UNIMIB @ DIACR-Ita: Aligning Distributional Embeddings with a Compass for Semantic Change Detection in the Italian Language

Federico Belotti
University of Milano-Bicocca
Viale Sarca 336, 20126
Milan, Italy
f.belotti8@campus.unimib.it

Federico Bianchi
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
f.bianchi@unibocconi.it

Matteo Palmonari
University of Milano-Bicocca
Viale Sarca 336, 20126
Milan, Italy
matteo.palmonari@unimib.it

Abstract

In this paper, we present our results related to the EVALITA 2020 challenge, DIACR-Ita, for semantic change detection for the Italian language. Our approach is based on measuring the semantic distance across time-specific word vectors generated with Compass-aligned Distributional Embeddings (CADE). We first generate temporal embeddings with CADE, a strategy to align word embeddings that are specific for each time period; the quality of this alignment is the main asset of our proposal. We then measure the semantic shift of each word, combining two different semantic shift measures. Eventually, we classify a word meaning as changed or not changed by defining a threshold over the semantic distance across time.

1 Introduction

Semantic change detection is the task of detecting if a word has shifted in meaning between different periods of time (Tahmasebi et al., 2018; Kutuzov et al., 2018). The DIACR-Ita (Basile et al., 2020a) challenge (at EVALITA (Basile et al., 2020b)) is meant to evaluate approaches for semantic change detection for the Italian Language.

The task is described as follows: for training, two corpora $t1$ and $t2$, consisting of text coming from different periods are given, for testing, a set of unlabeled target words is given, where for each of them a binary scores has to be predicted: 1 identifies lexical change between $t1$ and $t2$ while 0 does not.

In this paper, we present our approach to semantic change detection that is based on two components: 1) an alignment procedure to generate distributional vector spaces that are comparable for $t1$ and $t2$ and 2) the use of distance metrics to compute the degree of semantic change for a given word. Our alignment procedure is based on Compass Aligned Distributional Embeddings (CADE) proposed by Bianchi et al. (2020) (note the approach was introduced as Temporal Word Embeddings with a Compass by Di Carlo et al. (2019), but the name was changed to enforce the idea that the embeddings can be used to align more general corpora and not just diachronic ones). Given the aligned embeddings, we use two measures to compute the degree of change based on the similarities of the vectors in the embedded space. Our results show that our methodology for aligning spaces can be useful in detecting lexical semantic change.

2 Description of the System: Semantic Change Detection with Compass Aligned Embeddings

Our approach is based on measuring the semantic distance across time of time-specific word vectors generated with CADE and on the use of two measures for detecting semantic shifts i.e., the semantic distance between word vectors across time. This distance can be interpreted as a function of the words' self-similarity across time, where the similarity is measured by a linear combination of cosine and second-order similarity (Hamilton et al., 2016a).

Finally, a threshold over this self-similarity is used to classify a word as changed or not changed.

This methodology was applied also in the semantic shift detection challenge presented at SemEval2020 (Schlechtweg et al., 2020) (to which we participated after the end of the challenge). The challenge allowed us to explore and understand how the alignment and our self-similarity behaved. In the classification task of the SemEval2020 challenge (the one similar to this task),

"Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)."

we eventually achieved 0.703, 0.771, 0.725, 0.742, in accuracy for respectively the English, German, Latin and Swedish languages; these results have been obtained with extensive parameter search given the gold standard available in the post-evaluation.[1] In DIACR-Ita, the threshold and few other hyper parameters can be heuristically set to account for the limited number of possible submissions. In the next subsections we provide more details about the alignment methodology and the similarity function; more details about how we set the hyper parameters are provided in Section 3.

2.1 Aligning Embeddings

Word2vec (Mikolov et al., 2013) is a useful methodology to generate vectors of words allowing us to study word similarity through vector similarity. However, due to the stochasticity of the training procedure, running word2vec on different corpora creates word vectors that are not comparable. Thus, an alignment procedure that puts the temporal word vectors in the same space is needed.

