ISSN 2499-4553

IJCoL

Italian Journal
of Computational Linguistics

Rivista Italiana
di Linguistica Computazionale

Volume 1, Number 1
december 2015

Emerging Topics at the First Italian Conference on Computational Linguistics

aA
ccademia
university
press

registrazione in corso presso il Tribunale di Trento

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

direttore responsabile
Michele Arnese

ISSN 2499-4553
ISBN 978-88-99200-63-3

www.aAccademia.it/IJCoL_01

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it

IJCoL

Volume 1, Number 1
december 2015

Emerging Topics at the First Italian Conference on Computational Linguistics

a cura di
Roberto Basili, Alessandro Lenci,
Bernardo Magnini, Simonetta Montemagni

CONTENTS

Nota Editoriale
Roberto Basili, Alessandro Lenci, Bernardo Magnini, Simonetta Montemagni 7

Distributed Smoothed Tree Kernel
Lorenzo Ferrone, Fabio Massimo Zanzotto 17

An exploration of semantic features in an unsupervised thematic fit evaluation framework
Asad Sayeed, Vera Demberg, and Pavel Shkadzko 31

When Similarity Becomes Opposition: Synonyms and Antonyms Discrimination in DSMs
Enrico Santus, Qin Lu, Alessandro Lenci, Chu-Ren Huang 47

Temporal Random Indexing: A System for Analysing Word Meaning over Time
Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro 61

Context-aware Models for Twitter Sentiment Analysis
Giuseppe Castellucci, Andrea Vanzo, Danilo Croce, Roberto Basili 75

Geometric and statistical analysis of emotions and topics in corpora
Francesco Tarasconi, Vittorio Di Tomaso 91

Il ruolo delle tecnologie del linguaggio nel monitoraggio dell'evoluzione delle abilità di scrittura: primi risultati
Alessia Barbagli, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi 105

CLaSSES: a new digital resource for Latin epigraphy
Irene De Felice, Margherita Donati, Giovanna Marotta 125

Temporal Random Indexing: A System for Analysing Word Meaning over Time

Pierpaolo Basile[*]
Università di Bari, Aldo Moro

Annalina Caputo[*]
Università di Bari, Aldo Moro

Giovanni Semeraro[*]
Università di Bari, Aldo Moro

During the last decade the surge in available data spanning different epochs has inspired a new analysis of cultural, social, and linguistic phenomena from a temporal perspective. This paper describes a method that enables the analysis of the time evolution of the meaning of a word. We propose Temporal Random Indexing (TRI), a method for building WordSpaces that takes into account temporal information. We exploit this methodology in order to build geometrical spaces of word meanings that consider several periods of time. The TRI framework provides all the necessary tools to build WordSpaces over different time periods and perform such temporal 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. This analysis enables the detection of linguistic events that emerge in specific time intervals and that can be related to social or cultural phenomena.

1. Introduction

Imagine the Time Traveller of H.G. Wells' novel who takes a journey to year 2000 in a quest for exploring how the seventh art has evolved in the future. Nowadays, since looking for "moving picture" would produce no results, he would have probably come back to the past believing that the cinematography does not exist at all. A better comprehension of cultural and linguistic changes that accompanied the cinematography evolution might have suggested that "moving picture", within few years from its first appearance, was shorten to become just "movie" (Figure 1). This error stems from the assumption that language is static and does not evolve. However, this is not the case. Our language varies to reflect the shift in topics we talk about, which in turn follow cultural changes (Michel et al. 2011).

So far, the automatic analysis of language was based on datasets that represented a snapshot of a given domain or time period. However, since big data has arisen, making available large corpora of data spanning several periods of time, *culturomics* has emerged as a new approach to study linguistic and cultural trend over time by analysing these new sources of information. The term culturomics was coined by the research group who worked on the Google Book ngram corpus. The release of ngram frequencies spanning five centuries from 1500 to 2000 and comprising over 500 billion words (Michel et al. 2011) opened new venues to the quantitative analysis of changes in culture and linguistics. This study enabled the understanding of how some phenomena impact on written text, like the rise and fallen of fame, censorship, or evolution in

[*] Department of Computer Science, University of Bari Aldo Moro, Via, E. Orabona, 4 - 70125 Bari (Italy).
E-mail: {pierpaolo.basile, annalina.caputo, giovanni.semeraro}@uniba.it.

