Presented in the XXVII Spanish Applied Linguistics Association (AESLA) International Conference. Ciudad Real 2009.

DIFFERENT SENSES VISUALIZATION WITH LATENT SEMANTIC ANALYSIS

GUILLERMO JORGE-BOTANA, JOSÉ A. LEÓN, RICARDO OLMOS
Universidad Autónoma, Madrid

SUMMARY

Some psycholinguistic based algorithms have been proposed to emulate the management of the context that humans do, in the assumption that the value of a word is evanescent and take sense when interact with other structures. For example, predication algorithm (Kintsch, 2001), uses the vector representation of the words that produce LSA (Latent Semantic Analysis) to dynamically simulate the comprehension of predications and even, the comprehension of predicative metaphors. The objective of this presentation is to predict some unwanted effects that could be present in the vector-space models when extracting different senses to a polysemic word (Predominant sense inundation, Accurate-less and Low-level definition) using a Spanish corpus, and give some proposals based on such algorithms.

Palabras clave: Latent Semantic Analysis, predication algorithm, Context, Visualization.

RESUMEN

Se han propuesto algunos algoritmos para emular el manejo del contexto que hacen los humanos, en la asunción de que el valor que tiene una palabra es evanescente y cobra sentido cuando interactúa con otras estructuras. Por ejemplo, el algoritmo de predicación (Kintsch, 2001) emplea la representación vectorial de las palabras que se extraen de LSA (Latent Semantic Analysis) para simular la dinámicamente la comprensión de predicaciones e incluso de metáforas predicativas. El objetivo de esta presentación es, empleando un corpus en español, predecir algunos efectos indeseados que podrían presentarse en los modelos espacio-vectoriales al extraer sentidos a una palabra polisémica (Inundación de

sentido predominante, falta de certeza en la definición y definición de bajo nivel) y dar algunas propuestas basadas en esos algoritmos.

1. INTRODUCTION

LSA was first described as an information retrieval method (Deerwester et al.,1990) but Landauer & Dumais (1997) suggested that LSA is a good attempt to emulate the kind of human advantages that concern the capture of the deep relations of words. For instance, the problem called "stimulus poverty" or "platonic problem" which tends to formalize how people have more knowledge that they can possibly get, attending to the information they are generally exposed to. The solution is that a functional architecture such as LSA would allow to make an induction from the general environment, that is, that the reduced vectorial space representation of LSA permits the inferrence that some words have to do with another even if they have not been together in any sentence, paragraph, conversation, etc. Even more, Landauer & Dumais shows with a simulation, that for the acquisition of knowledge about a word is also very important the texts in which that word does not appear. And the more frequent a word is, the more benefit can be found in the texts in which it does not appear. These estimations agree with those studies that measure the capacity of relations of order-n (order more than one) to induce knowledge (Kontostathis & Pottenger, 2006; Lemaire & Denhière, 2006). But moreover to formalize acquisition and representation theories, LSA has been used as well to formalize word disambiguation in the frame of language comprehension, in the assumption that the value of a word is evanescent (Kintsch, 1998), and gains sense when interacting with other structures.There is no way to disambiguate some words (polisemy and homonmy) if we don´t take into account any context. Perhaps the mind triggers some automatic general senses but no full possibility is given to us to conclude such senses. But if we have the context, it will be formed an evanescent and temporary representation generated on-line with some linguistic and non linguistic contents. To represent this issue, ideally and parsimoniously, we need only two components, a flexible representation of words and a mechanism that can give the correct

shape in a given context (Kintsch, 2007), and take in account the representation biases. LSA can be a good base to represent words and texts in discrete values. It has a clear metric, is a data-driven technique and has a flexible vector representation, which supply to this statistical approach some advantages, for instance, over ontologies (Dumais, 2003), for implementing some efficient algorithms like predication algorithm (Kintsch, 2001).

