RESLA 22 (2009), 77-88

AN ATTEMPT TO FORMALIZE WORD SENSE DISAMBIGUATION: MAXIMIZING EFFICIENCY BY MINIMIZING COMPUTATIONAL COSTS

Pascual Cantos
Aquilino Sánchez
Moisés Almela[1]
Universidad de Murcia

ABSTRACT. *This paper presents an algorithm based on collocational data for word sense disambiguation (WSD). The aim of this algorithm is to maximize efficiency by minimizing (1) computational costs and (2) linguistic tagging/annotation. The formalization of our WSD algorithm is based on discriminant function analysis (DFA). This statistical technique allows us to parameterize each collocational item with its meaning, using just bare text. The parameterized data allow us to classify cases (sentences with an ambiguous word) into the values of a categorical dependent (each of the meanings of the ambiguous word). To evaluate the validity and efficiency of our WSD algorithm, we previously hand sense-tagged all the sentences containing ambiguous words and then cross-validated the hand sense-tagged data with the automatic WSD performance. Finally, we present the global results of our algorithm after applying it to a limited set of words in both languages: Spanish and English, highlighting the points we consider relevant for further analysis.*

KEY WORDS. *WSD, lexicology, computational linguistics, applied linguistics.*

RESUMEN. *En el presente artículo se expone la estructura de un algoritmo para la desambiguación automática de significados a partir de colocados. El objetivo de nuestro algoritmo es lograr la máxima eficiencia reduciendo al mínimo (1) los costes computacionales y (2) el recurso a los corpus anotados o etiquetados. La formalización del algoritmo se fundamenta en el análisis de funciones discriminantes. Esta técnica estadística nos permite parametrizar cada uno de los colocados con su correspondiente significado, valiéndonos solamente del texto plano. Los datos parametrizados nos permitirán clasificar cada caso (frases con una palabra ambigua) en una variable de valores de dependientes (es decir, cada uno de los significados de la palabra ambigua). Para comprobar la validez y eficiencia de nuestro algoritmo desambiguador, desambiguamos primero manualmente el significado de la palabra estudiada en cada una de las frases en que ésta* *aparecía, para luego validar los datos clasificados con la aplicación automática del desambiguador de sentidos. Finalmente, presentamos los resultados globales de nuestro algoritmo, tras aplicarlo a una muestra de limitada de oraciones de ambas lenguas, español e inglés. Al mismo tiempo ponemos de relieve algunos de los aspectos que consideramos relevantes de cara a investigaciones o trabajos futuros.*

PALABRAS CLAVE. *Desambiguación automática de significados, lexicología, lexicografía, lingüística computacional, lingüística del corpus, lingüística aplicada.*

1. Introduction

Word sense disambiguation (WSD) dates back to the 1950s, when natural language processing (NLP) became a field of research. Interest in this field has been steadily increasing. Succinctly, WSD is the process of identifying the meaning of words in context.

Most NLP applications include subtasks to identify the various senses of polysemous words. Wilks and Stevenson (1996), for example, define WSD as an "intermediate task" in NLP, like part-of-speech-tagging or syntactic parsing, which serves as a means to an end defined by the application in which it is to be used. There are at least three "final tasks" which would seem to benefit from access to reliable WSD technology: machine translation, information retrieval and grammatical analysis (Ide and Véronis 1998).

Machine translation (MT): WSD is essential for the proper translation; a system of automatic translation from English to Spanish needs to translate the English *bank* as *banco* (*financial institution*) or *orilla de un río* (*river bank*), depending on the context.

Information retrieval: searching for documents, for information within documents and for metadata about documents by means of specific keywords, requires the elimination documents where the specific keywords are used in an inappropriate sense; for example, when searching for medical references, it is desirable to get rid of documents where the word *cancer* is associated with a constellation, rather than with a malignant disease.

Grammatical analysis: WSD is useful for part-of-speech tagging; for instance, in the English sentence *Time flies like an arrow*, it is necessary to disambiguate the sense of *flies*, as it can mean a *sort of insects* or *action of flying*.

