DIACR-Ita @ EVALITA2020

Basile, Pierpaolo; Caputo, Annalina; Caselli, Tommaso; Cassotti, Pierluigi; Varvara, Rossella

Published in:
Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)

DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task

Pierpaolo Basile
Dept. of Computer Science
University of Bari, Italy
pierpaolo.basile@uniba.it

Annalina Caputo
ADAPT Centre
School of Computing, Dublin City University
annalina.caputo@dcu.ie

Tommaso Caselli
CLCG
University of Groningen, Netherlands
t.caselli@rug.nl

Pierluigi Cassotti
Dept. of Computer Science
University of Bari, Italy
pierluigi.cassotti@uniba.it

Rossella Varvara
DILEF
University of Florence, Italy
rossella.varvara@unifi.it

Abstract

English. This paper describes the first edition of the "Diachronic Lexical Semantics" (DIACR-Ita) task at the EVALITA 2020 campaign. The task challenges participants to develop systems that can automatically detect if a given word has changed its meaning over time, given contextual information from corpora. The task, at its first edition, attracted 9 participant teams and collected a total of 36 submission runs.

1 Background and Motivation

The Diachronic Lexical Semantics (DIACR-Ita) task focuses on the automatic recognition of lexical semantic change over time, combining together computational and historical linguistics. The aim of the task can be shortly described as follows: given contextual information from corpora, systems are challenged to detect if a given word has changed its meaning over time.

Word meanings can evolve in different ways. They can undergo *pejoration* or *amelioration* (when meanings become respectively more negative or more positive) or they can be object of *broadening* (also referred to as *generalization* or *extension*) or *narrowing* (also known as *restriction* or *specialization*). For instance, the English word *dog* is a clear case of broadening, since its more general meaning came from the late Old English "dog of a powerful breed" (Traugott, 2006). On the contrary, the Old English word *deor* with the general meaning of "animal" became *deer* in present-day English. Semantic changes can be further classified on the basis of the cognitive process that originated them, i.e. either from *metonymy* or *metaphor*. Lastly, it is possible to distinguish among changes due to language-internal or language-external factors (Hollmann, 2009). The latter usually reflects a change in society, as in the case of technological advancements (e.g. *cell*, from the meaning of "prisoner cell" to "cell phone").

The problem of the automatic analysis of lexical semantic change is gaining momentum in the Natural Language Processinng (NLP) and Computational Linguistics (CL) communities, as shown by the growing number of publications on the diachronic analysis of language and the organisation of related events such as the 1st International Workshop on Computational Approaches to Historical Language Change[1] and the project "Towards Computational Lexical Semantic Change Detection"[2]. Following this trend, SemEval 2020 has hosted for the first time a task on automatic recognition of lexical semantic change: the SemEval 2020 Task 1 - Unsupervised Lexical Semantic Change Detection[3] (Schlechtweg et al.,

[1] https://languagechange.org/events/2019-acl-lcworkshop/

[2] https://languagechange.org/

[3] https://competitions.codalab.org/competitions/20948

2020). While this task targets a number of different languages, namely Swedish, Latin, and German, Italian is not present.

Many are the existing approaches, data sets, and evaluation strategies used to detect semantic change, or drift. Most of the approaches rely on diachronic word embeddings, some of these are created as post-processing of static word embeddings, such as Hamilton et al. (2016); while others create dynamic word embeddings where vectors share the same space for all time periods (Del Tredici et al., 2016; Yao et al., 2018; Rudolph and Blei, 2018; Dubossarsky et al., 2019). Recent work exploits word sense induction algorithms to discover semantic shifts (Tahmasebi and Risse, 2017; Hu et al., 2019) by analyzing how induced senses change over time. Finally, Gonen et al. (2020) propose a simple approach based on the neighbors' intersection between two corpora. The neighborhood of a word is separately computed in each corpus, then the intersection is exploited to compute a measure of the semantic shift. The neighborhood in each corpus can be computed using the cosine similarity between word embeddings built on the same corpus without using vectors alignment. A more complete state of the art is described in a critical and concise way in the latest surveys (Tahmasebi et al., 2018; Kutuzov et al., 2018; Tang, 2018).

