QMUL-SDS @ DIACR-Ita: Evaluating Unsupervised Diachronic Lexical Semantics Classification in Italian

Rabab Alkhalifa[1,2], **Adam Tsakalidis**[1,3], **Arkaitz Zubiaga**[1], and **Maria Liakata**[1,3]

[1]Queen Mary University of London, United Kingdom
[2]Imam Abdulrahman bin Faisal University, Saudi Arabia
[3]Alan Turing Institute, United Kingdom

Abstract

In this paper, we present the results and main findings of our system for the DIACR-Ita 2020 Task. Our system focuses on using variations of training sets and different semantic detection methods. The task involves training, aligning and predicting a word's vector change from two diachronic Italian corpora. We demonstrate that using Temporal Word Embeddings with a Compass C-BOW model is more effective compared to different approaches including Logistic Regression and a Feed Forward Neural Network using accuracy. Our model ranked 3rd with an accuracy of 83.3%.

1 Introduction

The quantitative analysis of language evolution over time is a new emerging research area within the domain of Natural Language Processing (Turney and Pantel, 2010; Hamilton et al., 2016; Dubossarsky et al., 2017). The study of Diachronic Lexical Semantics (Tahmasebi et al., 2018; Kutuzov et al., 2018), which contributes towards detecting word-level language evolution, brings together researchers with broadly varying backgrounds from computational linguistics, cognitive science, statistics, mathematics, and historical linguistics, since the identification of words whose lexical semantics have changed over time has numerous downstream applications in various domains such as historical linguistics and NLP. Despite the increase in research interest, few tasks that track word meaning change over time have focused on non-English languages, while the comparison of different approaches in the same experimental and evaluation setting is still limited (Schlechtweg et al., 2020). The DIACR-Ita 2020 Task (Basile et al., 2020a; Basile et al., 2020b) aims to fill these gaps by focusing on the Italian language used during two different time periods and providing a single evaluation framework to researchers for testing their methods.

This work presents our approach towards detecting Italian words with altered lexical semantics during the two distinct time periods studied in the DIACR-Ita 2020 Shared Task. Our contribution focuses on evaluating findings from previous studies, exploring evaluation approaches for different methods and comparing their performance. We contrast several variants of training-testing words with different alignment approaches across two word embedding models, namely Skip-gram and Continuous Bag-of-Words (Mikolov et al., 2013). Our submission consisted of four models that showed the best average cosine similarity, calculated on the basis of their ability to accurately reconstruct the representations of Italian stop-words across the two periods of time under study. Our best performing model uses a Continuous Bag-of-Words temporal compass model, adapted from the model introduced by (Carlo et al., 2019). Our system ranked third in the task.

2 Related Work

Work related to unsupervised diachronic lexical semantics detection can be divided into different approaches depending on the type of word representations used in a diachronic model (e.g., based on graphs or probability distributions (Frermann and Lapata, 2016; Azarbonyad et al., 2017), temporal dimensions (Basile and McGillivray, 2018), frequencies or co-occurrence matrices (Sagi et al., 2009; Cook and Stevenson, 2010), neural- or Transformer-based (Hamilton

et al., 2016; Boleda et al., 2019; Shoemark et al., 2019; Schlechtweg et al., 2019; Giulianelli et al., 2020), etc.). In our work, we focus on dense word representations (Mikolov et al., 2013), due to their high effectiveness that has been demonstrated in prior work.

Systems operating on representations such as those derived from Skip-gram or Continuous Bag-of-Words leverage in most cases deterministic approaches using mathematical matrix transformations (Hamilton et al., 2016; Azarbonyad et al., 2017; Tsakalidis et al., 2019), such as Orthogonal Procrustes (Schönemann, 1966), or machine learning models (Tsakalidis and Liakata, 2020). The goal of these approaches is to learn a mapping between the word vectors that have been trained independently by leveraging textual information from two or more different periods of time. The common standard for measuring the level of diachronic semantic change of a word under this setting is to use a similarity measure (e.g., cosine distance) on the aligned space – i.e., after the mapping step is complete (Turney and Pantel, 2010).