There are different approaches to generate these aligned embeddings (see for example the work by (Hamilton et al., 2016b) and (Yao et al., 2018)). In this paper, we generate aligned embeddings with Compass Aligned Distributional Embeddings (CADE) (Bianchi et al., 2020) (See Figure 1 for a schematic description of the model). CADE is a strategy to align word embeddings that are specific for each time period that extends the word2vec Continuous Bag Of Word (CBOW) model proposed by Mikolov et al. (2013). CADE can be used to generate aligned temporal word embeddings (i.e., time-specific vectors of words, like "$amazon^{1974}$") from the different slices.

Given in input a set of slices of text, where each slice corresponds to text coming from a specific period of time, the alignment procedure is as follows:

First, the text from all the slices is concatenated and CBOW is run on this corpus in order to obtain a "compass" model, i.e., a model defining the embedding space. The CBOW model uses two matrices to generate the embeddings (**U** and **C** in Figure 1), one for the context words and one for the target words. The target word matrix of the compass is then used to initialize the target matrices for each new CBOW model fitted on each of the slices. During training, these new target matrices are frozen, i.e., they are not updated during the training on the slice. This ensures that at the end of the training process, the various temporal embeddings are all aligned in the same embedding space, making them comparable without losing their individual temporal distinctions. We use the publicly available online implementation of CADE.[2]

2.2 Computing Semantic Change

Once the embeddings are aligned, we need measures to evaluate the degree of semantic change. We compute the semantic shift of each word, i.e. the semantic distance between word vectors across time using the combination of two different measures: Local Neighbors (ln), introduced by Hamilton et al. (2016a) and cosine similarity (cos), merging them with a weighted linear combination into a new measure called *Move*.

Local Neighbors *ln* is based on the similarity between a word and its neighbor words in the two different time periods. Essentially we compute the degree of semantic change of the word w in two slices by first collecting the nearest neighbors (NNs) of $\mathbf{w}^t$ and $\mathbf{w}^{t+1}$ in the two respective slices, then given the embeddings at time t the similarities between the vector of $\mathbf{w}^t$ and the vectors of all the neighbors are computed.[3] The same process is run for time $t+1$ with $\mathbf{w}^{t+1}$, eventually giving us two vectors of similarity scores. These two vectors are again compared using cosine similarity. The higher the value of this measure the less the vector has changed with respect to its neighbors and thus the less the word should have shifted in meaning.

Cosine Similarity The second measure we use is simply the cosine similarity of the vectors of a word in two different time periods. Similarly as before , the higher the value the less the vector has changed and thus the less the word should have shifted in meaning.

The *Move* Measure We merge these measures together using a weighted linear combination, that is:

$$s(w^t, w^{t+1}) = (1-\lambda)\cdot \ln(w^t, w^{t+1}) + \lambda \cdot \text{cos-sim}(\mathbf{w}^t, \mathbf{w}^{t+1})$$

[1] Check the *belerico* entry in the challenge leaderboard at `https://competitions.codalab.org/competitions/20948#results`

[2] `http://github.com/vinid/cade`

[3] When a neighbor is missing in one time slice, we replace it with the average vector of the space.

