Analyzing Gaussian distribution of semantic shifts in Lexical Semantic Change Models

Pierluigi Cassotti *
Università di Bari A. Moro

Pierpaolo Basile **
Università di Bari A. Moro

Marco de Gemmis †
Università di Bari A. Moro

Giovanni Semeraro ‡
Università di Bari A. Moro

In recent years, there has been a significant increase in interest in lexical semantic change detection. Many are the existing approaches, data used, and evaluation strategies to detect semantic shifts. The classification of change words against stable words requires thresholds to label the degree of semantic change. In this work, we compare state-of-the-art computational historical linguistics approaches to evaluate the efficacy of thresholds based on the Gaussian Distribution of semantic shifts. We present the results of an in-depth analysis conducted on both SemEval-2020 Task 1 Subtask 1 and DIACR-Ita tasks. Specifically, we compare Temporal Random Indexing, Temporal Referencing, Orthogonal Procrustes Alignment, Dynamic Word Embeddings and Temporal Word Embedding with a Compass. While results obtained with Gaussian thresholds achieve state-of-the-art performance in English, German, Swedish and Italian, they remain far from results obtained using the optimal threshold.

1. Background and Motivation

The principal concern of Diachronic Linguistic is the investigation of language change across time. Language changes occur in different language levels: phonology, morphology syntax and semantics. With the growing availability of digitized diachronic corpora, the need for computational approaches able to deal with time annotated corpora becomes more pressing. Diachronic corpora include temporal features, such as the timestamp of the publication date that enables the study of word meaning change across time. The word meaning can be the object of several different types of change: 1) Polarity change, shifting in meaning from positive to negative (*pejoration*) or shifting from negative to positive meaning (*amelioration*); 2) Generalization and specialization refer to a meaning change in the lexical taxonomy. While the former implies a meaning broadening, the latter involves a meaning narrowing.

For example, the Italian verb *pilotare* (to drive) underwent a process of generalization, acquiring the figurative meaning "manipulate"[1]. Cognitive processes involved in language meaning change can be metaphors, metonymies and synecdoches. For example, the Italian word *lucciola* (firefly) acquired a new meaning. *Lucciola* also refers

* Dept. of Computer Science, Via E.Orabona 4, Bari, Italy. Email: pierluigi.cassotti@uniba.it
** Dept. of Computer Science, Via E.Orabona 4, Bari, Italy. Email: pierpaolo.basile@uniba.it
† Dept. of Computer Science, Via E.Orabona 4, Bari, Italy. Email: marco.degemmis@uniba.it
‡ Dept. of Computer Science, Via E.Orabona 4, Bari, Italy. Email: giovanni.semeraro@uniba.it
1 https://www.treccani.it/vocabolario/pilotare/

to a person that in the cinemas uses a portable torch to guide spectators to the seating position. [2] Lastly, it is possible to distinguish among changes due to language-internal or language-external factors, such as psychological, cultural or social causes (Culpeper et al. 2009). The latter usually reflects a change in society, as in the case of technological advancements (e.g. *cellulare* (cell), from the meaning of "compound of biologic cells" to "cell phone").

Lexical Semantic Change is gaining an increasing interest in Computational Linguistics. This is demonstrated by the growing number of publications on computational approaches for Lexical Semantic Change (LSC) and the organisation of related events such as the 1st International Workshop on Computational Approaches to Historical Language Change[3]. Moreover, SemEval 2020 hosted for the first time a task on automatic recognition of lexical semantic change: the SemEval-2020 Task 1 - Unsupervised Lexical Semantic Change Detection[4] (Schlechtweg et al. 2020) for the English, Latin, Swedish and German languages. After SemEval-2020, also EVALITA 2020 hosted the first task on Unsupervised Lexical Semantic Change Detection for the Italian language: DIACR-Ita[5] (Basile et al. 2020b).

In literature, several datasets and tasks are employed for the evaluation of LSC models. Common tasks against LSC models have evaluated are:

- Solving Temporal Analogies, which consists of detecting words analogies across time slices.
- Lexical Semantic Change Detection in a fixed target set requires to assign a label (stable or changed) to each word in a predefined set, as in DIACR-Ita and SemEval-2020 Task 1 Subtask 1.
- Lexical Semantic Change Ranking, rank a target set of words according to their semantic change degree, as in SemEval-2020 Task 1 Subtask 2.
- Lexical Semantic Change Detection in the overall vocabulary, given a list of attested semantic change.