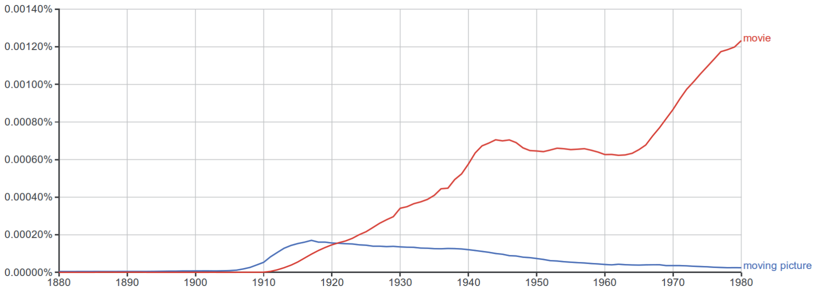

Figure 1
Trends from Google Books Ngram Viewer for words “movie” and “moving picture” over ten decades.

grammar and word senses. This paper focuses on senses, and proposes an algebraic framework for the analysis of word meanings across different epochs.

The analysis of word-usage statistics over huge corpora has become a common technique in many corpus-based 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. Recurrent 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 new perspective by optimizing the objective function of a neural network. 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, 2006), can be easily extended to allow the analysis of semantic changes of words over time (Jurgens and Stevens 2009). The ultimate aim is to provide a tool which enables the understanding of 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* for each different time periods and it makes all these spaces comparable to each other, actually enabling the analysis of word meaning changes over time by simple vector operations in *WordSpace*s.

The paper is structured as follows: Section 2 provides details about the adopted methodology and the implementation of our framework. Some examples that show the potentialities of our framework are reported in Section 3, while Section 4 describes previous work on this topic. Lastly, Section 5 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. Under this light, RI has the advantages of being very simple, since it is based on an incremental approach, and 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 in the same period of time with the target (*temporal*) word, i.e. the word under the temporal analysis. The idea behind RI has its origin in Kanerva work (Kanerva 1988) about Sparse Distributed Memory. RI assigns a random vector to each context unit, in our case represented by a word. The random 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 word as the sum of the random vectors representing the context in which the temporal element is observed. In our case the target element is a word, and contexts are the other co-occurring words that we observe analyzing a large corpus of documents.

Finally, we compute the cosine similarity between the vector representations of word pairs in order to compute their relatedness.

2.1 Random Indexing

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 $n \times m$ matrix A and an $m \times k$ matrix R, which contains random vectors, we define a new $n \times k$ matrix B as follows:

$$A^{n,m} \cdot R^{m,k} = B^{n,k} \quad k << m \tag{1}$$

The new matrix B has the property to preserve the distance between points, that is, if the distance between any two points in A is d; then the distance d_r between the corresponding points in B will satisfy the property that $d_r \approx c \times d$. A proof of that is reported in the Johnson-Lindenstrauss lemma (Dasgupta and Gupta 1999).

Specifically, RI creates the *WordSpace* in two steps:

1. A random vector is assigned to each word. This vector is sparse, high-dimensional and ternary, which means that its elements can take values in {-1, 0, 1}. A random vector contains a small number of randomly distributed non-zero elements, and the structure of this vector follows the hypothesis behind the concept of Random Projection;
2. Context vectors are accumulated by analyzing co-occurring words. In particular the semantic vector for any word is computed as the sum of the random vectors for words that co-occur with the analyzed word.

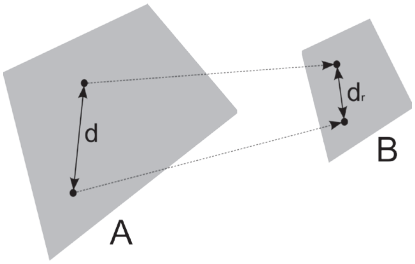

Figure 2
Random Projection.