(Dubossarsky et al., 2017) argue that using cosine distance introduces bias in the system triggered by word frequency variations. (Tan et al., 2015) only use the vectors of the top frequent terms to find the transformation matrix, and then they calculate the similarity for the remaining terms after applying the transformation to the source matrix. Incremental update (Kim et al., 2014; Boleda et al., 2019) used the intersection of words between datasets in each time frame by initializing the word embedding from the previous time slice to compare the word shift cross different years instead of using matrix transformation. Temporal Word Embeddings with a Compass (TWEC) (Carlo et al., 2019) approach uses an approach of freezing selected vectors based on model's architecture, it learn a parallel embedding for all time periods from a base embedding frozen vectors.

Our approaches, detailed in Section 4, follow and compare different methodologies from prior work based on (a) Orthogonal Procrustes alignment, (b) machine learning models and (c) aligned word embeddings across different time periods.

3 Task Description

The task was introduced by (Cignarella et al., 2020) and is defined as follows:

> Given two diachronic textual data, an unsupervised diachronic lexical semantics classifier should be able to find the optimal mapping to compare the diachronic textual data and classify a set of test words to one of two classes: 0 for stable words and 1 for words whose meaning has shifted.

We were provided with the two corpora in the Italian language, each from a different time period, and we developed several methods in order to classify a word in the given test set as "semantically shifted" or "stable" across the two time periods. The test set included 18 observed words – 12 stable and 6 semantically shifted examples.

4 Our Approach

Here we outline our approaches for detecting words whose lexical semantics have changed.

4.1 Generating Word Vectors

Word representations W_i at the period T_i were generated in two ways:

(a) IND: via Continuous Bag of Words (CBOW) and Skip-gram (SG) (Mikolov et al., 2013) applied to each year independently;

(b) CMPS: via the Temporal Word Embeddings with a Compass (TWEC) approach (Carlo et al., 2019), where a single model (CBOW or SG) is first trained over the merged corpus; then, SG (or CBOW) is applied on the representations of each year independently, by initialising and freezing the weights of the model based on the output of the first base model pass and learning only the contextual part of the representations for that year.

In both cases, we used gensim with default settings.[1] Sentences were tokenised using the simple split function for flattened sentences provided by the organisers, without any further pre-processing. Although there are many approaches to generate word representations (e.g., using syntactic rules), we focused on 1-gram rep-

[1] https://radimrehurek.com/gensim/

resentations using CBOW and SG, without considering words lemmas and Part-of-Speech tags.

4.2 Measuring Semantic Change

We employ the cosine similarity for measuring the level of semantic change of a word. Given two word vectors w^{T0}, w^{T1}, semantic change between them is defined as follows:

$$\cos(w^{T0}, w^{T1}) = \frac{w^{T0} \cdot w^{T1}}{\|w^{T0}\| \|w^{T1}\|} = \frac{\sum_{i=1} w_i^{T0} w_i^{T1}}{\sqrt{\sum_{i=1} w_i^{T0^2}} \sqrt{\sum_{i=1} w_i^{T1^2}}} \quad (1)$$

Though alternative methods have been introduced in the literature (e.g., neighboring by pivoting the top five similar words (Azarbonyad et al., 2017)), we opted for the similarity metric which is most widely used in related work (Hamilton et al., 2016; Shoemark et al., 2019; Tsakalidis et al., 2019).

4.3 Evaluation Sets

The challenge is expecting the lexical change detection to be done in an unsupervised fashion (i.e., no word labels have been provided). Thus, we considered stop words[2] (*SW*) and all of the other common words (*CW*) in T_0 and T_1 as our training and evaluation sets interchangeably.