2. PREDICATION ALGORITHM OPERATING OVER LSA.

As we saw above, the way to resolve polysemy is to retrieve the right contents and take the bad contents away. Instead of using the simple sum of vector of the terms of a predication, predication Algorithm (Kintsch, 2001) tends to do this following some simple principles based in previous models of discourse comprehension like Construction-Integration nets (Kintsch, 1998). Those principles are based in some rules of node activation that expose how the final vector that represents a predication has to be formed with the vectors of the most activated words in the net. This activation came from the two terms of the predication (*Predicate* and *Argument*) to all words of semantic space. It is assumed that the more activated nodes are the more pertinent words for the Predicate (P) and also for the Argument (A). The necessary steps to apply this algorithm are as following: 1) find *n* first terms with more similarity with the Predicate (P), being *n* an empiric parameter. 2) construct a net with connections among all that terms and the Predicate (P) and the argument (A). The connection weight of the net is the cosine (as similarity measure) between the words that are connected. 3) It can implement some inhibitory connection between terms in the same layer, 4) obtain stable activation nodes with a function that uses excitatory and inhibitory connections and prime bilateral excitation (before this, the net does not need any cycle of trial because the definitive weighs are those that impose the LSA matrix) 5) The final vector of the predication is calculated with the sum of Predicate-vector (P), Argument-Vector (A) and the *p* more activated vector-nodes (Again, *p* is an empirical value and $p < n$)

4. OBJETIVES

The main aim is to analyze the disambiguation of a polysemical word in a retrieval context and to visualize its semantic network. Although there has been some articles in which a bigger unit is taken to formalize the context, like sentences or paragraphs, the case of predication is the purest form in where, a simple context-word modulates a simple word. This small window of contexts has been called in other studies 'micro-contexts' (Ide & Veronis, 1998). We adapt the architecture of Kintsch´s algorithm to word-context structures (where the context is another word). The word "hoja" takes one meaning or another depending on their context: "tinta" or "rosal". Such a frame will be our starting point attending to the possible behaviour of a system as LSA or any vectorial space model in front of these kinds of structures.

3. SOME USUAL PROBLEMS IN THE EXTRACTION OF THE SENSE FROM POLYSEMICAL WORDS WITH EXPLICIT SCENT OF THE CONTEXT.

We consider three potential problems concerned with system managing about polysemic words and the extraction of semantic related neighbours:

(I) Predominant sense inundation**:** Usually, with the simple sum of vectors, it is possible that the sense that the context promotes never arises, because the word-context doesn't have enough representation to cope with the influx of the predominant sense of the word.
(II) Accurate-less definition: It is possible that the right senses arise but without an accurate definition.
(III) Low-level definition: It is possible that the right senses arise, but terms that represent such senses are restricted to local relations with the polysemic word. It seems that most of the neighbours extracted have a 1-contingent with the polisemic word (the one that we want to extract neighbours). Said otherwise, neighbours appear with the polysemic word, but not without it.

5.METHOD

4.1 LSA, corpus, and pre-process

LSA was developed with the Spanish corpus LEXESP (Sebastián, Cuetos, Carreiras & Martí, 2000) in a "by hand" lemmatized version (plural forms are transformed into their singular ones and feminine are transformed into their masculine ones and all the verbs are standardized in their infinitive form). We chose sentences as process units. We deleted words that appear in less than seven documents. This ensures a minimal representation of the terms analyzed. Then, we applied Entropy pre-process. Finally, we had a matrix with 18.174 terms in 107.622 documents where we applied SVD algorithm and reduced the three resulting matrices to 270 dimensions. For this purpose, we used Gallito ®, a LSA tool implemented in our research group with developments in .Net ® integrated with Matlab ®. To visualize the graphs, we used the Pajek package (*Kamada* –Kawai algorithm).

4.2 Methodology

Our methodology is as follows. We extracted (with simple cosine) a list of neighbours to each structure (word and word-context) in order to make a graph and draw it. To avoid the problem (III) (Low-level definition), we also extracted the terms that represent each sense, adjusting the simple cosine with the vector length of each term of the semantic space when compared with the vector of the query.

[Similarity = Cosine (A, I) x log (1 + Vector length (I))],

where A is the term from which we want to extract the neighbors and I is each of the terms in the semantic space.

The use of these two methods produced a list with terms that came from local relations and words that are less constrained for these

kind of relations. This method showed good results in a previous study in a specific domain corpus (Jorge-Botana, Olmos & León, accepted), but we wanted to investigate its effects in a general domain corpus like LEXESP. To avoid (I) Predominant sense inundation and (II) Low-level definition, we applied the Kintsch's algorithm, to filter the contents attending to the two terms of the query (even if they don't form a real predication as word-context pairs).

5. PROCEDURE

In this example, for the first graph, are extracted the 30 first neighbours of the word "hoja". For the second graph, the first 30 neighbours of the two word-context pairs ("hoja-rosal" and "hoja-tinta") are extracted, calculating the vectors of each structures using the sum of vectors. Finally, for the third graph, we used our suggested method, calculating the vectors of such structures with Kintsch's algorithm and extracting (for the each structures) the first 30 neighbours plus 30 more neighbours adjusting the simple cosine with the vector length.