Formally, given a corpus D of n documents, and a vocabulary V of m words extracted form D, we perform two steps: 1) assign a random vector r to each word w 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. The context is the set of c words that precede and follow w_i. The second step is defined by the following equation:

$$sv_i = \sum_{d \in D} \sum_{\substack{-c<j<+c \\ j \neq i}} r_j \tag{2}$$

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

For example, considering the following sentence: *"The quick brown fox jumps over the lazy dog"*. In the first step we assign a random vector[1] to each term as follows:

$$\begin{aligned}
r_{quick} &= (-1, 0, 0, -1, 0, 0, 0, 0, 0, 0) \\
r_{brown} &= (0, 0, 0, -1, 0, 0, 0, 1, 0, 0) \\
r_{fox} &= (0, 0, 0, 0, -1, 0, 0, 0, 1, 0) \\
r_{jumps} &= (0, 1, 0, 0, 0, -1, 0, 0, 0, 0) \\
r_{over} &= (-1, 0, 0, 0, 0, 0, 0, 0, 0, 1) \\
r_{lazy} &= (0, 0, -1, 1, 0, 0, 0, 0, 0, 0) \\
r_{dog} &= (0, 0, 0, 1, 0, 0, 0, 0, 1, 0)
\end{aligned}$$

In the second step we build a semantic vector for each term by accumulating random vectors of its co-occurring words. For example, fixing $c = 2$ the semantic vector for the word *fox* is the sum of the random vectors *quick, brown, jumps, over*. Summing these vectors, the semantic vector for *fox* results in $(0, 1, 0, -2, 0, -1, 0, 1, 0, 1)$. This operation is repeated for all

[1] The vector dimension is set to 10, while the number of non-zero element is set to 2.

the sentences in the corpus and for all the words in V. In this example, we used very small vectors, but in a real scenario the vector dimension ranges from hundreds to thousands of dimensions.

2.2 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. Specifically, given a document collection D annotated with metatada containing information about the year in which the document was written, we can split the collection in different time periods $D_1, D_2, \ldots, D_p$ we want to analyse. The first step in the classical RI is unchanged in Temporal RI: a random vector is assigned to each word in the whole vocabulary V. This represents the strength of our 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*s T_k is built for each time period D_k. Hence, the semantic vector for a word in a given time period 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 times 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 year $y_{k_{start}}$ to $y_{k_{end}}$, where $y_{k_{start}} < y_{k_{end}}$; then, to build the *WordSpace* T_k we consider only the documents d_k written during T_k as follows:

$$sv_{i_{T_k}} = \sum_{d_k \in D_k} \sum_{\substack{-m<j<+m \\ j \neq i}} r_j \tag{3}$$

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 separate 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.

For example, given the two sentences *“The quick brown fox jumps over the lazy dog”* and *“The Fox is an American commercial broadcast television”* belonging to the different periods of time T_k and T_h, we obtain for the word *fox* the semantic vectors *fox*$_{T_k}$ and *fox*$_{T_h}$. In the first step, we build the random vectors for the words: *american, commercial, broadcast, television*; in addition to those reported in Section 2.

$$\begin{aligned} r_{american} &= (1, -1, 0, 0, 0, 0, 0, 0, 0, 0) \\ r_{commercial} &= (0, 0, -1, 0, 0, 0, 0, 0, 0, 1) \\ r_{broadcast} &= (0, 0, 0, 0, 0, 0, 0, 1, -1, 0) \\ r_{television} &= (0, 0, 0, 1, 0, 0, 0, -1, 0, 0) \end{aligned}$$

The semantic vector for *fox*$_{T_k}$ is the same proposed in Section 2, while the semantic vector for *fox*$_{T_h}$ is $(1, -1, -1, 1, 0, 0, 0, -1, 1)$, which results from the sum of the random vectors of words: *american, commercial, broadcast, television*.

The idea behind this method is to separately accumulate the same random vectors in each time period. Then, the great potentiality of *TRI* lies on the use of the same random vectors to build different *WordSpace*s: semantic vectors in different time periods remain comparable because they are the linear combination of the same random vectors.

Since in the previous example the semantic vectors fox_{T_k} and fox_{T_h} are computed as the sum of different sets of random vectors their semantic similarity would result in a very low value. This low similarity highlights a change in semantics of the word under observation. This is the key idea behind our strategy to analyse change in word meanings over time. We adopt this strategy to perform some linguistic analysis described in Section 3.