4.4 Semantic Change Detection Methods

We employed the following approaches for detecting words whose lexical semantics have changed:

(a) Orthogonal Procrustes (OP): Due to the stochastic nature of CBOW/SG, the resulting word vectors W_0 and W_1 in *IND* were not aligned. Orthogonal Procrustes (Hamilton et al., 2016) tackles this issue by aligning W_1 based on W_0. The level of semantic shift of a word is calculated by measuring the cosine similarity between the aligned vectors. For evaluation purposes, we measured the cosine similarity of the stop words between the two aligned matrices. Higher values indicate a better model (i.e., stop words retain their meaning over time).

(b) Feed-Forward Neural Network (FFNN): We trained a FFNN that leverages *IND* to predict W_1 based on W_0. The level of semantic shift of a word in a test set is calculated by measuring the cosine similarity between the predicted W_1^* and W_1. For evaluation purposes, we measure the cosine similarity between the actual and predicted representations of words in T_1. Higher values for stop-words indicate a better model.

(c) Linear Regression (LR): We employed an ordinary linear mapping with least square error objective function.[3] The task and the evaluation setting was identical to FFNN.

(d) Temporal Word Embeddings with a Compass (TWEC) (Carlo et al., 2019): Working on the *CMPS* vectors, the level of semantic shift of a word is calculated by measuring the cosine similarity between T_0 and T_1 directly.

Notation In the rest of this paper, we denote a model M trained on CW (SW) as M_CW (M_SW). For the case of *OP*, the training process involves learning an alignment based on a specific word set (CW or SW). Note that this notation does not apply for $TWEC$, since the word vectors in the two time periods can be directly compared against each other – thus the level of semantic change can be calculated directly (i.e., there is no need to learn any mapping between W_0 and W_1). Finally, we add a subscript $_{CBOW}$ or $_{SG}$ to our models, denoting the type of algorithm that was used for generating the respective embeddings that are fed to our model.

Model Selection We select to apply the models on the test set providing high average cosine similarity with stop words.

4.5 Word Classification

As per the task guidelines (Cignarella et al., 2020), words can fall into one of the two categories: **0**: the target word does not change meaning between T_0 and T_1 and **1**: the target word changes its meaning between T_0 and T_1. For all of our submitted models, we considered all the words with cosine similarity below the mean as shifted words and labelled them with 1. We further investigate the model's ability to detect words laying two standard deviations below the mean ($\mu - 2\sigma$), a.k.a variance. Interestingly, some of the models including LR and FFNN_CW$_{CBOW}$ showed an increase in accuracy.

5 Results

The results are shown in Table 1, where we split our results based on model #M ar-

[2] https://github.com/stopwords-iso/stopwords-it

[3] https://scikit-learn.org/stable/

IND		SG						C-BOW					
		Accuracy			Ranking			Accuracy			Ranking		
train.	M	CS^{SW}_{avg}	$\%\mu$	$\%\mu-2\sigma$	$\%\mu_{rank}$	R_{p50}	$R_{\downarrow 6}$	CS^{SW}_{avg}	$\%\mu$	$\%\mu-2\sigma$	$\%\mu_{rank}$	R_{p50}	$R_{\downarrow 6}$
SW	***OP***	0.748	**0.778**	**0.667**	**0.222**	**1.000**	**0.667**	0.784	**0.778**	0.667	**0.270**	**1.000**	**0.833**
	LR	**0.854**	0.333	0.389	0.373	0.833	0.500	**0.795**	0.500	**0.778**	**0.278**	0.833	0.500
	FFNN	**0.769**	0.333	0.333	0.373	0.833	0.500	0.709	0.556	**0.722**	0.341	0.833	0.500
CW	***OP***	0.464	0.389	**0.778**	0.381	0.667	0.500	0.289	0.611	0.667	0.397	0.833	0.333
	LR	0.409	0.333	0.444	0.508	0.500	0.333	0.146	0.333	0.444	0.381	0.667	0.667
	FFNN	0.658	0.333	0.389	**0.317**	**1.000**	0.500	0.621	0.333	**0.722**	0.317	0.833	0.500
	TWEC	0.722	**0.722**	**0.667**	**0.317**	0.833	**0.667**	**0.833**	**0.833**$_*$	0.667	0.286	**1.000**	**0.667**

