10.12871/CLICIT201418

Analysing Word Meaning over Time by Exploiting Temporal Random Indexing

Pierpaolo Basile and **Annalina Caputo** and **Giovanni Semeraro**
Department of Computer Science
University of Bari Aldo Moro
Via, E. Orabona, 4 - 70125 Bari (Italy)
{firstname.lastname}@uniba.it

Abstract

English. This paper proposes an approach to the construction of *WordSpace*s which takes into account temporal information. The proposed method is able to build a geometrical space considering several periods of time. This methodology enables the analysis of the time evolution of the meaning of a word. Exploiting this approach, we build a framework, called Temporal Random Indexing (*TRI*) that provides all the necessary tools for building *WordSpace*s and performing such linguistic analysis. We propose some examples of usage of our tool by analysing word meanings in two corpora: a collection of Italian books and English scientific papers about computational linguistics.

Italiano. *In questo lavoro proponiamo un approccio per la costruzione di WordSpaces che tengano conto di informazioni temporali. Il metodo proposto costruisce degli spazi geometrici considerando diversi intervalli temporali. Questa metodologia permette di studiare l'evoluzione nel tempo del significato delle parole. Utilizzando questo approccio abbiamo costruito uno strumento, chiamato Temporal Random Indexing (TRI), che permette la costruzione dei WordSpaces e fornisce degli strumenti per l'analisi linguistica. Nell'articolo proponiamo alcuni esempi di utilizzo del nostro tool analizzando i significati delle parole in due corpus: uno relativo a libri nella lingua italiana, l'altro relativo ad articoli scientifici in lingua inglese nell'ambito della linguistica computazionale.*

1 Introduction

The analysis of word-usage statistics over huge corpora has become a common technique in many corpus linguistics tasks, which benefit from the growth rate of available digital text and computational power. Better known as Distributional Semantic Models (DSM), such methods are an easy way for building geometrical spaces of concepts, also known as *Semantic* (or *Word*) *Spaces*, by skimming through huge corpora of text in order to learn the context of usage of words. In the resulting space, semantic relatedness/similarity between two words is expressed by the closeness between word-points. Thus, the semantic similarity can be computed as the cosine of the angle between the two vectors that represent the words. DSM can be built using different techniques. One common approach is the Latent Semantic Analysis (Landauer and Dumais, 1997), which is based on the Singular Value Decomposition of the word co-occurrence matrix. However, many other methods that try to take into account the word order (Jones and Mewhort, 2007) or predications (Cohen et al., 2010) have been proposed. Recursive Neural Network (RNN) methodology (Mikolov et al., 2010) and its variant proposed in the *word2vect* framework (Mikolov et al., 2013) based on the continuous bag-of-words and skip-gram model take a complete new perspective. However, most of these techniques build such *SemanticSpaces* taking a *snapshot* of the word co-occurrences over the linguistic corpus. This makes the study of semantic changes during different periods of time difficult to be dealt with.

In this paper we show how one of such DSM techniques, called Random Indexing (RI) (Sahlgren, 2005; Sahlgren, 2006), can be easily extended to allow the analysis of semantic changes of words over time. The ultimate aim is to provide a tool which enables to understand how words change their meanings within a document corpus as a function of time. We choose RI for two main reasons: 1) the method is incremental and requires few computational resources while still retaining good performance; 2) the methodology for building the space can be easily expanded to integrate temporal information. Indeed, the disadvantage of classical DSM approaches is that *WordSpace*s built on different corpus are not comparable: it is always possible to compare similarities in terms of neighbourhood words or to combine vectors by geometrical operators, such as the tensor product, but these techniques do not allow a direct comparison of vectors belonging to two different spaces. Our approach based on RI is able to build a *WordSpace* on different time periods and makes all these spaces comparable to each another, actually enabling the analysis of word meaning changes over time by simple vector operations in *WordSpaces*.

