Temporal Random Indexing: a Tool for Analysing Word Meaning Variations in News

Pierpaolo Basile
Dept. of Computer Science
Univ. of Bari Aldo Moro
name.surname@uniba.it

Annalina Caputo
Dept. of Computer Science
Univ. of Bari Aldo Moro
name.surname@uniba.it

Giovanni Semeraro
Dept. of Computer Science
Univ. of Bari Aldo Moro
name.surname@uniba.it

Abstract

The availability of data spanning different epochs has inspired a new analysis of cultural, social, and linguistic phenomena from a temporal perspective. This paper describes the application of Temporal Random Indexing (TRI) to the news context. TRI is able to build geometrical spaces of word meanings that consider several periods of time. Hence, TRI enables the analysis of the evolution in time of the meaning of a word. We propose some examples of application of TRI to the analysis of word meanings in the news scenario; this analysis enables the detection of linguistic variations that emerge in specific time intervals and that can be related to particular events.

1 Introduction

The analysis of word meaning variations over time periods is a crucial task for identifying changes in social and cultural phenomena. The diachronic analysis of a language allows to discover linguistic variations over time. Generally, a diachronic analysis is performed on a large time interval since linguistic variations happen quite slowly. However, this is not the case for fast data-streaming scenarios like the Web, and in particular social media such as Twitter or Facebook, where socio-cultural and linguistic phenomena quickly rise and fall. Although the news scenario is generally characterized by the use of a regular language, the large number of events that occur along the time line causes sudden topic shifts, making the analysis of this data similar to the data-streaming scenario.

In this paper we describe a technique called Temporal Random Indexing (TRI) that we have successfully applied to several diachronic analyses of the language [BCS15]. TRI is able to build several geometrical spaces of word meanings, called Distributional Semantic Models (DSM), one for each time interval, by skimming through huge corpora of text in order to learn the context of usage of words over time. In the resulting spaces, semantic similarity between 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. We show how to adopt TRI as a tool to discover particular phenomena in news data-streaming and how to link these linguistic changes to interesting events reported in the news content.

2 Methodology

TRI is based on Random Indexing (RI) [Sah05], a dimensionality reduction methodology and computational framework for distributional semantics. Given a term-term co-occurrence matrix A, RI builds a new matrix B where the Euclidean distance between points is preserved. Formally, given a corpus D of n documents, and a vocabulary V of m words extracted from D, we perform two steps: 1) assign a random vector r_i to each word $w_i \in V$; 2) compute a semantic vector sv_i for each word w_i as the sum of all random vectors assigned to words co-occurring with w_i in a given context. The context is the set of c words that precede and follow w_i. The second step is defined by the following equation:

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

$$sv_i = \sum_{d \in D} \sum_{\substack{-c<j<+c \\ j \neq i}} r_j \quad (1)$$

The set of semantic vectors assigned to words in V represents the *WordSpace*.

The classical RI does not take into account temporal information, but it can be easily adapted to our purposes by applying the methodology proposed in [JS09]. Specifically, if the corpus of n documents D is annotated with metadata containing information about the publication date, we can split the collection in p subsets $D_1, D_2, \ldots, D_p$, where p is the number of different time periods we want to analyse. The first step in the classical RI is unchanged in TRI: a random vector is assigned to each word in the whole vocabulary V. This represents the strength of this approach: the use of the same random 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 information: a different *WordSpace* T_k is built for each time period D_k. Hence, the semantic vector for a word in a given time interval is the result of its co-occurrences with other words in the same time interval, but the use of the same random vectors for building the word representations over different time spans guarantees their comparability along the timeline. This means that a vector in the *WordSpace* T_1 can be compared with vectors in the space T_2.

Let T_k be a period that ranges from $t_{k_{start}}$ to $t_{k_{end}}$, where $t_{k_{start}} < t_{k_{end}}$. In order to build the *WordSpace* T_k we consider only the documents d_k whose publication date falls within the time interval T_k as follows:

$$sv_{i_{T_k}} = \sum_{d_k \in D_k} \sum_{\substack{-c<j<+c \\ j \neq i}} r_j \quad (2)$$

Using this approach we can build a *WordSpace* for each time period T_k over a corpus D tagged with information about the publication year. The word w_i has a distinct semantic vector $sv_{i_{T_k}}$ for each time period T_k built by accumulating random vectors according to the co-occurring words in that period. The great potentiality of TRI lies on the use of the same random vectors to build different *WordSpaces*: semantic vectors in different time periods remain comparable because they are the linear combination of the same random vectors.

3 Case study

The main goal of this case study is to show how to adopt TRI[1] to discover interesting phenomena in the news scenario. Specifically, we can analyse the similarity between the vector representations of a term across different time periods in order to detect changes in the usage of the term. Then, we can scrutinise both the neighbour terms and the news related to such a term during the period of time when the similarity has changed in order to understand if a specific event occurred.

Table 1: Neighbour terms of the word "scandal" in the two time periods 14-20 and 21-27 September 2015.

14-20 September 2015		21-27 September 2015	
allegations	0.60	cheating	0.86
called	0.60	volkswagen	0.83
corruption	0.59	rigging	0.80
made	0.59	automaker	0.79
apology	0.59	tests	0.79
met	0.58	carmaker	0.77
became	0.58	deception	0.77
case	0.58	german	0.76
initially	0.58	diesel	0.76
forced	0.58	emissions	0.76

We adopt the Signal Media One-Million News Articles dataset that consists of 1 million articles scraped during the time interval 1-30 September 2015. News are extracted from Reuters, in addition to local news sources and blogs. We split the dataset in five time periods of about one week: 1-6, 7-13, 14-20, 21-27, and 28-30. The split reflects the start and end of weeks in the month of September 2015. Then, for each period we build a *WordSpace* exploiting TRI. In particular, we analyse the 150,000 most frequent words in the whole corpus and we set the vector dimension to 500 using two non-zero elements in the random vector.

