A Web Interface for Diachronic Semantic Search in Spanish

Pablo Gamallo, Iván Rodríguez-Torres
Universidade de Santiago de Compostela
Centro Singular de Investigación
en Tecnoloxías da Información (CiTIUS)
15782 Santiago de Compostela, Galiza
pablo.gamallo@usc.es,
ivan.rodriguez.torres@usc.es

Marcos Garcia
Universidade da Coruña
LyS Group
Faculty of Philology
15701 A Coruña, Galiza
marcos.garcia.gonzalez@udc.gal

Abstract

This article describes a semantic system which is based on distributional models obtained from a chronologically structured language resource, namely Google Books Syntactic Ngrams. The models were created using dependency-based contexts and a strategy for reducing the vector space, which consists in selecting the more informative and relevant word contexts. The system allows linguists to analize meaning change of Spanish words in the written language across time.

1 Introduction

Semantic changes of words are as common as other linguistic changes such as morphological or phonological ones. For instance, the word 'artificial' originally meant 'built by the man' (having a positive polarity), but this word has recently changed its meaning if used in contrast with 'natural', aquiring in this context a negative polarity.

Search methods based on frequency (such as Google Trends or Google Books Ngram Viewer search engines) are useful for finding words whose use significantly increases —or decreases— in a specific period of time, but these systems do not not allow language experts to detect semantic changes such as the above mentioned one.

Taking the above into account, this demonstration paper describes a distributional-based system aimed at visualizing semantic changes of words across historical Spanish texts.

The system relies on yearly corpora distributional language models, learned from the Spanish Google Books Ngrams Corpus, so as to obtain word vectors for each year from 1900 to 2009. We compared the pairwise similarity of word vectors for each year and inserted this information in a non structured database. Then, a web interface was designed to provide researchers with a tool, called *Diachronic Explorer*, to search for semantic changes. Diachronic Explorer offers different advanced searching techniques and different visualizations of the results.

Even if there exist similar distributional approaches for languages such as English, German, or French, to the best of our knowledge our proposal is the first work using this technique for Spanish. In addition, we provide the Hispanic community with a useful tool to do linguistic research. The system is open-source[1] and includes a web interface[2] which will be used in the Software Demonstration.

2 The Diachronic Explorer

The Diachronic Explorer relies on a set of distributional semantic models for Spanish language for each year from 1900 to 2009. This semantic resource is stored in a NoSQL database which feeds a web server used for searching and visualizing lexical changes on dozens of thousands of Spanish words along the time.

2.1 Architecture

Figure 1 shows the architecture of Diachronic Explorer. It takes as input a large corpus of dependency-based ngrams to build 110 distributional models, one per year since 1900 to 2009. Then, Cosine similarity is computed for all words in each model in order to generate CSV files consisting of pairs of similar words. Each word is associated with its N (where N = 20) most similar words according to the Cosine coefficient. These files are stored in MongoDB to be accessed by web

[1] https://github.com/citiususc/explorador-diacronico

[2] https://tec.citius.usc.es/explorador-diacronico/index.php

Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017, pages 45–48

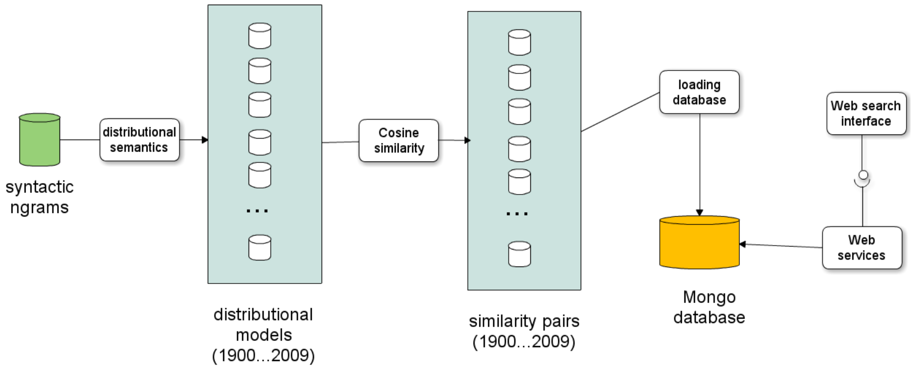


Figure 1: Architecture of Diachronic Explorer

services which generate different types of word searching.