WSD has been recognized as one of the most important problems in MT (Bar-Hillel, 1960). Bar Hillel stated that machines can only achieve fully automatic high quality machine translation (FAHQT) only if they succeed in processing meaning. He proclaimed that "sense ambiguity could not be resolved by electronic computer either current or imaginable", and used the following example, containing the polysemous word *pen*, as evidence: *Little John was looking for his toy box. Finally he found it. The box was in the* pen. *John was very happy*. The word *pen* may have two meanings (*something to write with* and *a container of some kind*). But how could the machine decide on the right meaning of *pen* in this sentence? What humans found so easy to discriminate, machines would never be able to 'understand'. Analysis of the example shows that this is a case where context and selectional restrictions fail to disambiguate *pen*. It is particularly Bar-Hillel's views on FAHQT –no doubt an influence on the deliberations of the *Automatic Language Processing Advisory Committee*-report (ALPAC, 1966)– that have had most impact, causing the U.S. Government to reduce its funding of MT dramatically.

In the 1970s, most attempts to solve the problem of WSD were based on artificial intelligence (AI) approaches, such as preference semantics. Nevertheless, the unavailability, at that time, of large machine-readable knowledge repositories was a mayor drawback. This changed dramatically in the 1980s, with the creation of large-scale knowledge sources and subsequent automatic knowledge extraction methods. In the 1990s, we find a further turning point with an increasing application of statistics to WSD and the periodic evaluation exercises for WS programs (SENSEVAL)[2].

2. Motivation and Intuition

The growing concern about WSD can be linked with the generalized feeling in the WSD community that change is necessary. New issues should guide the discussion in forthcoming research. Agirre and Edmonds (2006) compare two different "routes forward." The first direction concentrates on the role of WSD in computational linguistics; the second direction focuses on the application of WSD to specific NLP tasks. The present paper follows the first direction rather than the second. Consequently, it is a must for the model to reconcile computational tractability with linguistic-theoretical adequacy.

Our WSD approach elaborates upon the Firthian postulate of *meaning by collocation*, according to which the actual sense of a word is lexically codified in the forms of its syntagmatic environment. Thus, collocation-based semantic analysis provides an access to meaning via surface text. Computationally, the axiom of *meaning by collocation* has the advantage of minimizing the dimensions of the feature space. The search for disambiguating clues in context relies only on surface co-occurrence data, hence it dispenses with any kind of "deep" linguistic knowledge or enriched feature representation.

The linguistic underpinnings of our approach can be illustrated as follows: assume we have a polysemous word *w* with three different meanings m_1, m_2 and m_3. If we take for granted that each actual sense of a word is lexically codified in the forms of its syntagmatic environment, we find that each meaning (m_1, m_2 and m_3) has accordingly a number of associated collocates. That is, for meaning 1 (m_1) we find two collocates ($m_{1\text{-}c1}$ and $m_{1\text{-}c2}$); *for* meaning 2 (m_2), three collocates ($m_{2\text{-}c1}$, $m_{2\text{-}c2,}$ and $m_{2\text{-}c3}$); and for meaning 3 (m_3), seven collocates ($m_{3\text{-}c1}$, $m_{3\text{-}c2,}$ $m_{3\text{-}c3}$, $m_{3\text{-}c4}$… $m_{3\text{-}c7}$).

This results into three main meanings, defined by their respective sets of collocates, as illustrated in Figure 1.

