is an ontology too [9], is possible to know the meaning of a word and at the same time go to other words using ontological relations like synonymy, hypernymy, meronymy. In [18] it is explained the use of WordNet to expand and rewrite a query. Other work where WordNet has been used can be read in [21]. The way of calculate semantic distance can be consulted in [5]. Finally the work developed in GUMse has been used as inspiration to perform the word sense disambiguation process, and to rewrite the queries [16].

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 1076–1083
Torremolinos (Málaga), June 22–27, 2008

This paper is organized as follows: In section 2 the paper explains the architecture of the searcher. In the section 3 it is described the representation model, where an expanded VSM model is used to build an index and how an automatic disambiguation process is applied to the texts. In section 4 it is explained how to make a query and how this query can be expanded. Finally, in the last sections the experiments performed and the conclusions obtained are shown.

2. Bianca Architecture.

Bianca is a search engine based on meanings. Bianca reads the query introduced by the user and performs the search of these words in a set of documents. The difference between commercial search engines and Bianca arises when Bianca transforms the words into meanings and uses them to find relevant documents. Previously Bianca processes the documents and builds a conceptual document index based on.

Like any search engine, the main functionalities of Bianca are two. Bianca is able to build an index using a set of documents. Next, the system can perform searches into this index. In both of them, Bianca transforms the words into meanings using the original Princeton WordNet and the European WordNet [2] and performs a word sense disambiguation process.

Given a word, WordNet identifies a set of words with a similar meaning and groups them in a structure called synset. Each synset has an identity number in such way that when WordNet is asked about a word, it returns this number [11]. Bianca uses these numbers, the synset, to build the index and in the same way to perform the search.

In order to build a multi-lingual index based on meanings, Bianca used EuroWordNet. When Bianca process documents or queries not in English, it is used a component called InterLingual Index (ILI) [2] to get the equivalent meanings between different languages.

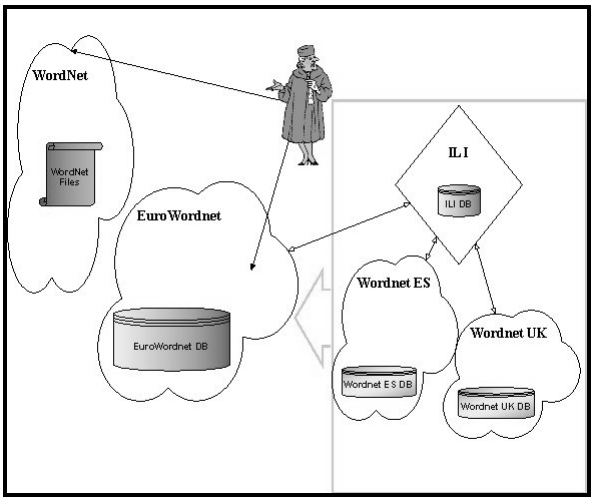

Figure 1: Bianca Architecture

Bianca applies an automatic disambiguation process to the terms using WordNet in order to control feature generation and reduction. The main purpose of this process is to perform a more efficient search.

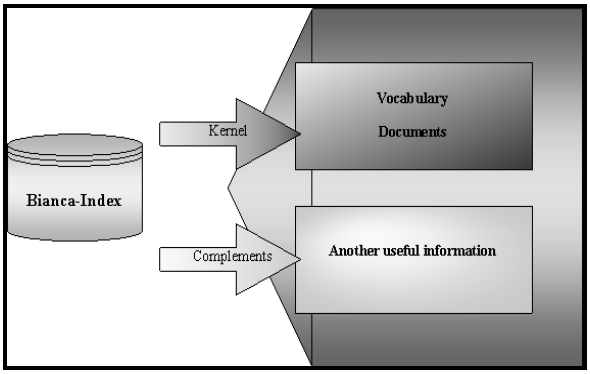

Figure 2: Index Structure

The index is composed by a few components grouped in two parts (Figure 2): the kernel and the complements. In the kernel is stored the vocabulary of the corpus and VSM vectors. Using the vocabulary component is possible to find the relations between a word and the different synsets found in WordNet or EuroWordNet. This task is necessary to know the synsets used in each document.

The complements are used in different tasks in the application. One of these tasks is building snippets [12]. When the result of a search is shown the user can view parts of the texts where the word has been located. Another common task is storing non-existent words in WordNet. WordNet does not grow as fast as the vocabulary; this is one of the reasons because of some words are not possible to be found in it [8]. Bianca detects these words and stores them.

3. A Representation Model using meanings

The use of a representation model based on meanings in a search engine has two main purposes:

- Eliminate some words that do not contribute to the meaning of queries and documents.
- Find fast and easily the documents more adequate to the query.