2.2 Yearly Models from Syntactic-Ngrams

To build the distributional semantic models along the time, we made use of the dataset based on the Google Books Ngrams Corpus, which is described in detail in Michel et al. (2011). The whole project contains over 500 billion words (45B in Spanish), being the majority of the content published after 1900. More precisely, we used the Spanish syntactic n-grams from the Version 2012/07/01.[3] Lin et al. (2012) and Goldberg and Orwant (2013) describe the method to extract syntactic ngrams. Each syntactic ngram is accompanied with a corpus-level occurrence count, as well as a time-series of counts over the years. The temporal dimension allows inspection of how the meaning of a word evolves over time by looking at the contexts the word appears in within different time periods.

We transformed the syntactic ngrams into distributional 'word-context' matrices by using a filtering-based strategy that selects for relevant contexts (Gamallo, 2016; Gamallo and Bordag, 2011). A matrix was generated for each year, where each word is represented as a context vector. The final distributional-based resource consists of 110 matrices (one per year from 1900 to 2009). The size of the whole resource is 2,8G, with 25M per matrix in average. Then, the Cosine similarity between word vectors was calculated and, for each word, the 20 most similar ones were selected by year. In total, a data structure with more than 300 million pairs of similar words was generated, giving rise to 110 CSV files (i.e., one file per year).

2.3 Data Storage

We use two ways for storing data. First, we store the distributional models and the similarity pairs in temporal CSV files. These temporal files are generated from offline processes which can be executed periodically as the linguistic input (ngrams) is updated or enlarged with new language sources. Second, the similarity files are imported in a database to make the data access more efficient and easier.

The database system that we chose was MongoDB, which is a NoSQL DB. This type of database system fits well with the task because of two main reasons:

- Our data do not follow a relational model, so we do not need some of the features that make the relational databases powerful, for example, reference keys or table organization.

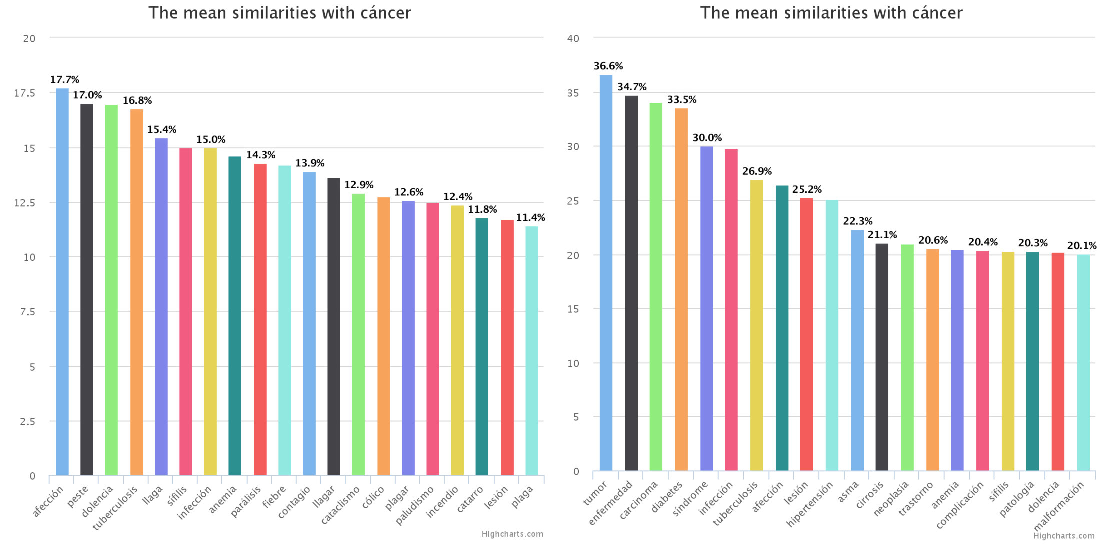
- NoSQL databases scale really well. Unlike relational databases they implement automaric sharding, which enables us to share and distribute data among servers in the most efficient way possible. So, as the data increase it will be easy to rise up a new instance without any additional engineering.