In this work, we focus on the Lexical Semantic Change Detection, using the data provided by both SemEval-2020 Task 1 Subtask 1 and DIACR-Ita. We compare several approaches: Temporal Random Indexing (TRI) (Basile, Caputo, and Semeraro 2016), Temporal Word Embeddings with a Compass (TWEC) (Carlo, Bianchi, and Palmonari 2019), Orthogonal Procrustes Alignment (OP) (Hamilton, Leskovec, and Jurafsky 2016), Temporal Referencing (TR) (Dubossarsky et al. 2019)and Dynamic Word Embeddings (DWE) (Yao et al. 2018). We evaluate all the models against both DIACR-Ita and SemEval-2020 Task 1 since some of these models, currently, have been evaluated in only one of the two tasks.

All the models evaluated in this paper are graded, which means that they output a *degree* of semantic change. The degree of semantic change is typically expressed as the cosine between word vectors (embeddings) computed at different time, assuming that the lowest value of cosine similarity corresponds to the highest degree of change. A common strategy to map the degree of change to discrete stable/change label is:

2 https://www.treccani.it/vocabolario/lucciola/
3 https://languagechange.org/events/2019-acl-lcworkshop/
4 https://competitions.codalab.org/competitions/20948
5 https://diacr-ita.github.io/DIACR-Ita/

- Compute the degree of change δ (cosine similarities) for each target word in the target set T, $\Sigma = \{\delta | w \in T\}$
- Compute the Gaussian $\mathcal{N}(\mu, \sigma)$ parameters of Σ
- Use μ, σ to assign a label to the target words (e.g. target words with degree of change less than $\mu - \sigma$ are labeled as change)

This work aims to get an overview of how thresholds based on the Gaussian parameters (e.g. $\mu - \sigma, \mu, \mu + \sigma$) work over different Lexical Semantic Change models and languages.

The paper is structured as follows: Section 2 reports a review of the approaches used to detect semantic changes in both SemEval-2020 Task 1 and DIACR-Ita tasks, while Section 3 describes the Lexical Semantic Change models under analysis. Section 5 reports details about the evaluation setting used in our work, while results of the evaluation are reported and discussed in Section 6.

2. Related Work

DIACR-Ita and SemEval-2020 Task 1 Subtask 1 require to assign a label (stable or changed) to each word in a predefined set. Most Lexical Semantic Change models produce graded scores that need to be labeled in one of the two classes. Choose a threshold is a crucial phase in binary classification since we need a strategy that should be independent by different Lexical Semantic Change models and languages. Systems that participated in SemEval-2020 Task 1 and DIACR-Ita employed several strategies to label graded scores (e.g. cosine similarities) obtained by Lexical Semantic Change Models.

The simplest approach is based on the idea that stable and changed words are equally distributed. In this case, it is possible to sort the words by the cosine similarity (in ascending order) and the first portion of the set is labelled as change. However, this is a weak approach since the equal distribution assumption does not fit the real-world.

Another common solution is to use an empirically chosen threshold, that, however, could be model-dependent. For instance, models such as DWE or TR produce smoothness changes than OP applied to vectors computed with Skip-grams with Negative Sampling (Mikolov et al. 2013). In (Belotti, Bianchi, and Palmonari 2020), authors use TWEC to compute word vectors and the *move* measure that is a linear combination of the cosine similarity and the similarity of local neighbourhoods. The authors empirically set the *move* threshold to 0.7. The system ranked 3rd in the DIACR-Ita task.

More advanced solutions involve unsupervised approaches to compute the threshold. In (Cassotti et al. 2020), target words are clustered using Gaussian Mixture Clustering (Huang, Peng, and Zhang 2017) to form two clusters: the cluster of change targets and the cluster of stable targets. TRI with Gaussian Mixture Clustering ranked 1st in SemEval-2020 Task 1 Subtask 1 for the Swedish language. In (Zhou and Li 2020) authors hypothesize that the target words cosine distances follow a Gamma distribution. Target words at the peak are classified as stable, while those at the tail are classified as change.

