Diachronic Analysis of the Italian Language exploiting Google Ngram

Pierpaolo Basile[1] and **Annalina Caputo**[1] and **Roberta Luisi**[2] and **Giovanni Semeraro**[1]
Department of Computer Science
University of Bari Aldo Moro
Via, E. Orabona, 4 - 70125 Bari (Italy)
[1]{firstname.surname}@uniba.it
[2]{roby.luisi}@gmail.com

Abstract

English. In this paper, we propose several methods for the diachronic analysis of the Italian language. We build several models by exploiting Temporal Random Indexing and the Google Ngram dataset for the Italian language. Each proposed method is evaluated on the ability to automatically identify meaning shift over time. To this end, we introduce a new dataset built by looking at the etymological information reported in some dictionaries.

Italiano. *In questo lavoro proponiamo alcuni metodi per l'analisi diacronica della lingua italiana. Abbiamo costruito differenti modelli utilizzando la tecnica del Temporal Random Indexing e Google Ngram per l'italiano. Ciascun metodo proposto è stato valutato rispetto alla capacità di identificare automaticamente i cambi di significato nel tempo. A tale scopo introduciamo uno nuovo dataset costruito mediante le informazioni etimologiche presenti in alcuni dizionari.*

1 Motivation and Background

Languages can be studied from two different and complementary viewpoints: the *diachronic* perspective considers the evolution of a language over time, while the *synchronic* perspective describes the language rules at a specific point of time without taking its history into account (De Saussure, 1983). In this work, we focus on the diachronic approach, since language appears to be unquestionably immersed in the temporal dimension. Language is subject to a constant evolution driven by the need to reflect the continuous changes of the world. The evolution of word meanings has been studied for several centuries, but this kind of investigation has been limited by the low amount of data on which to perform the analysis. Moreover, in order to reveal structural changes in word meanings, this analysis has to explore long periods of time.

Nowadays, the large amount of digital content opens new perspectives for the diachronic analysis of language. This large amount of data needs efficient computational approaches. In this scenario, Distributional Semantic Models (DSMs) represent a promising solution. DSMs are able to represent words as points in a geometric space, generally called WordSpace (Schiitze, 1993; Sahlgren, 2006) simply analysing how words are used in a corpus. However, a WordSpace represents a snapshot of a specific corpus and it does not take into account temporal information.

Since its first release, the Google Ngram dataset (Michel et al., 2011) has inspired a lot of works on the analysis of cultural trends and linguistic variations. Moving away from mere frequentist approaches, Distributional Semantic Models have proved to be quite effective in measuring a meaning shift through the analysis of variation of word co-occurrences. One of the earlier attempts can be dated to Gulordava and Baroni (2011), where a co-occurrence matrix is used to model the semantics of terms. In this model, similarly to ours, the cosine similarity between vectors representing a term in two different periods is exploited as a predictor of the meaning shift: low values suggest a change in the words that co-occur with the target. The co-occurrence matrix is computed with local mutual information scores and the context elements are fixed with respect to the different time

periods, hence the spaces are directly comparable. However, this kind of direct comparison does not hold when the vector representation is manipulated, like in reduction methods (SVD), or learning approaches (word2vec). In these cases, each space has its own coordinate axis. Then, some kind of alignment between spaces is required. To this end, Hamilton et al. (2016) use orthogonal Procrustes, while Kulkarni et al. (2015a) learn a transformation matrix.

In this paper, we propose an evolution of our previous work (Basile et al., 2014; Basile et al., 2015) for analysing word meanings over time. This model, differently from those of Hamilton et al. (2016) and Kulkarni et al. (2015a), creates different WordSpaces for each time period in terms of the same common random vectors; then, the resulting word vectors are directly comparable with one another. In particular, we propose an efficient method for building a DSM model which takes into account temporal information relying on a very large corpus: the Google Ngram for the Italian language. Moreover, for the first time, we provide a dataset for the evaluation of word meaning change points detection specifically set up for the Italian language.