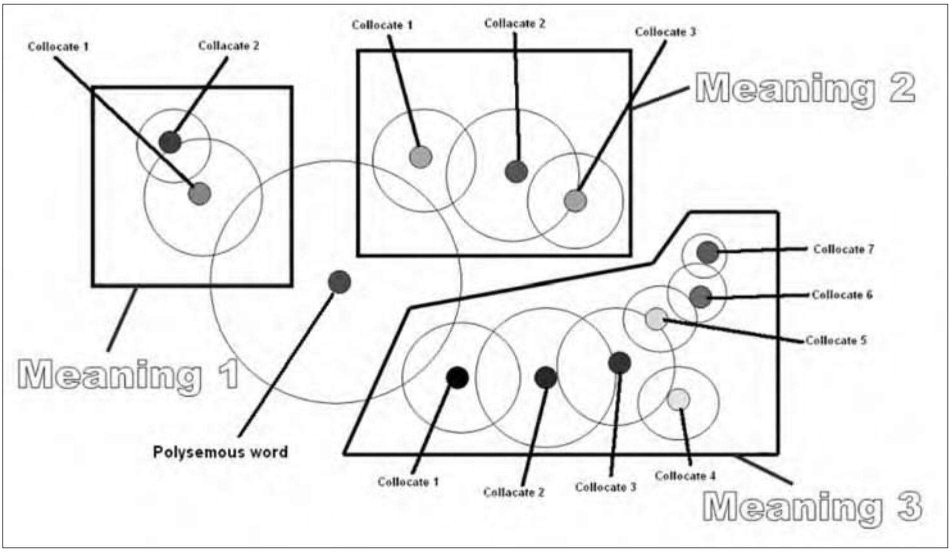


Figure 1. *Meanings and Collocates.*

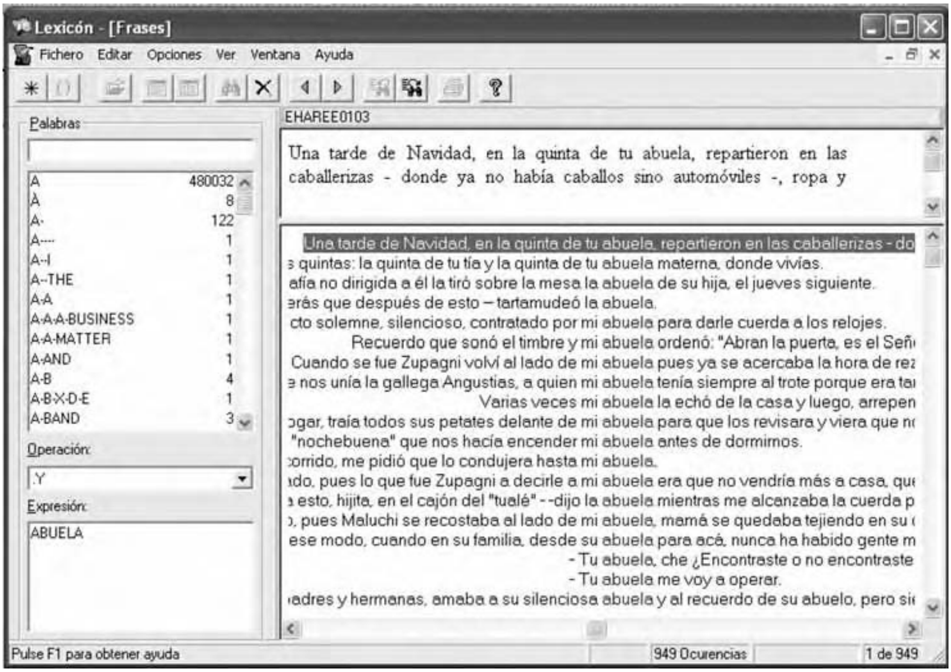
3. Modelling the Idea

As already mentioned above, distinct meanings of the same word attract different co-occurrence data. Elsewhere, we have analysed the distribution of co-occurrences of the Spanish noun *abuela* (grandmother) in a sense-tagged sub-corpus (Almela et al. 2006). Now, to model this idea, we need a formal method that involves the predicting of a categorical dependent variable (meaning) by one or more continuous or binary independent variables (collocates). One possible method is using discriminant function analysis (DFA). DFA will be used to determine which variables (collocates) discriminate between two or more naturally occurring groups (meanings).

We applied the DFA to a polysemous word *w* starting from a set of collocational data with *n* entries. The number of entries is determined by the number of sentences containing *w*, that is, as many entries as sentences containing *w*, irrespective of the meaning of *w*. For instance, if we take the Spanish noun *abuela* (*w*) and extract its concordances in a corpus (*Cumbre 20*), we get 949 concordance sentences (*n = 949*).

For each of the *n* entries, we extracted *p* numeric independent variables (collocational data), defining the profile of features of each *n.* Consequently, *p* becomes the window-span, that is, the span around the node word (*w*) which we will take as collocational data for our DFA.

An additional quantitative dependent variable is considered with as many categories as word senses *w* has. This variable is used to assign group membership (meaning *m*) and to define the group to which each sentence (or item) belongs to. The resulting table is of size table *n*(p+1)*, where each case appears with its profile and a group membership assignment.