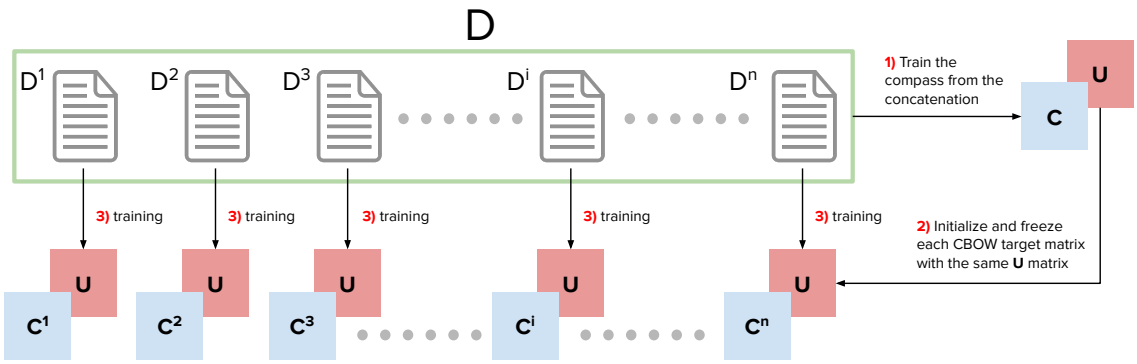

Figure 1: An high level overview of the Compass Aligned Distributional Embeddings model.

with $\lambda \in [0, 1]$. In particular λ express the usage strength of the two measures: a high λ will shift *Move* towards the cosine similarity, while a low one towards the *ln* measure. As introduced before we classify if the meaning has changed by defining a threshold over s (more details about this are presented in the next Section).

3 Experimental Evaluation

The dataset provided by the challenge's organizers (Basile et al., 2020a) is a collection of documents extracted by newspapers written in the Italian language labeled with temporal information. Participants must train their models only on the data provided, so a pre-processed corpus is given: tab separated, with one token per line, where for each token there are its corresponding part-of-speech (POS) tag and lemma, with sentences separated by empty lines. The corpus is split into two slices, each belonging to a specific period of time, $t1$ and $t2$, where $t1 < t2$.

3.1 Dataset

For the training data we used the *flat* version with only the lemmas, obtained by the organizers' script (Basile et al., 2020a); in addition we applied a pre-processing step, in which we removed punctuation and non alpha-numeric symbols and we kept only those sentences with at least two tokens.

3.2 Models Considered

We use the embeddings aligned with CADE and the *move* measure. The parameters of the moving average we need to consider are: the number of nearest neighbors (NNs) to be collected by *ln*, λ for the moving average and the threshold for the similarity. We set the threshold to decide if a word is stable or not is set to 0.7, with the decision given by:

$$\begin{cases} 0 & \text{if } s(w^t, w^{t+1}) \geq 0.7 \\ 1 & \text{otherwise} \end{cases}$$

Essentially, the less changed are the two vectors of the words (for *cos*) and the neighbors (for *ln*) the more the word has been stable between the two time periods. As heuristics we chose $\lambda \in \{0.3, 0.5, 0.7\}$ to evaluate the relationship between the two measures used to build *move*, and we set to 22 the number of nearest neighbors to be considered by the *ln*; this is the general setup that gave the results that have been submitted to the challenge.

We trained CADE for 10 epochs to learn 100-dimensional vectors, with the window size set to 5, 10 negative examples for every positive one, with the initial learning rate set to 0.025 and decreased linearly during training.

As other models, in the post evaluation we also considered one that only uses the *cos* (CADE (cos)) similarity measure and one that uses only the *ln* metric CADE (ln)) (again with 0.7 as threshold and with the number of NNs for *ln* set to 22).

As baselines, the authors propose to use *baseline-freq*, that is the absolute value of the difference between the words' frequencies and *baseline-colloc*, where the Bag-of-Collocations of the two words in the two different periods is built and then cosine similarity is applied. A threshold is used on both metrics to define semantic change (Basile et al., 2020a). We report also the results of the other participants.

	λ	Acc.
team$_1$	/	0.944
team$_2$	/	0.944
team$_3$	/	0.889
CADE (move)†	0.3	**0.833**
team$_4$	/	0.833
team$_5$	/	0.833
team$_6$	/	0.778
team$_7$	/	0.722
team$_8$	/	0.667
team$_9$	/	0.611
baseline-colloc	/	0.611
baseline-freq	/	0.500
CADE (move)†	0.5	0.722
CADE (move)†	0.7	0.722
CADE (cos)	/	0.722
CADE (ln)	/	**0.889**

Table 1: Accuracy scores for the binary classification w.r.t. the other participants to the challenge. † identifies our submitted results.