The paper is structured as follows: Section 2 provides details about our methodology, while Section 3 describes the dataset that we have developed and the results of a preliminary evaluation. Section 4 reports final remarks and future work.

2 Methodology

Our method has its roots in a previous model based on Temporal Random Indexing (TRI) (Basile et al., 2014; Basile et al., 2015). In particular, we evolve the TRI approach in two directions: 1) we improve the system in order to manage very large datasets, such as Google Ngram; 2) we introduce a new approach based on Reflective Random Indexing (RRI) (Cohen et al., 2010) with the aim of identifying indirect inferences that can lead to the discovery of implicit connections between word meanings.

The idea behind TRI is to build different WordSpaces for each time period that we want to analyse. The peculiarity of TRI is that word vectors over different time periods are directly comparable because they are built using the same random vectors. In particular TRI works as follows:

1. Given a corpus C of documents and a vocabulary V of terms[1] extracted from C, the method assigns a random vector r_i to each term $t_i \in V$. A random vector is a vector that has values only in {-1, 0, 1} and it is sparse with few non-zero elements distributed randomly along its dimensions. The set of random vectors assigned to all terms in V are near-orthogonal;

2. The corpus C is split in different time periods T_k using temporal information, for example the year of publication;

3. For each period T_k, a WordSpace WS_k is built. All the terms of V occurring in T_k are represented by a semantic vector. The semantic vector sv_i^k for the i-th term in T_k is built as the sum of all the random vectors of the terms co-occurring with t_i in T_k. When computing the sum, we weigh the random vector; in this case we adopt a formula based on inverse document frequency. Formally, the weight is computed as $w(r_i) = log\left(\frac{C_k}{\#t_i^k}\right)$, where C_k is the total number of occurrences in T_k and $\#t_i^k$ is the occurrences of the term t_i in T_k. The idea is to give less weight to the most frequent words.

In this way, the semantic vectors across all time periods are comparable since they are the sum of the same random vectors.

RRI can be implemented by repeating the steps 2 and 3 several times. Where at each iteration random vectors are replaced by the semantic vectors built in the previous step. The idea is to model implicit connections between terms that never co-occur together, but that could occur frequently with other shared terms.

The next two sub-sections provide details about the Google Ngram dataset and the method used to automatically detect word meanings shift.

2.1 Google Ngram

Google Ngram is a very large dataset containing all the n-grams (up to five) extracted from Google Books. It is built by analysing over five millions books spanning the years from 1500 to 2012, but the developers estimate that the most reliable period is from 1800 to 2012. The dataset covers several languages including Italian. For each

[1] The terms that we want to analyse. Usually, the most n frequent terms are extracted.

language, several compressed files are released. Each file contains for each line the following information: `Ngram <TAB> year <TAB> match_count <TAB> volume_count`. For example, the line "`analysis is often described as 1991 104 5`" means that the ngram "*analysis is often described as*" occurs 104 times in 5 books in the 1991 .

We modify TRI for building the WordSpaces directly from the Google Ngram dataset. In particular, we need a pre-processing step in which we split the n-grams in several files according to the time periods we want to analyse. For example, if we fix the dimension of a time period to ten years from 1850 to 2012, we build several files for each period: T_1 = 1850-1859, T_2 = 1860-1869, ..., T_{16} = 2000-2009, T_{17} = 2010-2012. Each file contains only the n-grams that occur in the specific time period. We remove information about the year and the book count since they are not useful in the subsequent steps. Considering the previous example, the line "`analysis is often described as 104`" will be stored in the file 1990-1999.

After this pre-processing step, we can easily run TRI and RRI, where RRI can be repeated multiple times.

2.2 Change point detection

To track the word meaning change over time, for each term t_i we build a time series $\Gamma(t_i)$ taking into account several methods. A time series is a sequence of values, one value for each time period, that indicates the semantic shift of that term in the specific period. We adopt several strategies for building time series. The first strategy is based on term log-frequency; each value in the series is defined as: $\Gamma_k(t_i) = log(\frac{\#t_i^k}{C_k})$.