EHAREE0103

Una tarde de Navidad, en la quinta de tu abuela, repartieron en las caballerizas - donde ya no había caballos sino automóviles -, ropa y

Una tarde de Navidad, en la quinta de tu abuela, repartieron en las caballerizas - do
s quintas: la quinta de tu tía y la quinta de tu abuela materna, donde vivías.
afía no dirigida a él la tiró sobre la mesa la abuela de su hija, el jueves siguiente.
erás que después de esto – tartamudeó la abuela.
cto solemne, silencioso, contratado por mi abuela para darle cuerda a los relojes.
Recuerdo que sonó el timbre y mi abuela ordenó: "Abran la puerta, es el Señ
Cuando se fue Zupagni volví al lado de mi abuela pues ya se acercaba la hora de rez
e nos unía la gallega Angustias, a quien mi abuela tenía siempre al trote porque era ta
Varias veces mi abuela la echó de la casa y luego, arrepen
ogar, traía todos sus petates delante de mi abuela para que los revisara y viera que n
"nochebuena" que nos hacía encender mi abuela antes de dormimos.
corrido, me pidió que lo condujera hasta mi abuela.
ado, pues lo que fue Zupagni a decirle a mi abuela era que no vendría más a casa, qu
a esto, hijita, en el cajón del "tualé" --dijo la abuela mientras me alcanzaba la cuerda p
), pues Maluchi se recostaba al lado de mi abuela, mamá se quedaba tejiendo en su
ese modo, cuando en su familia, desde su abuela para acá, nunca ha habido gente m
- Tu abuela, che ¿Encontraste o no encontraste
- Tu abuela me voy a operar.
adres y hermanas, amaba a su silenciosa abuela y al recuerdo de su abuelo, pero si

Figure 2. *Concordance extraction;* w = *abuela and* n = *949.*

And this is how it works:

1. Extract *n* entries (concordances) containing the ambiguous word *w* under investigation.
2. Determine the window-span; for instance, 10 words, 5 on each side of the node word (-5 +5), giving *p* = 10.

n	-5	-4	-3	-2	-1	w	+1	+2	+3	+4	+5
1	en	la	quinta	de	tu	ABUELA	repartieron	en	las	caballerizas	donde
2	y	la	quinta	de	tu	ABUELA	materna	donde	vivías		
...	...	...	...	...	...	...	...	...	...	...	...
18					Tu	ABUELA	me	voy	a	operar	

Table 1. *Sample of tabulated data.*

3. Next, to all instances *n* a further variable was added, *m*, assigning the meaning of the word *w* in each instance *n.*

The discriminant mathematical model will be obtained out of the table *n*(p+1)* and might allow us to examine the profile of new items (sentences containing *w*) and assign them to the most likely group (meaning).

Intuitively, this is how the algorithm performs: once the concordance sentences are extracted and tabulated (see *Figure 2* and *Table 1*), the algorithm transforms all collocates found in positions -5 to +5 into numerical values and parameterizes all the data. By doing this, it determines that n_1 and n_2 are quite similar, whereas n_{18} is different to the two previous ones. Consequently, our DFA model establishes a single *discriminant function* that allows the categorization of two possible meanings found:

1. n_1 and n_2 conform to the same meaning: m_1
2. n_{18} conforms a meaning on its own: m_2

4. DFA FOR WSD

Additionally, some previous data preparation is needed, such as:

1. Removing repeated items, since they are just duplicated sentences with no additional information and would increase the volume of the data and slow down the algorithm performance.
2. Normalizing the data of the classifying variables by means of a logarithmic transformation.
3. Computing the mean of all classifying variables; and in order to relate all means we computed the grand mean. In addition, the value of the grand mean will be negative signed in those instances where no collocate occurs (empty position). This is done as a centralization measure and to diminish the dispersion of the data.
4. Removing outliers from the data set; outliers can be a major source of skewness in the data set. Therefore, it is important to exclude outliers so that they do not introduce possible bias into our analysis.

In what follows we shall illustrate the algorithm performance on an example, the Spanish common noun: *abuela*. The five meanings analysed are:

1. *La madre del padre o de la madre de una persona* (The mother of one's father/mother)
2. *INFML (frec des) Mujer anciana o de avanzada edad* (Elderly lady)
3. *INFML indica incredulidad o duda por parte del oyente* (Something that produces doubts or incredulity on the part of the listener)
4. *INFML se dice irónicamente de una persona que se alaba a sí misma en exceso* (Used ironically of a person who praises her/himself in excess)

5. *VULG indica, irónicamente, el aumento inoportuno de personas o cosas cuando ya hay muchas o demasiadas en un lugar* (Used ironically to express the inopportune increase of people or things when there are already many or too many in a place) (From Sánchez 2001)

From the *Cumbre* Corpus (20 million version), we extracted all concordance sentences with the noun *abuela* and classified them according to the meanings above; the resulting sense distribution was the following:

Sense	Counts	%
1	893	94,10
2	32	3,37
3	4	0,42
4	14	1,48
5	6	0,63
	949	100,00

Table 2. *Sense distribution for* abuela.

