Learning Diachronic Analogies to Analyze Concept Change

Matthias Orlikowski
Digital Humanities
Paderborn University
morlikow@mail.upb.de

Matthias Hartung
CITEC
Bielefeld University
mhartung@cit-ec.uni-bielefeld.de

Philipp Cimiano
CITEC
Bielefeld University
cimiano@cit-ec.uni-bielefeld.de

Abstract

We propose to study the evolution of concepts by learning to complete diachronic analogies between lists of terms which relate to the same concept at different points in time. We present a number of models based on operations on word embedddings that correspond to different assumptions about the characteristics of diachronic analogies and change in concept vocabularies. These are tested in a quantitative evaluation for nine different concepts on a corpus of Dutch newspapers from the 1950s and 1980s. We show that a model which treats the concept terms as analogous and learns weights to compensate for diachronic changes (weighted linear combination) is able to more accurately predict the missing term than a learned transformation and two baselines for most of the evaluated concepts. We also find that all models tend to be coherent in relation to the represented concept, but less discriminative in regard to other concepts. Additionally, we evaluate the effect of aligning the time-specific embedding spaces using orthogonal Procrustes, finding varying effects on performance, depending on the model, concept and evaluation metric. For the weighted linear combination, however, results improve with alignment in a majority of cases. All related code is released publicly.

1 Introduction

Research on the evolution of concepts is a long-standing topic within philosophy, history and linguistics. However, recent work on the computational analysis of semantic change based on word embeddings has surprisingly little to offer in this regard. Most work focuses on the meaning of individual words. In comparison, there are only few contributions which analyze concepts and the changing vocabularies which are used to express them (Kenter et al., 2015; Recchia et al., 2016).

The goal of this paper is to provide insights into how distributional semantics (Harris, 1954; Firth, 1957), in particular word embeddings (Mikolov et al., 2013a; Levy et al., 2015), can be used to analyze concept change. We propose to model concept change in terms of analogies between concept vocabularies at different points in time. This extends well-established synchronic models of analogy based on word embeddings (Mikolov et al., 2013b) to the diachronic case. We build on the underlying *parallelogram model* of analogy (Rumelhart and Abrahamson (1973), cf. Chen et al. (2017)), assuming that analogies of the type of "*a* is to *b* as *c* is to *d*" can be described by linear relationships between distributional representations of the four words. While parallelogram relationships can be found in other vector representations as well (Levy and Goldberg, 2014), embeddings derived with skip-gram can be considered a robust baseline for analogy tasks (Levy et al., 2015).

We detail our approach (Section 3) and propose a number of simple models to learn diachronic analogies (Section 4) which are evaluated quantitatively (Section 6) on a corpus of historical Dutch newspapers (Section 5). We report on two related experiments which are motivated by the intuition that diachronic analogies should be *coherent* in regard to the represented concept and *discriminative* in regard to the vocabulary of other concepts. All related code is released publicly[1].

[1] https://gitlab.com/morlikowski/diachronic-analogies-code

2 Related work: Distributional approaches to semantic change

Word-level semantic change. There is a great body of work which uses word vector representations to study changes in word semantics. Examples include Gulordava and Baroni (2011) or Radinsky et al. (2011), while Kim et al. (2014) are among the first to use neural word embeddings to analyze changes in word meanings. Kulkarni et al. (2015) use a similar approach, but automatically identify points of meaning changes based on shifts in the mean of the respective time series. Hamilton et al. (2016) and Dubossarsky et al. (2017) present methods to more precisely quantify trends in changes of word meaning and address a central problem in using word embeddings for diachronic analyses: They align the axes of the vector spaces from neighboring time periods using a mapping derived with orthogonal Procrustes (Hamilton et al., 2016). This method will be presented in more detail in Section 4.3.

Concept-level semantic change. There are a number of approaches to studying lexical semantic change above the level of individual words. These include simple co-occurrence statistics or topic modeling (Blei and Lafferty, 2006) which are only loosely related to our work based on word embeddings. For example, Tan et al. (2017) present a topic modeling approach to study relations between ideas that helps to detect gradual substitution and prevalence, or mutual fostering and coexistence.

