Domain-Specific Word Embeddings with Structure Prediction

David Lassner[1,2*]   **Stephanie Brandl**[1,2,3*]   **Anne Baillot**[4]   **Shinichi Nakajima**[1,2,5]

[1]TU Berlin, Germany   [2]BIFOLD, Germany   [3]University of Copenhagen, Denmark

[4]Le Mans Université, France   [5]RIKEN Center for AIP, Japan

{lassner@tu-berlin.de,brandl@di.ku.dk}

[*]Authors contributed equally.

Abstract

Complementary to finding good general word embeddings, an important question for representation learning is to find dynamic word embeddings, for example, across time or domain. Current methods do not offer a way to use or predict information on structure between sub-corpora, time or domain and dynamic embeddings can only be compared after post-alignment. We propose novel word embedding methods that provide general word representations for the whole corpus, domain-specific representations for each sub-corpus, sub-corpus structure, and embedding alignment simultaneously. We present an empirical evaluation on New York Times articles and two English Wikipedia datasets with articles on science and philosophy. Our method, called *Word2Vec with Structure Prediction (W2VPred)*, provides better performance than baselines in terms of the general analogy tests, domain-specific analogy tests, and multiple specific word embedding evaluations as well as structure prediction performance when no structure is given a priori. As a use case in the field of Digital Humanities we demonstrate how to raise novel research questions for high literature from the German Text Archive.

1 Introduction

Word embeddings (Mikolov et al., 2013b; Pennington et al., 2014) are a powerful tool for word-level representation in a vector space that captures semantic and syntactic relations between words. They have been successfully used in many applications such as text classification (Joulin et al., 2016) and machine translation (Mikolov et al., 2013a). Word embeddings highly depend on their training corpus. For example, technical terms used in scientific documents can have a different meaning in other domains, and words can change their meaning over time—''apple'' did not mean a tech company before Apple Inc.

was founded. On the other hand, such local or *domain-specific* representations are also not independent of each other, because most words are expected to have a similar meaning across domains.

There are many situations where a given target corpus is considered to have some *structure*. For example, when analyzing news articles, one can expect that articles published in 2000 and 2001 are more similar to each other than the ones from 2000 and 2010. When analyzing scientific articles, uses of technical terms are expected to be similar in articles on similar fields of science. This implies that the structure of a corpus can be a useful side resource for obtaining better word representation.

Various approaches to analyze semantic shifts in text have been proposed where typically first individual static embeddings are trained and then aligned afterwards (e.g., Kulkarni et al., 2015; Hamilton et al., 2016; Kutuzov et al., 2018; Tahmasebi et al., 2018). As most word embeddings are invariant with respect to rotation and scaling, it is necessary to map word embeddings from different training procedures into the same vector space in order to compare them. This procedure is usually called *alignment*, for which orthogonal Procrustes can be applied as has been used in Hamilton et al. (2016).

Recently, new methods to train diachronic word embeddings have been proposed where the alignment process is integrated in the training process. Bamler and Mandt (2017) propose a Bayesian approach that extends the skip-gram model (Mikolov et al., 2013b). Rudolph and Blei (2018) analyze dynamic changes in word embeddings based on exponential family embeddings. Yao et al. (2018) propose Dynamic Word2Vec where word embeddings for each year of the New York Times corpus are trained based on individual positive point-wise information matrices and aligned simultaneously.

320

We argue that apart from diachronic word embeddings there is a need to train dynamic word embeddings that not only capture temporal shifts in language but for instance also semantic shifts between domains or regional differences. It is important that those embeddings can be trained on small datasets. We therefore propose two generalizations of Dynamic Word2Vec. Our first method is called *Word2Vec with Structure Constraint (W2VConstr)*, where domain-specific embeddings are learned under regularization with any kind of structure. This method performs well when a respective graph structure is given a priori. For more general cases where no structure information is given, we propose our second method, called *Word2Vec with Structure Prediction (W2VPred)*, where domain-specific embeddings and sub-corpora structure are learned at the same time. W2VPred simultaneously solves three central problems that arise with word embedding representations:

1. Words in the sub-corpora are embedded in the same vector space, and are therefore directly comparable without post-alignment.

2. The different representations are trained simultaneously on the whole corpus as well as on the sub-corpora, which makes embeddings for both general and domain-specific words robust, due to the information exchange between sub-corpora.