Next, we evaluated and interpreted the following output data:

1. Wilks' λ tests the null hypothesis, showing whether the variables used discriminate positively or not. *Table 3* indicates that all variables discriminate significantly (all *sig. values* are < *0,05*) that is, among the sentences (population) the meanings (groups) do not differ from one another on the mean for any of the discriminant functions. This Wilks' λ is evaluated with a chisquare approximation (values of λ close to 0 are statistically significant and indicate that the variables discriminate).

Test of function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 4	,307	367,165	40	,000
2 through 4	,678	120,521	27	,000
3 through 4	,847	51,516	16	,000
4	,945	17,635	7	,014

Table 3. *Wilks' λ. The* sig. values *are all statistically significant.*

2. Eigenvalues also called the *characteristic roots* of each discriminant function, reflect the ratio of importance of the dimensions which classify cases of the

dependent variable. The greater the values, the more discrimination power the function has. *Table 4* reflects that Function 1 has the most discrimination power (1,213), explaining 74,2 % of the whole variance.

Function	Eigenvalue	% of Variance	Cumulative %	Canonical correlation
1	1,213	74,2	74,2	,740
2	,249	15,2	89,4	,446
3	,115	7,0	96,4	,322
4	,058	3,6	100,0	,235

Table 4. *Eigenvalues and canonical correlation.*

a. The canonical correlations show that all functions discriminate, being Function 1 the most powerful discriminator of all, with a score of 0,74.
b. The cross validation reveals a high accuracy percentage of the DFA model: 96,9%.

3. One of the most positive and powerful contributions of DFA is that once the functions are known, we can construct a model that allows us prediction of membership (meaning). This is done by means of the resulting discriminant function coefficients (*Table 5*).

	Function			
	1	2	3	4
pre5log	,033	,023	,040	-,077
pre4log	-,123	,113	,232	,064
pre3log	,028	,068	-,165	,043
pre2log	,036	-,226	-,004	-,137
pre1log	-,118	,258	-,038	,168
pos1log	,070	-,025	,169	,079
pos2log	,181	-,170	,000	,030
pos3log	,126	-,121	-,037	,192
pos4log	,139	,208	,108	,017
pos5log	,242	,131	-,082	-,212
(Constant)	-4,153	-1,114	-,535	-,666

Table 5. *Discriminant function coefficients.*

4. Finally, we get centroids, that is, the mean discriminant scores for each of the dependent variable categories for each of the discriminant functions. We want the means to be well apart to show that the discriminant function is clearly discriminating. The closer the means, the more errors of classification there likely will be. To illustrate its usefulness, consider the centroids for the noun *heart*. *Heart* has four 'core' meanings (1-4): the centroids of meanings 1, 2 and 3 are clearly distinct, whereas meaning 4, clearly overlaps with meanings 1 and 2. The visual representation of centroids might allow us to evaluate the effectiveness of the DFA model and see which senses are more likely to be positively modelled and which are more likely to present problems for automatic WSD.