2.3 The TRI System

We develop a system, called *TRI*, able to perform Temporal RI using a corpus of documents with temporal information. *TRI* provides a set of features to:

1. Build a *WordSpace* for each year, provided that a corpus of documents with temporal information is available. In particular, given a set of documents with publication year metadata, *TRI* extracts the co-occurrences and builds a *WordSpace* for each year applying the methodology described in Section 2;
2. Merge *WordSpace*s that belong to a specific time period, the new *WordSpace* can be saved on disk or stored in memory for further analysis. Using this feature is possible to build a *WordSpace* that spans a given time interval;
3. Load a *WordSpace* and fetch vectors from it. Using this option is possible to load in memory word vectors from different *WordSpace*s in order to perform further operations on them;
4. Combine and sum vectors in order to perform semantic composition between terms. For example, it is possible to compose the meaning of the two words *big+apple*;
5. Retrieve similar vectors using the cosine similarity. Given an input vector, it is possible to find the most similar vectors which belong to a *WordSpace*. Through this functionality it is possible to analyse the neighbourhood of a given word;
6. 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 online[2] 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 meanings 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. Given two time period intervals and a word w, we can build two *WordSpaces* (T_k and T_h) by summing the *WordSpace*s assigned to the years that belong to each time period interval. Due to the fact that *TRI* makes *WordSpace*s comparable, we can extract the vectors assigned to w in T_k and in T_h, and compute the cosine

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

similarity between them. The similarity shows how the semantics 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. Specifically, we can analyse how the meaning of a word has changed in an interval spanning several periods of time. We study the semantics related to a word by analysing its nearest words in the *WordSpace*. 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. Moreover, it is possible to analyse changes in the semantic relatedness between two words. Given two vector representations of terms, we compute their cosine similarity time-by-time. Since the cosine similarity is a measure of the semantic relatedness between the two term vectors, through this analysis we can detect changes in meanings that involves two words.

3.1 Gutenberg Dataset

The first collection consists of Italian books with publication year by the Project Gutenberg[3] 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 created 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)*[4] related to fascism propaganda occur in $T_{post900}$, while in T_{pre900} we can observe some concepts *(nostra, causa, italia)*[5] probably more related to independence movements in Italy.

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**

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 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

[3] http://www.gutenberg.org/
[4] In English: *(law/order, virtue, honour)*.
[5] In English: *(our, reason, Italy)*.

to *cinematografo* in $T_{post900}$. This phenomenon can be ascribed to the sound introduction after 1900.

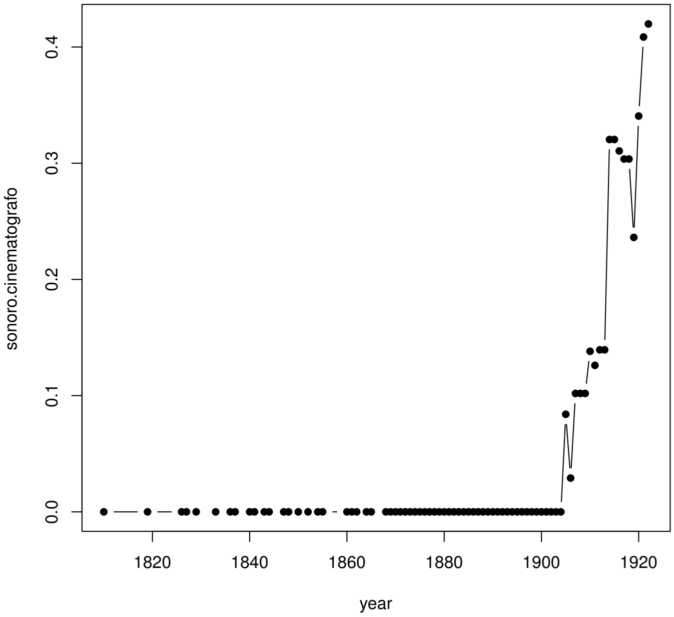

Figure 3
Word-to-word similarity variation over time for Sonoro (*sound*) and Cinematografo (*cinema*) in the Gutenberg dataset.

This behaviour is highlighted in Figure 3 in which we plot the cosine similarity between *cinematrografo* and *sonoro* over the time. This similarity starts to increase in 1905, but only in 1914 we observe a substantial level of similarity between the two terms. We report in Figure 4 a similar case between the words *telefono* (*telephone*) and *chiamare* (*call*, as verb). Their similarity starts to increase in 1879, while a stronger level of similarity is obtained after 1895.