Kenter et al. (2015) is a central reference point for our work, because they explicitly attempt to model changes in concept meaning, use word embeddings and also adress their method at use cases from the (digital) humanities. Most importantly, Kenter et al. (2015) also published a ground truth dataset for quantitative evaluation. The authors present a method to trace concept vocabularies through a time-stamped corpus based on a set of seed terms (few words, typically one or two) which is tailored towards ad-hoc use over broad time periods. This method is also at the heart of the related systems presented in Martinez-Ortiz et al. (2016b), Martinez-Ortiz et al. (2016a) and Wevers et al. (2015). The authors create vector space models for overlapping periods of the corpus. They then use a number of graph-based unsupervised algorithms that they combine in different ways to generate the final vocabularies. In their evaluation, all their methods beat a baseline that outputs the n most related terms per queried time period using a vector space trained on that period.

Recchia et al. (2016) present a variation of this method, but do not give evaluation results as they report on work in progress of constructing more extensive ground truth data. Their method selects a fully connected graph of k nodes that must contain all words in the previous time period (or seed terms for the first time slice) and have the highest possible minimum edge weight.

3 Concept change as diachronic analogies

Following Kenter et al. (2015), we denote concepts as a set of terms, the *concept vocabulary*. Each term in the concept vocabulary is represented using time-period-specific word embeddings which are derived from training on slices of a time-stamped corpus. We distinguish concept terms which make up the conceptual core (*core concept terms*) from the rest of the vocabulary (*characterizing concept terms*), carrying forward Kenter et al.'s distinction of the core and the margin of concepts. For example, for the concept of ECONOMIC EFFICIENCY, core terms might be *efficiency* and *efficient*, while characterizing terms might be *robotization*, *automatization* or *labor productivity*. In our notion of *diachronic concept change*, the characterizing terms are expected to change over time, while the surface forms of the core terms are assumed to stay the same.

In previous approaches (Kenter et al., 2015; Recchia et al., 2016), the role of the time-specific vector spaces is limited to providing a similarity metric that is used in constructing a weighted semantic graph. Taking the concept vocabulary of the previous time period as input, similar terms are added to the graph which is subsequently pruned based on a centrality measure to generate the new concept vocabulary. In contrast, we are interested in utilizing the features of the time-specific vector spaces directly when predicting concept change to allow for more detailed analyses and comparisons, using models which are based on vector operations.

We reduce the problem of concept change to the problem of predicting valid characterizing terms for a core concept term given a respective characterizing term at an earlier point in time. More formally, given the embedding of a core concept term $\vec{a}_{t_0}$ for a time period t_0, the embedding for the *same* core concept term $\vec{a}_{t_1}$ for a later time period t_1 and the embedding of a characterizing term for the earlier period $\vec{b}_{t_0}$, our goal is to predict the embedding for the missing characterizing term $\vec{b}_{t_1}$ with some function f:

$$\vec{b}_{t_1} = f(\vec{a}_{t_0}, \vec{b}_{t_0}, \vec{a}_{t_1}) \tag{1}$$

In the following section, we will present and discuss a number of possible instantiations of f, which are motivated by an inductive learning perspective on analogy (Cornuéjols and Ales-Bianchetti, 1998) and methods of analogy recovery used in connection with word embeddings (Mikolov et al., 2013b). Consequently, we view a 4-tuple $(\vec{a}_{t_0}, \vec{b}_{t_0}, \vec{a}_{t_1}, \vec{b}_{t_1})$ as constituting a loose diachronic analogy between concept terms, for ECONOMIC EFFICIENCY e.g. "*efficiency* is to *robotization* at one point in time as *efficiency* was to *mechanization* at an earlier point in time".

Analogies between two word pairs are based on highly similar semantic relations among the words in each pair (Turney, 2006). In our adaptation of analogies between concept terms, relational similarity is implied by the assumption that both term pairs relate to the same concept. Note that the type of semantic relation underlying our notion of diachronic concept analogies is rather generic, as it only describes *membership* of a characterizing term in a concept.

4 Learning diachronic analogies

4.1 Baselines and models

This section describes two baselines and two preliminary models to learn to complete diachronic analogies. Each model is based on different intuitions and assumptions about the characteristics of diachronic analogies which will be detailed and subsequently tested in the quantitative evaluation.

No transfer baseline. As a naive baseline we set $\vec{b}_{t_1} = \vec{a}_{t_1}$, which predicts the embedding of the core concept term at t_1 by effectively ignoring the previous time period. We refer to this model as the NO baseline. It is intended to provide a minimal benchmark for model performance.