Almost all of the previously mentioned methods use English as the target language for the diachronic analysis, leaving the other languages still under-explored. To date, only one evaluation has been carried out on Italian using the Kronos-it dataset (Basile et al., 2019).

The DIACR-Ita task at the EVALITA 2020 campaign (Basile et al., 2020b) fosters the implementation of new systems purposely designed for the Italian language. To achieve this goal, a new dataset for the evaluation of lexical semantic change on Italian has been developed based on the "L'Unità" corpus (Basile et al., 2020a). This is the first Italian dataset manually annotated with semantic shifts between two different time periods.

2 Task Description

The goal of DIACR-Ita is to establish if a set of *target* words change their meaning across two time periods, T_1 and T_2, where T_1 precedes T_2.

Following the SemEval 2020 Task 1 settings, we focus on the comparison of two time periods. In this way, we tackle two issues:

1. We reduce the number of time periods for which data has to be annotated;

2. We reduce the task complexity, allowing for the use of different models' architectures, and thus widening the range of potential participants.

During the test phase, participants have been provided with two corpora C_1 and C_2 (for the time periods T_1 and T_2, respectively), and a list of target words. For each target word, systems have to decide whether the word changed or not its meaning between T_1 and T_2, according to its occurrences in sentences in C_1 and C_2. For instance, the meaning of the word "imbarcata" is known to have expanded[4], i.e, it has acquired a new sense, from T_1 to T_2. This will be reflected in different occurrences of the word usage in sentences between C_1 and C_2.

The task is formulated as a closed task, i.e. participants must train their model only on the data provided in the task. However, participants may rely on pre-trained word embeddings, but they cannot train embeddings on additional diachronic Italian corpora, they can use only synchronic corpora.

3 Data

This section provides an overview of the datasets that were made available to the participants in the two different stages of the evaluation challenge, namely **trial** and **test**.

3.1 Trial data

The trial phase corresponds to the evaluation window in which the participants have to build their systems before the official test data are release. The following data were provided:

- An example of 5 trial target words for which predictions are needed;

- An example of gold standard for the trial target words;

- A sample submission file for the trial target words;

[4]The word originally referred to an acrobatic manoeuvre of aeroplanes. Nowadays, it is also used to refer to the state of being deeply in love with someone.

- Two trial corpora that participants could use to develop their models and check the compliance of the generated output to the required format;
- An evaluation and some additional utility scripts for managing corpora.

Trial data do not reflect the actual data from C_1 and C_2. The sample training corpora and target words were artificially built just to provide an example of the data format for developing their systems. Since the training corpus is publicly available on the Internet, we decided not to release these data during the trial phase to prevent participants from identifying the source data and consequently potential set of target words.

3.2 Test data

For the test phase, the following data were provided:

- A diachronic split of the "L'Unità" corpus into the two sub-corpora, C_1 and C_2, each belonging to a specific time period;
- 18 target words, among which 6 were identified as target of semantic meaning change between the two time periods.

Corpus Creation The "L'Unità" diachronic corpus (Basile et al., 2020a) is a collection of documents extracted from the digital archive of the newspaper "L'Unità".[5]

For the task, the corpus has been initially split into two sub-corpora, C_1, corresponding to the time period $T_1 = [1945 - 1970]$, and C_2, corresponding to the time period $T_2 = [1990 - 2014]$.

To facilitate participants in the closed-task formulation, the corpora were provided in a pre-processed format. In particular, we adopted a tab separated format, with one token per line. For each token, we provided its corresponding part-of-speech and lemma. Sentences are separated by empty lines. Data were pre-processed with UDPipe[6] using the ISDT-UD v2.5 model. An example of the data format is illustrated below.

```
Questa  PRON questo
è  AUX essere
una DET uno
frase  NOUN frase
. PUNCT .

Questa  PRON questo
è  AUX essere
un' DET uno
altra  ADJ altro
frase NOUN frase
. PUNCT .
```

Participants are free to combine the available information as they want. Furthermore, to facilitate the generation of word embeddings, we made available a script for generating a format containing one sentence per line.

