Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 623–632
Marseille, 20-25 June 2022

On the Impact of Temporal Representations on Metaphor Detection

Giorgio Ottolina, Matteo Palmonari, Manuel Vimercati, Mehwish Alam
University of Milano-Bicocca, Milan, Italy
FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Karlsruhe, Germany
g.ottolina1@campus.unimib.it, {matteo.palmonari, manuel.vimercati}@unimib.it, mehwish.alam@fiz-karlsruhe.de

Abstract

State-of-the-art approaches for metaphor detection compare their literal - or core - meaning and their contextual meaning using metaphor classifiers based on neural networks. However, metaphorical expressions evolve over time due to various reasons, such as cultural and societal impact. Metaphorical expressions are known to co-evolve with language and literal word meanings, and even drive, to some extent, this evolution. This poses the question of whether different, possibly time-specific, representations of literal meanings may impact the metaphor detection task. To the best of our knowledge, this is the first study that examines the metaphor detection task with a detailed exploratory analysis where different temporal and static word embeddings are used to account for different representations of literal meanings. Our experimental analysis is based on three popular benchmarks used for metaphor detection and word embeddings extracted from different corpora and temporally aligned using different state-of-the-art approaches. The results suggest that the usage of different static word embedding methods does impact the metaphor detection task and some temporal word embeddings slightly outperform static methods. However, the results also suggest that temporal word embeddings may provide representations of the core meaning of the metaphor even too close to their contextual meaning, thus confusing the classifier. Overall, the interaction between temporal language evolution and metaphor detection appears tiny in the benchmark datasets used in our experiments. This suggests that future work for the computational analysis of this important linguistic phenomenon should first start by creating a new dataset where this interaction is better represented.

Keywords: Metaphor Detection, Temporal Word Embeddings, Static Word Embeddings

1. Introduction

Accounting for figurative language is one of the key challenges in Natural Language Processing (NLP) (Reforgiato Recupero et al., 2019; Shutova, 2015). Figurative language often contains metaphorical expressions which map one concept from a source domain to another concept in a target domain. For instance, in the sentence "*The wheels of Stalin's regime were well-oiled and already turning*", a political system (target concept) is viewed in terms of a mechanism (source concept) that can function, break, have wheels, etc. This association allows us to transfer knowledge from the domain of *mechanical engineering* to that of *politics*. Therefore, political systems are thought about in terms of mechanisms, leading to multiple metaphorical expressions. The phenomenon of source-target domain mapping was first introduced by George Lakoff known as Conceptual Metaphor Theory (Lakoff and Johnson, 1980). Due to previously defined characteristics, the presence of metaphorical expression in text causes misinterpretation in the algorithms such as machine translation or sentiment analysis (Mohammad et al., 2016).

Recent studies addressing the metaphor detection problem are based on machine learning and exploit word embeddings (Shutova, 2015; Leong et al., 2020), often relying on pre-trained models as linguistic resources. The key intuition is to recognize that words are used in a context that is different from their usual context. In the previous example, the term "*wheels*" is collocated, in the sentence, close to "*Stalin*" and "*regime*", thus defining a context different from the contexts in which it usually appears, i.e., in the domain of mechanical engineering. Most of the recent approaches have therefore combined non-contextual and contextual word embeddings to provide signals for this comparison (Mao et al., 2019; Swarnkar and Singh, 2018). For example, (Mao et al., 2019; Gulordava and Baroni, 2011; Mikolov et al., 2013a) combine non-contextual GloVe embeddings (Pennington et al., 2014) with contextual ELMo embeddings (Peters et al., 2018) within a BiLSTM neural network for sequence labeling. GloVe embeddings account for literal word meanings, while ELMo embeddings account for contextual word meanings.

An important linguistic phenomenon that is not considered in state-of-the-art methods for metaphor detection is language evolution. The trait of the evolution of meaning over time is also shared by metaphorical expressions, which can be due to various reasons such as cultural and societal impact. Metaphorical expressions are known to co-evolve with language and literal word meanings drive this evolution to some extent (Smith and Höfler, 2015; Aitchison, 2010). This leads to the question of whether different, possibly time-specific, representations of literal meanings impact the task of metaphor detection. In conclusion, if metaphor detection approaches tend to compare a sentence-specific and a literal meaning, we must be aware that literal meaning as accounted for in static word embeddings 1) depends on the corpus (the reference linguistic resource) and method used to train the embeddings, and 2) evolves over time.

To this end, this empirical study focuses on analyzing the impact of different word embeddings accounting

for literal word meaning on the task of metaphor detection. Special attention should be paid to possible interactions between metaphor detection and time-specific (non–contextual) word representations used to account for literal meanings at different times. The empirical study discussed in this paper aims to make a first step towards addressing the co-evolution of metaphors and language evolution which is known to be an important factor in language evolution itself (Smith and Höfler, 2015; Aitchison, 2010).

The methodology adopted in this study consists of the following protocol. First, a state-of-the-art Recurrent Neural Networks (RNN)-based model (Gao et al., 2018), which uses static word embeddings to account for literal word meaning, is selected for metaphor detection. This model performs metaphor detection as a sequence classification task where each word occurrence is labeled as either a metaphor usage or a literal usage. Second, three widely used benchmark datasets are selected to evaluate the performance of the models. Third, the RNN-based model is fed with literal meaning vectors obtained from different (non-contextual) word embeddings including temporal word embeddings computed for different decades and aligned with state-of-the-art alignment methods, such as *Procrustes* (Grave et al., 2018) and *Compass* (Bianchi et al., 2020) (first referred to as Temporal Word Embeddings with a Compass (Di Carlo et al., 2019) - TWEC).