Linear combination baseline. The baseline described in equation (2) performs a linear combination of the known term vectors to recover the fourth, unknown vector. This corresponds to the analogy recovery method used by Mikolov et al. (2013b) without a search for the closest word vector in the vocabulary. We refer to this model as the ADD baseline. It assumes a direct linear relationship between the analogy vectors, even though the source and target vectors belong to vector spaces computed from two distinct subsets of the corpus.

$$\vec{b}_{t_1} = \vec{b}_{t_0} - \vec{a}_{t_0} + \vec{a}_{t_1} \tag{2}$$

Transformation. This model amounts to explicitly encoding the relation between the source terms as a function and reapplying it to the target. The model encodes the assumption that the two vector spaces are structurally similar, so that the same (geometric) relation holds in both instances. In the following, it is referred to as TRANS. The model learns a transformation between the concept term vectors at t_0 and applies the same function to the core concept term at t_1 to approximate the unknown term vector. This means that we learn

$$\vec{b}_{t_0} = \mathbf{A}_{t_0} \cdot \vec{a}_{t_0} \tag{3}$$

and then predict by reusing $\mathbf{A}_{t_0}$ on t_1 as

$$\vec{b}_{t_1} = \mathbf{A}_{t_0} \cdot \vec{a}_{t_1} \tag{4}$$

Weighted linear combination. The weighted linear combination model (equation 5) is equivalent to the ADD baseline with weights attached to each word vector before combining them. We refer to this system as the WEIGHTS model. In contrast to ADD, this model is based on the intuition that for diachronic analogies the model has to compensate for the displacement of vectors due to semantic change when trying to complete the analogy based on the parallelogram assumption.

$$\vec{b}_{t_1} = \mathbf{B}_{t_0} \cdot \vec{b}_{t_0} - \mathbf{A}_{t_0} \cdot \vec{a}_{t_0} + \mathbf{A}_{t_1} \cdot \vec{a}_{t_1} \tag{5}$$

4.2 Training method

We view the equations in Section 4.1 as describing shallow neural networks which we implemented using the framework PyTorch.[2] The models are trained for 10 iterations over the training data using the Adam optimization method (Kingma and Ba, 2014). The model's error is measured using the cosine distance between the predicted vector and a gold vector. Weights are initialized with an identity matrix, which makes the output of the untrained TRANS model equal to the NO baseline and the untrained WEIGHTS model equal to the ADD baseline. Other initialization strategies were tested (in particular random weights and an identity matrix combined with small random values), but were not found to improve results.

4.3 Vector space alignment

For low-dimensional vector representations, specifically derived with skip-gram as used in the reported experiments, vectors for the same word from different spaces can be arbitrary orthogonal transformations (Hamilton et al., 2016). To counteract this problem, the authors frame the alignment of two matrices of word embeddings as an orthogonal Procrustes problem and solve it using the closed form solution from Schönemann (1966). As Bamler and Mandt (2017) point out, this method conceptualizes the differences of diachronic word vectors as the result of a global rotation (introduced by the rotation-invariant cost functions used in deriving the embeddings) and some semantic drift that becomes available for analysis after alignment.

Note that aligned embeddings can only be computed for the intersection of the vocabularies of the two time periods, discarding all words that occur only in one of the time periods. Thus, while the reasoning of Hamilton et al. (2016) is convincing, we will empirically assess the effects of alignment in our task.

5 Data

In our experiments we use a dataset of Dutch newspaper articles digitized by the National Library of the Netherlands with related ground truth data published by Kenter et al. (2015). In the following, we will describe the data in more detail and will outline the performed data collection and preprocessing.

5.1 Ground truth data

Kenter et al. (2015) provide evaluation data for predictions of diachronic changes in concept vocabularies, which we adapt slightly to generate diachronic analogies between concept terms. In their dataset, for every 5-year interval, two domain experts (historians of contemporary history) rate the relevance of a number of words in relation to a given set of concept terms (*seed terms*). This is done for 21 sets of seed terms in total, each corresponding to a distinct concept. The rating scale goes in integer steps from -1 (not related) up to 2 (perfect match).

For each concept, we first choose a time interval t_0 for the analogy source and an interval t_1 for the analogy target. We treat the seed terms as core concept terms and use them to derive the vectors $\vec{a}_{t_0}$ and $\vec{a}_{t_1}$. The vectors $\vec{a}_{t_0}$ and $\vec{a}_{t_1}$ are derived from any term in the concept vocabulary with an average score greater than a threshold τ.