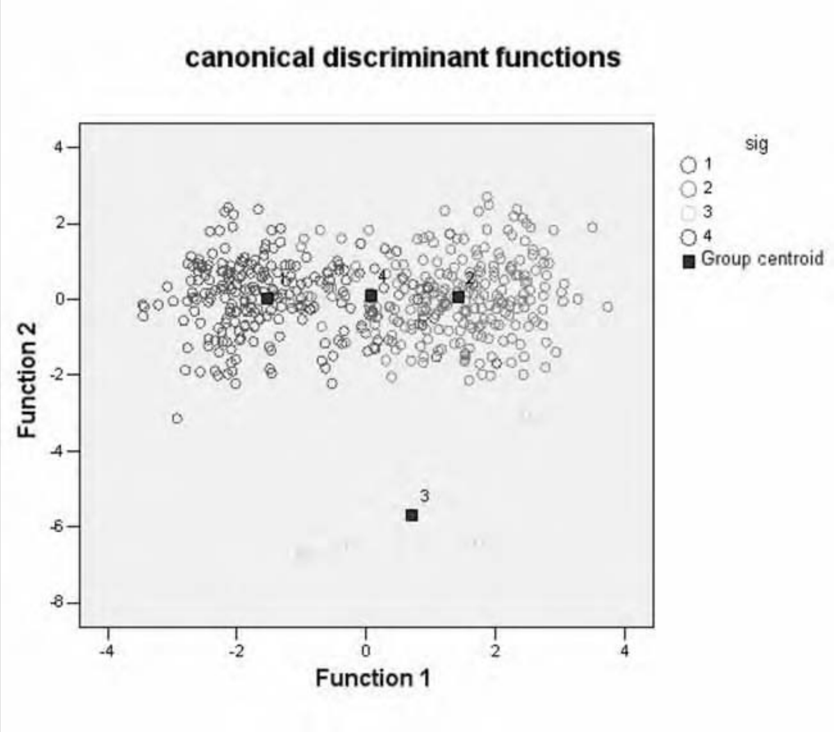


Figure 3. *Centroids.*

5. Sample and global results

The model was applied to a reduced sample of items, in English and Spanish. The Spanish words disambiguated were 46 (*familia, familias, abuela, abuelas, abuelo, abuelos,* *hermana, hermanas, hermano, hermanos, hija, hijas, hijo, hijos, padre, padres, madre, madres, prima, primas, primo, primos, tía, tías, tío, tíos, boca, bocas, brazo, brazos, corazón, corazones, dedo, dedos, mano, manos, ojo, ojos, pie, pies, piel, pieles, pierna, piernas, sangre, sangres*), and the English ones, 44 (*family, families, grandfather, grandfathers, grandmother, grandmothers, father, fathers, mother, mothers, son, sons, daughter, daughters, cousin, cousins, aunt, aunts, uncle, uncles, sister, sisters, brother, brothers, mouth, mouths, arm, arms, heart, hearts, finger, fingers, hand, hands, eye, eyes, skin, skins, leg, legs, foot, feet, blood, bloods*). The results, taken as a whole, were highly satisfactory, and in many instances above the average in similar studies, as Figure 4 reveals:

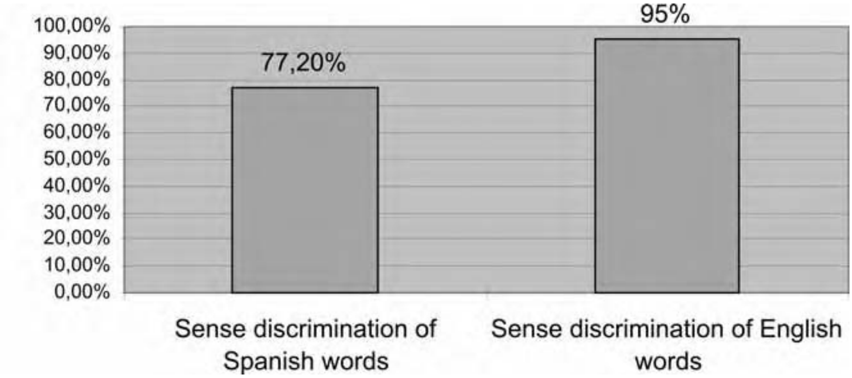


Figure 4. *WSD of Spanish and English sentences.*

The analysis of results offers interesting data for further comments and future research on the performance of our WSD algorithm in each one of the languages and in each one of the words of the sample. It is obvious, for example, that the efficiency reached is higher in English than in Spanish. At the same time, the rate of success differs among the words disambiguated: it ranges between 60% (the lowest in the scale) and 98% (the highest in the scale). The degree of semantic granularity of the senses also affects the success rate. Those data deserve however, a more detailed and thorough study and comment.

6. Conclusions

It is precisely the initial robustness of the different distribution of co-occurrence data (Almela et al. 2006) that has motivated the present study, on the assumption that distinct meanings of the same word attract different co-occurrence data.

Our first goal was to try to model this behaviour in a most economical way. That is, low computer cost and raw corpus data. The starting point was extracting full concordance sentences, all containing the same ambiguous word and hand-sense-tagged the sentences according to the meaning of that word, according to the sense definitions of a standard paper dictionary. This supervised method gave us valuable data on sense distributions and co-occurrence data around the sense distributions.