The experimental results indicate that different word embeddings impact the metaphor detection task and some temporal word embeddings slightly outperform classic methods on some performance measures. These quantitative results are then explained with the help of a qualitative analysis of the predictions made by the models. An example that illustrates our findings is given in the following figurative sentence coming from a state-of-the-art dataset (see Section 3.3), which has been mistakenly classified as literal by a model exploiting an atemporal embedding, and correctly detected as metaphorical by the same model exploiting a temporal word embedding: *"The virus attacked Argonne National Laboratory outside Chicago starting at 11.54 pm EST Wednesday and throughout the night*". If we investigate the ten nearest neighbors of "*virus*", in a temporal embedding we find words such as "*infection", "respiratory"* and *"organism*", while in the atemporal one we find, for example, "*malware"* and *"spyware*", which diverge from the original literal core meaning and are related to a modern connotation of the word. When exploiting the temporal word embedding, the model is able to understand that the "*virus*" in this sentence is a computer virus, and therefore is used metaphorically along with the verb "*attacked*". However, the analysis provided in this paper also suggests that temporal word embeddings may provide representations of words' core meaning too close to their metaphorical meaning, thus confusing the classifier.

The paper is organized as follows: Section 2 discusses the related work about metaphor detection as well as temporal word embeddings. Section 3 discusses the methodology which has been followed, while Section 4 shows the experimental results of the paper. Finally, Section 5 concludes the paper.

2. State of the Art

This section discusses the state-of-the-art approaches for metaphor detection and the studies related to temporal language evolution.

2.1. Early Approaches for Metaphor Detection

In (Wilks, 2015), the author proposes an approach for metaphor detection based on preferential semantics, which affirms that metaphors are "a violation of semantic constraints put by verbs onto their arguments". In (Fass, 1997), the author proposes an approach for processing metonyms as well as metaphors that take into account the distinction between literalness, metonymy, metaphoricity, and anomaly. This work uses hand-coded patterns for testing sentences containing metonymic relations. The drawback of this approach was that the interpretations were always context-dependent. In (Peters et al., 1998), the authors use WordNet hierarchy to group senses and to find hyponymy relations. If two words are not included in the same synset and/or in hierarchically related synsets, then they are most likely part of a metaphorical phrase. CorMet (Mason, 2004) was the first system to automatically discover source-target domain mappings. A survey on these approaches has been given in (Shutova, 2015).

2.2. Neural Network Based Approaches for Metaphor Detection

Numerous approaches based on BiLSTM take advantage of both contextualized and pre-trained embeddings in the classification layer (Mao et al., 2019; Swarnkar and Singh, 2018). In particular, the *Di-LSTM* Contrast system (Swarnkar and Singh, 2018) encodes the left and right side context of a target word through forward and backward LSTMs. The classification is based on a concatenation of the target word representation and its difference with the encoded context. (Mao et al., 2019) combined GloVe and BiLSTM hidden states for sequence labeling. Some of the most recent systems fine-tune pre-trained contextual language models such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). For example, (Dankers et al., 2020) fine-tuned a BERT model, fed with a discourse fragment as input. Hierarchical attention computes both token and sentence level attention after the encoded layers, leading to better results. A more detailed discussion on methods for metaphor detection is given in this dedicated survey (Rai and Chakraverty, 2020). Another recent approach (Li et al., 2021) uses the hierarchical contextualized representation to extract more information from both sentence-level and discourse-level. For

our study we tested the approaches (Gao et al., 2018) for two main reasons: they explicitly model the interaction between literal and contextual meaning (and thus they support the replacement of embeddings accounting for literal word meaning with different corpus and time-specific embeddings) and they achieved state-of-the-art performance on several metaphor detection datasets when we started our study.

2.3. Temporal Language Evolution

According to (Aitchison, 2010), theories often come as a formalization of metaphors, which "*can populate history with new objects and kinds, and provide both access to interesting new worlds and great field-internal success*". Based on the observations that language is always changing (Beckner et al., 2009; La Mantia et al., 2017; Massip-Bonet and Bastardas Boada, 2013), linguists have formulated different theories and models searching for rules and regularities in semantic change, such as the "*Diachronic Prototype Semantics*" (Geeraerts, 1997; Geeraerts et al., 1999), the "*Invited Inference Theory of Semantic Change*" (Traugott and Dasher, 2001), and "*semantic change based on metaphor and metonymy*" (Heine et al., 1991).

Historically, much of the theoretical work on semantic shifts has been devoted to documenting and categorizing various types of semantic shifts (Breal, 1897; Stern, 1975). Semantic shifts are separated into two important classes: "*linguistic drifts*" (slow and steady changes in core meaning of words) and "*cultural shifts*" (changes in associations of a given word determined by cultural influences). In (Gulordava and Baroni, 2011), the authors showed that distributional models capture cultural shifts, like the word "*sleep*" acquiring more negative connotations related to the sleep disorders domain, when comparing its 1960s contexts with its 1990s contexts. Researchers studying semantic change from a computational point of view have empirically shown the existence of this distinction (Hamilton et al., 2016a).