3.1. Index Structure

The chosen structure to store the index is based on the Vector Space Model (VSM). The VSM model builds a vector for each document where each feature´s element is a word. The classical approach says that for each word, the vector stores its frequency, but in this work it is used the meaning instead the words [21]. The traditional VSM model is expanded trying to store information about meanings.

The real process of building the index begins reading the documents in the chosen corpus; Bianca reads the documents written in a specific format (html, plain text). It is easy to develop new parsers, so the number of possible corpus can be extended.

In all the information retrieval process one of the most important task is detecting the language of the text. The stop-word technique helps to reduce the number of non-useful words, like articles, prepositions, etc.; the stemming process is necessary to avoid "lexical noise" [17] and it's necessary to use the right WordNet (Princeton or EuroWordNet) to get the meanings of a word.

When a word is found in a text, WordNet is asked about the meanings. The result of this question is a list of all the synsets.

Each synset is an entry in the vector that represents the document. The relevance of each entry is calculated using this formula [20] when word includes the meaning m1:

$$relevance\ (m_1) = \sum freq(word) / meanings(word)$$

Where m1 is the meaning of a word and *meanings (word)* returns the number of meanings of a word. If this meaning would appear in another word this would be promoted, increasing its value according by the meanings of this new word.

The way in which meanings are promoted is explained by the example shown in Table 1. The next words, word1, word2 and word3, share the ma, mb, mc, md, me, mf, and mg meanings like the Table 1 shows.

Table 1: Meanings Distribution

WORDS	**MEANINGS**						
	ma	**mb**	**mc**	**md**	**me**	**mf**	**mg**
W1	x	x	x				
W2	x			x			
W3	x	x			x	x	x
Relev.	**1.03**	**0.55**	**0,33**	**0,5**	**0,2**	**0,2**	**0,2**

Following the previous formula the more promoted meanings will be the shared ones as it can be seen in Figure 3.

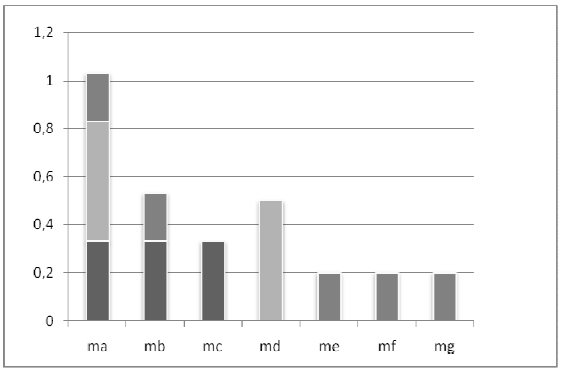

Figure 3: Meanings promotion

3.2. Feature reduction using disambiguation.

The next process tries to reduce the number of meanings. The automatic word sense disambiguation in this task is very important because of two reasons:

- the number of elements in a vector can be too large if it is considered all the meanings for each word.
- not all the meanings are relevant in the text where the word is found. If the non useful meanings were considered there would be some risks to consider wrong information as a good one [7].

Bianca uses WordNet trying to disambiguate automatically the meaning of a word. The

methods applied to achieve this objective are the following:

- two words are related because of the shared synsets. Imagine two words where some elements in its synsets are common, this means that these two words have some meanings in common.
- two words are related using the WordNet relations; synonym, hypernymy, meronymy. If WordNet determines that there is any of these relations between the words, the meanings not related will be set to null value
- two words are related if one of them is found in the definition of the another The word "car" and the word "engine" are related because "engine" appears in the definition of the word "car" (synset {02929975})

As it can be seen in the next tables, the disambiguation process can reduce drastically the features of document vectors and improve the relevancy of the results of a search.

In the example, there are two documents each one with two words ({w1, w2}; {w1, w3}) one of these is shared (w1). In the table 2 is only considered the number of meanings. Each vector has 14 positions.

Table 2: Meanings Distribution and Promotion

MEANINGS	WORDS			DOCS	
	W1	W2	W3	D1	D2
M1	X		X	0,33	0,44
M2	X	X		0,53	0,33
M3	X	X		0,53	0,33
M4		X		0,2	
M5		X		0,2	
M6		X		0,2	
M7			X		0,11
M8			X		0,11
M9			X		0,11
M10			X		0,11
M11			X		0,11
M12			X		0,11
M13			X		0,11
M14			X		0,11

The common word (w1) between the documents has some meanings relations with the others. In this case the document 1 will be returned when a word with the meaning *m1* is searched although there are meanings relations between the word *w1* and the word *w2*.