Table 1: Performance of our models using different evaluations methods. *$(_*)$ best submission.*

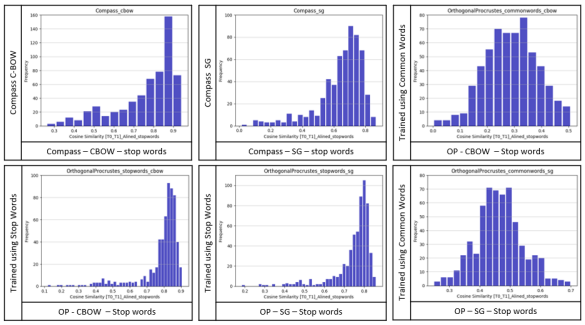

Figure 1: Frequency of stop words by their cosine similarity scores, where each subfigure pertains to a different model.

chitecture, $_{SG}$ and $_{CBOW}$ and model's training word sets, Stop-Words (SW) and Common-Words (CW). For models based on linear transformation, our top performing models scored below average cosine similarity, TWEC$_{CBOW}$ (0.833), OP_SW$_{SG}$(0.778), OP_SW$_{CBOW}$ (0.778), TWEC$_{SG}$(0.722). As shown in Figure 5, we observe that these models tend to have skewed distributions for stop words, where the vast majority of stop words are assigned high cosine similarity scores. However, other models did not show this skewness, e.g. OP_CW$_{SG}$(0.389) and OP_CW$_{CBOW}$(0.611). When labeling the change based on variance ($\mu-2\sigma$), as in outlier detection, some models showed an increase from the dummy classifier's performance. For instance, OP_CW$_{sg}$ showed an increase on performance from (0.389) to (0.778) showing that those with low average cosine similarity lay out in the tail from majority similarity. Similarly, models based on reducing the similarity error between the predicted and actual vectors, e.g. LR and FFNN considering the outlier detection methodology, tend to achieve better performance, including LR_SW$_{CBOW}$, FFNN_SW$_{CBOW}$ and FFNN_CW$_{CBOW}$ where LR_SW$_{CBOW}$ showed an increase from frequency classifier's baseline (0.500) to (0.778), and LR_SW$_{CBOW}$ showed an increase from dummy classifier performance (0.333) to (0.722).

Ranking methods, average ranking (μ_{rank}) and Recall (R), expect prior knowledge about the evaluation labels to make them useful for evaluating the reliability of the model of interest. For that, we further investigate the reliability of our experiment models, using μ_rank and R at %50 (R$_{p50}$) and %30 (R$_{\downarrow 6}$). Although using (R$_{p50}$) signal OP_SW$_{SG}$, OP_SW$_{CBOW}$, FFNN_CW$_{SG}$, TWEC$_{CBOW}$ as equalliy good, μ_{rank} ranked top models as OP_SW$_{SG}$, OP_SW$_{CBOW}$, LR_SW$_{CBOW}$ then TWEC$_{CBOW}$ with (0.222, 0.270, 0.278 and 0.286), respectively. Additionally, under extreme conditions, OP_SW$_{CBOW}$ ranked better than all including TWEC$_{CBOW}$. This shows that under extreme conditions, a good method is the one which keeps providing out of distribution signals to changing words and that needs to take a careful consideration about the distribution of the words before and after the alignments as in OP. In general, CBoW-based models showed better performance than SG-based models with average accuracy of ($\%\mu$ 0.564 and $\%\mu-2\sigma$ 0.667) compared to ($\%\mu$ 0.460 and $\mu-2\sigma$ 0.524) for words labelled by mean and variance, respectively. Further, alignment using non-changing words (e.g. *stop-words*) yields higher performance than using all common words with average cosine similarity for stop words as (CS^{SW}_{avg} 0.777) compared to (CS^{SW}_{avg} 0.431), which is expected because SW-based models learns the optimal mapping with less noise than CW-based models.