For the ground truth dataset used in the reported experiments, we chose $\tau = 0$, $t_0 = (1955, 1959)$ and $t_1 = (1985, 1989)$. Note that an example can only be used in training and evaluation if there are representations for every term in the analogy in the respective vector space. To be able to use as much of the validation data as possible, we tried to obtain a corpus for training the embeddings as similar to the one used by Kenter et al. (2015) as possible (see Section 5.2). Table 1 gives an overview of the concepts that were used in the experiments after taking into account the effective number of analogies for each and filtering out concepts with only few examples. In the following, we will mostly refer to individual concepts by the last word of the concept core words, e.g. to the DUTCH CITIES concept by *utrecht*.

Besides presumably changing concepts, the evaluation set also includes concepts which are expected to have stayed semantically stable over the evaluation period. These concepts are used to evaluate whether the models are able to predict both semantic change and semantic constancy. In detail, concepts with a tendency towards stability are *utrecht*, *violen*, *boekje*, *beethoven* and *koeien*. While diachronic analogies

[2] PyTorch 0.2.0, https://github.com/pytorch/pytorch/tree/v0.2.0

Concept core words	Description	N	N_{emb}
amsterdam, rotterdam, utrecht	Dutch cities	7350	5940
neger, negers, negerin, kleurling	(Discriminating) terms for black people	312	160
efficiency, efficiëntie	Economic efficiency	1008	684
viool, violen	Musical instruments	1682	1512
boek, boeken, boekje	Writings and books	5472	4773
mozart, brahms, beethoven	Famous composers	720	720
waterstofbom, waterstofbommen, atoombom, atoombommen	Nuclear weapons	1440	540
koe, koeien	Cattle farming	1380	900
jodenvervolging, deportatie, deportaties	Persecution of jews	264	150

Table 1: Overview of all conceptual diachronic analogies in the dataset for 1955–1959 and 1985–1989 with concept core words and number of examples with untrimmed (N) and embedding vocabulary (N_{emb})

between concept term embeddings seem to be primarily suitable to describe semantic change in the strict sense, constancy can be expressed equally well in terms of smaller differences between vectors. In the edge case of identical semantic spaces for two time periods, the assumed parallelogram between analogy vectors becomes a line.

Note that while the number of examples is reduced by the embeddings' vocabulary as influenced by hyperparameters (cf. N_{emb} in Table 1), the number of examples already varies notably per concept on the basis of the untrimmed corpus-specific vocabulary (N). Also, in the original dataset the number of annotated words differs between the concepts. However, no systematic relation could be established between the number of original annotations and the number of examples in the generated analogies, so that the influence of the (effective) vocabulary seems to be decisive.

5.2 Koninklijke Bibliotheek Historical Newspaper Corpus

Kenter et al. (2015) train their embeddings on a subset (1950–1994) of the historical collection of digitized Dutch newspapers which are archived by the Koninklijke Bibliotheek (KB), the National Library of the Netherlands. Unfortunately, the exact corpus used by the authors is not available as there is no self-contained dataset available for the KB historical newspapers corpus. The individual articles have to be crawled using a set of related APIs which are only available after signing a contract with the KB due to copyright restrictions of the newspapers' content.

First, the complete set of article identifiers for a time period was crawled. Then a 10% sample was drawn for full-text retrieval, resulting in about 500.000 articles for both selected periods. Tokenization, sentence detection and normalization (lowercasing, stop word and punctuation removal) were applied using the spaCy library[3] with a pretrained[4] Dutch model. No lemmatization or stemming was applied, following Kenter et al. (2015). Note that the resulting dataset is nevertheless different from Kenter et al. (2015), as the collection changed in the meantime and also was not sampled in their work. Our final corpus for $t_0 = (1955, 1959)$ contains 8.527.393 sentences with 63.556.890 tokens in total. The corpus for $t_1 = (1985, 1989)$ contains 15.025.711 sentences with 113.813.461 tokens in total. Note that for the later period the number of tokens is almost twice as much. Exemplary analysis showed that articles in the 1980s tend to be longer, so that the number of sentences and tokens is higher.

For each period, word embeddings were trained using the implementations by the Gensim library[5]. In correspondence with Kenter et al. (2015), we used the skip-gram architecture to train embeddings with 300 dimensions. We use a slightly different configuration, however, in particular negative sampling, a subsampling threshold of 10^{-5}, a context width of 4 and a minimum word count of 10.