3.2 AAN Dataset

The ACL Anthology Network Dataset (Radev et al. 2013)[6] 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, 2000-2009, 2010-2014), and for each decade we build a different *WordSpace* with *TRI*. 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

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

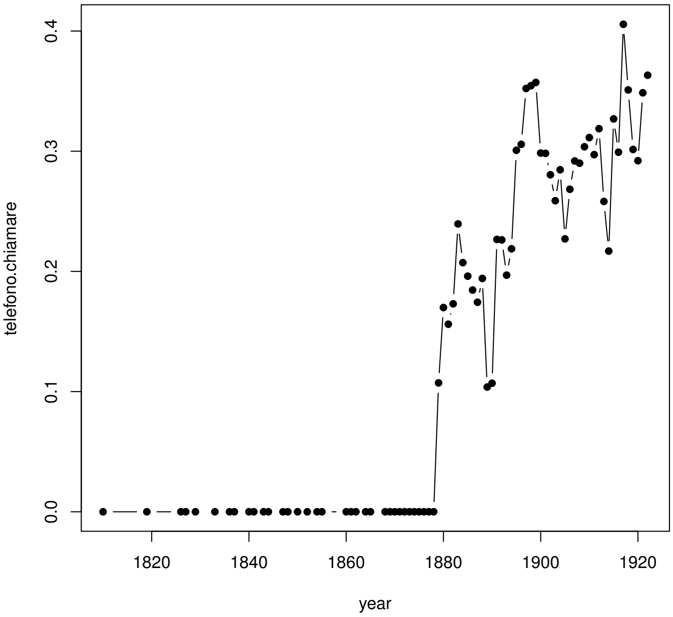

Figure 4
Word-to-word similarity variation over time for Telefono (*telephone*) and Chiamare (*call*) in the Gutenberg dataset.

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.

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

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 with *supervised*, *disambiguation*, *technique*, *probabilistic*, *algorithms*, *statistical*. Finally, the word *logic* has also changed 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 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.

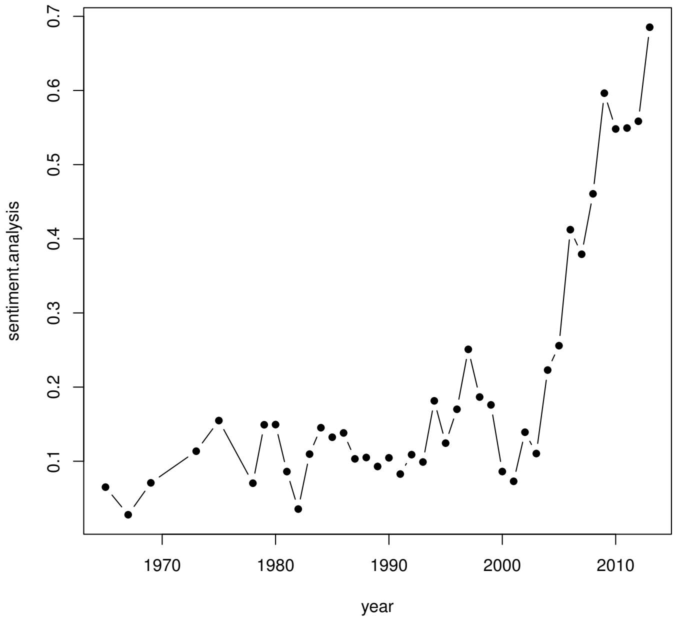

Figure 5
Word-to-word similarity variation over time for Sentiment and Analysis in the AAN dataset.

Figures 5 and 6 show the variation in similarity values between pairs of words: an upsurge in similarity reflects the increment of co-occurrences between the two words in similar contexts. Figure 5 shows the plot of the cosine similarity between the words *sentiment* and *analysis*. We note that in 2004 the similarity is very low (0.22), while only two years later, in 2006, the similarity achieves the value 0.41. This pinpoints the growing interest of the linguistic community about the topic *sentiment analysis* during those years. Analogously, we can plot the similarity values for the words *distributional* and *semantics*. Analysing Figure 6 we can note that these two words have started to show some correlations around the early 70s, followed by a drop of interest until 1989; whereupon, although with a fluctuating trend, the interest in this topic has started to increase more and more.