In (Prazák et al. 2020) and (Prazák, Pribán, and Taylor 2020) SGNS vectors are aligned by exploiting Canonical Analysis (Hardoon, Szedmak, and Shawe-Taylor 2004) and Orthogonal Procrustes (Hamilton, Leskovec, and Jurafsky 2016) as Post-alignment models. The authors exploit two different thresholds over the cosine distances: the binary-threshold and the global threshold. The former is computed averaging the target cosine distances, while the latter is computed averaging over the binary-threshold

computed for each language. The system based on the binary-threshold ranked 1st in both SemEval-2020 Task 1 Subtask 1 and DIACR-Ita. The experiments in (Kaiser, Schlechtweg, and im Walde 2020), following the same approach in DIACR-Ita, confirm the results obtained by (Prazák, Pribán, and Taylor 2020).

In general, we can distinguish three different approaches used by systems proposed in SemEval-2020 Task 1 Subtask 1 and DIACR-Ita to compute thresholds by exploiting the change degree Gaussian distribution:

- Approach 1: Compute the Gaussian parameters over the target set.
- Approach 2: Compute the Gaussian parameters over all the dictionary.
- Approach 3: Compute the Gaussian parameters over the targets and get final thresholds averaging across different languages.

3. Models

Most of the Natural Language Processing algorithms that deal with semantics rely on the distributed hypothesis, as Firth puts it, "you shall know a word by the company it keeps " (Firth 1957). In Distributed Semantic Models (DSM), words are mapped to high dimensional vectors in a geometric space. The first DSMs were count-based, they compute word vectors by counting how many times a word appears in a context, sentence, paragraph or document, according to the chosen granularity. The main drawback of count-based models is that they create very high sparse vectors. Dimensionality reduction techniques, such as LSA, helped to overcome this problem, although these techniques require a large computational effort to construct spaces and use them. On the other hand, prediction-based models use a continuous representation of word embeddings to predict the probability distribution $P = (w_t|context) \quad \forall t \in V$ of a target word w_t given the context words $context$, for all the words in the vocabulary V. An example of prediction-based model is the Word2Vec Model Skip-grams with Negative Sampling (SGNS) (Mikolov et al. 2013).

In general, DSMs approaches produce word vectors that are not comparable across time due to the stochastic nature of low-dimensional reduction techniques or sampling techniques. To overcome this issue a widely adopted approach is to align the spaces produced for each time period, based on the assumption that only few words change their meaning across time. Words that turn out to be not aligned after the alignment, changed their semantics.

Alignment models can be classified in post-alignment and jointly alignment models. *Post-alignment* models first train static word embeddings for each time slice and then align them. *Jointly Alignment* models train word embeddings and jointly align vectors across all time slices. Further, *Jointly Alignment* models can be distinguished in *Explicit alignment* models and *Implicit alignment* models. The objective function of *explicit* alignment models involves constraints on word vectors. Typically those constraints require that the distance of two-word vectors in two consecutive periods is the smallest possible. In the *implicit* alignment, there is no need for *explicit* constraint since the alignment is automatically performed by sharing the same word context vectors across all the time spans.

Orthogonal Procrustes (OP) (Hamilton, Leskovec, and Jurafsky 2016) is a Post-alignment model, which aligns word embeddings with a rotation matrix. Word embeddings are computed using traditional approaches such as Singular Value Decomposition (SVD) of Positive Point-wise Mutual Information (PPMI) matrices, FastText (Joulin et al.

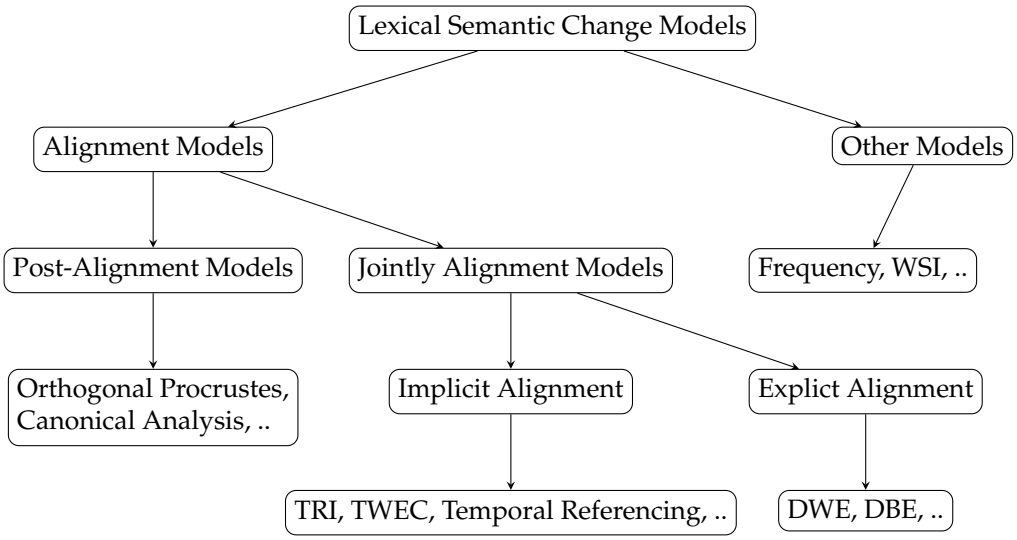

Figure 1
A classification of Lexical Semantic Change models.