[3] https://github.com/explosion/spaCy/tree/v2.0.5

[4] Dutch multi-task CNN trained on the Universal Dependencies and WikiNER corpus, version 2.0.0, https://github.com/explosion/spacy-models/releases/tag/nl_core_news_sm-2.0.0

[5] Gensim 3.2.0, https://github.com/RaRe-Technologies/gensim/tree/3.2.0

To align the coordinate axes of two word embeddings spaces as proposed by Hamilton et al. (2016), we use a port of their code by Ryan Heuser that provides an interface to the Gensim library[6].

6 Experiments and results

Despite relying on their data, we cannot compare our results directly to Kenter et al. (2015), due to the differences in approach and evaluation period. Therefore, we evaluate the models presented in Section 4 intrinsically in two separate experiments: The first experiment tests how well the different systems predict the missing vector $\vec{b}_{t_1}$. Subsequently, we report on an experiment that evaluates how well these vectors can be mapped back onto the vocabulary to receive human-readable lists of terms. These experiments are designed to assess as to what extent the diachronic transformations applied result in *semantically coherent* as well as *discriminative* concepts at t_1.

Results will be discussed and compared with and without alignment (cf. Section 4.3). Note that drawing conclusions from this comparison is difficult, because the alignment inevitably changes the vocabulary and thereby the composition of the dataset, so that we effectively compare across two different datasets. However, given the decisive stance of Hamilton et al. (2016) on alignment for diachronic embedding spaces, it seems important to evaluate its effects on the task.

We train and evaluate all four models separately for each concept listed in Table 1. As some concepts only have few examples, all experiments are run in a k-fold cross-validation setting. We use $k = 5$ and report scores averaged over all folds along with their standard deviation (in plots indicated by error bars).

6.1 Experiment 1: Predicting the missing analogy vector

To evaluate the prediction of the missing concept term, we use the cosine similarity between the predicted and the label vector and average over all examples. For a set of analogies A, the predicted vector $\vec{b}_{t_{1_i}}$ and the label vector $\vec{b}^*_{t_{1_i}}$ for the i-th analogy, the average cosine similarity (COS) score is defined as

$$COS = \frac{1}{|A|} \sum_{i=1}^{|A|} \frac{\vec{b}_{t_{1_i}} \cdot \vec{b}^*_{t_{1_i}}}{||\vec{b}_{t_{1_i}}||_2 ||\vec{b}^*_{t_{1_i}}||_2} \tag{6}$$

Hence, this experiment addresses the *coherence* aspect in evaluating the diachronic transformations in the sense that higher COS values indicate smaller distances between predicted and expected terms, thus resulting in more coherent concept representations at t_1.

Figure 1 shows the average cosine similarity scores of a 5-fold cross-validation for both aligned and non-aligned vector spaces per concept for each of the models and baselines for the 1955–1959 and 1985–1989 time intervals on the subset of the KB historical newspapers corpus described in Section 5.2.

For all concepts and models as well as irrespective of alignment, the standard deviation is very low, indicating stable performance across folds.

The ADD baseline generally has a lower score than the NO baseline except for *koeien* and *boekje* with alignment. Apparently a naive search in the neighborhood of the core concept term for the later period is better than assuming a simple linear relationship between the terms across time periods. Alignment improves the scores for the ADD baseline, but we see hardly any differences for NO. As the NO baseline ignores the aligned space t_0, the small differences in performance might be caused by the differences in the dataset due to the altered vocabulary.

The results for TRANS are strongly influenced by alignment. With non-aligned spaces, the model is worse than the NO baseline for all concepts. With aligned spaces, the score of the TRANS model is dramatically higher than without, sometimes more than twice as high (*boekje*, *utrecht*), and consistently beats both baselines. Apparently it is only beneficial to assume that the same transformation is applicable to both vector spaces across time when their dimensions are aligned, so that the transformation will have a similar effect.

WEIGHTS is the best performing model, sometimes with a notable margin. When using alignment, we see hardly any improvements for WEIGHTS. Probably, in the non-aligned case the weights manage