The whole "L'Unità" diachronic corpus has been built, cleaned and annotated automatically. This process consisted of several steps, namely:

Step 1: Downloading All PDF files are downloaded from the source site and stored into a folder structure that mimics the publication year of each article.

Step 2: Text extraction The text is extracted from the PDF files by using the Apache Tika library.[7] First, the library tries to extract the embedded text if present in the PDF. If this process fails, the internal OCR system is used. It is important to notice that during this step several OCR errors may occur due to different reasons. The processing of the early years of publications, i.e., between 1945–1948, represented a non trivial challenge for the extraction of the textual data. In particular, we noticed that the page format had a major impact on the quality of the OCR. In these period, the newspaper has quite an unconventional format where a few large pages contain many articles scattered into several columns. This affected the performance of the OCR due to its failure in properly identifying the column boundaries.

Step 3: Cleaning In this step, we try to fix some text extraction issues. We identified two lines of actions, the first dealing with paragraph splits and the second with noisy text. In the text extraction process, paragraphs are separated by means of an empty line. However, word hyphenation can trigger errors in the paragraph segmentation phase by wrongly adding empty lines. We addressed this issue by reconstructing the paragraph on a single text line, thus ensuring that empty lines are

[5] https://archivio.unita.news/

[6] http://lindat.mff.cuni.cz/services/udpipe/run.php

[7] https://tika.apache.org/

only used to delimit the actual paragraphs. In our case, noisy text corresponds to tokens whose composing characters are wrongly interpreted by the OCR mixing together alphabetical characters with numbers or symbols. Two heuristics were implemented to limit the amount of noisy text. The first heuristic requires that paragraphs must contain at least five tokens composed by only alphabetical characters. The second heuristic requires that at least 60% of each paragraph must contain words that are attested in a dictionary. For this, we did not use a reference dictionary, but we automatically created it by extracting tokens from the Paisà corpus (Lyding et al., 2014). Numbers were excluded and only alphabetical strings were retained. The output of the cleaning process is a plain text file for each year where each paragraph is separated by an empty line.

Step 4: Processing All plain text files produced by the cleaning step are processed by a Python script that splits each paragraph into sentences and analyses each sentence with UDPipe [8] ISDT-UD v2.5 model. In this way, we obtain tokens, part-of-speech tags, and lemmas. The processed data are then stored in a vertical format as illustrated is Section 3.

After these preparation steps, the valid and retained data for the task span over a temporal period between 1948 and 2014. We revised the initial split of the two sub-corpora as follows: C_1 ranges between $T_1 = [1948 - 1970]$, and C_2 between $T_2 = [1990 - 2014]$. Table 1 illustrates the distributions of the tokens across the two time periods for the sub-corpora. The difference in the number of tokens between C_1 and C_2 reflects differences in the trends in the number of daily published articles, due to cheaper printing costs and the availability of new technologies such as the World Wide Web.

Corpus	Period	#Tokens
L'Unità	1948-1970	52,287,734
L'Unità	1990-2014	196,539,403

Table 1: Official Training Corpora: Occurrence of Tokens.

Creation of the Gold Standard The selection of the target words that compose the Gold Standard data required a manual annotation. Identifying words that have undergone a semantic change is not an easy task. To boost the identification of candidate target words, we adopted a semi-automatic method. In the following paragraphs we illustrate in detail our approach.

Step 1: Selection of candidate words. The initial selection of potential candidate words was based on Kronos-IT (Basile et al., 2019). Kronos-IT is a dataset for the evaluation of semantic change point detection algorithms for the Italian language automatically built by using a web scraping strategy. In particular, it exploits the information presents on the online dictionary "Sabatini Colletti"[9] to create a pool of words that have undergone a semantic change. In the dictionary, some lemmas are tagged with the year of the first attestation of its sense. In some cases, associated with the lemma there are multiple years attesting the introduction of new senses for that word. Kronos-IT uses this information to identify the set of semantic changing words. We retained those words that were predicted to have changed their meaning after 1970, so as to match the temporal periods of the sub-corpora. In this way, we obtained 106 candidate lemmas.