Diachronic corpora provide empirical resources for semantic change analysis. The availability of large corpora enabled the development of new methodologies for the study of lexical-semantic shifts within general linguistics (Traugott and Dasher, 2001). A key assumption is that changes in a word's collocational patterns reflect changes in word meaning, thus providing a usage-based account of semantic shifts. Semantic changes are often reflected in large corpora that can be sliced into time-specific chunks (e.g., texts coming from a same decade), which account for changes in the contexts of a word that is affected by the shift. Most recent approaches to studying diachronic semantic change are based on temporal word embeddings. These approaches are based on (1) slicing a corpus into time-specific slices (e.g., one slice per decade), and (2) generating slice-specific representations by solving the cross-slice alignment problem (Hamilton et al., 2016a). Other novel approaches include (Giulianelli et al., 2020) those based on contextualized word embeddings and (Tsakalidis and Liakata, 2020) those based on sequential modeling. The current study uses temporal word embeddings for decade-specific slices generated with two alignment methods: HistWords (SGNS) embeddings aligned with the *Procustes* method (Hamilton et al., 2016a; Hamilton et al., 2018) and CADE word embeddings aligned with the *Compass* method (Bianchi et al., 2020; Di Carlo et al., 2019). HistWords embeddings are used in this study because they are used in the key studies about semantic change and are available as pre-trained embeddings; CADE embeddings are used because they achieved state-of-the-art performance on different tasks at the time this study was started (Di Carlo et al., 2019).

3. Methodology

This paper considers the metaphor detection task in its more general settings: the task consists in *detecting all the occurrences of words used metaphorically in an input sentence, independently from their POS tags*. To analyze the effect of different word embeddings (i.e., temporal or static) on the metaphor detection task, the metaphor detection approaches proposed in (Mao et al., 2019) were selected. These approaches use both (non-contextual) word embeddings and contextual word embeddings within a neural network with a final classification layer.

The goal is to train and test different instances of the same architecture with different (non-contextual) word embeddings that account for literal meanings in the metaphor detection algorithms, to verify whether using word representations derived from different linguistic resources leads to different classification outcomes.

3.1. Metaphor Detection Approach

In (Mao et al., 2019), two end-to-end metaphor identification models for detecting metaphors are proposed, both performing better than the previous state-of-the-art baseline (Gao et al., 2018). The two proposed state-of-the-art models are: *(i) Recurrent Neural Network Hidden GloVe (RNN HG)*, based on the interaction between literal and contextual word meanings; *(ii) Multi-Head Context Attention (RNN MHCA)*, based on multi-head context attention. It was observed that the RNN HG and RNN MHCA models achieve comparable results on state-of-the-art metaphor detection datasets. After some preliminary experiments, RNN HG was found to be more suitable to our concerns, since the static embedding is explicitly digested by the network and compared with the contextual embedding.

RNN HG. Figure 1 shows the overall architecture of RNN HG as described in the original study. The RNN HG model can be represented through the following equations:

$$t = f_{BiLSTM}\left([g_t; e_t], \vec{h}_{t-1}, \overleftarrow{h}_{t+1}\right)$$

$$p\left(\hat{y}_t \mid h_t, g_t\right) = \sigma\left(w^\top \left(h_t; g_t\right) + b\right)$$

where: h_t is the hidden state; g_t is the input GloVe (Pennington et al., 2014) literal representation; e_t is the input ELMo (Peters et al., 2018) representation; w is trained parameters; σ is the softmax function; $\hat{y}$ is the the probability of a label prediction for a target word at position t; t is the target word.

In the original architecture/model GloVe embeddings serve as literal (non-contextual) representations of a word (g_t) and are concatenated with the representation from the hidden layer (h_t) of a BiLSTM. These embeddings, located in two different encoding spaces, are concatenated feeding the BiLSTM network and fulfilling the MIP (Group, 1997) requirement. The literal and contextual representations then get compared in the so-called comparison stage. This last step consists of a softmax function σ, which computes the probability of a label prediction $\hat{y}$ for a target word at position *t*, conditioned on both its contextual and literal meaning representations.

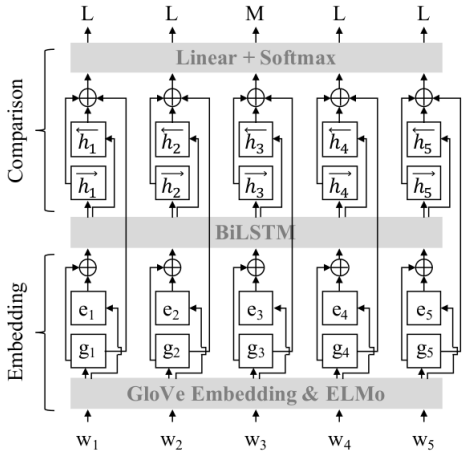

Figure 1: RNN HG model architecture based on MIP procedure.

In RNN HG, both static and contextual representations are used to account for the differences (or the similarities) between them. Also, from the equation it can be easily seen that the vector of the static representation can be modified without touching the model. The different static representations used in this paper are explained in section 3.2.

3.2. Word Embeddings

Word embeddings providing literal word meanings in the RNN HG network used in our experiments are of two main kinds: temporal word embeddings and static word embeddings. Both kinds of embeddings are trained with different approaches and corpora to account for several variables that are expected to have an impact on the final representations (especially: corpus, word embedding algorithm, and alignment method).

3.2.1. Temporal Word Embeddings

HistWords - SGNS[1] provides a set of pre-trained temporal word embeddings generated using the Skip-gram variant of Word2vec (Mikolov et al., 2013a) trained with negative sampling (also referred to as SGNS) on different sliced diachronic corpora (Hamilton et al., 2016b; Hamilton et al., 2018). Decade-specific embeddings obtained from the same corpus are aligned using the Procrustes method, one of the most used in the literature. It solves the task of aligning two sets of points in high dimensions (which has many applications in NLP), through the joint estimation of an orthogonal matrix and a permutation matrix (Grave et al., 2018). A stochastic algorithm is proposed to minimize the cost function on large-scale problems. In our study we consider HistWords embeddings obtained from different corpora. **CoHa Word SGNS** (1900-2010) is a set of eleven decade-specific models covering the time span 1900-2010 (with "1900" we refer to the "1900-1910" slice); they are trained on a genre-balanced subset of the Corpus of Historical American English (CoHa) (Davies, 2015), the largest structured corpus of historical English, which contains more than 400 million words and text published between 1820 and 2000s. **CoHa Lemma SGNS** (1900-2010) is a set of eleven decade-specific models trained on CoHa after applying lemmatization. **NGrams English All** and **NGrams English Fiction** (1900-2000) are two sets of ten decade-specific models each trained on a subset of Google N-Grams that considers, respectively, all genres or fiction only. Observe that we consider a total of 42 models (each one containing decade-specific word embeddings) based on Procrustes alignment. In the experiments, we may filter out some models that do not achieve the best performance for space limitations.