CONDITION	STRUCTURE	NUMBER OF NEIGHBORS		
		COS	COS (VL)	
[1] ISOLATED	WORD	30	X	to graph
[2] SUM OF VECTORS	WORD(CONTEXT WORD 1)	30	X	to graph
	WORD(CONTEXT WORD 2)	30	X	
[3] PREDICATION ALGORITHM	WORD(CONTEXT WORD 1)	30	30	to graph
	WORD(CONTEXT WORD 2)	30	30	

Table 1. From the three conditions, several lists of 30 neighbours are extracted. [1] one list from isolated condition. [2] from the two structures formed with the Sum of vector, two lists of 30 are extracted. [3] from the two structures formed with the Kintsch's algorithm, two lists of 30 are extracted plus two lists of 30 neighbours adjusted with the vector length (VL).

6. EXAMPLES

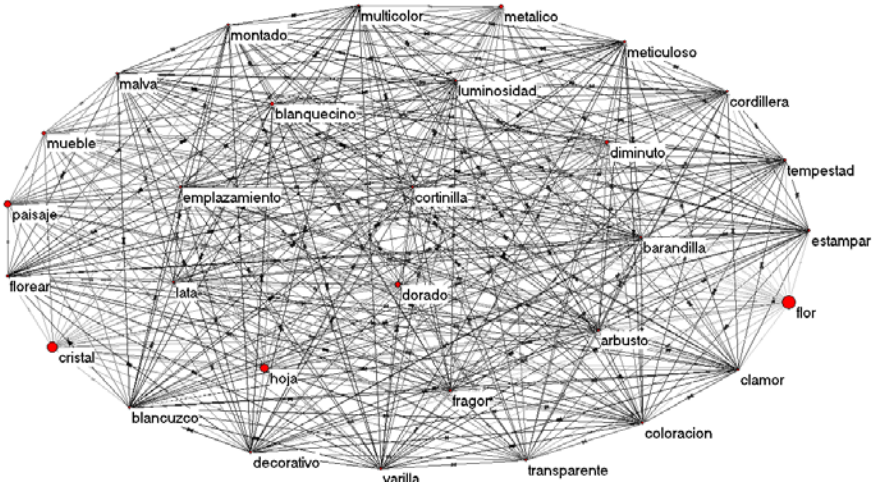

Figure 1. Visual representation of the word “hoja” without any scent of context. The 30 first neighbours are represented.

As was explained above, first, we drew a condition in which the word is alone without any of its contexts and we have seen the base line of the polysemical word and its definition-biases (figure 1). The nodes were usually filled with words that do not configure a real definition. It is an amalgam of contents plus the predominant contents. Next, when we tested the word followed by its contexts, we made it using two methods to solve the two structures (“hoja-tinta” and “hoja-rosal”). The first (figure 2), with the simple sum of vectors which kept some of the definition-biases of the first one. The contents are more or less the same but it seems that the predominant contents have flooded a great majority of the nodes.

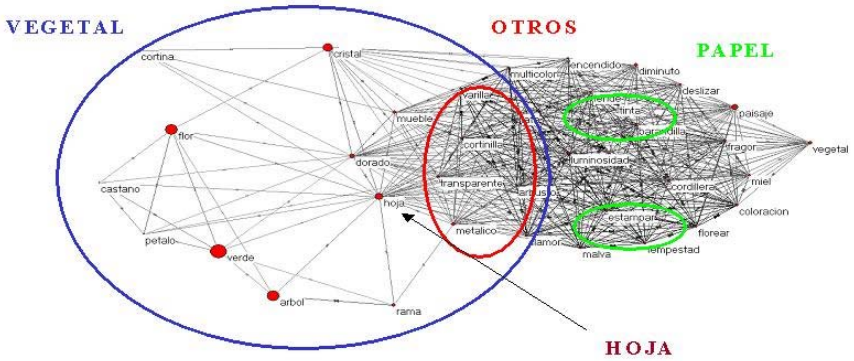

Figure 2. Visual representation of the word “hoja” in the domain of two contexts, “rosal” and “tinta”. The simple sum of vectors is applied. 30 neighbours of hoja-rosal, 30 neighbours of hoja-tinta.