6 Discussion

Our work provides a comprehensive analysis for Italian lexical diachronic methods introduced from previous work. For models that are based on matrix linear transformation including TWEC and OP, we find a relation between high average stop words similarity and accuracy. Further, C-BOW tends to achieve better results than the SG architecture for most experiments. Visually, we find that a visibly skewed distribution showing the tendency of stop words to have high cosine similarity scores leads to effective means for capturing semantic shift. We also showed that by evaluating the models using different methods, TWEC_{CBOW} achieved top performance. Followed by OP_SW and OP_CW_{SG}, and LR using outlier detection methodology. Further, FFNN showed high recall (R_{p50}) by ranking changed words with lowest cosine similarity on testing set similar to OP_SW and TWEC_{CBOW}. This provides promising insights encouraging further investigation of neural network models using different languages and larger datasets.

7 Conclusions

In this report, we describe and compare our models submitted to the DIACR-Ita 2020 shared task, which assessed the ability to classify semantic-shift of words in Italian. We show that the TWEC model yields better performance than Orthogonal Procrustes, labelling all words scored below average cosine similarity as semantically shifted words, i.e. words with altered semantics over the two time periods. Additionally, we showed that using an outlier detection methodology yields better results in prediction-based models such as Linear Regression and Feed-Forward Neural Network, boosting the performance significantly compared to the baselines and dummy classifier.

In the future we aim to focus on fine tuning SoTa pre-trained language models such as ELMo and BERT for word level semantics-shift detection as well as investigating the ability of dynamic graph models on capturing word evolution.

8 Acknowledgments

This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT.

References

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. *International Conference on Information and Knowledge Management, Proceedings*, Part F1318(3):1509–1518.

Pierpaolo Basile and Barbara McGillivray. 2018. Exploiting the web for semantic change detection. In *International Conference on Discovery Science*, pages 194–208. Springer.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020a. DIACR-Ita @ EVALITA2020: Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020b. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Gemma Boleda, Marco Del Tredici, and Raquel Fernández. 2019. Short-term meaning shift: a distributional exploration. *Proceedings of the 2019; 2019 Jun 2-7; Minneapolis, United States of America. Stroudsburg (PA): ACL; 2019. p. 2069–75.*

Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training temporal word embeddings with a compass. *CoRR*, abs/1906.02376.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Overview of the EVALITA 2020 Task on Stance Detection in Italian Tweets (SardiStance). In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Paul Cook and Suzanne Stevenson. 2010. Automatically Identifying Changes in the Semantic Orientation of Words. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on*

empirical methods in natural language processing, pages 1136–1145.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July. Association for Computational Linguistics.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Yoon Kim, 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.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111. Association for Computational Linguistics.

Dominik Schlechtweg, 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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *arXiv preprint arXiv:2007.11464*.

Peter H Schönemann. 1966. A Generalized Solution of the Orthogonal Procrustes Problem. *Psychometrika*, 31(1):1–10.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint arXiv:1811.06278*.

Luchen Tan, Haotian Zhang, Charles Clarke, and Mark Smucker. 2015. Lexical comparison between wikipedia and twitter corpora by using word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 657–661.

Adam Tsakalidis and Maria Liakata. 2020. Autoencoding word representations through time for semantic change detection. *arXiv preprint arXiv:2004.13703*.

Adam Tsakalidis, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile, and Barbara McGillivray. 2019. Mining the UK web archive for semantic change detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1212–1221.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.