CADE - Compass Aligned Embeddings are temporal word embeddings trained with Word2vec (Mikolov et al., 2013b) and aligned with the Compass method (Di Carlo et al., 2019; Bianchi et al., 2020), which can be summarized as follows. Word2vec is trained over the entire corpus (all the slices). One of the two-weight matrices obtained after this step is used as a *compass* when training Word2vec again on each slice: the compass matrix is frozen, while the other matrix is initialized and trained again over each slice, thus obtaining slice-specific word embeddings (the word embeddings referring to the slice period). We use CADE with the CBOW architecture as in the original paper (Di Carlo et al., 2019), thus using the target matrix as a compass and the weights in the context matrix as final embeddings. CADE embeddings used in the study are trained using the code and implementation details available online[2]. To obtain embeddings as comparable as possible to CoHa Word SGNS, we trained CADE embeddings using the CoHa corpus.[3] **CoHa Word**

[1] https://stanford.io/3txN0Hd

[2] https://github.com/vinid/cade

[3] Unfortunately, the genre-balanced subset of CoHA used

CBOW (1900-2010) is the set of eleven slice-specific models trained with this approach.

3.2.2. Static Word Embeddings

Three static word embeddings obtained from as many corpora are also considered to account for the impact of the corpus and embedding algorithm on metaphor detection.

Common Crawl GloVe is the model that is based on the embeddings used in the original RNN HG network. These embeddings are trained using the Common Crawl [4] corpus, which is expected to contain relatively recent content extracted from the web. **Wikipedia CBOW** consists of the embeddings trained over the English Wikipedia using Word2vec with the CBOW architecture. It accounts for a relatively recent text covering encyclopedic knowledge. **Full CoHa CBOW** is derived from the embeddings trained with CoHa using Word2vec with the CBOW architecture. It consists of the embeddings (i.e., the context matrix) obtained after the first pass of the CADE approach over the CoHa corpus. It supports the comparison between static and temporal word embeddings trained with a common corpus and algorithm.

3.3. Metaphor Detection Datasets

Three datasets are used to show the feasibility of the proposed claims. Table 1 shows the main characteristics of the datasets.

MOH-X (Mohammad et al., 2016) is derived from the subset of MOH dataset used by (Shutova et al., 2016). Mohammad et al. annotated different senses of WordNet verbs for metaphoricity. They extracted verbs that had between three and ten senses in WordNet along with their glosses. The verbs were annotated for metaphoricity with the help of crowd-sourcing. Ten annotators were recruited for each sentence and only those verbs were selected that were annotated positive for metaphoricity by at least 70% of the annotators. The final dataset consisted of 647 verb-noun pairs, 316 metaphorical, and 331 literal.

VUA consists of 117 fragments sampled across four genres from the British National Corpus, i.e., Academic, News, Conversation, and Fiction (Leong et al., 2018). The data was annotated using the MIP-VU procedure (Steen et al., 2010) based on the MIP procedure (Group, 1997). The tagset is rich and hierarchically organized, detecting various types of metaphors, words that flag the pre-sense of metaphors, etc. The majority of sentences in this dataset have the timestamp for the decade 1985-1994.

TroFi contains feature lists consisting of the stemmed nouns and verbs in a sentence, with target or seed words. After a first collection phase, the final TroFi dataset is obtained by filtering out some "frequent words" (common words in the British National Corpus along with contractions, single letters, and numbers from 0 to 10). The target set is built using the '88-'89 Wall Street Journal Corpus (WSJ) tagged using the (Ratnaparkhi, 1996) tagger and the (Joshi, 1999) SuperTagger. 10-fold cross-validation was adopted on MOH-X and TroFi datasets because of their small sizes (k was set equal to 10). More details regarding the models' hyperparameters can be found in the GitHub repository of (Mao et al., 2019)'s work. [5]

4. Experiments

4.1. Experimental Design

All data and source code related to our experiments are publicly released.[6] A first part of the experimentation consists in evaluating the performance of the RNN HG classifier (evaluated using well-known Precision, Recall, F1-Score and Accuracy measures) when different word embeddings are used instead of GloVe embeddings. In particular, we address the following research questions:

- **RQ1:** *Can the results of the state-of-the-art metaphor detection algorithms be improved by using different word embeddings, especially temporal embeddings such as HistWords - SGNS?*
- **RQ2:** *Are there any observable patterns that can lead to the assumption that metaphor detection tasks performed with temporal embeddings and representations impact datasets with known temporal connotations more than others?*
- **RQ3:** *Are the representations obtained through Compass alignment more effective for metaphor detection than the embeddings aligned with traditional methods (e.g.: Procrustes)?*
- **RQ4:** *Are specific word embeddings' architectures more effective for metaphor detection tasks than others?*