The second one (figure 3), with the predication algorithm and corrected with the vector length. This picture shows two different areas. One for one context and one for the other, and also with words that came from local relations and as well as words that are less constrained for these kind of relations, which are located in the extremes. Now, there is not what we called Predominant sense inundation effect. This time, the definitions of the exemplars were better adjusted to the topic, avoiding the Accurate-less definition effect. Also, we managed to avoid the Low-level definition.

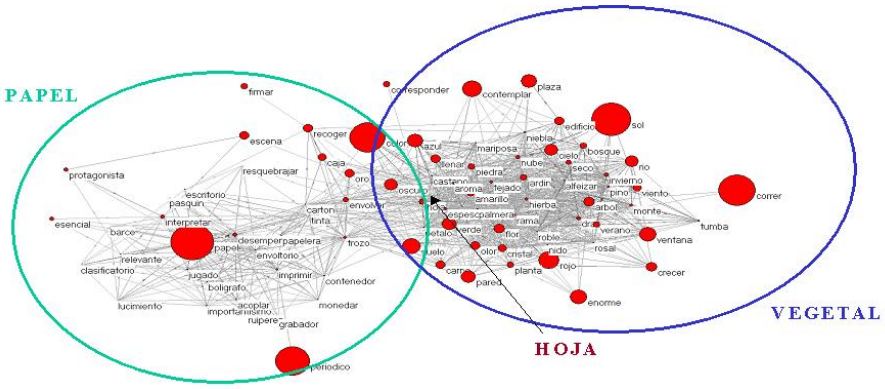

Figure 3. Visual representation of the word “hoja” in the domain of two contexts, “rosal” and “tinta”. The Kintsch’s algorithm is applied and also adjusting with the vector length.. 30 neighbours of hoja-rosal, 30 neighbours of hoja-tinta plus 30 adjusted neighbours of hoja-rosal , 30 adjusted neighbours of hoja-tinta. Note the big vertices that means well represented terms.

CONCLUSION

In summary, the storage and the retrieval are not independent processes. It does not matter what kind of shape, a word has been stored. It does not matter if representation has an aberrant shape in storage (a very atypical one as claim Deerwester et al., 1990). What is important is that, like things in nature, linguistic structures that have the correct properties depend on retrieval context. The goal is to get a good management from flexible representation. This is the same as what Kintsch (1998) claimed when he said that knowledge is relatively permanent (the representation that supplies LSA) but the meaning, that is, the portion of the net that is activated, is flexible, changeable and temporary.

In this presentation, we have presented a protocol to visualize the contexts that a word can have, and we have followed a sequence to show the contents of a word in origin; the contents of a word with contexts but without using predication algorithm and the contents of a word with arguments, but this time, using predications algorithm. The visual net has demonstrated that good management of the context ensures a good representation of the meanings that we want to retrieve. These kinds of human-based methods could be used in retrieval applications or in indexing or tagging machines.
Also, the nets produced by these methods could be used as a visual information retrieval interface (VIRI) allowing the users to visually recognize the information that is needed, instead of writing a query in the search boxes or providing an overview of a semantic domain, helping the user to know what information can be retrieved using the interface.

BIBLIOGRAPHY

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., y Harshman, R. (1990). Indexing By Latent Semantic Analysis. *Journal of the American Society For Information Science, 41*, 391-407.

Dumais, S. T. (2003). Data-Driven approaches to information access. *Cognitive Science, 2*, 491-524.

Jorge-Botana, G., Olmos, R., León, J.A. Using LSA and the predication algorithm to improve extraction of meanings from a diagnostic corpus. Spanish Journal of Psychology. Accepted December 2008.

Ide, N., & Véronis, J. (1998). Word sense Disambiguation: The state of the Art. *Computacional Linguistics, 24* (1), 1-41.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W. (2001). Predication. *Cognitive Science, 25*, 173-202.

Kintsch, W. (in press). Symbols systems and perceptual representations. In M. de Vega, A. Glenberg & A. Graesser (Eds.), Symbols, Embodiment, and Meaning. Oxford: UniversityPress.

Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding LSI performance. *Information Processing and Management, 42 (1)*, 56-73.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Lemaire, B., & Denhière, G. (2006). Effects of High-Order Co-occurrences on Word Semantic. Similarity. *Current Psychology Letters*, 18, 1.Web site: http://cpl.revues.org/document471.html.

Sebastián-Gallés, N., Martí, M.A., Carreiras, M., & Cuetos, F. (2000). *LEXESP: Una base de datos informatizada del español. Barcelona.* Universitat de Barcelona.