3. The estimated graph structure can be used for confirmatory evaluation when a reasonable prior structure is given. W2VPred together with W2VConstr identifies the cases where the given structure is not ideal, and suggests a refined structure which leads to an improved embedding performance; we call this method *Word2Vec with Denoised Structure Constraint*. When no structure is given, W2VPred provides insights on the structure of sub-corpora, for example, similarity between authors or scientific domains.

All our methods rely on static word embeddings as opposed to currently often used contextualized word embeddings. As we learn one representation per slice such as year or author, thus considering a much broader context than contextualized embeddings, we are able to find a meaningful structure-

between corresponding slices. Another main advantage comes from the fact that our methods do not require any pre-training and can be run on a single GPU.

We test our methods on 4 different datasets with different structures (sequences, trees, and general graphs), domains (news, wikipedia, high literature), and languages (English and German). We show on numerous established evaluation methods that W2VConstr and W2VPred significantly outperform baseline methods with regard to general as well as domain-specific embedding quality. We also show that W2VPred is able to predict the structure of a given corpus, outperforming all baselines. Additionally, we show robust heuristics to select hyperparameters based on proxy measurements in a setting where the true structure is not known. Finally, we show how W2VPred can be used in an explorative setting to raise novel research questions in the field of Digital Humanities. Our code is available at `https://github.com/stephaniebrandl/domain-word-embeddings`.

2 Related Work

Various approaches to track, detect, and quantify semantic shifts in text over time have been proposed (Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016; Zhang et al., 2016; Marjanen et al., 2019).

This research is driven by the hypothesis that semantic shifts occur, for example, over time (Bleich et al., 2016) and viewpoints (Azarbonyad et al., 2017), in political debates (Reese and Lewis 2009), or caused by cultural developments (Lansdall-Welfare et al., 2017). Analysing those shifts can be crucial in political and social studies but also in literary studies, as we show in Section 5.

Typically, methods first train individual static embeddings for different timestamps, and then align them afterwards (e.g., Kulkarni et al., 2015; Hamilton et al., 2016; Kutuzov et al., 2018; Devlin et al., 2019; Jawahar and Seddah, 2019; Hofmann et al., 2020; and a comprehensive survey by Tahmasebi et al., 2018). Other approaches, which deal with more general structure (Azarbonyad et al., 2017; Gonen et al., 2020) and more general applications (Zeng et al., 2017; Shoemark et al., 2019), also rely on post-alignment

321

of static word embeddings (Grave et al., 2019). With the rise of larger language models such as BERT (Devlin et al., 2019) and, with that, contextualized embeddings, a part of the research question has shifted towards detecting language change in contextualized word embeddings (e.g., Jawahar and Seddah, 2019; Hofmann et al., 2020).

Recent methods directly learn dynamic word embeddings in a common vector space without post-alignment: Bamler and Mandt (2017) proposed a Bayesian probabilistic model that generalizes the skip-gram model (Mikolov et al., 2013b) to learn dynamic word embeddings that evolve over time. Rudolph and Blei (2018) analyzed dynamic changes in word embeddings based on exponential family embeddings, a probabilistic framework that generalizes the concept of word embeddings to other types of data (Rudolph et al., 2016). Yao et al. (2018) proposed Dynamic Word2Vec (DW2V) to learn individual word embeddings for each year of the New York Times dataset (1990-2016) while simultaneously aligning the embeddings in the same vector space. Specifically, they solve the following problem for each timepoint $t = 1, \ldots, T$ sequentially:

$$\min_{U_t} L_{\mathrm{F}} + \tau L_{\mathrm{R}} + \lambda L_{\mathrm{D}}, \quad \text{where} \quad (1)$$

$$L_{\mathrm{F}} = \left\| Y_t - U_t U_t^\top \right\|_F^2, L_{\mathrm{R}} = \left\| U_t \right\|_F^2,$$

$$L_{\mathrm{D}} = \left\| U_{t-1} - U_t \right\|_F^2 + \left\| U_t - U_{t+1} \right\|_F^2 \quad (2)$$

represent the losses for data fidelity, regularization, and diachronic constraint, respectively. $U_t \in \mathbb{R}^{V \times d}$ is the matrix consisting of d-dimensional embeddings for V words in the vocabulary, and $Y_t \in \mathbb{R}^{V \times V}$ represents the positive pointwise mutual information (PPMI) matrix (Levy and Goldberg, 2014). The diachronic constraint L_{D} encourages alignment of the word embeddings with the parameter λ controlling how much the embeddings are allowed to be dynamic ($\lambda = 0$: no alignment and $\lambda \to \infty$: static embeddings).