The first experiments have been performed using Histwords as new static embeddings. The datasets discussed in Section 3.3 were used to train a RNN HG model and evaluate it. We are interested in evaluating the effectiveness of the new corpus, so we compared the result obtained with HistWords with results obtained using the atemporal Word2vec embeddings trained on the entire Wikipedia corpus. Finally, experiments have been performed using word embeddings obtained by aligning all different decade slices of the CoHa corpus (ranging from 1820 to 2000) with *Compass* alignment method. Slices of the CoHa corpus for

in HistWords could not be retrieved to train the embeddings on the very same data. Also, the CBOW architecture has been used because it is reported to generate temporal word embeddings of better quality with the compass (Di Carlo et al., 2019)

[4] https://commoncrawl.org/

[5] https://bit.ly/324ZcUI

[6] https://bit.ly/3qS7NCu

Table 1: Dataset Characteristics

Dataset	#sentences	Train/Test Splits	Temporal Annotation?	Creation Detail
MOH-X	646	No	No	Derived from MOH dataset. The verbs are used as metaphors.
VUAsequence	5323	Yes	Yes (1985-1994)	117 fragments sampled across 4 genres from British National Corpus (academic, news, conversation, and fiction)
TroFi	3737	No	Yes (1987-1989)	The sentences (each one with a single annotated target verb) are taken from '87-'89 Wall Street Journal Corpus.

each decade needed to be aligned with Compass in order to perform equivalent experiments to the ones with other embeddings. The first step consisted in concatenating all CoHa text slices, and obtaining a final corpus for all the decades. The pre-processing steps included stripping HTML tags, removing text between square brackets and stop words, and replacing all contractions. The following steps were carried out to perform the alignment using CADE and Compass: (i) Creating the main Compass file by concatenating all the processed CoHa decade slices; (ii) Training the obtained compass-aligned embeddings; (iii) Training all the different slices from the compass obtaining their respective models; (iv) Converting the CoHa compass models in Word2Vec format, so that they could have the same architecture of the HistWords - SGNS and Wikipedia embeddings, and be exploited inside the modified Recurrent Neural Network model. Only the aligned models of the decade slices ranging from 1900 to 2000 are kept so that the results could be comparable to the previous ones. A *Full CoHa CBOW* (CADE) model was also obtained by training the compass on all the aforementioned decade slices which were used for qualitative analyses.

Qualitative Analysis. To get more insights into our results, we also investigate the characteristics of the words that are correctly or mistakenly identified as metaphors, by 1) controlling for linguistic features such as topic and genre, and 2) checking the nearest neighbors of target words in the embeddings used to account for their literal meaning (which account for a more in-depth characterization of word meaning). Due to the large number of experiments performed in this study, we could not inspect all the results. We therefore focused on MOH-X, VUA, and TroFi datasets' predictions obtained with four embeddings: **(i)** Full CoHa CBOW (CADE); **(ii)** GloVe (state-of-the-art-representation); **(iii)** CoHa Word CBOW 1990 Decade Slice; **(iv)** CoHa Word SGNS 1990 Decade Slice. Embeddings based on the 1990 decade slice were chosen because of VUA and TroFi datasets' sentences temporal connotations (see Table 1) and the good performance achieved with these models. Therefore, the four selected embeddings allowed us to look at predictions made by the RNN HG model with both temporal and atemporal representations and with different embedding and alignment algorithms. In order to check all the predictions made by our model for MOH-X and TroFi, we combined each one of their 10-folds intermediate results, since these two datasets are not split into the train, validation, and test sets like VUA. For each one of the three state-of-the-art datasets, the analysis considered correctly identified metaphors and mistakenly identified metaphors. For MOH-X and TroFi only verbs are considered.

4.2. Quantitative Results

Tables 2, 3, and 4 provide the overall quantitative performances and scores for each dataset (the best results are highlighted in bold)[7]. Combining the observations gathered from all the performed experiments, the following conclusions can be drawn:

1. Word2vec architecture (HistWords - SGNS, CADE, and Wikipedia embeddings) works better than GloVe architecture for metaphor detection;
2. HistWords - SGNS temporal embeddings perform better on the datasets with known temporal connotations (TroFi and VUA) compared to the atemporal Wikipedia embeddings;
3. Results on MOH-X are generally good, but they do not show a clear pattern. VUA and TroFi do not show clear patterns either;
4. Procrustes alignment (HistWords - SGNS) and CADE - Compass alignment methods (CoHa corpus) lead to similar performances and results. Although, while the latter performs better on the TroFi dataset (data extracted from Wall Street Journal corpus), the first one impacts slightly more than the VUA dataset (data extracted from the British National Corpus).

4.3. Qualitative Results

This analysis confirmed several expected patterns and revealed some new ones. Among the confirmed patterns, we found that topics related to *economics*, *politics*, and *emotions* are the most recurring ones in sentences containing correctly identified metaphors. Verbs

[7]We only reported a few slices (the most representative ones) for some corpora listed in the tables, such as *Lemma, English All* and *English Fiction*, due to lack of space.