The paper is structured as follows: Section 2 provides details about the adopted methodology and the implementation of our framework. Some examples of the potentiality of our framework are reported in Section 3. Lastly, Section 4 closes the paper.

2 Methodology

We aim at taking into account temporal information in a DSM approach, which consists in representing words as points in a *WordSpace*, where two words are similar if represented by points close to each other. Hence, this *Temporal WordSpace* will be suitable for analysing how word meanings change over time. Under this light, RI has the advantages of being very simple, since it is based on an incremental approach, and is easily adaptable to the *temporal* analysis needs. The *WordSpace* is built taking into account words co-occurrences, according to the distributional hypothesis (Harris, 1968) which states that words sharing the same linguistic contexts are related in meaning. In our case the linguistic context is defined as the words that co-occur with the *temporal* word, i.e. the word under the temporal analysis. The idea behind RI has its origin in Kanerva work (Kanerva, 1988) about the Sparse Distributed Memory. RI assigns a context vector to each unit; in our case, each word represents a context. The context vector is generated as a high-dimensional random vector with a high number of zero elements and a few number of elements equal to 1 or -1 randomly distributed over the vector dimensions. Vectors built using this approach generate a nearly orthogonal space. During the incremental step, a vector is assigned to each temporal element as the sum of the context vectors representing the context in which the temporal element is observed.The mathematical insight behind the RI is the projection of a high-dimensional space on a lower dimensional one using a random matrix; this kind of projection does not compromise distance metrics (Dasgupta and Gupta, 1999).

Formally, given a corpus C of n documents, and a vocabulary V of m words extracted form C, we perform two steps: 1) assigning a context vector c_i to each word in V; 2) generating for each word w_i a semantic vector sv_i computed as the sum of all the context vectors assigned to the words co-occurring with w_i. The context is the set of m words that precede and follow w_i. The second step can be defined by the equation:

$$sv_i = \sum_{d \in C} \sum_{-m<i<+m} c_i \qquad (1)$$

After these two steps, we obtain a set of semantic vectors assigned to each word in V representing our *WordSpace*.

2.1 Temporal Random Indexing

The classical RI does not take into account temporal information, but it can be easily adapted to the methodology proposed in (Jurgens and Stevens, 2009) for our purposes. In particular, we need to add a metadata containing information about the year in which the document was written, to each document in C. Then, Temporal RI can build several *WordSpace*s T_k for different time periods, with these spaces being comparable to each other. This means that a vector in the *WordSpace* T_1 can be compared with vectors in the space T_2. The first step in the classical RI is unchanged in Temporal RI, and represents the strength of our approach: the use of the same context vectors for all the spaces makes them comparable. The second step is similar to the one proposed for RI but it takes into account the temporal period. Let T_k be a period that ranges between years $y_{k_{start}}$ and $y_{k_{end}}$, where $y_{k_{start}} < y_{k_{end}}$; then, for building the *WordSpace* T_k we consider only the documents d_k written during T_k.

$$sv_{i_{T_k}} = \sum_{d_k \in C} \sum_{-m<i<+m} c_i \qquad (2)$$

Using this approach we can build a *WordSpace* for each time period over a corpus C tagged with information about the publication year.

2.2 The TRI System

We build a system, called TRI, able to perform Temporal RI using a corpus of documents with temporal information. TRI provides a set of features: 1) to build a *WordSpace* for each year, provided that a corpus of documents with temporal information is available; 2) to merge *WordSpace*s that belong to a particular time period (the new *WordSpace* can be saved on disk or stored in memory for further analysis); 3) to load a *WordSpace* and fetch vectors; 4) to combine and sum vectors; 5) to retrieve similar vectors using the cosine similarity; 6) to extract the neighbourhood of a word or compare neighbourhoods in different spaces for the temporal analysis of a word meaning. All these features can be combined to perform linguistic analysis using a simple shell. Section 3 describes some examples. The TRI system is developed in JAVA and is available on-line[1] under the GNU v.3 license.