3 Methods

By generalizing DW2V, we propose two methods, one for the case where sub-corpora structure is given as prior knowledge, and the other for the case where no structure is given a priori. We also argue that combining both methods can improve the performance in cases where some

prior information is available but not necessarily reliable.

3.1 Word2Vec with Structure Constraint

We reformulate the diachronic term in Eq. 1 as

$$L_{\mathrm{D}} = \sum\nolimits_{t'=1}^{T} W_{t,t'}^{\mathrm{diac}} \left\| U_t - U_{t'} \right\|_F^2$$

$$\text{with } W_{t,t'}^{\mathrm{diac}} = \mathbb{1}(\{|t - t'| = 1\}), \quad (3)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. This allows us to generalize DW2V for different neighborhood structures: Instead of the chronological sequence (3), we assume $W \in \mathbb{R}^{T \times T}$ to be an arbitrary affinity matrix representing the underlying semantic structure, given as prior knowledge.

Let $D \in \mathbb{R}^{T \times T}$ be the pairwise distance matrix between embeddings such that

$$D_{t,t'} = \left\| U_t - U_{t'} \right\|_F^2, \quad (4)$$

and we impose regularization on the distance, instead of the norm of each embeddings. This yields the following optimization problem:

$$\min_{U_t} L_{\mathrm{F}} + \tau L_{\mathrm{RD}} + \lambda L_{\mathrm{S}}, \quad \text{where} \quad (5)$$

$$L_{\mathrm{F}} = \left\| Y_t - U_t U_t^\top \right\|_F^2, L_{\mathrm{RD}} = \left\| D \right\|_F,$$

$$L_{\mathrm{S}} = \sum\nolimits_{t'=1}^{T} W_{t,t'} D_{t,t'}. \quad (6)$$

We call this generalization of DW2V *Word2Vec with Structure Constraint* (W2VConstr).

3.2 Word2Vec with Structure Prediction

When no structure information is given, we need to estimate the similarity matrix W from the data. We define W based on the similarity between embeddings. Specifically, we initialize (each entry of) the embeddings $\{U_t\}_{t=1}^{T}$ by independent uniform distribution in $[0, 1)$. Then, in each iteration, we compute the distance matrix D by Eq. (4), and set $\widetilde{W}$ to its (entry-wise) inverse, that is,

$$\widetilde{W}_{t,t'} \leftarrow \begin{cases} D_{t,t'}^{-1} & \text{for } t \neq t', \\ 0 & \text{for } t = t' \end{cases} \quad (7)$$

and normalize it according to the corresponding column and row:

$$W_{t,t'} \leftarrow \frac{\widetilde{W}_{t,t'}}{\sum_{t''} \widetilde{W}_{t,t''} + \sum_{t''} \widetilde{W}_{t'',t'}}. \quad (8)$$

The structure loss (6) with the similarity matrix W updated by Eqs. 7 and 8 constrains the

distances between embeddings according to the similarity structure that is at the same time estimated from the distances between embeddings. We call this variant *Word2Vec with Structure Prediction* (W2VPred). Effectively, W serves as a weighting factor that strengthens connections between close embeddings.

3.3 Word2Vec with Denoised Structure Constraint

We propose a third method that combines W2VConstr and W2VPred for the scenario where W2VConstr results in poor word embeddings because the a priori structure is not optimal. In this case, we suggest applying W2VPred and consider the resulting structure as an input for W2VConstr. This procedure needs prior knowledge of the dataset and a human-in-the-loop to interpret the predicted structure by W2VPred in order to add or remove specific edges in the *new* ground truth structure. In the experiment section, we will condense the predicted structure by W2VPred into a sparse, denoised ground truth structure that is meaningful. We call this method *Word2Vec with Denoised Structure Constraint* (W2VDen).