			Metrics and Scores			
Main Corpus	Alignment	Slice	Precision	Recall	F1 Score	Accuracy
Common Crawl GloVe	NA	All	0.77	0.81	0.78	0.79
Wikipedia CBOW	NA	All	**0.81**	0.79	0.80	**0.81**
CoHa Word SGNS	Procrustes	1900	0.79	0.81	0.80	0.80
CoHa Word SGNS	Procrustes	1910	0.77	0.83	0.80	0.79
CoHa Word SGNS	Procrustes	1920	0.79	0.80	0.79	0.80
CoHa Word SGNS	Procrustes	1930	0.78	0.81	0.79	0.79
CoHa Word SGNS	Procrustes	1940	0.79	0.81	0.80	0.80
CoHa Word SGNS	Procrustes	1950	**0.81**	0.80	0.80	**0.81**
CoHa Word SGNS	Procrustes	1960	0.79	0.80	0.80	0.80
CoHa Word SGNS	Procrustes	1970	0.80	0.80	0.80	0.80
CoHa Word SGNS	Procrustes	1980	0.78	0.81	0.79	0.80
CoHa Word SGNS	Procrustes	1990	0.78	0.80	0.79	0.79
CoHa Word SGNS	Procrustes	2000	0.80	0.82	**0.81**	**0.81**
CoHa Lemma SGNS	Procrustes	1900	0.79	0.81	0.80	0.80
CoHa Lemma SGNS	Procrustes	1950	0.77	0.81	0.79	0.79
CoHa Lemma SGNS	Procrustes	1990	0.76	0.83	0.79	0.79
NGrams English All	Procrustes	1900	0.78	0.81	0.79	0.80
NGrams English All	Procrustes	1950	0.80	0.78	0.79	0.80
NGrams English All	Procrustes	1990	0.76	**0.84**	0.80	0.79
NGrams English Fiction	Procrustes	1900	0.77	0.82	0.79	0.79
NGrams English Fiction	Procrustes	1950	0.77	0.82	0.79	0.79
NGrams English Fiction	Procrustes	1990	0.80	0.81	0.80	**0.81**
Full CoHa CBOW	Compass	All	**0.81**	0.80	0.79	0.78
CoHa Word CBOW	Compass	1900	0.77	0.81	0.79	0.79
CoHa Word CBOW	Compass	1910	0.80	0.78	0.78	0.79
CoHa Word CBOW	Compass	1920	0.78	0.79	0.78	0.79
CoHa Word CBOW	Compass	1930	0.78	0.81	0.80	0.80
CoHa Word CBOW	Compass	1940	**0.81**	0.78	0.79	0.80
CoHa Word CBOW	Compass	1950	0.77	0.80	0.78	0.78
CoHa Word CBOW	Compass	1960	0.79	0.80	0.79	0.80
CoHa Word CBOW	Compass	1970	0.78	0.81	0.79	0.79
CoHa Word CBOW	Compass	1980	0.79	0.81	0.80	0.80
CoHa Word CBOW	Compass	1990	0.80	0.79	0.79	0.80
CoHa Word CBOW	Compass	2000	0.79	0.79	0.78	0.79

Figure 2: Results related to MOH-X dataset, with every single embedding.

			Metrics and Scores			
Main Corpus	Alignment	Slice	Precision	Recall	F1 Score	Accuracy
Common Crawl GloVe	NA	All	0.72	0.76	**0.74**	0.93
Wikipedia CBOW	NA	All	0.75	0.69	0.72	0.93
CoHa Word SGNS	Procrustes	1900	0.76	0.72	**0.74**	**0.94**
CoHa Word SGNS	Procrustes	1950	0.76	0.71	0.73	**0.94**
CoHa Word SGNS	Procrustes	1990	0.76	0.71	0.73	**0.94**
CoHa Lemma SGNS	Procrustes	1900	**0.77**	0.70	0.73	**0.94**
CoHa Lemma SGNS	Procrustes	1910	0.76	0.71	0.73	**0.94**
CoHa Lemma SGNS	Procrustes	1920	0.76	0.70	0.73	**0.94**
CoHa Lemma SGNS	Procrustes	1930	**0.77**	0.70	0.73	**0.94**
CoHa Lemma SGNS	Procrustes	1940	**0.77**	0.68	0.72	**0.94**
CoHa Lemma SGNS	Procrustes	1950	**0.77**	0.69	0.73	**0.94**
CoHa Lemma SGNS	Procrustes	1960	0.75	0.73	**0.74**	**0.94**
CoHa Lemma SGNS	Procrustes	1970	0.76	0.70	0.73	0.93
CoHa Lemma SGNS	Procrustes	1980	0.76	0.71	0.73	**0.94**
CoHa Lemma SGNS	Procrustes	1990	**0.77**	0.71	**0.74**	**0.94**
CoHa Lemma SGNS	Procrustes	2000	0.76	0.70	0.73	**0.94**
NGrams English All	Procrustes	1900	**0.77**	0.69	0.73	**0.94**
NGrams English All	Procrustes	1950	0.74	0.74	**0.74**	**0.94**
NGrams English All	Procrustes	1990	**0.77**	0.70	0.73	**0.94**
NGrams English Fiction	Procrustes	1900	0.75	0.71	0.73	**0.94**
NGrams English Fiction	Procrustes	1950	0.75	0.73	**0.74**	**0.94**
NGrams English Fiction	Procrustes	1990	0.76	0.70	0.73	**0.94**
Full CoHa CBOW	Compass	All	0.70	**0.80**	0.74	**0.94**
CoHa Word CBOW	Compass	1900	0.73	0.67	0.70	0.93
CoHa Word CBOW	Compass	1910	0.70	0.69	0.69	0.92
CoHa Word CBOW	Compass	1920	0.72	0.69	0.70	0.93
CoHa Word CBOW	Compass	1930	0.72	0.68	0.70	0.93
CoHa Word CBOW	Compass	1940	0.51	**0.80**	0.63	0.88
CoHa Word CBOW	Compass	1950	0.74	0.67	0.70	0.93
CoHa Word CBOW	Compass	1960	0.73	0.67	0.70	0.93
CoHa Word CBOW	Compass	1970	0.72	0.68	0.70	0.93
CoHa Word CBOW	Compass	1980	0.72	0.66	0.69	0.93
CoHa Word CBOW	Compass	1990	0.68	0.76	0.72	0.91
CoHa Word CBOW	Compass	2000	0.69	0.72	0.70	0.92

Figure 3: Results related to VUA dataset, with every single embedding.

having a literal meaning characterized by physical connotations often assume metaphorical/figurative meanings when used in sentences related to the contexts listed before. This suggests that embeddings derived from these corpora and slices maintain as the core meaning the one related to the physical connotation. This pattern is at first observed especially in TroFi predictions, but with the help of the nearest neighbors analysis, the same pattern is detected even in the other datasets.