2017) or Word2vec. The assumption of the OP method is that each word space has axes similar to the axes of the other word spaces, and two-word spaces are different due to a rotation of the axes. In this work, we use Skip-grams with Negative Sampling (SGNS) (Mikolov et al. 2013) to compute word embeddings and align them using Orthogonal Procrustes (OP-SGNS). In order to align SGNS word emebddings we compute the orthogonal matrix

$$R = \arg\min_{Q^T Q = I} \left\| QW^t - W^{t+1} \right\|_F$$

where W^t and W^{t+1} are two word spaces for time slices t and $t+1$, respectively. We normalize the length of the matrices W^t and W^{t+1} and mean centre them. Q is an orthogonal matrix that minimizes the Frobenius norm of the difference between W^t and W^{t+1}. The aligned matrix is computed as

$$W^{align} = RW^t$$

Dynamic word embeddings (DWE) (Yao et al. 2018) is a Jointly Alignment Model. DWE is based on the PPMI matrix factorization. In a unique optimization function, DWE produces embeddings and tries to align explicitly them according to the following equation:

$$\min_{U(t)} \frac{1}{2} \left\| Y(t) - U(t)U(t)^T \right\|_F^2 + \frac{\lambda}{2} \left\| U(t) \right\|_F^2 +$$
$$\frac{\tau}{2} \left(\left\| U(t-1) - U(t) \right\|_F^2 + \left\| U(t) - U(t+1) \right\|_F^2 \right)$$

where the terms are, respectively, the factorization of the PPMI matrix $Y(t)$, a regularization term and the alignment constraint that keeps the word embeddings similar to the previous and the next word embeddings.

Temporal Word Embedding with a Compass (TWEC), Temporal Referencing (TR) and Temporal Random Indexing (TRI) are instances of *Jointly Implicit Alignment Models*.

TWEC (Carlo, Bianchi, and Palmonari 2019) relies on the two Word2Vec models SGNS and CBOW. TWEC freezes the target and the context embeddings, respectively in CBOW and SGNS model across time, initializing them with the atemporal compass, i.e. word embeddings trained on the whole corpus. TWEC learn temporal specific word embeddings, training only the context or the target embeddings, respectively in CBOW and SGNS models across time.

TR (Dubossarsky et al. 2019) replace a subset of words in the dictionary (target words) with time-specific tokens. Temporal referencing is not performed when the word is considered a context word. Since TR is a generic framework, authors in (Dubossarsky et al. 2019) applied TR to both low-dimensional embeddings learned with SGNS and high-dimensional sparse PPMI vectors. In this work, we focus on the implementation based on SGNS (TR-SGNS). TR requires to fix a set of target words, this makes it impossible to compare words that are not in the target words set.

Finally, we investigate Temporal Random Indexing (TRI) (Basile, Caputo, and Semeraro 2016) that is able to produce aligned word embeddings in a single step. Unlike previous approaches, TRI is a count-based method. TRI is based on Random Indexing (Sahlgren 2005), where a word vector (word embedding) $sv_j^{T_k}$ for the word w_j at time T_k is the sum of random vectors r_i assigned to the co-occurring words taking into account only documents $d_l \in T_k$. Co-occurring words are defined as the set of m words that precede and follow the word w_j. Random vectors are vectors initialized randomly and shared across all time slices so that word spaces are comparable.

4. Data

In this work, we consider data coming from both SemEval and EVALITA.

SemEval-2020 Task 1 (Schlechtweg et al. 2020) comprises two tasks and covers corpora written in four different languages, namely German (Zeitung 2018; Textarchiv 2017), English (Alatrash et al. 2020), Latin (McGillivray and Kilgarriff 2013), and Swedish (Borin, Forsberg, and Roxendal 2012). Corpus statistics are reported in Table 1. Given two corpora C_1 and C_2 for two periods t_1 and t_2, Subtask 1 requires participants to classify a set of target words in two categories: words that have lost or gained senses from t_1 to t_2 and words that did not, while Subtask 2 requires participants to rank the target words according to their degree of lexical semantic change between the two periods.