3.4 Optimization

We solve the problem (5) iteratively for each embedding U_t, given the other embeddings $\{U_{t'}\}_{t' \neq t}$ are fixed. We define one epoch as complete when $\{U_t\}$ has been updated for all t. We applied gradient descent with Adam (Kingma and Ba, 2014) with default values for the exponential decay rates given in the original paper and a learning rate of 0.1. The learning rate has been reduced after 100 epochs to 0.05 and after 500 epochs to 0.01 with a total number of 1000 epochs. Both models have been implemented in PyTorch. W2VPred updates W by Eqs. 7 and 8 after every iteration.

4 Experiments on Benchmark Data

We conducted four experiments starting with well-known settings and datasets and incrementally moving to new datasets with different structures. The first experiment focuses on the general embedding quality, the second one presents results on domain-specific embeddings, the third one evaluates the method's ability to predict structure and the fourth one shows the method's performance on various word similarity tasks. In the following subsections, we will

Category	#Articles
Natural Sciences	8536
Chemistry	19164
Computer Science	11201
Biology	10988
Engineering & Technology	20091
Civil Engineering	17797
Electrical & Electronic Engineering	6809
Mechanical Engineering	4978
Social Sciences	17347
Business & Economics	14747
Law	13265
Psychology	5788
Humanities	15066
Literature & Languages	24800
History & Archaeology	16453
Religion & Philosophy & Ethics	19356

Table 1: Categories and the number of articles in the WikiFoS dataset. One cluster contains 4 categories (rows): The top one is the main category and the following 3 are subcategories. Fields joined by & originate from 2 separate categories in Wikipedia[3] but were joined, according to the OECD's definition.[2]

first describe the data, preprocessing, and then the results. Further details on implementation and hyperparameters can be found in Appendix A.

4.1 Datasets

We evaluated our methods on the following three benchmark datasets.

New York Times (NYT): The New York Times dataset[1] (NYT) contains headlines, lead texts, and paragraphs of English news articles published online and offline between January 1990 and June 2016 with a total of 100,945 documents. We grouped the dataset by years with 1990-1998 as the train set and 1999-2016 as the test set.

Wikipedia Field of Science and Technology (WikiFoS): We selected categories of the OECD's list of Fields of Science and Technology[2] and downloaded the corresponding

---

[1]https://sites.google.com/site/zijunyaorutgers/.
[2]http://www.oecd.org/science/inno/38235147.pdf.

323

Category	#Articles
Logic	3394
Concepts in Logic	1455
History of Logic	76
Aesthetics	7349
Philosophers of Art	30
Literary Criticism	3826
Ethics	5842
Moral Philosophers	170
Social Philosophy	3816
Epistemology	3218
Epistemologists	372
Cognition	8504
Metaphysics	1779
Ontology	796
Philosophy of Mind	976

Table 2: Categories and the number of articles in the WikiPhil dataset. One cluster contains 3 categories: The top one is the main category and the following are subcategories in Wikipedia.

articles from the English Wikipedia. The resulting dataset Wikipedia Field of Science and technology (WikiFoS) contains four clusters, each of which consists of one main category and three subcategories, with 226,386 unique articles in total (see Table 1). We published the data set at `https://huggingface.co/datasets/millawell/wikipedia_field_of_science`. The articles belonging to multiple categories[3] were randomly assigned to a single category in order to avoid similarity because of overlapping texts instead of structural similarity. In each category, we randomly chose $1/3$ of the articles for the train set, and the remaining $2/3$ were used as the test set.

Wikipedia Philosophy (WikiPhil): Based on Wikipedia's definition of categories in *philosophy*, we selected 5 main categories and their 2 largest subcategories each (see Table 2). Categories and subcategories are based on the definition given by Wikipedia. We downloaded 41,603 unique articles in total from the English Wikipedia. Similarly to WikiFoS, the articles belonging to multiple categories were randomly assigned to a single category, and the articles in each category were divided into a train set ($1/3$) and a test set ($2/3$).