Step 2: Filtering candidate targets. A challenging issue is the attestation of the potential candidate words in both sub-corpora with a relatively high number of occurrences to account for different contexts of use. Frequency, indeed, plays a quite relevant role for the task: infrequent tokens must be discarded because they affect the quality of word representations. The initial list of candidate targets has been further cleaned by removing all tokens that occur less than 20 times in each corpora. Moreover, we conducted a further analysis by manually inspecting some randomly sampled lemma contexts. The aim of this analysis was to remove targets for which the lemmas occurrences are affected by OCR errors. This analysis was performed by the means of the Sketch Engine[10], in particular we analyze concordances of the target word in order to discover OCR errors. One of such words was "toro" derived from the mistaken

[8] http://lindat.mff.cuni.cz/services/udpipe/run.php

[9] https://dizionari.corriere.it/dizionario_italiano/

[10] https://www.sketchengine.eu/

OCR of "loro". At the end of this process, we obtained a list of 27 candidate targets for the annotation.

Step 3: Manual Annotation. For each target, we randomly extracted up to 100 sentences from each of the sub-corpus[11]. Each sentence was then annotated by two annotators: they were asked to assign each occurrence to one of the meaning of the lemma according to those reported in the Sabatini-Coletti dictionary. In case the meaning of the word in a sentence was not present in the list of senses reported in the reference dictionary, the annotators were allowed to add the sense to the word. In total, we annotated 2,336 occurrences of the candidate target words.

Step 4: Annotation check. All cases of disagreement were collectively discussed among all of the annotators to reach a final decision. We observed that some disagreements were also due to a biased interpretation of the context of occurrence by one of the annotators. These cases mainly concerned short ambiguous sentences that prevented a clear identification of the word meaning. As a result of this step, a few candidates were removed from the pool of candidates because occurring in too ambiguous context.

Step 5: Creation of the gold standard. We retained as valid instances of lexical semantic change all those targets that had occurrences of one specific sense only in T_2, and never in T_1. In other words, in the context of this task, a valid lexical semantic change corresponds to the acquisition of a new meaning by a target word. Out of the 23 candidate target words, only 6 of them show a semantic change in T_2. All the other targets did not show a diachronic meaning change. In the final Gold Standard, we kept 12 candidate target words that did not change meaning obtaining a final set of 18 target words.

The Gold Standard contains 18 targets listed as lemmas, one lemma per line, with an accompanying label to mark whether the lemmas has undergone semantic change (label 1) or not (label 0).

[11]This means that in case a target words occurs less than 100 times, all occurrences were annotated.

Participants were given a file containing the 18 target lemmas, one per each line, without annotation. The expected system output is a modification of this file where the participant had to annotate each target lemma with the system prediction (0 or 1).

4 Evaluation

The task is formulated as a binary classification problem. Systems predictions are evaluated against the change labels annotated in the Gold Standard by using accuracy.

The test set (G) contains both positive (P) and negative (N) examples, i.e. $G = P \cup N$. For example:

$$P = \{pilotato, lucciola, ape, rampante\}$$

$$N = \{brama, processare\}$$

Negative words are those that did not undergo a change in their meaning. Systems' predictions involve both positive and negative classified targets $Pr = Pr_{pos} \cup Pr_{neg}$. Then, true positives (positive targets classified as positive) are $TP = P \cap Pr_{pos}$, true negatives (negative targets classified as negative) are $TN = N \cap Pr_{neg}$, false negatives (positive targets classified as negative) are $FN = P \cap Pr_{neg}$ and false positives (negative targets classified as positive) are $FP = N \cap Pr_{pos}$. We can then compute the accuracy as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4.1 Baselines

We provided two baseline models:

- Frequencies: The absolute value of the difference between the word frequencies in the two sub-corpora;
- Collocations: For each word, we build two vector representations consisting of the Bag-of-Collocations related to the two different time periods (T_0 and T_1). Then, we compute the cosine similarity between the two BoCs. It is the same approach evaluated in (Basile et al., 2019).