DIACR-Ita focuses on the Unsupervised Lexical Semantic Change Detection for the Italian language. DIACR-Ita exploits the "L'Unità" corpus (Basile et al. 2020a) that consist of two corpora C_1 and C_2. C_1 covers the period 1945-1970, while C_2 covers the period 1990-2014. An important aspect that distinguishes DIACR-Ita from SemEval is the annotation method. While, SemEval uses the DUREL framework for the annotation, DIACR-Ita relies on a sense-aware method guided by annotation retrieved by the Sabatini Coletti Dictionary (Basile, Semeraro, and Caputo 2019). The method consists of a selection and filtering of candidate words followed by manual annotation. The gold standard is obtained by checking that attested semantic change in the Sabatini Coletti dictionary is present in the training corpus.

Table 1
SemEval-2020 Task 1 statistics.

Language	Corpus	Period	#Tokens
English	CCOHA	1810-1860	6.5M
English	CCOHA	1960-2010	6.7M
German	DTA	1800-1899	70.2M
German	BZ+ND	1946-1990	72.3M
Swedish	Kubhist	1790-1830	71.0M
Swedish	Kubhist	1990-2014	110.0M
Latin	LatinISE	-200-0	1.7M
Latin	LatinISE	0-2000	9.4M

Table 2
DIACR-Ita statistics.

Corpus	Period	#Tokens
L'Unità	1948-1970	52.2M
L'Unità	1990-2014	196.5M

5. Experimental setting

In order to estimates results, avoiding errors due to stochastic parameters initialization, we bootstrap ten runs for each model and language, respectively, averaging the results across the runs. We set the hyper-parameters according to the findings of works proposed for DIACR-Ita and SemEval. For all the models, we set the number of iterations over the data to 5. In particular, for TWEC we set the number of static iterations to 3 and the number of dynamic iterations to 2.

We use a *context-window* of 5 for all the models. We set the number of *negatives* to 5 in all the models that use negative sampling. We set the vector dimension (*dim*) to 300 in all the models, except that for DWE. In DWE, we set the vector dimension *dim* to 100. We use a down-sampling (*sampling*) of 0.001 for all the models: TRI, TWEC, OP-SGNS and TR-SGNS. Table 3, reports models and hyper-parameters values. Where not specified, we adopt default values used by the authors of the models reported in SemEval or DIACR-Ita reports.

In particular, in DWE we specify the number of the alignment weight τ, the regularization weights λ and γ as reported in Table 3. In TRI, we set the number of *seeds* to the default value 10.

6. Results

In SemEval-2020 Task 1, systems are evaluated against three baselines. The Frequency Distance Baseline is based on the absolute difference of the normalized frequency in the two corpora as a measure of change. The Count Baseline implements the model described in (Schlechtweg et al. 2019), while the Majority Baseline always predicts the majority class. DIACR-Ita, as SemEval, provides the frequency distance baseline. Moreover, DIACR-Ita proposes the Collocations baseline. Collocations baseline, introduced

Table 3
Models hyper-parameters.

DWE		TRI		TWEC		OP-SGNS		TR-SGNS	
Param.	Value	Param.	Value	Param.	Value	Param.	Value	Param.	Value
dim	100	dim	300	dim	300	dim	300	dim	300
window	5	window	5	window	5	window	5	window	5
iter	5	iter	5	iter	5	iter	5	iter	5
λ	10	sampling	0.001	sampling	0.001	sampling	0.001	sampling	0.001
γ	100	seeds	10	negatives	5	negatives	5	negatives	5
τ	50								

Table 4
Target words cosine similarities mean and standard deviation across different models and languages, computed on the target set.

Model	English	German	Swedish	Latin	Italian	Model Avg.
TRI	.51±.18	.48±.16	.49±.17	.63±.18	.55±.24	**.53±.18**
DWE	.86±.07	.56±.17	.66±.13	.80±.08	.56±.17	**.69±.13**
TWEC	.65±.10	.54±.12	.56±.12	.61±.10	.59±.15	**.59±.12**
OP-SGNS	.55±.14	.41±.16	.45±.14	.50±.13	.43±.21	**.47±.16**
TR-SGNS	.48±.10	.42±.11	.42±.11	.41±.08	.50±.15	**.45±.11**
Language Avg.	**.61±.12**	**.48±.15**	**.52±.13**	**.59±.12**	**.53±.18**	

Table 5
Target words cosine similarities mean and standard deviation across different models and languages, computed on the overall dictionary.