In the second case (Table 3) the non useful meanings are set to null. Each vector has only 3 positions. The relevant information can be found due to the relations among the meanings not because of the number of them. If the last search is done in this moment, only the document 2 will be returned because the meanings promotion has been done correctly using the disambiguation.

Table 3: Meanings Distribution and Promotion via word sense disambiguation process

MEANINGS	WORDS			DOCS	
	W1	W2	W3	D1	D2
M1	X		X		2
M2	X	X		1	
M3	X	X		1	

An example of the real effects of the disambiguation process in Bianca can be seen in the meanings promotion in the document 62 of ADI-SMART collection [4]. It has been used the disambiguation method where it has only been considered the WordNet relations. The text of the analyzed query is "*Computerized information retrieval systems. Computerized indexing systems*". For what Bianca returns after the stemming process and the stop words elimination the next lists of words, "*computerize*, *information*, *retrieval*, *system*, *index*". The disambiguation process returns the results shown in the Table 4.

If it had been considered the WordNet definition of the word "system" it is possible to observe that some meanings such as "*system#2 -- (instrumentality that combines interrelated interacting artifacts designed to work as a coherent entity; "he bought a new stereo system"; "the system consists of a motor and a small computer")*", are not considered.

Table 4: WordNet Meanings Extraction

Word	Mean
index	**<verb.social> index#1 -- (list in an index)**
information	**<noun.communication> information#1, info#1 -- (a message received and understood)**
information	**<noun.cognition> information#3 -- (knowledge acquired through study or experience or instruction)**
retrieval	**<noun.cognition> retrieval#2 -- (the cognitive operation of accessing information in memory; "my retrieval of people's names is very poor")**
system	**<noun.cognition> system#3, system of rules#1 -- (a complex of methods or rules governing behavior; "they have to operate under a system they oppose"; "that language has a complex system for indicating gender")**
system	**<noun.group> system#1, scheme#3 -- (a group of independent but interrelated elements comprising a unified whole; "a vast system of production and distribution and consumption keep the country going")**

4. Searching by meanings

The next step when the index has been stored, it is to realize a search. The process to do this task is very similar to the index building. When the user introduces a query, the following process is carried on: stop words elimination, stemming process, and word sense disambiguation process.

4.1. Query Expansion

As a meanings searcher, Bianca uses some relations between the words to improve the results of the searches. The main important relations used are synonymy, hypernymy, and meronymy. These relationships are also used in the disambiguation process. The query expansion method in Bianca is used to get more results than using a common query and not to improve the quality of the search results. The user can specify to the searcher which of this relations he wants to use in his search. WordNet is asked about the words related by means these relations and the possible synsets found are added to the original query. In this way the user can find documents speaking about "oaks" although the text introduced only had the word "plant".

Also, Bianca use disambiguation to rewrite the original user query, trying to improve the results of a current query. As it is said in [18], this technique means an interesting technique to improve the search results.

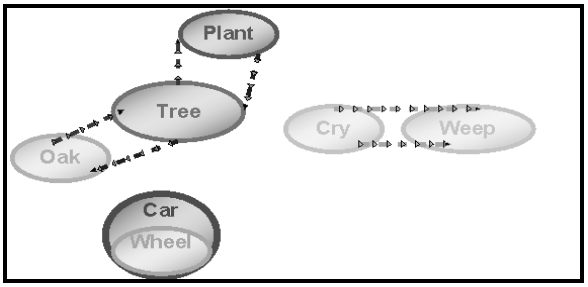

Figure 4: WordNet Relations

The Figure 4 shows an example of the query expansion process using the synonymy [10] relation between "cry" and "weep". Following this definition about synonym, "cry" and "weep" can be used in the same context, so when the system expands the query using synonyms of the first word, the synsets of the second word will be added to the query. Bianca will find documents with the two words, not only with the original one although the user does not know what words have been added to the original query.

In the same figure it is shown the relation among three words, "plant", "tree" and "oak"; an Oak is a tree, and a tree is a plant (hypernymy [1]). In the opposite side exists the other relation (hyponymy). If the user wants to find texts about an "oak" and choose in Bianca to find the hypernymy relations, the searches may return texts not only with the word "oak" but also with the words "tree" and "plant" or another that WordNet detects as related.

Also, the figure 4 shows the relation between "car" and "wheel". The "wheel" is a part of the car as could be the roof or the engine (meronymy [1]). As in the other examples, if the user chooses this relation there could be documents in the results with the word "wheel" when the searched word is "car".

4.2. Results

The documents returned will be ordered using the relevance of each of the synsets found in them using the Jaccard coefficient as similarity function.