The nearest neighbor analysis of the target metaphorical words in the sentences extracted from state-of-the-art datasets leads to comparing meanings of target words in embeddings generated from different resources, especially, in non-temporal vs temporal word embeddings. When exploiting the temporal word embedding, the model could correctly understand that the words (in our examples: "*apple, virus, attack, hearts* and *glow*") were used in a figurative way, thus correctly classifying them as metaphors. Furthermore, the entire

			Metrics and Scores			
Main Corpus	Alignment	Slice	Precision	Recall	F1 Score	Accuracy
Common Crawl GloVe	NA	All	0.68	0.76	0.71	0.74
Wikipedia CBOW	NA	All	0.70	0.71	0.71	0.74
CoHa Word SGNS	Procrustes	1900	0.69	0.73	0.71	0.74
CoHa Word SGNS	Procrustes	1950	0.69	0.74	0.71	0.74
CoHa Word SGNS	Procrustes	1990	0.70	0.72	0.71	0.74
CoHa Lemma SGNS	Procrustes	1900	0.69	0.73	0.71	0.74
CoHa Lemma SGNS	Procrustes	1950	0.69	0.73	0.71	0.74
CoHa Lemma SGNS	Procrustes	1990	0.70	0.72	0.71	0.74
NGrams English All	Procrustes	1900	0.71	0.71	0.71	**0.75**
NGrams English All	Procrustes	1910	**0.72**	0.70	0.71	**0.75**
NGrams English All	Procrustes	1920	0.70	0.72	0.71	0.74
NGrams English All	Procrustes	1930	0.70	0.71	0.71	0.74
NGrams English All	Procrustes	1940	0.71	0.71	0.71	**0.75**
NGrams English All	Procrustes	1950	0.68	0.75	0.71	0.74
NGrams English All	Procrustes	1960	0.69	0.73	0.71	0.74
NGrams English All	Procrustes	1970	0.70	0.72	0.71	0.74
NGrams English All	Procrustes	1980	0.71	0.72	0.71	0.74
NGrams English All	Procrustes	1990	0.70	0.73	0.71	0.74
NGrams English Fiction	Procrustes	1900	0.69	0.73	0.71	0.74
NGrams English Fiction	Procrustes	1950	0.68	0.75	0.71	0.73
NGrams English Fiction	Procrustes	1990	0.70	0.73	0.71	0.74
Full CoHa CBOW	Compass	All	**0.72**	0.76	**0.74**	0.74
CoHa Word CBOW	Compass	1900	0.69	0.77	0.72	**0.75**
CoHa Word CBOW	Compass	1910	0.69	0.76	0.72	0.74
CoHa Word CBOW	Compass	1920	0.70	0.75	0.72	**0.75**
CoHa Word CBOW	Compass	1930	0.68	0.77	0.72	0.74
CoHa Word CBOW	Compass	1940	0.69	0.76	0.72	0.74
CoHa Word CBOW	Compass	1950	0.68	0.77	0.72	0.74
CoHa Word CBOW	Compass	1960	0.68	0.78	0.72	0.74
CoHa Word CBOW	Compass	1970	0.68	0.77	0.72	0.74
CoHa Word CBOW	Compass	1980	0.68	0.78	0.73	**0.75**
CoHa Word CBOW	Compass	1990	0.70	**0.80**	0.73	0.74
CoHa Word CBOW	Compass	2000	0.69	0.76	0.72	**0.75**

Figure 4: Results related to TroFi dataset, with every single embedding.

sentences were also correctly classified as metaphorical, since the core meanings of the word were closer to their literal core meaning. This also explains the fluctuations across slices, because corpora are never fully representative, and some contexts may be represented more than others in one specific decade.

Different results related to the domains of the sentences have been observed in the VUA dataset's predictions. With *Full CoHa CADE* embedding, only one sentence belonging to the *academic* genre was correctly classified, whereas as far as *CoHa SGNS 1990 slice* is concerned, no sentences belonging to the *news* genre are correctly predicted. The latter result could indicate that for that specific time period, SGNS words' representations of the *news* genre are biased towards their metaphorical meaning (words are used in metaphorical contexts much more than in literal ones). This would prevent the proposed models from correctly identifying the words as metaphors.

5. Conclusions

This study can be considered a first attempt to investigate the interaction between metaphorical word usage and semantic change using computational metaphor detection methods and corpus-specific word embeddings, including temporal word embeddings.

The results suggest that temporal word embeddings can improve the performance of the task of metaphor detection, even though their overall impact on three benchmark datasets is rather limited. However, independently from the absolute performance on the considered datasets, the interaction between the specificity of the embeddings (especially their temporal specificity) and metaphor detection is found in the experiments conducted in this study. In fact, these experiments verify that if the core meaning of the words of interest in a sentence is too similar to their figurative meaning in the word embedding, a metaphorical sentence could get misclassified as literal. Moreover, when temporal word embeddings provide the representations of the words that are more inclined towards their literal core meaning, exploited models end up correctly identifying metaphors more easily. Word embeddings belonging to some language domains in specific time periods can be biased towards their metaphorical meaning, leading to words being used in metaphorical contexts much more than in literal ones. This would prevent neural models from correctly identifying the words as metaphors. To improve the experimental framework, both temporal and atemporal representations could be built on the same corpora with temporal slices. Furthermore, when building atemporal embeddings, the corpora could be subsampled to obtain a comparable size.