[6] https://gist.github.com/quadrismegistus/ 09a93e219a6ffc4f216fb85235535faf

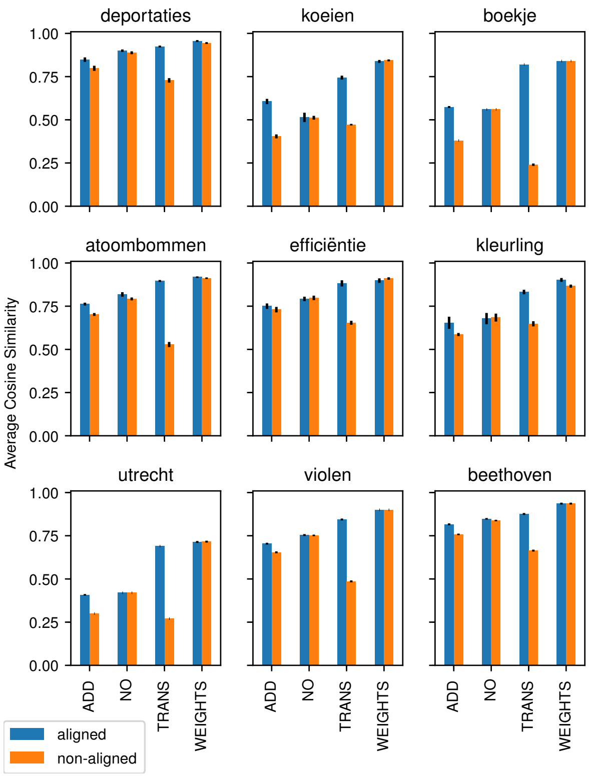

Figure 1: Averaged (5-fold cross-validation) cosine similarities between predicted and label vector per model and concept for aligned and non-aligned vector spaces

to compensate for the missing alignment to a high degree, so that the alignment does not add much. Comparing these results to ADD, it seems to be beneficial to include the t_0 embeddings for predicting the missing vector at t_1 only if weights are learned to account for the differences between the two vector spaces. Comparing to the simple TRANS model, the performance without alignment is clearly superior, but with alignment WEIGHTS is only slightly better.

6.2 Experiment 2: Using predicted vectors for vocabulary retrieval

In the following, we will report on experiments that evaluate how well the predicted vectors can be used to retrieve meaningful word lists that help to study concept change. To perform the mapping from an embedding to a term, the system takes a predicted vector and performs an n-nearest neighbor search over all word embeddings of the t_1 vector space using cosine similarity as proximity metric. In the reported experiments, $n = 10$ is used. Note that this evaluation setting is comparatively hard for the proposed models, as the vocabulary retrieval performance is not only determined by the quality of the prediction, but also by the other vectors in the t_1 space outside the set of vectors that are part of the ground truth dataset. As these vectors were not seen by the model during training, performance inevitably is influenced to an arbitrary degree by the specific concepts and embedding spaces used.

We evaluate the vocabulary retrieval performance in terms of the mean reciprocal rank (MRR) (Voorhees, 1999) of the label terms in the ranked lists of vocabulary terms. For a set of analogies A and the rank of the label term in the list of vocabulary terms for the i-th analogy $rank_i$, the MRR is defined as

$$MRR = \frac{1}{|A|} \sum_{i=1}^{|A|} \frac{1}{rank_i} \tag{7}$$

When for any analogy the label term is not in the list of returned vocabulary terms (i.e., $rank_i$ is not defined), the reciprocal rank $\frac{1}{rank_i}$ is set to 0. Hence, this experiment aims at assessing whether the diachronic transformations applied yield a semantic space at t_1 that effectively *discriminates* between

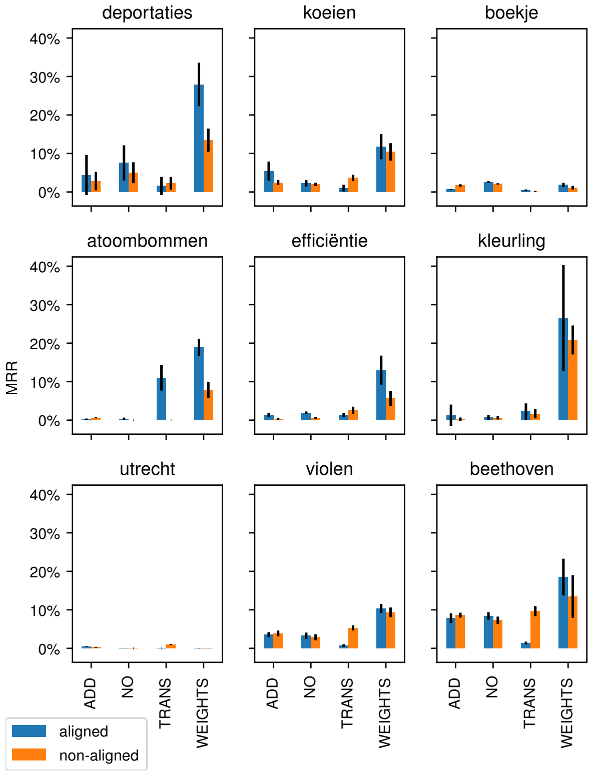

Figure 2: Averaged (5-fold cross-validation) mean reciprocal rank scores (in percent) per model and concept for aligned and non-aligned vector spaces

different concepts such that terms included in the vocabulary of a concept should be consistently ranked higher than confounders from the vocabulary of other concepts, thus resulting in higher MRR scores.