Model	English	German	Swedish	Latin	Italian	Model Avg.
TRI	.24±.22	.34±.23	.30±.20	.25±.22	.46±.26	**.32±.22**
DWE	.72±.12	.51±.16	.47±.15	.69±.11	.50±.17	**.58±.14**
TWEC	.69±.09	.56±.10	.54±.11	.64±.10	.63±.11	**.61±.10**
OP-SGNS	.51±.16	.40±.15	.36±.17	.44±.15	.43±.17	**.43±.16**
Language Avg.	**.54±.15**	**.45±.16**	**.42±.16**	**.51±.14**	**.50±.18**	

in (Basile, Semeraro, and Caputo 2019), computes the time-dependent representation of targets words using Bag-of-Collocations related to the two different periods. In this work, we use only the frequency baseline. In both SemEval and DIACR-Ita systems are evaluated using the Accuracy.

Tables 4 and 5 report, respectively, the statistics about cosine similarity over the target set and the overall dictionary[6]. The language average cosine computed on the

6 TR-SGNS temporal-aware representations are available only for target words, for this reason it is not possible to compute the cosine similarities for the overall dictionary.

target set is greater than the language average cosine computed on the overall dictionary, even when the target set consists of a greater number of change words, as in the Latin language. It appears that the language average cosine computed on the target set is not correlated with the class balance reported in Table 8.

We test three Gaussian thresholds: $\mu - \sigma$, μ, $\mu + \sigma$ computed over the target set for each language and for each model, as reported in Table 4. We plot the Accuracy obtained by each model, averaging over all the languages in Figure 2. The $\mu - \sigma$ threshold outperforms in every case the μ and $\mu + \sigma$ thresholds. We report results obtained in SemEval in Table 6, while results obtained in DIACR-Ita in Table 7 using the $\mu - \sigma$ threshold. Moreover, to test the efficacy of the Gaussian threshold, we compute the optimal threshold, maximising the accuracy, of λ for each model and language. In particular, we test different values of λ in order to find the optimal value that maximize the accuracy.

In DIACR-Ita task, all models outperform the baseline when $\mu - \sigma$ threshold is used. In SemEval, for the German and Swedish languages, the baseline obtains an accuracy very close to the considered models. This fact is more evident if we consider the optimal threshold. The accuracy obtained by any of the considered models with the Gaussian threshold remains above the accuracy obtained by the Baseline with the optimal threshold. In SemEval, the baseline with the optimal threshold outperforms all the models in combination with the Gaussian threshold in both Swedish and Latin languages.

An important consideration is that the target set of the DIACR-Ita task is smaller than about 50% of the English, German, Swedish and Latin target sets. On the other hand, the class balance of DIACR-Ita is very close to the class balance of German and Swedish languages in SemEval.

The class balance, reported in Table 8, may have affected the effectiveness of the used threshold. The $\mu - \sigma$ threshold never fits the optimal threshold. In particular, the accuracy of all the models using the $\mu - \sigma$ threshold decreases dramatically for the Latin language. We can hypothesize that the $\mu - \sigma$ threshold is affected by the unbalancing of the target set for the Latin language. The Latin language target set consists of only 35% of stable words. Some considerations for the Latin language:

- The target set for the Latin language consists of a greater number of change words rather than stable words, but most of the models rely on the assumption that only few words change their meaning, while the majority remain stable.
- The Latin dataset is challenging, since the first corpus refers to the ancient Latin, while the second one refers to the Latin of the Catholic Church.

These peculiarities make it challenging to compare the results obtained in the Latin language against the other languages.

7. Conclusions

In this work, we evaluated graded Lexical Semantic Change Models using thresholds based on the Gaussian distribution of the cosine similarity. We considered several models: Dynamic Word Embeddings, Temporal Random Indexing, Temporal Referencing, OP-SGNS and Temporal Word Embeddings with a Compass. The evaluation is performed using datasets coming from SemEval-2020 Task 1 Subtask 1 and DIACR-Ita.