Among the different types of NoSQL databases, we chose MongoDB because it is document oriented, which fits very well with our data structure.

2.4 Web Interface

The Diachronic Explorer is provided with an web interface that offers different ways of making a diachronic search. Here, we describe just two types of search:

[3] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html



(a) (b)

Figure 2: Similar words to *cáncer* (cancer) using a simple search in 1900 (a) and 2009 (b)

Simple: The user enters a target word and selects for a specific period of time (e.g. from 1920 to 1939), and the system returns the 20 most similar terms (in average) within the selected period of type. The user can select a word cloud image to visualize the output. Figure 2 shows the similar words to *cáncer* ('cancer') returned with a simple search in two different periods: 1900 (a) and 2009 (b). Notice that this word is similar to *peste* and *tuberculosis* at the begining of the 20th century, while nowadays it is more similar to technical words, such as *tumor* or *carcinoma*.

Track record: As in the simple search, the user inserts a word and a time period. However, this new type of search returns the specific similarity scores for each year of the period, instead of the similarity average. In addition, the system allows the user to select any word from the set of all words (and not just the top 20) semantically related to the target one in the searched time period.

3 Conclusions

In this abstract we described a system, *Diachronic Explorer*, that allows linguists to analize meaning change of Spanish words in the written language across time. The system is based on distributional models obtained from a chronologically structured language resource, namely Google Books Syntactic Ngrams. The models were created using dependency-based contexts and a strategy for reducing the vector space, which consists in selecting the more informative and relevant word contexts.

As far as we know, our system is the first attempt to build diachronic distributional models for Spanish language. Besides, it uses NoSQL storage technology to scale easily as new data is processed, and provides an interface enabling useful types of word searching across time. Other similar works for English are reported in different papers (Wijaya and Yeniterzi, 2011; Gulordava and Baroni, 2011; Jatowt and Duh, 2014; Kim et al., 2014; Kulkarni et al., 2015).

An interesting direction of research could involve the use of the system infrastructure for holding other types of language variety, namely diatopic variation. Distributional models can be built from text corpora organized, not only with diachronic information, but also with dialectal features. For instance, we could adapt the system to search meaning changes across different diatopic varieties: Spanish from Argentina, Mexico, Spain, Bolivia, and so on. The structure of our system is generic enough to deal with any type of variety, not only that derived from the historical axis.

Acknowledgments

This work has received financial support from a 2016 BBVA Foundation Grant for Researchers and Cultural Creators, TelePares (MINECO, ref:FFI2014-51978-C2-1-R), *Juan de la Cierva formación* Grant (Ref: FJCI-2014-22853), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF),

References

Pablo Gamallo and Stefan Bordag. 2011. Is singular value decomposition useful for word simalirity extraction. *Language Resources and Evaluation*, 45(2):95–119.

Pablo Gamallo. 2016. Comparing Explicit and Predictive Distributional Semantic Models Endowed with Syntactic Contexts. *Language Resources and Evaluation*, First online: 13 May 2016. doi:10.1007/s10579-016-9357-4.

Yoav Goldberg and Jon Orwant. 2013. A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (SEM 2013)*, pages 241–247, Atlanta, Georgia.

Kristina Gulordava and Marco Baroni. 2011. A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, Scotland. ACL.

Adam Jatowt and Kevin Duh. 2014. A Framework for Analyzing Semantic Change of Words Across Time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2014*, pages 229–238, Piscataway, NJ. IEEE Press.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, Maryland. ACL.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International World Wide Web Conference (WWW 2015)*, pages 625–635, Florence, Italy. ACM.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic Annotations for the Google Books Ngram Corpus. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 169–174. ACL.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding Semantic Change of Words over Centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web (DETECT 2011)*, pages 35–40, Glasgow, Scotland. ACM.