3.3 Results

The evaluation metric used in this challenge is the accuracy, that is, the number of correct predictions over the target data. Table 1 shows the results. Our model was the third most accurate. However, in the post-evaluation we discovered that just using the *ln* metric and ignoring the use of *cos* (this is equivalent to using $\lambda = 0$ in our *move* measure) improves the performance leading to the second best accuracy score in the leaderboard.

4 Discussion

Our results show that CADE (Bianchi et al., 2020) is an effective method to generate aligned embeddings for the Italian language. This result, together with those obtained on the SemEval2020 data, suggest that CADE can support models of semantic shift detection in several languages. Indeed, we show that in combination with some simple semantic change measures it is possible to provide a good model for semantic change detection that can be subsequently extended with more features. Appendix A contains some more detailed examples of the words that CADE (ln) and CADE (move), with lambda set to 0.3, could not classify correctly. Also, we show the neighborhood for some of those words to give more context on why we get those errors. A more precise use of pre-processing techniques with the combination of other metrics to compute semantic change might help in reducing these errors.

References

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020a. 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 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.

Federico Bianchi, Valerio Di Carlo, Paolo Nicoli, and Matteo Palmonari. 2020. Compass-aligned distributional embeddings for studying semantic differences across corpora. *arXiv preprint arXiv:2004.06519*.

Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training temporal word embeddings with a compass. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6326–6334.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas, November. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

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, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

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.

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*.

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

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining*, pages 673–681.

A CADE Misclassifications

We report in Tables 2 and 3 CADE's misclassifications with the two best metrics, namely CADE (move) with $\lambda = 0.3$ and CADE (ln). Eventually, we also show in Tables 4 and 5 some examples of neighborhood for the target words.

Word	Pred	True
trasferibile	changed	not changed
pacchetto	changed	not changed
piovra	changed	not changed

Table 2: Wrong predictions done by CADE (move) with $\lambda = 0.3$.

Word	Pred	True
pacchetto	changed	not changed
rampante	not changed	changed

Table 3: Wrong predictions done by CADE (ln).

Table 4 shows the top 10 nearest neighbors of the target word "pacchetto" and we think CADE classifies its meaning as changed because during time t_1 the meaning is more focused in the economic area, as one can see from neighbors like "azionario", "obbligazione" or "contante" (translated to "stock" as referred to the market, "bond" and "cash" resp.); while at time t_2 shifts to a more political sense, as shown by words such as "decreto" or "emendamento" ("decree" and "amendment" resp.).

t_1	t_2
azionario	maxiemendamento
obbligazione	finanziaria
azionista	decretone
azionano	decreto
edison	ddl
casseforte	emendamento
contante	liberalizzazioni
siap	decretere
shell	maxidecreto
prestire	ecobonus

Table 4: First 10 nearest neighbors by cosine similarity of the word "pacchetto" from t_1 and t_2

The same it seems to happen for the target word "piovra", as one can see from Table 5, where at time t_1 CADE gathers senses from both considering it as the animal, for example from the word "tentacle", or as someone tied to crime in general, given words such as "profittatore" or "ruberia" ("profiteer" and "robbery" resp.); while at time t_2 captures a shift towards the Italian crime TV series "La piovra", as emerge from words such as "fiction", "camorra" or "retequattro", which is an Italian television channel.

t_1	t_2
tentacolo	fiction
ingordigia	sceneggiato
profittatore	tentacolo
somaro	camorrere
feudatario	retequattro
insaziabile	raidue
impere	puntato
ruberia	camorra
zanne	gomorra
putrido	miniserie

Table 5: First 10 nearest neighbors by cosine similarity of the word "piovra" from t_1 and t_2