Presented in the Twenty-third Annual Meeting of the Society for Text and Discourse, Valencia from 16 to 18, July 2013

Gallito 2.0: a Natural Language Processing tool to support Research on Discourse

Guillermo Jorge-Botana
Universidad Nacional de Educación a Distancia (UNED)

gdejorge@psi.uned.es

Ricardo Olmos
Universidad Autónoma de Madrid

ricardo.olmos@uam.es

Alejandro Barroso
Universidad Nacional de Educación a Distancia (UNED)

abarroso@plusnetsolutions.com

ABSTRACT

Gallito 2.0. is a tool designed to allow both production of and experimentation with vector space models based on Latent Semantic Analysis (LSA). There is a freely (but time-limited) version available (http://www.elsemantico.es/gallito20/index-eng.html). The tool supports creation and evaluation of semantic spaces generated from middle-scale to huge corpora (notice that Gallito 2.0 uses sparse matrices), as well as several related tasks, such as the extraction of term entropy indices, familiarity measured through vector length, similarity between terms, lists of semantic neighbors, K-means cluster, interpretation of topics, Change of Basis, Gram-Schmitdt re-ortogonalization, Construction-Integration representations, textual coherence, essay evaluation, etc. The present poster shows some uses of the tool using a small-scale corpus taken from several newspapers.

1. INTRODUCTION

There are several computational linguistic models for developing semantic technology to process texts, one of them being LSA, a well-established technique descended from earlier vector space models. LSA was originally described by Deerwester, et al. (1990) as a method for Information Retrieval, but beyond the original conception of LSA, some authors have "reinvented" it as a model of knowledge acquisition and representation (Landauer & Dumais,1997) . In fact, it has been one of the most productive models in cognitive science for mathematically modeling language representation and cognitive process (Tonta Y & Darvish, 2010). Various software packages for LSA modeling have been in use for many years. Many implementations can be found by Googling the term "LSA tool". This paper describes a recent addition to the set of available tools, the Gallito 2.0 NLP Tool. A first implementation of this tool with minimal functionality was created in 2005. Over the next years a series of new versions were made available to our research group, which uses them to support simulation investigations. During these years, much of the current functionality was being tested. But in 2012, a considerably better release was delivered to the Technological Transfer Office of UNED University (OTRI), which licenses it. Compared to existing tools, Gallito 2.0 offers a user interface and an extensible set of a very interesting features based on psycholinguistics claims: for example, changing the perspective of the latent semantic space and converting the abstract dimensions into real words, or the well-known Construction-Integration mechanism. We will describe most of this functionality in the following paragraphs.

2. FUNCTIONALITY

The main purpose of Gallito 2.0 is to process language corpora and transform them into semantic space representations. Then a semantic space can be saved in the hard disc and reloaded when necessary with no need to creating it again. Semantic spaces can be generated under various parameters: dimensions, pruning, additions, word class tagging, entropy mechanisms, normalizations, paragraph or sentence separation, minimum document sizes, and minimum word occurrences. When a semantic space is available in the memory, the vector norms and the entropy for every term, the lists of semantic neighbors of a term, the similarities between terms, the similarity between existing documents, the similarity between documents that do not exist in space (pseudodocuments), coherence inside a text, Construction-Integration representation, etc. can be directly extracted. Some batch process can also be run as neighbors, similarity by pairs, similarity matrices, text coherence and essay evaluation. For interchangeable purposes, you can obtain the various matrices in plain text and even generate cluster nets in Pajek format (Batagelj, & Mrvar, 2003) for visualization (figure 1).

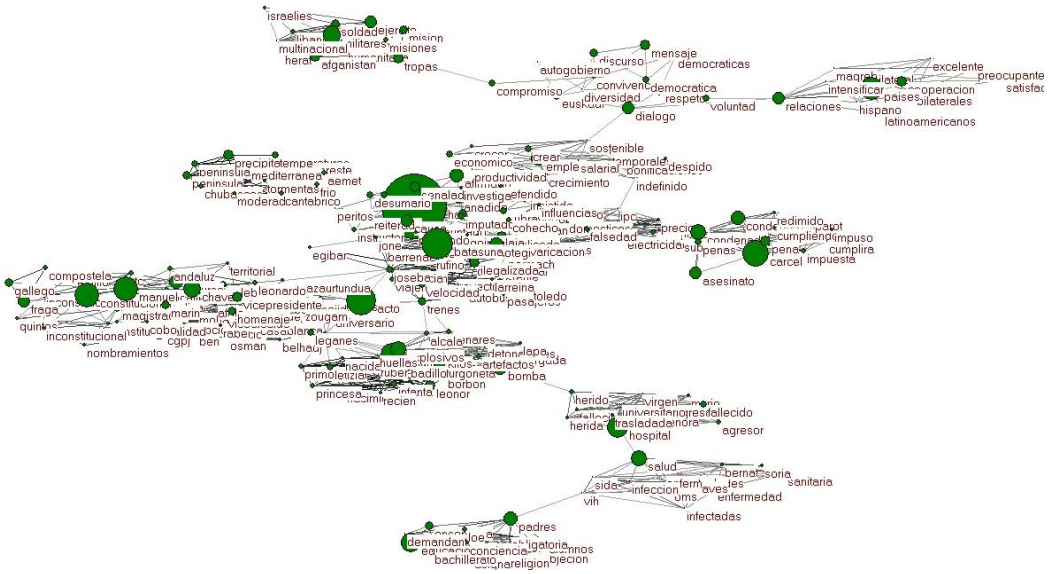

Figure 1

Recently, the latest release of Gallito 2.0 has provided users with a very interesting feature: changing the perspective of the semantic space by means of a change of basis and converting the abstract dimensions into real words (figure 2). This can be done even with the possibility of using the k-means procedure to find the best words to compound such a new basis and the Gram-Schmidt mechanism to re-orthogonalize it, in order to preserve 100% of the distances of the former latent semantic space in the new latent semantic space (see Olmos, et al. in this same proceedings).