Figure 2 shows MRR results per concept for each of the models and baselines using aligned and non-aligned vector spaces computed for the same dataset and 5-fold cross-validation setting as in Section 6.1. Overall, compared to Experiment 1, the results are much less uniform and stable. The ADD baseline performs worse or comparable to the NO baseline. An exception is *koeien* with alignment. These overall results for ADD are expected, since NO has higher COS scores than ADD for almost all concepts, independent of alignment. Alignment has no clear effect on MRR scores for ADD. For some concepts (e.g., *boekje*) the baseline performs worse than non-aligned, for others (e.g., *koeien*) it performs better.

Surprisingly, without alignment TRANS often yields more relevant predictions than the two baselines, even though its COS scores are mostly lower (cf. Figure 1). With aligned embedding spaces, the MRR performance of TRANS shows very variable behavior. For some concepts, we see small improvements with alignment, in the case of *atoombommen* it is even very large. For other concepts, we see a drop in performance, sometimes very sharp as for *beethoven* or *violen*. This is a puzzling result, since we see notable improvements in the cosine similarity score for TRANS when the vector spaces are aligned.

For most concepts, WEIGHTS performs best and often does so with a large margin – with or without alignment. Exceptions are *utrecht* and *boekje* with a score below the baselines. Interestingly, while applying rotational alignment only leads to negligible improvements in COS for WEIGHTS, the MRR score is always higher with alignment than without, although for some concepts (e.g., *kleurling*) the standard deviation increases exceptionally.

7 Discussion

Taken together, the vector transformations applied in our experiments in order to solve diachronic analogies tend to produce robust and promising results with regard to local conceptual coherence (cf. Experiment 1); however, the resulting conceptual spaces barely exhibit the property of discriminability between concepts (cf. Experiment 2).

Comparing the individual models' performance, it is notable that while the differences between TRANS and WEIGHTS in terms of COS are small with aligned axes, WEIGHTS has much higher MRR scores for both aligned and non-aligned embeddings. This result is consistent for all concepts, so that the WEIGHTS model is the best- performing system overall.

Due to its comparable performance with non-aligned embeddings, when training WEIGHTS, differences between the vocabularies of the two time periods could be included in the data (see section 4.3). Because of this, WEIGHTS can also be applied to cases of concept change where words disappear from the concept vocabulary or new ones are added. On the other hand, the TRANS model requires less complex training data which only needs to contain conceptual word pairs for one time period. This allows for more exploratory use cases where expert knowledge about concepts only exists for t_0. Taking into account the often negative effect of alignment on MRR for TRANS, the embeddings should be non-aligned. While we admittedly would expect less relevant term lists when using this model instead of WEIGHTS, it should, according to the evaluation, nevertheless give better results than the baseline approaches.

As discussed in Section 5.1, our evaluation data contains concepts that exhibit semantic change as well as ones that tend towards semantic stability. The results observed in Experiment 1 are largely unaffected by this difference, which may suggest that the models are applicable to cases of diachronic change and stability as well. This hypothesis is only partially corroborated in Experiment 2, though. From our current analyses, we conjecture that the variance in performance across concepts is mostly explained by concept size (cf. Table 1) rather than the difference between changing vs. stable concepts. A more detailed investigation of these effects is left to future work.

8 Conclusions and outlook

We have introduced the task of completing diachronic analogies to analyze concept change. We have presented two learned models to recover diachronic analogies and tested them in a quantitative evaluation. The experiments showed that, for most of the evaluated concepts, a model based on a weighted linear combination of the analogous words' embeddings is able to more accurately predict the missing vector which also corresponds to more relevant terms than a learned transformation and two related baselines. More specifically, we have evaluated the effect of a rotational alignment of the time-period-specific embedding spaces, finding varying effects on performance, depending on the model, concept and evaluation metric. For the weighted linear combination, however, results improve with alignment in the majority of cases. In sum, it is beneficial for prediction of diachronic changes in concept vocabularies to treat the concept terms as analogous when weights are learnt to compensate for diachronic drift. However, while all models tend to be coherent in relation to the represented concept, they are only to some degree discriminative in regard to the vocabulary of other concepts.