In both baselines, we use a threshold to predict if the word has changed its meaning. While for the frequencies, a change is detected when the difference is higher than the average. For the collocations a semantic change occurs when the similarity between the two time periods drops under the average plus the variance. Both the average and the variance are computed on the set of target words.

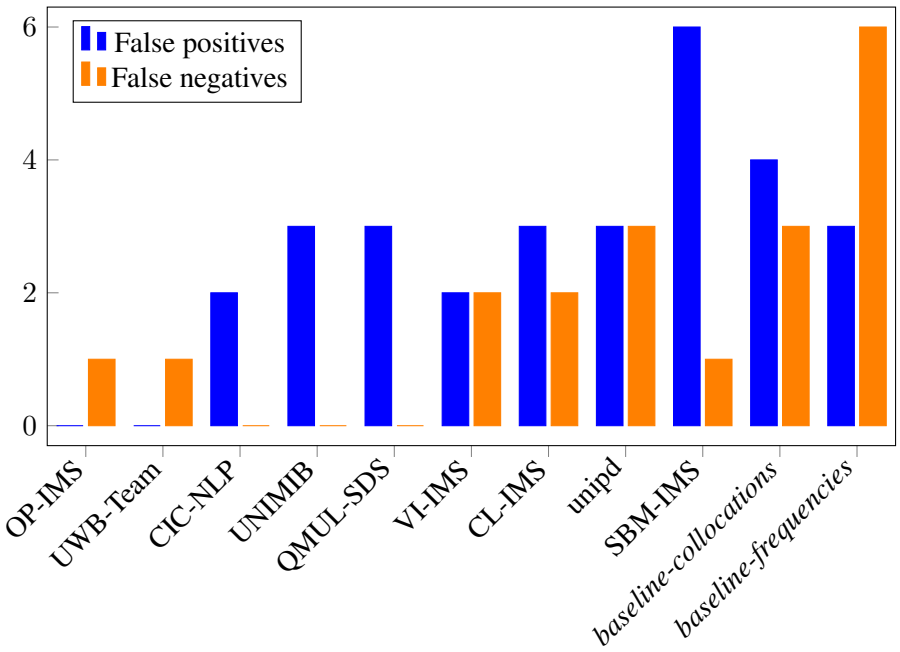

Figure 1: Number of false positives and false negatives for each system.

System	Type
OP-IMS	Post-alignement
UWB Team	Post-alignement
CIC-NLP	PoS tag features
UNIMIB	Jointly alignment
QMUL-SDS	Jointly alignment
VI-IMS	Jointly alignment
CL-IMS	Contextual Embeddings
unipd	Contextual Embeddings
SBM-IMS	Graph

Table 2: Systems types.

5 Systems

21 teams registered to the DIACR-Ita task. However, 9 teams participated in the final task for a total of 36 submitted runs. Based on the algorithms employed, we can group systems into four categories: Post-alignment, Joint Alignment, Contextual Embeddings, Graph-based and PoS tag features (see Table 2). The first two classes are characterised by the type of alignment used. Post-alignment systems first train static word embeddings for each time periods, and then align them. Joint Alignment systems train word embeddings and jointly align vectors across all time slices. Contextual Embeddings systems use contextualized embeddings, such as BERT (Devlin et al., 2019); while Graph-based systems rely on graph algorithms. PoS tag features system rely on the distribution of targets PoS tags across the two time periods. The majority of participating systems use cosine distance as a measure of semantic change, i.e. compute the cosine distance between the vectors of the target lemmas among time periods. Other systems use the Average Pairwise Cosine Distance or the Average Canberra Distance, since the cosine distance does not fit contextual embeddings representations. The last group of systems uses graph-based measures.