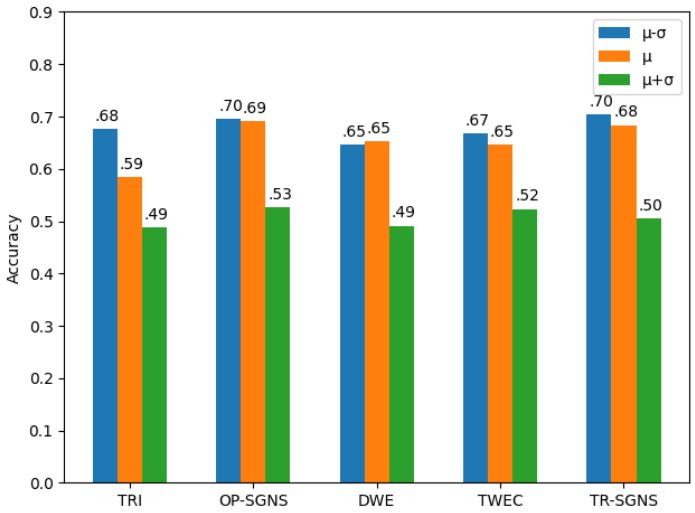

Figure 2
Models accuracy with different Gaussian thresholds: $\mu - \sigma$, μ, $\mu + \sigma$ computed over the target set for each language and for each model. Accuracy is averaged across English, German, Swedish, Latin and Italian language.

Table 6
Accuracy obtained in SemEval-2020 Task 1 Subtask 1.

	English		**German**		**Swedish**		**Latin**	
Model	$\mu - \sigma$	λ	$\mu - \sigma$	λ	$\mu - \sigma$	λ	$\mu - \sigma$	λ
TRI	.65±.03	.67±.02	.65±.02	.70±.04	**.80±.02**	.83±.02	.48±.01	.66±.01
DWE	.66±.03	.69±.01	.69±.02	.73±.03	.74±.02	.81±.02	.40±.02	.67±.01
TWEC	.65±.02	.67±.01	.74±.02	.78±.02	.74±.01	.77±.00	**.49±.03**	.70±.01
OP-SGNS	.64±.02	.66±.02	.75±.02	.80±.01	.75±.03	.79±.02	.44±.02	.69±.01
TR-SGNS	**.71±.01**	.73±.02	**.80±.01**	.87±.02	.73±.02	.79±.02	.45±.02	.70±.02
Baseline	.62±.00	.68±.00	.65±.00	.65±.00	.74±.00	.81±.00	.35±.00	.62±.00

For each dataset and approach, we compute statistics about the Gaussian distribution of the cosine similarity and the optimal threshold for each model to perform a comparison. Results obtained with Gaussian thresholds achieve state-of-the-art performance in English, German, Swedish and Italian. Moreover, results showed that the distribution of the cosine similarities is not correlated with the classes balance in the target set. We plan to investigate how the findings of this work can be used in a completely unsupervised setting, where the evaluation is not limited to a fixed target set but rely on the overall dictionary. Further, we plan to investigate the role of specific word features such as PoS tags and frequency in evaluating performance.

Table 7
Accuracy obtained in DIACR-Ita.

	Italian	
Model	$\mu - \sigma$	λ
TRI	.81±.04	.83±.02
DWE	.76±.04	.84±.02
TWEC	.73±.02	.88±.02
OP-SGNS	**.91±.04**	.96±.02
TR-SGNS	.83±.00	.95±.02
Baseline	.67±.00	.67±.00

Table 8
Classes balance for each language.

Language	Stable	Changed
English	43%	57%
German	67%	33%
Swedish	74%	26%
Latin	35%	65%
Italian	67%	33%

Acknowledgment

This research has been partially funded by ADISU Puglia under the post-graduate programme "Emotional city: a location-aware sentiment analysis platform for mining citizen opinions and monitoring the perception of quality of life".

References

Alatrash, Reem, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean Corpus of Historical American English. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020*, pages 6958–6966, Marseille, France, May. European Language Resources Association.

Basile, Pierpaolo, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020a. A Diachronic Italian Corpus based on "L'Unità". In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769 of *CEUR Workshop Proceedings*, Bologna, Italy, March. CEUR-WS.org.

Basile, Pierpaolo, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020b. DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online event, December. CEUR-WS.org.