Future work should carry out more in-depth evaluations, annotating task-specific ground truth data and exploring evaluation settings like zero-shot learning which has been show to obtain promising results in related problems (cf. Hartung et al. (2017)). We also expect benefits from training with an objective function which includes negative examples and relates more closely to MRR. Beyond this, we are interested in designing more complex and task-specific models. Not last, we plan to explore use cases based on cooperation with scholars from the humanities. For example, we see potential in analysing how an author's use of specific concepts changes across works using a combination of both interpretative and automatic methods of diachronic analogy recovery.

Acknowledgments

We thank Steven Claeyssens, Curator of Digital Collections at the Koninklijke Bibliotheek, for providing access to the data as well as further documentation and support. This work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

Robert Bamler and Stephan Mandt. 2017. Dynamic Word Embeddings. In *PMLR*, pages 380–389, July.

David M. Blei and John D. Lafferty. 2006. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA. ACM.

Dawn Chen, Joshua C. Peterson, and Thomas L. Griffiths. 2017. Evaluating vector-space models of analogy. *arXiv:1705.04416 [cs]*, May. arXiv: 1705.04416.

Antoine Cornuéjols and Jacques Ales-Bianchetti. 1998. Analogy and Induction: Which (missing) link? In *Proceedings of Workshop on Analogy*, Sofia, Bulgaria, July.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1147–1156, Copenhagen, Denmark. Association for Computational Linguistics.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pages 1–31. Blackwell, Oxford.

Kristina Gulordava and Marco Baroni. 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*, GEMS '11, pages 67–71, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Zellig S. Harris. 1954. Distributional Structure. *Word*, 10(2-3):146–162, August.

Matthias Hartung, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano. 2017. Learning compositionality functions on word embeddings for modelling attribute meaning in adjective-noun phrases. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 54–64, Valencia, Spain, April. Association for Computational Linguistics.

Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. 2015. Ad Hoc Monitoring of Vocabulary Shifts over Time. In *Proceedings of the 24th ACM InternationalConference on Information and Knowledge Management (CIKM'15).*

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, Baltimore, MD, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December. arXiv: 1412.6980.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 625–635, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Omer Levy and Yoav Goldberg. 2014. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the 18th Conference on Computational Language Learning (CoNLL)*, pages 171–180. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3(0):211–225, May.

Carlos Martinez-Ortiz, Tom Kenter, Melvin Wevers, Pim Huijnen, Jaap Verheul, and Joris van Eijnatten. 2016a. Design and implementation of ShiCo: Visualising shifting concepts over time. In *Proceedings of the 3rd International Workshop on Computational History (HistoInformatics 2016).*

Carlos Martinez-Ortiz, Tom Kenter, Melvin Wevers, Pim Huijnen, Jaap Verheul, and Joris van Eijnatten. 2016b. ShiCo: A Visualization Tool for Shifting Concepts Through Time. In *Proceedings of the 3rd DH Benelux Conference (DH Benelux 2016).*

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, January. arXiv: 1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A Word at a Time: Computing Word Relatedness using Temporal Semantic Analysis. In *Proceedings of the 20th International World Wide Web Conference*, pages 337–346, Hyderabad, India, March.

Gabriel Recchia, Ewan Jones, Paul Nulty, John Regan, and Peter de Bolla. 2016. Tracing Shifting Conceptual Vocabularies Through Time. In *Knowledge Engineering and Knowledge Management*, Lecture Notes in Computer Science, pages 19–28. Springer, Cham, November.

David E Rumelhart and Adele A Abrahamson. 1973. A model for analogical reasoning. *Cognitive Psychology*, 5(1):1–28, July.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, March.

Chenhao Tan, Dallas Card, and Noah A. Smith. 2017. Friendships, Rivalries, and Trysts: Characterizing Relations between Ideas in Texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 773–783, Vancouver, Canada. Association for Computational Linguistics.

Peter D. Turney. 2006. Similarity of Semantic Relations. *Computational Linguistics*, 32(3):379–416, August.

Ellen M. Voorhees. 1999. The TREC-8 Question Answering Track Report. In *Proceedings of TREC-8*, pages 77–82.

Melvin Wevers, Tom Kenter, and Pim Huijnen. 2015. Concepts Through Time: Tracing Concepts in Dutch Newspaper Discourse (1890-1990) using Word Embeddings. *Digital Humanities 2015 (DH2015).*