Future work may stem from our last exploratory analysis. Searching for words known to undergo semantic change across time, we retrieved a suitable list from the *SemEval 2020 Task 1: Unsupervised Lexical Semantic Change Detection Competition*[8]. We searched for all the occurrences of these words in the three datasets used in our study and in the competition data, to classify the word usage as metaphorical or not, but we could not find enough metaphorical statements. This suggests that more work is needed to collect more data better accounting for the interaction between metaphorical word usage and semantic change along time, a phenomenon that is advocated by many scholars as a very important driver of language evolution.

[8] https://bit.ly/3Gz9vPU

6. References

Aitchison, J. (2010). *Motives for Language Change*, volume abs/10.1111.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., et al. (2009). Language is a complex adaptive system: Position paper. *Language learning*, 59:1–26.

Bianchi, F., Di Carlo, V., Nicoli, P., and Palmonari, M. (2020). Compass-aligned distributional embeddings for studying semantic differences across corpora.

Breal, M. (1897). *Essai de semantique (Science des significations)*. Hachette Paris.

Dankers, V., Malhotra, K., Kudva, G., Medentsiy, V., and Shutova, E. (2020). Being neighbourly: Neural metaphor identification in discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 227–234, Online, July. Association for Computational Linguistics.

Davies, M. (2015). Corpus of Historical American English (COHA).

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Di Carlo, V., Bianchi, F., and Palmonari, M. (2019). Training temporal word embeddings with a compass. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6326–6334, Jul.

Fass, D. (1997). Processing metonymy and metaphor. *Contemporary Studies in Cognitive Science & Technology*.

Gao, G., Choi, E., Choi, Y., and Zettlemoyer, L. (2018). Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Geeraerts, D., Grondelaers, S., and Speelman, D. (1999). *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kledingen voetbaltermen.*

Geeraerts, D. (1997). *Diachronic Prototype Semantics: A Contribution to Historical Lexicology*. Oxford University Press.

Giulianelli, M., Del Tredici, M., and Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online, July. Association for Computational Linguistics.

Grave, E., Joulin, A., and Berthet, Q. (2018). Unsupervised alignment of embeddings with wasserstein procrustes. May.

Group, P. (1997). Mip: A method for identifying metaphorically used words in discourse.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (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. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016b). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proc. Assoc. Comput. Ling. (ACL)*.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2018). Diachronic word embeddings reveal statistical laws of semantic change.

Heine, B., Claudi, U., and HÃ¼nnemeyer, F. (1991). Grammaticalization: A conceptual framework. *Bibliovault OAI Repository, the University of Chicago Press*.

Joshi, A. K. (1999). Supertagging: An approach to almost parsing.

La Mantia, F., Licata, I., and Perconti, P. (2017). Language in complexity. *Lecture Notes in Morphogenesis*.

Lakoff, G. and Johnson, M. (1980). *LaTeX: Metaphors we live by.* Online.

Leong, C. W. B., Beigman Klebanov, B., and Shutova, E. (2018). A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.

Leong, C. W. B., Beigman Klebanov, B., Hamill, C., Stemle, E., Ubale, R., and Chen, X. (2020). A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Li, S., Yang, L., He, W., Zhang, S., Zeng, J., and Lin, H. (2021). Label-enhanced hierarchical contextualized representation for sequential metaphor identification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3533–3543, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Mao, R., Lin, C., and Guerin, F. (2019). End-to-end

sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Mason, Z. J. (2004). CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System. *Computational Linguistics*, 30(1):23–44, 03.

Massip-Bonet, and Bastardas Boada, A. (2013). *Complexity perspectives on language, communication and society*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, Workshop Track Proceedings.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems*, 26:3111–3119.

Mohammad, S., Shutova, E., and Turney, P. (2016). Metaphor as a medium for emotion: An empirical study. *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, page 23–33, August.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peters, W., Peters, I., and Vossen, P. (1998). Automatic sense clustering in eurowordnet. *Artificial Intelligence*.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: human language technologies, Association for Computational Linguistics, New Orleans, Louisiana*, 1 (Long Papers):2227—-2237.

Rai, S. and Chakraverty, S. (2020). A survey on computational metaphor processing. *ACM Comput. Surv.*, 53(2):24:1–24:37.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. *Conference on Empirical Methods in Natural Language Processing*.

Reforgiato Recupero, D., Alam, M., Buscaldi, D., Grezka, A., and Tavazoee, F. (2019). Frame-based detection of figurative language in tweets [application notes]. *IEEE Comput. Intell. Mag.*, 14(4):77–88.

Shutova, E., Kiela, D., and Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. *Proceedings of NAACL-HLT 2016, pages 160–170, San Diego, California, June 12-17, 2016. c 2016 Association for Computational Linguistics*, pages 160–170, June.

Shutova, E. (2015). Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41, December.

Smith, A. and Höfler, S., (2015). *The pivotal role of metaphor in the evolution of human language*, pages 123–.

Steen, G., Dorst, L., Kaal, A., and Herrmann, J. B. (2010). A method for linguistic metaphor identification: From mip to mipvu.

Stern, G. (1975). *Meaning and Change of Meaning: With Special Reference to the English Language*. Greenwood Press.

Swarnkar, K. and Singh, A. K. (2018). Di-LSTM contrast : A deep neural network for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 115–120, New Orleans, Louisiana. Association for Computational Linguistics.

Traugott, E. C. and Dasher, R. B. (2001). *Regularity in Semantic Change*. Cambridge Studies in Linguistics. Cambridge University Press.

Tsakalidis, A. and Liakata, M. (2020). Sequential modelling of the evolution of word representations for semantic change detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8485–8497, Online, November. Association for Computational Linguistics.

Wilks, Y. (2015). Making preferences more active. *Artificial Intelligence*, 11(3):197–223, December.