One of the revealing findings was the little overlapping of co-occurrences among senses, which is very much in favour for continuing experimenting with Lesk's based algorithms (Lesk 1986, Cowie et al. 1992, Stevenson and Wilks 2001, etc.), using real co-occurrence and/or collocational data extracted from a corpus (Cantos 1996), instead of sets of dictionary entries

NOTES

1. This research was financially supported by the Spanish Ministry of Education and Science and by the Murcian Government. The first research project (Ref.: HUM2004-00080/FILO) was funded by the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica. The second project (00481/PI/04) was funded by Fundación Séneca, Comunidad Autónoma de la Región de Murcia.
2. The interested reader can refer to Ide and Véronis (1998) for an in-depth early history of WSD.

REFERENCES

Agirre, E. and Edmonds, P., eds. 2006. *Word Sense Disambiguation. Algorithms and Applications*. Dordrecht: Springer.

Agirre, E. and Edmonds, P. 2006. "Introduction". Eds. E. Agirr, y P. Edmonds. 1-28.

Almela, M. 2006. *From Words to Lexical Units: A Corpus-Driven Account of Collocation and Idiomatic Patterning in English and English Spanish*. Frankfurt am Main: Peter Lang.

Almela, M., Sánchez, A. and Cantos, P. 2006. "Lexico-Semantic Mapping of Meanings in English and Spanish: A Model of Analysis". *Aspects of Translation*. Ed. J. M. Bravo. Universidad de Valladolid (2006). 11-43.

ALPAC. *Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences*. Washington D.C.: National Academy of Sciences, National Research Council.

Bar-Hillel, Y. 1960. "The present status of automatic translation of languages". *Advances in Computers*. 191-163.

Bruce, R. and Wiebe, J. 1994. "Word-Sense Disambiguation Using Decomposable Models". *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM, 1994*. 139-146.

Cantos, P. 1996. *Lexical Ambiguity, Dictionaries, Corpora*. Servicio de Publicaciones de la Universidad de Murcia.

Cantos, P. and Sánchez, A. 2001. "Lexical Constellations: What Collocates Fail to Tell". *International Journal of Corpus Linguistics* 6 (2): 199-228.

Cowie, J., Guthrie, A. and Guthrie, L. 1992. "Lexical Disambiguation. Using Simulated Annealing". *Proceedings of the 14th International Conference on Computational Linguistics (COLING), Nantes, France, 1992*. 359–365.

Duda, R. O., Hart, P. E. and Stork, D. G. 2001. *Pattern Classification*. New York: John Wiley & Sons.

Gale, W. A., Church, K. W. and Yarowski, D. 1992. "Work on Statistical Methods for Word Sense Disambiguation". *AAAI Fall Symposium Series: Probabilistic Approaches to Natural Language* (*Working Notes*). Cambridge, MA. 54-60.

Ide, N. and Véronis, J. 1998. "Word Sense Disambiguation: The State of the Art". *Computational Linguistics* 24 (1). 1–40.

Ide, N. and Wilks, Y. 2006. "Making Sense about Sense". Eds. E. Agirre, y P. Edmonds. 47-74.

Kilgarriff, A. 1993. "Dictionary Word Sense Distinctions: An Enquiry into their Nature". *Computers and the Humanities* 26. 365-387.

Kilgarriff, A. 1997. "I Don't Believe in Word Senses". *Computers in the Humanities* 31 (2). 91–113.

Kilgarriff, A. 2006. "Word Senses". Eds. E. Agirre y P. Edmonds. 29-46.

Leacock, C., Miller, G., Randee, T. and Bunker, R. 1993. "A Semantic Concordance". *Proceedings of the 3rd DARPA Workshop on Human Language Technology, Plainsboro, New Jersey, 1993*. 303-308.

Lesk, M. 1986. "Automated Sense Disambiguation. Using Machine-Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone". *Proceedings of the 1986 ACM SIGDOC Conference, Toronto, Canada, 1986*. 24-26.

Pustejovski, J. and Boguraev, B. 1996 (1996). *Lexical Semantics: The Problem of Polysemy*. Clarendon, Oxford.

Sánchez, A., Sarmiento, R., Cantos, P. and Simón, J., eds. 1995. *CUMBRE. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*. Madrid: SGEL.

Sánchez, A., ed. 2001. *Gran Diccionario de Uso del Español Actual*. Madrid: SGEL.

Stevenson, M. and Wilks, Y. 2001. "The Interaction of Knowledge Sources in Word Sense Disambiguation". *Computational Linguistics* 27 (3). 321-349.

Wilks, Y. and Stevenson, M. 1996. *The grammar of sense: Is word-sense tagging much more than part-of-speech tagging?* Sheffield Department of Computer Science, Research Memoranda, CS-96-05 (1996).