In order to exploit the ability of our methods in computing vectors similarity over time periods, we define two strategies for building the time series:

point-wise: $\Gamma_k(t_i)$ is defined as the cosine similarity between sv_i^k and sv_i^{k-1}. In this way, we want to capture vector changes between two time periods;

cumulative: we build a cumulative vector $sv_i^{C_{k-1}} = \sum_{j=0}^{k-1} sv_i^j$ and compute the cosine similarity with respect to the vector sv_i^k. The idea is that the semantics at point $k-1$ depends on the semantic of all the previous time periods.

Given a time series we need a method for finding significant change points in the series. We adopt the strategy proposed in (Kulkarni et al., 2015b) based on the *Mean shift model* (Taylor, 2000). According to this model, we define a mean shift of a general time series Γ pivoted at time period j as:

$$K(\Gamma) = \frac{1}{l-j}\sum_{k=j+1}^{l}\Gamma_k - \frac{1}{j}\sum_{k=1}^{j}\Gamma_k \quad (1)$$

In order to understand if a mean shift is statistically significant at time j, a bootstrapping (Efron and Tibshirani, 1994) approach under the null hypothesis that there is no change in the mean is adopted. In particular, statistical significance is computed by first constructing B bootstrap samples by permuting $\Gamma(t_i)$. Second, for each bootstrap sample P, $K(P)$ is calculated to provide its corresponding bootstrap statistic and statistical significance (p-value) of observing the mean shift at time j compared to the null distribution. Finally, we estimate the change point by considering the time point j with the minimum p-value score. Since multiple words can have the same p-value, we sort them according to their frequency. The output of this process is a ranking of words that potentially have changed meaning.

3 Evaluation

The goal of the evaluation is twofold: 1) to build a standard benchmarking for meaning shift detection for the Italian language; 2) to evaluate the performance of the proposed methods and compare them with the baseline model based on the word frequency.

A list of meaning shifts for the Italian language is not available, then we build a new dataset using a pooling strategy. In particular, we retrieve the list of meaning shifts, as explained in Section 2.2, using the cumulative strategy for each of the following methods: word frequency, TRI, TRRI with one iteration and TRRI with two iterations.

Taking into account the first 50 words for each system, we manually check for each word if a meaning shift occurs by exploiting some dictionaries. We use two dictionaries: the "Sabatino Coletti" available on-line[2] and the "Dizionario Eti-

[2] `http://dizionari.corriere.it/dizionario_italiano/`

mologico Zanichelli" available on CD-ROM. Finally, we obtain a gold standard that consists of 40 words and their corresponding change points.

All the methods, with exception of word frequency, are built using co-occurrences information extracted from 5-grams in the Google Ngram dataset for the Italian. The vector dimension is set to 1,000 for all the approaches based on Random Indexing using two non-zero elements in the random vector.

We adopt accuracy as evaluation metric. Given a list of n change points returned by the system, we compute the ratio between the number of change points correctly identified in the gold standard[3] and n. In order to identify the correct change points, we consider not only the word[4], but also the year of the change point. In particular, the year predicted by the system must be equal or greater than one of the years reported in the gold standard. We compute the accuracy using different values of n (10, 100, ALL). Results of the evaluation are reported in Table 1. In particular, we evaluate 7 systems: $logfreq$ is the baseline based on word frequency; TRI is the Temporal Random Indexing method, $TRRI1$ is the Temporal Reflective Random Indexing with one iteration, while $TRRI2$ adopts two iterations. For the methods based on Random Indexing, we investigate both the point-wise and the cumulative strategy to compute the change points.

Table 1: Results of the evaluation.