Before the document is shown to the user is necessary to build a summary of the document in order to help the user to take a decision about the text to consult. These little texts are called snippets [12]. Therefore, the user reads all the texts extracted and selects the more relevant document.

5. Experiments

Bianca has been tested to verify the performance of the representation model. Bianca works as a usual searcher, the user introduces the query and returns the documents, although some tools have been implemented and integrated in Bianca in order to realize some experiments and measures with the results returned.

5.1. The experimental mode

There is an experimental mode to perform the queries in order to compare theoretical optimal results with the results obtained. The theoretical optimal results are integrated by two documents; the first one describes the queries. The second one describes the documents related with the queries. This relation between the queries and the optimal results are set by experts in the corpus domain. Not only is it possible to compare the theoretical optimal results with the ones Bianca returns but also the same experimental search can be done using the searcher as a usual literal-word searcher..

5.2. Metrics

The SMART collections [4] have been chosen to test the Bianca methods. The document collections have been indexed and it has been assessed some quality measures over the results. The next parameters have been considered: n_i: number of documents returned by the searcher; n_j: number of relevant documents in the query; n_{ij}: common documents, between returned and theoretical.

The measures that can be calculated with this information are Precision and Recall [19]. These measures are related in an inverse way, high levels of precision can be achieved by keeping recall low and vice versa.

$$p_{ij} = \frac{n_{ij}}{n_i} \qquad r_{ij} = \frac{n_{ij}}{n_j}$$

F-measure [13] has been used to test the effectiveness of the searcher and the representation model. This measure combines the precision and recall measures. A higher value of the F-measure means high quality level.

$$F(i,j) = \frac{2 * r_{ij} * p_{ij}}{r_{ij} + p_{ij}}$$

5.3. The experimental mode results

A summary of the results of Bianca can be seen in the Figure 5 where is compared the searcher using meanings versus using words. It has been achieved more precision using meanings than using words and at least in one search has been obtained the maximal precision.

Table5: Meanings vs. Words Metrics Results

	Precision	Recall	Max. P.	F-Meas
Meanings	35	67	100	88
Words	28	82	41	56

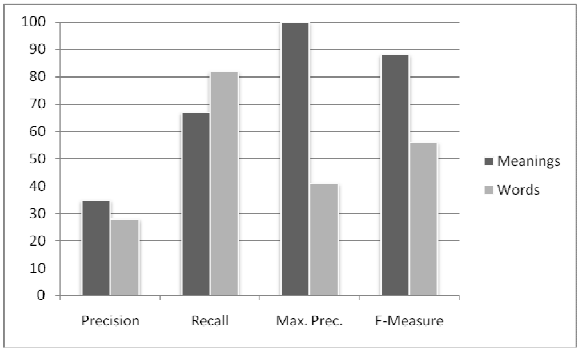

Figure 5 Meanings versus Words Metrics

It is possible to study the process carried out by Bianca with some queries from the ADI collection belonging to the SMART collection.

The maximal precision case is the following: For the eighteenth ADI query whose text is, "What methods are there for encoding, automatically matching, and automatically drawing structures extended in two dimensions, like the structural formulas for chemical compounds?", Bianca works with the next words before the stop words and stemmer process: *method, encode, automatically, match, draw, structure, extend, dimension, structural, formula, chemical, compound.*

The table 6 shows the based-meanings versus based-words results in Bianca.

Table 6: Meanings vs. Words Results

	Theoretical Docs.	Bianca Docs.	Precision	Recall
Meanings	3	2	100	67
Words	3	21	14	100

The optimal searcher may return the documents 2, 9 and 70 and Bianca returns only the 2 and the 70. According the words are in the documents index (Table 7), in the document 9 there is only a word that co-occurs, "chemical", this is the reason because Bianca does not return this document. This fact is not enough important, apparently.

Table 7: Words in Documents

Doc	**2**	**70**	**9**
Word	**chemical** **draw** **match** **method** **structure**	**automatically** **encode** **formula** **Method** **normal** **structural**	**chemical**

6. Conclusions and Future Works.

In this work, it is proposed a representation model based on meanings. This model is used in a searcher called Bianca. This approach uses Wordnet to get the meaning of a word and some possible relations with other words. It is supposed that these relations will get more semantic power to the search in the aim of understanding the texts to improve the results compared to a usual searcher. In general terms, it is possible to say that the based-meanings search, that Bianca performs, improve the returned results if these are compared to a usual based-word search. It returns results with better precision and recall.

In the future the structure of Bianca's index will be more efficient, to improve the index construction and the time to perform a query. This has to be done avoiding the generic structures [3] and using other ones more specialized.