Basile, Pierpaolo, Annalina Caputo, and Giovanni Semeraro. 2016. Temporal Random Indexing: a Tool for Analysing Word Meaning Variations in News. In Miguel Martinez-Alvarez, Udo Kruschwitz, Gabriella Kazai, Frank Hopfgartner, David P. A. Corney, Ricardo Campos, and Dyaa Albakour, editors, *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016)*, volume 1568 of *CEUR Workshop Proceedings*, pages 39–41, Padua, Italy, March. CEUR-WS.org.

Basile, Pierpaolo, Giovanni Semeraro, and Annalina Caputo. 2019. Kronos-it: a Dataset for the Italian Semantic Change Detection Task. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics*, volume 2481 of *CEUR Workshop Proceedings*, Bari, Italy, November. CEUR-WS.org.

Belotti, Federico, Federico Bianchi, and Matteo Palmonari. 2020. UNIMIB @ DIACR-Ita: Aligning Distributional Embeddings with a Compass for Semantic Change Detection in the Italian Language (short paper). In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online event, December. CEUR-WS.org.

Borin, Lars, Markus Forsberg, and Johan Roxendal. 2012. Korp - the corpus infrastructure of Spräkbanken. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 474–478, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Carlo, Valerio Di, Federico Bianchi, and Matteo Palmonari. 2019. Training Temporal Word Embeddings with a Compass. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI*, pages 6326–6334, Honolulu, Hawaii,USA, January. AAAI Press.

Cassotti, Pierluigi, Annalina Caputo, Marco Polignano, and Pierpaolo Basile. 2020. GM-CTSC at SemEval-2020 Task 1: Gaussian Mixtures Cross Temporal Similarity Clustering. In Aurélie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING2020*, pages 74–80, Barcelona (online), December. International Committee for Computational Linguistics.

Culpeper, Jonathan, Francis X. Katamba, P. Kerswill, R. Wodak, and T. McEnery. 2009. *English Language: Description, Variation and Context*. Palgrave USA.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 457–470, Florence, Italy, July. Association for Computational Linguistics.

Firth, J. R. 1957. A synopsis of linguistic theory 1930-55. *Studies in linguistic analysis*, 1952-59:1–32.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Volume 1: Long Papers*, Berlin, Germany, August. The Association for Computer Linguistics.

Hardoon, David R., Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Huang, Tao, Heng Peng, and Kun Zhang. 2017. Model selection for gaussian mixture models. *Statistica Sinica*, 27(1):147–169.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Volume 2: Short Papers*, pages 427–431, Valencia, Spain, April. Association for Computational Linguistics.

Kaiser, Jens, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020. OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still Rocks Semantic Change Detection. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop*

(EVALITA 2020), volume 2765 of *CEUR Workshop Proceedings*, Online event, December. CEUR-WS.org.

McGillivray, Barbara and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of Latin. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics, Tübingen. Narr*.

Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, Arizona, USA, May.

Prazák, Ondrej, Pavel Pribán, and Stephen Taylor. 2020. UWB @ DIACR-Ita: Lexical Semantic Change Detection with CCA and Orthogonal Transformation. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online event, December. CEUR-WS.org.

Prazák, Ondrej, Pavel Pribán, Stephen Taylor, and Jakub Sido. 2020. UWB at SemEval-2020 Task 1: Lexical Semantic Change Detection. In Aurélie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING2020*, pages 246–254, Barcelona (online), December. International Committee for Computational Linguistics.

Sahlgren, Magnus. 2005. An introduction to random indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International conference on Terminology and Knowledge Engineering, TKE 2005*, Copenhagen, Denmark, August.

Schlechtweg, Dominik, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 732–746, Florence, Italy, July. Association for Computational Linguistics.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In Aurélie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING2020*, pages 1–23, Barcelona (online), December. International Committee for Computational Linguistics.

Textarchiv, Deutsches. 2017. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften. `http://www.deutschestextarchiv.de/`.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic Word Embeddings for Evolving Semantic Discovery. In Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek, editors, *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018*, pages 673–681, Marina Del Rey, CA, USA, February. ACM.

Zeitung, Berliner. 2018. Diachronic newspaper corpus published by Staatsbibliothek zu Berlin. `http://zefys.staatsbibliothek-berlin.de/index.php?id=155`.

Zhou, Jinan and Jiaxin Li. 2020. TemporalTeller at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection with Temporal Referencing. In Aurélie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING2020*, pages 222–231, Barcelona (online), December. International Committee for Computational Linguistics.