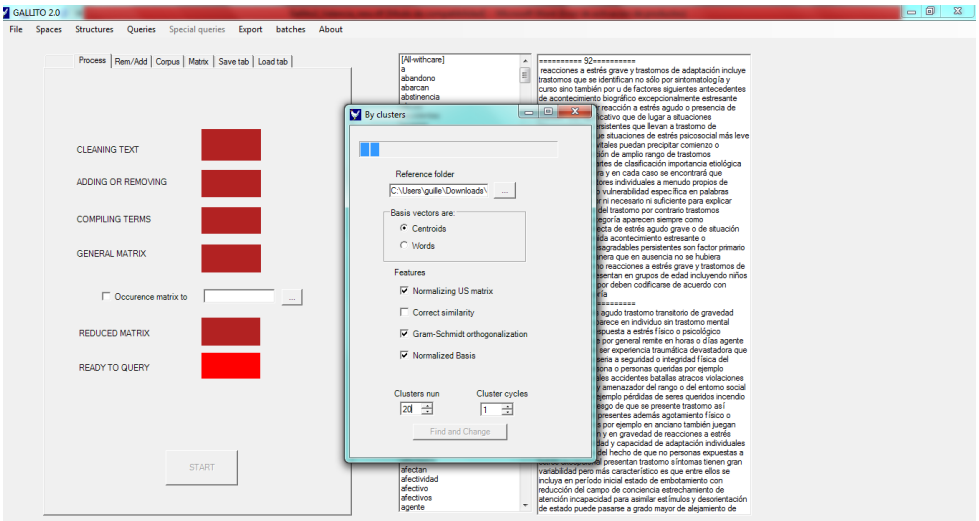

Figure 2

3. SOME EXAMPLES OF REAL USES

Gallito has been used in several studies in order to ensure experimental control. For instance, Duñabeitia, Avilés, Afonso, Scheepers & Carreiras (2009) used similarities measures from a semantic space trained with the LexEsp corpus to make the words of the experimental conditions semantically homogeneous. They also found correlations between Gallito measures and a different set of semantic similarity measures: English translation equivalents of their materials on the LSA of Boulder, English translation of HAL (Hyperspace Analogue to Language), and Spanish human similarity ratings, were significant. They conclude that all of them measure the same in a reasonably consistent manner. Nieuwland (2013) also controlled the critical words in his experimental material using the LexEsp semantic space. He investigated the responses to sentences about unrealistic counterfactual worlds that require people to construct novel conceptual combinations. In a novel study currently under review by Nieuwland, Martin, & Carreiras, the critical words were matched again using the LexEsp semantic space, this time in order to exercise experimental control. In addition, Gallito was used to visualize conceptual nets (Jorge-Botana, G., León, J.A.,Olmos, R. & Hassan-Montero, Y.(2010), monitor the consequences of lexical ambiguity in a vector space model (Olmos, Jorge-Botana, León & Escudero, 2010) or even categorize calls in a real call-routing application. Recently, it was used to change the references in the conceptual interpretations (see Olmos, Jorge-Botana, León & Escudero in this same proceedings).

4. REFERENCES

Batagelj, & Mrvar. Pajek - Analysis and Visualization of Large Networks. in Jünger, M., Mutzel, P., (Eds.) Graph Drawing Software. Springer, Berlin 2003. p. 77-103

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. &Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society For Information Science, 4, 39-407.

Duñabeitia, J.A., Avilés, A., Afonso, O., Scheepers, C., & Carreiras, M. (2009) Qualitative differences in the representation of abstract versus concrete words: Evidence from the visual-world paradigm. Cognition, 110, 284-292

Jorge-Botana, G., León, J.A.,Olmos, R. & Hassan-Montero, Y.(2010). Visualizing polysemy using LSA and the predication algorithm. Journal of the American Society for Information Science and Technology. Vol 61, Issue 8, pp. 1706–1724

Jorge-Botana, G. Olmos, R. & León, J. A. Monitoring the penalization/advantage of lexical ambiguity in vector model representations. 1st Joint Conference of the EPS and SEPEX, Granada 2010

Jorge-Botana, G.,Olmos, R., & Barroso, A.(2012) The Construction-Integration framework: A means to disminish bias in LSA-based Call Routing. International Journal of Speech Technology. Vol.15 Number 2 pag. 151–164

Landauer, T. K., y 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.

Nieuwland, M.S. (2013). "If a lion could speak …": Online sensitivity to propositional truth-value of

unrealistic counterfactual sentences. Journal of Memory and Language, 68(1), 54-67

Nieuwland, M.S., Martin, A.E., & Carreiras, M. (under revision). Event-related brain potential evidence for animacy processing asymmetries during sentence comprehension

Olmos, R., Jorge-Botana, G., León, J.A., Escudero, I. (2012). Giving an interpretation for the semantic dimensions in Latent Semantic Analysis. In Proccedings of Twenty-third Annual Meeting of the Society for Text and Discourse. Valencia, 16-18 July 2013.

Tonta Y, Darvish H.R. (2010). Diffusion of Latent Semantic Analysis as a Research Tool: A Social Network Analysis Approach. Journal of Informetrics 4 (2010) 166–174