Finally the disambiguation process has to be encouraged by means the use of more techniques.

Acknowledgements

Partially supported by PAC06-0059 SCAIWEB project, JCCM, Spain, and TIN2007-67494 F-META project, MEC-FEDER, Spain.

References

[1] R. Beckwith, G.A. Miller. (1990) Implementing a lexical network. In International Journal of Lexicography vol. 3 nº 4, pages 302 - 312, 1990.

[2] L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, W. Peters (1998). The EuroWordNet Base Concepts and Top Ontology en Piek Vossen (eds) EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht. 1998.

[3] S. Brin; L. Page (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7 / Computer Networks vol. 30(1-7), pages 107-117, 1998.

[4] C. Buckley (1995). Implementation of the smart information retrieval system, Technical Report TR85-686, Universidad de Cornell, 1985.

[5] A. Budanitsky, G. Hirst (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2001), Pittsburgh, PA, 2001.

[6] L. Codina (2003). Internet invisible y web semántica: ¿el futuro de los sistemas de información en línea?. Tradumática. Número 2, 2003.

[7] J. De la Mata, J.A. Olivas, J. Serrano-Guerrero (2004). Desambiguación de significados en GUMSe basada en el refinamiento del sentido. Actas del XII Congreso Español sobre Tecnologías y Lógica Fuzzy ESTYLF'04, Universidad de Jaén, España, pp. 315-320, 2004.

[8] J. De la Mata, J.A. Olivas, J. Serrano-Guerrero (2005) The Gaps of the Thesaurus Wordnet Used in Information Retrieval. R. Moreno Díaz et al. (Eds.): EUROCAST 2005, LNCS 3643, pp 229-234, 2005.

[9] J. De la Mata, J.A. Olivas, J. Serrano-Guerrero (2006). Mejorando la búsqueda Web mediante la adaptación de consultas en GUMSe. Actas del XIII Congreso Español sobre Tecnologías y Lógica Fuzzy ESTYLF'06, Universidad de Castilla-La Mancha, España, pp 261-266, 2006.

[10] C. Fellbaum (1990). English verbs as a semantic net. In: International Journal of Lexicography, vol. 3, nº 4, pp. 278 - 301, 1990.

[11] C. Fellbaum (eds) (1998). WordNet: An Electronic Lexical Database. The MIT Press, 1998.

[12] C. Lin and E. Hovy (2000). The automatic acquistion of topic signatures for text summarization. In Proc. of COLING, 2000.

[13] B. Larsen, C. Aone (1999). Fast and Effective Text Mining Using Linear-time Document Clustering, Proceedings of the KDD-99, San Diego, California, 1999.

[14] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross., K. Miller (1990). Introduction to WordNet: an on-line lexical database. In International Journal of Lexicography vol. 3 nº 4, pages 235 – 244, 1990.

[15] J.A. Olivas, P. Garcés, F.P. Romero (2003) An application of the FIS-CRM model to the FISS metasearcher: Using fuzzy synonymy and fuzzy generality for representing concepts in documents. International Journal of Approximate Reasoning (Soft Computing in Recognition and Search) 34, pp.201- 219, 2003.

[16] J.A. Olivas, J. De la Mata, J. Serrano-Guerrero (2004) Ontology Constructor Agent for Improving Web Search with GUMSe. Proceedings of the Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU'04, Perugia, Italy, pp. 1341 - 1348, 2004

[17] J.A. Olivas, J. De la Mata, J. Serrano-Guerrero, P. Garcés, F.P. Romero (2006) Desarrollo de motores inteligentes de búsqueda en Internet en el marco del grupo de investigación SMILe. En Olivas, J. A., Sobrino, A. (eds.): Recuperación de información textual, Text Information Retrieval. Universidade de Santiago de Compostela, pp. 89 – 102, 2006

[18] D. Parapar, A. Barreiro, A. (2005) Query Expansion using Wordnet with a logical model of information retrieval. IADIS International Conference, IADIS'2005, Algarve (Portugal), 2005.

[19] C.J. Van Rijsbergen (1979). Information retrieval, Butterworth, 1979.

[20] A. Soto, J.A. Olivas, M.E. Prieto (2006) Fuzzy Approach of Synonymy and Polysemy for Information Retrieval. Proceedings of the International Symposium on Fuzzy and Rough Sets, ISFUROS-2006, Santa Clara, Cuba, 2006.

[21] L.A. Ureña; M. De Buenaga, (1999). Utilizando WordNet para Complementar la Información de Entrenamiento en la Identificación del Significado de las Palabras. Inteligencia Artificial. Revista Iberoamericana de IA, Número 7, pag. 13-20, 1999.