4. Related Work

The release of Google Book ngram in 2009 has sparked several research fields in the area of computational linguistics, sociology, and diachronic systems. Up until that moment, *"most big data*" were "*big but short*" (Aiden and Michel 2013), leaving little room for massive study of cultural, social, and lexicographic changes during different epochs. Instead, the publication of

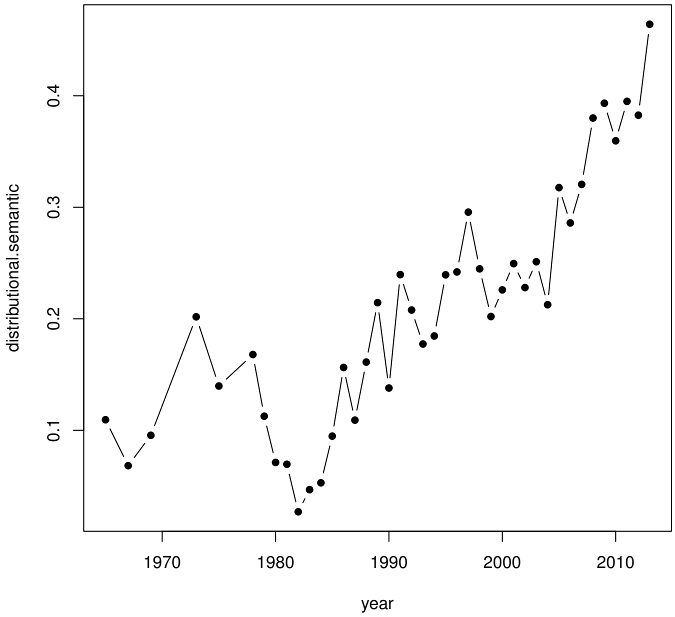

Figure 6
Word-to-word similarity variation over time for Distributional and Semantics in the AAN dataset.

this huge corpus enabled many investigation of both social (Michel et al. 2011) and linguistic trends (Mihalcea and Nastase 2012; Mitra et al. 2014; Popescu and Strapparava 2014).

Through the study of word frequencies across subsequent years, Michel et al. (Michel et al. 2011) were able to study: grammar trends (low-frequency irregular verbs replaced by regular forms), memory of past events, rise and fall in fame, censorship and repression, or historical epidemiology. Moreover, the study of the past enabled prediction for the future. For example, the burst of illness-related word frequencies was studied to predict outbreak in pandemic flu or epidemic (Ritterman, Osborne, and Klein 2009; Culotta 2010).

Some work has tried to detect the main topics or peculiar word distributions of a given time period in order to characterize an epoch. Popescu and Strapparava (Popescu and Strapparava 2014) explored different statistical tests to trace significant changes in word distributions. Then, analysing emotion words associated to terms, they were able to associate an *emotional blueprint* to each epoch. Moreover, they proposed a task (Popescu and Strapparava 2015) to analyse epoch detection on the basis of (1) explicit reference to time anchors, (2) language usage, and (3) expressions typical of a given time period.

Mihalcea and Nastase (Mihalcea and Nastase 2012) introduced the new task of word epoch disambiguation. The authors queried Google Book with a predefined set of words in order to collect snippets for each epoch considered in the experiment. Then, they extracted from the snippets a set of local and topical features for the task of disambiguation. Results suggested that words with highest improvement with respect to the baseline are good candidate for delimiting epochs. Wijaya and Yeniterzi (Wijaya and Yeniterzi 2011) proposed a method to understand changes in word semantics. They proposed a methodology that outdoes the simple observation of word frequencies. They queried Google Books Ngram in order to analyse a predefined set of words, on which they performed two methods for detecting semantic changes. The first method was based on Topics-Over-Time (TOT), a variation of Latent Dirichlet Allocation (LDA) that captures changes in topic. The latter method consisted in retrieving ngrams for a given word by treating all ngrams belonging to a year as a document. Then, they clustered the whole set: a change in meaning occurs if two consecutive years (documents) belong to two different clusters. LDA was also at the heart of the method proposed in (Anderson, McFarland, and Jurafsky 2012). Authors analysed ACL papers from 1980-2008, LDA served to extract topics from the corpus that were assigned to documents, and consequently to people that authored them. This enabled some analysis, like the flow of authors between topics, and the main epochs in ACL history.