3 Evaluation

The goal of this section is to show the usage of the proposed framework for analysing the changes of word meaning over time. Moreover, such analysis supports the detection of linguistics events that emerge in specific time intervals related to social or cultural phenomena. To perform our analysis we need a corpus of documents tagged with time metadata. Then, using our framework, we can build a *WordSpace* for each year. We study the semantics related to a word by analysing the nearest words in the *WordSpace*. For example, we can analyse how the meaning of word has changed in an interval spanning several periods of time. Given two time period intervals and a word w, we can build two *WordSpace*s (T_1 and T_2) by summing the *WordSpace*s assigned to the years that belong to each time period interval. Then using the cosine similarity, we can rank and select the nearest words of w in the two *WordSpace*s, and measure how the semantics of w is changed. Due to the fact that TRI makes *WordSpace*s comparable, we can extract the vectors assigned to w in T_1 and in T_2, and compute the cosine similarity between them. The similarity shows how the semantic of w is changed over time; a similarity equals to 1 means that the word w holds the same semantics. We adopt this last approach to detect words that mostly changed their semantics over time and analyse if this change is related to a particular social or cultural phenomenon. To perform this kind of analysis we need to compute the divergence of semantics for each word in the vocabulary.

Gutenberg Dataset. The first collection consists of Italian books with publication year by the Project Gutenberg[2] made available in text format. The total number of collected books is 349 ranging from year 1810 to year 1922. All the books are processed using our tool TRI creating a *WordSpace* for each available year in the dataset. For our analysis we create two macro temporal periods, before 1900 (T_{pre900}) and after 1900 ($T_{post900}$). The space T_{pre900} contains information about the period 1800-1899, while the space $T_{post900}$ contains information about all the documents in the corpus. As a first example, we analyse how the neighbourhood of the word *patria* (*homeland*) changes in T_{pre900} and $T_{post900}$. Table 1 shows the ten most similar words to *patria* in the two time periods; differences between them are reported in bold. Some words *(legge, virtù, onore)*[3] related to fascism propaganda occur in $T_{post900}$, while in T_{pre900} we can observe some concepts *(nostra, causa, italia)*[4] probably more related to independence movements in Italy. As an example, analysing word meaning evolution over time, we observed that the word *cinematografo* (*cinema*) clearly changes its semantics: the similarity of the word *cinematrografo* in the two spaces

Table 1: Neighbourhood of *patria* (*homeland*).

T_{pre900}	$T_{post900}$
libertà	libertà
opera	gloria
pari	giustizia
comune	comune
gloria	**legge**
nostra	pari
causa	**virtù**
italia	**onore**
giustizia	opera
guerra	**popolo**

[1] https://github.com/pippokill/tri

[2] http://www.gutenberg.org/

[3] In English: *(law/order, virtue, honour).*

[4] In English: *(our, reason, Italy).*

Table 2: Neighbourhoods of *semantics* across several decades.

1960-1969	1970-1979	1980-1989	1990-1999	2000-2010	2010-2014
linguistics	**natural**	syntax	syntax	syntax	syntax
theory	linguistic	natural	theory	theory	theory
semantic	semantic	**general**	interpretation	interpretation	interpretation
syntactic	**theory**	theory	general	description	description
natural	syntax	semantic	linguistic	**meaning**	complex
linguistic	language	syntactic	description	linguistic	meaning
distributional	processing	linguistic	**complex**	logical	linguistic
process	syntactic	**interpretation**	natural	complex	logical
computational	description	**model**	representation	representation	structures
syntax	**analysis**	**description**	**logical**	**structures**	representation

is very low, about 0.40. To understand this change we analysed the neighbourhood in the two spaces and we noticed that the word *sonoro* (*sound*) is strongly related to *cinematografo* in $T_{post900}$. This phenomenon can be ascribed to the sound introduction after 1900.