Method	**acc@10**	**acc@100**	**ALL**
TRI_{point}	0.0247	0.1111	0.3086
TRI_{cum}	0.0123	0.0247	0.2963
$TRRI1_{point}$	0.0000	0.0247	0.2716
$logfreq$	0.0247	0.1111	0.2346
$TRRI2_{point}$	0.0000	0.0370	0.1728
$TRRI1_{cum}$	0.0000	0.0000	0.1605
$TRRI2_{cum}$	0.0000	0.0000	0.1235

The analysis of the results shows that TRI generally provides better results than $TRRI$. Moreover, the point-wise strategy always outperforms the cumulative one. With respect to the baseline, it has the same accuracy of TRI for both acc@10 and acc@100, while it performs worse than TRI and $TRRI1$ when the accuracy is computed over the whole list of terms (ALL). These results suggest that, while there are not too many differences between the two methods considering smaller lists of results, TRI is actually able to detect more meaning shifts on a larger set of terms. $TRRI2$ always provides the worst results; we speculate that two iterations introduce too much noise in the model. A closer scrutiny to the list of words provided by $TRRI2$ highlights the presence of many foreign words: a simplistic conclusion may suggest that this approach is able to identify foreign terms that are introduced in the Italian language. However, we think that the output of this method deserves more investigations carried out by designing an ad-hoc evaluation.

The evaluation is based on the predicted year, which has to be equal or greater than one of the years reported in the gold standard, we conduct a further analysis to measure how far the prediction is from the exact value. In particular, we compute the mean and the standard deviation taking into account the differences between the predicted and the exact year. The results of this analysis are reported in Table 2. We observe the both $TRRI1_{cum}$ and $TRRI2_{cum}$ produce the best results despite their low accuracy, while TRI_{cum} reports the best trade-off between accuracy and precision in detecting the correct year. It is important to underline that the size of the time interval influences this kind of analysis since if the algorithm predicts 1900, the change point could happen in the interval 1900-1909[5]. We plan to design a more accurate analysis by exploring a time interval set to one year as future work.

Table 2: Mean and standard deviation of the differences between the predicted and the exact year.

Method	**Mean**	**Std.Deviation**
TRI_{point}	38.04	34.90
TRI_{cum}	26.45	19.60
$TRRI1_{point}$	65.86	49.96
$logfreq$	24.15	16.19
$TRRI2_{point}$	54.50	52.70
$TRRI1_{cum}$	16.61	14.62
$TRRI2_{cum}$	19.40	19.85

[3]The gold standard adopted in this evaluation is available here: `https://dl.dropboxusercontent.com/u/16026979/data/TRI_CLIC_2016_change_word`.

[4]The word matching is performed taking into account also the inflected forms.

[5]In our experiment, the size of the time interval is set to ten years.

4 Conclusions

In this work we proposed several methods based on Random Indexing for the diachronic analysis of the Italian language. We built a dataset for the evaluation of meaning shift by exploiting etymological information taken from two Italian dictionaries. We compared our approaches with respect a baseline based on word frequency obtaining promising results. In particular, the TRI method showed its better capability in retrieving more meaning shifts on a longer list of terms. As future work, we plan to extend the dataset with further words and to investigate other methods based on word-embeddings.

Acknowledgement

This work is partially supported by the project "Multilingual Entity Liking" funded by the Apulia Region under the program FutureInResearch.

References

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. Analysing word meaning over time by exploiting temporal random indexing. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *First Italian Conference on Computational Linguistics CLiC-it 2014*. Pisa University Press.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2015. Temporal random indexing: A system for analysing word meaning over time. *Italian Journal of Computational Linguistics*, 1(1):55–68, 12.

Trevor Cohen, Roger Schvaneveldt, and Dominic Widdows. 2010. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of biomedical informatics*, 43(2):240–256.

Ferdinand De Saussure. 1983. *Course in general linguistics*. La Salle, Illinois: Open Court.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.

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, UK, July. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *CoRR*, abs/1605.09096.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015a. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 625–635, New York, NY, USA. ACM.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015b. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. ACM.

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, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Magnus Sahlgren. 2006. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.

Hinrich Schiitze. 1993. Word space. *Advances in neural information processing systems*, 5:895–902.

Wayne A Taylor. 2000. *Change-point analysis: a powerful new tool for detecting changes*. Taylor Enterprises, Inc.