Most similar to the method proposed here are those works that avoid the frequentist analysis of a predefined set of words, but rather build a semantic space of words that takes into account also the temporal axis. In such a space, words are not just a number, but have a semantics defined by the context of usage. Kim et al. (Kim et al. 2014) used a vector representation of words by training a Neural Language Model, one for each year from 1850-2009. The comparison between vectors of the same word across different time periods indicates when the word changed its meaning. Such a comparison was performed through cosine similarity. Jatowt and Duh (Jatowt and Duh 2014) exploited three different distributional spaces based on normal co-occurrences, positional information, and Latent Semantic Analysis. The authors built a space for each decade, in order to compare word vectors and detect when a difference between the word contexts has occurred. Moreover, they analysed the sentiment expressed in the context associated to the word over time. Mitra et al. (Mitra et al. 2014) built a distributional thesaurus (DT) for each period of time they wanted to analyse. Then, they applied a co-occurrence graph based clustering algorithm in order to cluster words according to senses in different time periods: the difference between clusters is exploited to detect changes in senses. All these works have in common the fact that they build a different semantic space for each period taken into consideration; this approach does not guarantee that each dimension bears the same semantics in different spaces (Jurgens and Stevens 2009), especially when reduction techniques are employed. In order to overcome this limitation, Jurgens and Stevens (Jurgens and Stevens 2009) introduced Temporal Random Indexing technique as a means to discover semantic changes associated to different events in a blog stream. Our methodology relies on the technique introduced by (Jurgens and Stevens 2009) but with a different aim. While Jurgens and Stevens exploit TRI for the specific task of event detection, in this paper we built a framework on TRI for the general purpose of analysing linguistic phenomena, like changes in semantics between pairs of words and neighbourhood analysis over time.

5. Conclusions

The analysis of cultural, social, and linguistic phenomena from a temporal perspective has gained a lot of attention during the last decade due to the availability of large corpora containing temporal information. In this paper, we proposed a method for building *WordSpace*s taking into account information about time. In a *WordSpace*, words are represented as mathematical points whose proximity reflects the degree of semantic relatedness between the terms involved. The proposed system, called *TRI*, is able to build several *WordSpace*s, which represent words in different time periods, and to compare vectors belonging to different spaces to understand how the meaning of a word has changed over time.

We reported some examples of the temporal analysis that can be carried out by our framework on an Italian dataset about books and an English dataset of scientific papers on computational linguistics. Our investigation shows the ability of our system to (1) capture changes in word usage over time, and (2) analyse changes in the semantic relationship between two words.

This analysis is useful to detect linguistic events that emerge in specific time intervals and that can be related to social or cultural phenomena.

As future work we plan a thoroughly temporal analysis on a bigger corpus like Google ngram and an extensive evaluation on a temporal task, like SemEval-2015 Diachronic Text Evaluation Task (Popescu and Strapparava 2015).

References

Aiden, Erez and Jean-Baptiste Michel. 2013. *Uncharted: Big data as a lens on human culture*. Penguin.

Anderson, Ashton, Dan McFarland, and Dan Jurafsky. 2012. Towards a computational history of the acl: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL '12, pages 13–21, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cohen, Trevor, 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.

Culotta, Aron. 2010. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA. ACM.

Dasgupta, Sanjoy 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.

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

Jatowt, Adam 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 '14, pages 229–238, Piscataway, NJ, USA. IEEE Press.

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

Jurgens, David 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.

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

Kim, Yoon, 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, MD, USA, June. Association for Computational Linguistics.

Landauer, Thomas K. 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.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, The Google Book Team, 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.

Mihalcea, Rada and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 259–263, Jeju Island, Korea, July. Association for Computational Linguistics.

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

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

Mitra, Sunny, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029, Baltimore, Maryland, June. Association for Computational Linguistics.

Popescu, Octavian and Carlo Strapparava. 2014. Time corpora: Epochs, opinions and changes. *Knowledge-Based Systems*, 69:3 – 13.

Popescu, Octavian and Carlo Strapparava. 2015. Semeval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages

870–878, Denver, Colorado, June. Association for Computational Linguistics.

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

Ritterman, Joshua, Miles Osborne, and Ewan Klein. 2009. Using prediction markets and twitter to predict a swine flu pandemic. In *1st International Workshop on Mining Social Media*, volume 9, pages 9–17.

Sahlgren, Magnus. 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.

Sahlgren, Magnus. 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.

Wijaya, Derry Tanti 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 '11, pages 35–40, New York, NY, USA. ACM.