ANN Dataset. The ACL Anthology Network Dataset (Radev et al., 2013)[5] contains 21,212 papers published by the Association of Computational Linguistic network, with all metadata (authors, year of publication and venue). We split the dataset in decades (1960-1969, 1970-1979, 1980-1989, 1990-1999, 2000-2010, 2010-2014), and for each decade we build a different *WordSpace* with TIR. Each space is the sum of *WordSpace*s belonging to all the previous decades plus the one under consideration. In this way we model the whole word history and not only the semantics related to a specific time period. Similarly to the Gutenberg Dataset, we first analyse the neighbourhood of a specific word, in this case *semantics*, and then we run an analysis to identify words that have mostly changed during the time. Table 2 reports in bold, for each decade, the new words that entered in the neighbourhood of *semantics*. The word *distributional* is strongly correlated to *semantics* in the decade 1960-1969, while it disappears in the following decades. Interestingly, the word *meaning* popped up only in the decade 2000-2010, while *syntax* and *syntactic* have always been present.

Regarding the word meaning variation over time, it is peculiar the case of the word *bioscience*. Its similarity in two different time periods, before 1990 and the latest decade, is only 0.22. Analysing its neighbourhood, we can observe that before 1990 *bioscience* is related to words such as *extraterrestrial* and *extrasolar*, nowadays the same word is related to *medline*, *bionlp*, *molecular* and *biomedi*. Another interesting case is the word *unsupervised*, which was related to *observe*, *partition*, *selective*, *performing*, before 1990; while nowadays has correlation of *supervised*, *disambiguation*, *technique*, *probabilistic*, *algorithms*, *statistical*. Finally, the word *logic* changes also its semantics after 1980. From 1979 to now, its difference in similarity is quite low (about 0.60), while after 1980 the similarity increases and always overcomes the 0.90. This phenomenon can be better understood if we look at the words *reasoning* and *inference*, which have started to be related to the word *logic* only after 1980.

4 Conclusions

We propose a method for building *WordSpace*s taking into account information about time. In a *WordSpace*, words are represented as mathematical points and the similarity is computed according to their closeness. The proposed framework, called TRI, is able to build several *WordSpace*s in different time periods and to compare vectors across the spaces to understand how the meaning of a word has changed over time. We reported some examples of our framework, which show the potential of our system in capturing word usage changes over time.

Acknowledgements

This work fullfils the research objectives of the project PON 01 00850 ASK-Health (Advanced System for the interpretation and sharing of knowledge in health care).

[5] Available on line: http://clair.eecs.umich.edu/aan/

References

Trevor Cohen, Dominique Widdows, Roger W. Schvaneveldt, and Thomas C. Rindflesch. 2010. Logical Leaps and Quantum Connectives: Forging Paths through Predication Space. In *AAAI-Fall 2010 Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes*, pages 11–13.

Sanjoy Dasgupta and Anupam Gupta. 1999. An elementary proof of the Johnson-Lindenstrauss lemma. Technical report, Technical Report TR-99-006, International Computer Science Institute, Berkeley, California, USA.

Zellig S. Harris. 1968. *Mathematical Structures of Language*. New York: Interscience.

Michael N. Jones and Douglas J. K. Mewhort. 2007. Representing Word Meaning and Order Information in a Composite Holographic Lexicon. *Psychological Review*, 114(1):1–37.

David Jurgens and Keith Stevens. 2009. Event Detection in Blogs using Temporal Random Indexing. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 9–16. Association for Computational Linguistics.

Pentti Kanerva. 1988. *Sparse Distributed Memory*. MIT Press.

Thomas K. Landauer and Susan T. Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological review*, 104(2):211–240.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent Neural Network based Language Model. In *INTERSPEECH*, pages 1045–1048.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL Anthology Network Corpus. *Language Resources and Evaluation*, pages 1–26.

Magnus Sahlgren. 2005. An Introduction to Random Indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, volume 5.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm: Stockholm University, Faculty of Humanities, Department of Linguistics.