We report a short description of each team (best submission) as follows:

OP-IMS (Kaiser et al., 2020) This team uses Skipgram model with Negative sampling (SGNS) to compute word embeddings, the resulting matrices are mean-centred. Word embeddings are aligned using Orthogonal Procrustes. They choose cosine similarity to compare vectors of different word spaces and a threshold based on mean and standard deviation to classify target words.

UWB Team (Pražák et al., 2020) The team maps semantic spaces using linear transformations, such as Canonical Correlation Analysis and Orthogonal Transformation and cosine similarity as a measure to decide if a target word is stable or not. They use a threshold based on mean.

CIC-NLP (Angel et al., 2020) This team analyses the Part-Of-Speech distribution over the

two corpora and create vectors with information about the most common word POS-tags. Then, they obtain a score using pairs of vectors of the two time periods and the sum of Euclidean, Manhattan and cosine distance. They rank targets in discerning order. Finally, they label first upper-third targets as changed words.

UNIMIB (Belotti et al., 2020) The team creates temporal word embeddings using Temporal Word Embeddings with a Compass (TWEC) (Di Carlo et al., 2019). They use the move measure, i.e. a weighted linear combination of the cosine and Local Neighbors, introduced by (Hamilton et al., 2016). They label targets as stable if the move measure is greater than 0.7.

QMUL-SDS (Alkhalifa et al., 2020) The team uses TWEC (Di Carlo et al., 2019) to compute temporal word embeddings with TWEC C-BoW model (Continuous Bag of Words) default settings. They use a cosine similarity as measure of change and a threshold based on mean.

VI-IMS The team uses SGNS to create word embeddings exploiting Vector Initialization (Kim et al., 2014). They use cosine distance as a measure of semantic change and a threshold based on the mean and the standard deviation to classify targets words.

CL-IMS (Laicher et al., 2020) The team creates word vectors using different combinations of the first and last four layers of BERT. They rank targets according to Average Pairwise Cosine Distance, and label the first 7 targets as changed words.

unipd (Benyou et al., 2020) This team uses contextualised word embeddings and an linear combination of distances metrics to measure semantic change, namely Euclidean Distance, Average Canberra distance, Hausdorff distance, as well as Jensen–Shannon divergence between cluster distributions. They rank targets according to the score obtained, and label the first half as changed words.

SBM-IMS The team compute token vectors using BERT. They create a graph where the vertices are the vectors extracted from BERT, while the edges are the cosine distance between word vectors. They cluster the graph with Weighted Stochastic Block Model. Then, they consider the number of incoming edges from the first and second period as a measure of semantic change.

Team	**Accuracy**
OP-IMS	0.944
UWB Team	0.944
CIC-NLP	0.889
UNIMIB	0.833
QMUL-SDS	0.833
VI-IMS	0.778
CL-IMS	0.722
unipd	0.667
SBM-IMS	0.611
baseline-collocations	0.611
baseline-frequencies	0.500

Table 3: Results.

6 Results

Table 3 reports the final results. The best result has been achieved by two systems: *OP-IMS* and *UWB-Team*. Both systems exploit post-alignment strategy. The second system *CIC-NLP* uses an approach based on PoS tag features. QMUL-SDS and VI-IMS are based on joint alignment, while *unipd* and *SBM-IMS* use contextual embeddings. The last system *SBM-IMS* is the only graph-based approach. Moreover, we report both false negative and false positives in Figure 1. Both post-alignment systems share the same unique false negative: the target "tac", while *CIC-NLP* detects two false positives. Joint-alignment systems have a number of false positives higher or at least equal to the number of false negatives. *CL-IMS* and *unipd* produce respectively 2 and 3 false negatives and both misclassify three stable words. The only graph-based approach, *SBM-IMS*, reports the highest number of false positives. In conclusion, the results show that systems based on post/joint alignment and PoS tag features achieve the best performance, while contextual embeddings do not perform as good in this type of task. However all the systems outperform both the baselines.