In each time interval, we try to discover terms that change their semantics with respect to the previous periods. Formally, given two time periods T_h and T_k, where T_h precedes T_k, and a term t_i, we can simple compute the cosine similarity between the semantic vector of t_i in T_h and the semantic vector of t_i in T_k ($sim(sv_{i_{T_h}}, sv_{i_{T_k}})$). The similarity is a good indicator of the variation of semantics of the term t_i: a low similarity suggests a meaning shift. Using this approach we can rank all terms according to their similarity in ascending order. Top terms in the rank are good candidates for further analysis. However, in order to limit our analysis to those terms that frequently occur in the whole collection, the similarity scores have been multiplied by the term document frequency. By looking to such ranks, we discover that the word "scandal" had a semantic shift between the 3rd and the 4th week as showed in Table 1.

Another interesting analysis is the variation in similarity values between pairs of words over time: an

[1] TRI is available as an open-source project at: https://github.com/pippokill/tri

upsurge in similarity reflects the increment of co-occurrences between the two words in similar contexts. Figure 1 reports the similarity between "scandal" and "Volkswagen" over time. The plot shows a spike in the similarity value starting from the fourth time interval (21-27 September), which corresponds to the scandal about the Volkswagen diesel emission.

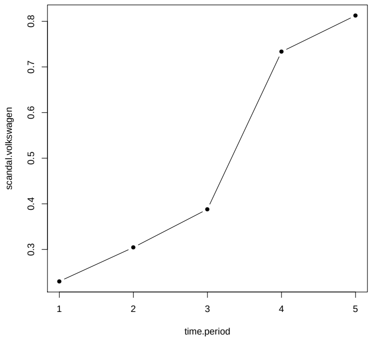

Figure 1: Word-word similarity between the terms: "scandal" and "Volkswagen".

Semantic vectors can be exploited to implement a semantic information retrieval system [BCS11]. The idea is to provide a vector representation for both documents and queries. In particular given a text W (e.g. a document or a query) composed of k terms we can build the vector representation of W as the vector sum of the k semantic vectors occurring in W. Formally, given $W = t_1 t_2 \dots t_k$ the sequence of k terms in W, its vector representation is $\mathbf{w} = sv_{t_1} + sv_{t_2} + \cdots + sv_{t_k}$. Using the same approach we can build the vector representation $\mathbf{q}$ for a query Q. Then the similarity between a query Q and a document D is given by the cosine similarity between $\mathbf{q}$ and $\mathbf{d}$. TRI provides different *WordSpaces* for each time period T_k. Then, the vector representation of a document published during the period T_k is built by exploiting only the semantic vectors of the corresponding *WordSpace*. At query time, the query representation is built by taking into account the semantic vectors of the time period we want to search.

As showed in Figure 1, in the third time period the similarity between "scandal" and "Volkswagen" starts to increase. We try to investigate this phenomenon from the information retrieval point of view. Table 2 reports the first three snippets retrieved by the query "scandal" and "Volkswagen" in the third time period. The column VSM reports results obtained with a classical vector space model implemented by Lucene[2], while the column TRI reports results obtained by TRI.

The VSM model gives more importance to documents that contain both terms, this is the case of the first two documents, while the third document is not relevant at all. The first three documents retrieved by TRI are all relevant for the given query since they all talk about events related to the Volkswagen diesel emission scandal. However, it is interesting to notice that no document contains explicitly the word "scandal". These results can be explained by the nature of the semantic search, which does not rely on string matching, but rather assigns a rank to documents on the basis of their proximity to the semantic vector $scandal + Volkswagen$ taken from the third time period. Semantic search based on TRI opens new opportunities for implementing effective semantic search engines that take into account word meaning variation over time. We plan to deeply investigate this aspect in future research.

Table 2: Search results for query "scandal Volkswagen" in the third time interval: 14th Sept.-20th Sept. 2016

VSM	**TRI**
Volkswagen multi billion pollution coverup *scandal...*	*Volkswagen* to recall 500,000... a device that disguises pollution levels...
Volkswagen emissions cheating... investigations over an emissions *scandal...*	EPA, California investigate *Volkswagen* for clean air violations...
The reinvention of *Volkswagen.* In the *Volkswagen* Group, there is a sense...	*Volkswagen* Ordered to Recall Half a Million Cars After It Cheated on Smog Checks...

References

[BCS11] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. Integrating sense discrimination in a semantic information retrieval system. In *Information Retrieval and Mining in Distributed Environments*, volume 324, pages 249–256. Springer, 2011.

[BCS15] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. Temporal random indexing: A system for analysing word meaning over time. *IJCL*, 1(1):55–68, 12 2015.

[JS09] David Jurgens and Keith Stevens. Event Detection in Blogs using Temporal Random Indexing. In *Proc. of the Workshop on Events in Emerging Text Types*, pages 9–16, 2009.

[Sah05] Magnus Sahlgren. An Introduction to Random Indexing. In *Methods and Applications of Semantic Indexing Workshop at TKE 2005*, volume 5, 2005.

[2] `https://lucene.apache.org/core/`