7 Conclusions

We proposed for the first time the "Diachronic Lexical Semantics" (DIACR-Ita) task. The goal

of the task is to develop systems able to automatically detect if a given word has changed its meaning over time, given contextual information from corpora. We created two corpora for two different time periods T_1 and T_2, and we manually annotated a set of target words that change/do not change meaning across these two periods. This is the first Italian dataset of this type. 9 teams participated in the task for a total of 36 submitted runs. All the systems are able to outperform the two baselines. The results suggests that methods based on post-alignment are the most suitable for this type of task, resulting in better performance even when compared to contextual embedding methods, such as BERT.

References

Rabab Alkhalifa, Adam Tsakalidis, Arkaitz Zubiaga, and Maria Liakata. 2020. QMUL-SDS @ DIACR-Ita: Evaluating Unsupervised Diachronic Lexical Semantics Classification in Italian. 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.

Jason Angel, Carlos A. Rodriguez-Diaz, Alexander Gelbukh, and Sergio Jimenez. 2020. CIC-NLP @ DIACR-Ita: POS and Neighbor Based Models for Lexical Semantic Change in Diachronic Italian Corpora. 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.

Pierpaolo Basile, Giovanni Semeraro, and Annalina Caputo. 2019. Kronos-it: A dataset for the Italian semantic change detection task. In *CEUR Workshop Proceedings*, volume 2481.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Casotti, and Rossella Varvara. 2020a. A Diachronic Italian Corpus based on "L'Unità". In *CEUR Workshop Proceedings*.

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 Belotti, Federico Bianchi, and Matteo Palmonari. 2020. UNIMIB @ DIACR-Ita: Aligning Distributional Embeddings with a Compass for Semantic Change Detection in the Italian Language. 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.

Wang Benyou, Emanuele Di Buccio, and Massimo Melucci. 2020. University of Padova at DIACR-Ita. 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.

Marco Del Tredici, Malvina Nissim, and Andrea Zaninello. 2016. Tracing metaphors in time through self-distance in vector spaces. In *CEUR Workshop Proceedings*. 3rd Italian Conference on Computational Linguistics, CLiC-it 2016 and 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2016 ; Conference date: 05-12-2016 Through 07-12-2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. In *57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470. Association for Computational Linguistics (ACL), sep.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online, July. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, volume 3, pages 1489–1501, may.

Willem Hollmann. 2009. Semantic change. In *English Language: Description, Variation and Context*, pages 301–313. Basingstoke: Palgrave.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View. In *57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908.

Jens Kaiser, 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 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

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. *27th International Conference on Computational Linguistics*.

Severin Laicher, Dominik Schlechtweg, Gioia Baldissin, Enrique Castaneda, and Sabine Schulte Im Walde. 2020. CL-IMS @ DIACR-Ita: Volente o Nolente: BERT does not outperform SGNS on Semantic Change Detection. 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.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The paisa' corpus of italian web texts. In *9th Web as Corpus Workshop (WaC-9)@ EACL 2014*, pages 36–43. EACL (European chapter of the Association for Computational Linguistics).

Ondřej Pražák, Pavel Přibáň, , 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 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Maja Rudolph and David Blei. 2018. Dynamic Embeddings for Language Evolution. In *WWW '18: Proceedings of the 2018 World Wide Web Conference*, pages 1003–1011. Association for Computing Machinery (ACM).

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 and Thomas Risse. 2017. Finding IndividualWord Sense Changes and their Delay in Appearance. In *International Conference Recent Advances in Natural Language Processing*, pages 741–749. Assoc. for Computational Linguistics Bulgaria, nov.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Lexical Semantic Change. *1st International Workshop on Computational Approaches to Historical Language Change 2019*.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676, sep.

Elizabeth Closs Traugott. 2006. Semantic change: Bleaching, strengthening, narrowing, extension. In *Encyclopedia of Language and Linguistics*. Elsevier.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, volume 2018-Febua, pages 673–681.