[3] `https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories`.
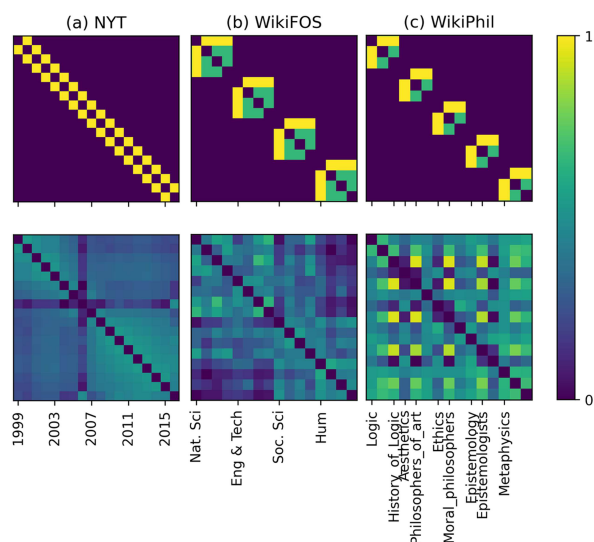

Figure 1: Prior affinity matrix W used for W2VConstr (upper), and the estimated affinity matrix by W2VPred (lower) where the number indicates how close slices are (1: identical, 0: very distant). The estimated affinity for NYT implies the year 2006 is an outlier. We checked the corresponding articles and found that many paragraphs and tokens are missing in that year. Note that the diagonal entries do not contribute to the loss for all methods.

4.2 Preprocessing

We lemmatized all tokens, that is, assigned their base forms with spaCy[4] and grouped the data by years (for NYT) or categories (for WikiPhil and WikiFoS). For each dataset, we defined one individual vocabulary where we considered the 20,000 most frequent (lemmatized) words of the entire dataset that are also within the 20,000 most frequent words in at least 3 independent slices, that is, years or categories. This way, we filtered out ''trend'' words that are of significance only within a very short time period/only a few categories. The 100 most frequent words were filtered out as stop words. We set the symmetric context window (the number of words before and after a specific word considered as context for the PPMI matrix) to 5.

4.3 Ex1: General Embedding Performance

In our first experiment, we compare the quality of the word embeddings trained by W2VConstr and W2VPred with the embeddings trained by baseline methods, GloVe, Skip-Gram, CBOW and DW2V. For GloVe, Skip-Gram and CBOW, we computed one set of embeddings on the entire dataset. For DW2V, W2VConstr, and W2VPred,

[4] `https://spacy.io`.

324

domain-specific embeddings $\{U_t\}$ were averaged over all domains. We use the same vocabulary for all methods. For W2VConstr, we set the affinity matrix W as shown in the upper row of Figure 1, based on the a priori known structure, that is, diachronic structure for NYT, and the category structure in Tables 1 and 2 for WikiFoS and WikiPhil. The lower row of Figure 1 shows the learned structure by W2VPred.

Specifically, we set the ground-truth affinity $W^*_{t,t'}$ as follows: for NYT, $W^*_{t,t'} = 1$ if $|t - t'| = 1$, and $W^*_{t,t'} = 0$ otherwise; for WikiFoS and WikiPhil, $W^*_{t,t'} = 1$ if t is the parent category of t' or vice versa, $W^*_{t,t'} = 0.5$ if t and t' are under the same parent category, and $W^*_{t,t'} = 0$ otherwise (see Tables 1 and 2 for the category structure of WikiFoS and WikiPhil, respectively, and the top row of Figure 1 for the visualization of the ground-truth affinity matrices).

We evaluate the embeddings on general analogies (Mikolov et al., 2013b) to capture the general meaning of a word. Table 3 shows the corresponding accuracies averaged across 10 runs with different random seeds.

For NYT, W2VConstr performs similarly to DW2V, which has essentially the same constraint term—L_S in Eq. (6) for W2VConstr is the same as L_D in Eq. (2) for DW2V up to scaling when W is set to the prior affinity matrix for NYT—and significantly outperforms the other baselines. W2VPred performs slightly worse then the best methods. For WikiFoS, W2VConstr and W2VPred outperform all baselines by a large margin. In WikiPhil, W2VConstr performs poorly (worse than GloVe), while W2VPred outperforms all other methods by a large margin. Standard deviation across the 10 runs are less than one for NYT (all methods and all n), slightly higher for WikiFoS, and highest for WikiPhil W2VPred and W2VConstr (0.28-3.17).

These different behaviors can be explained by comparing the estimated (lower row) and the a priori given (upper row) affinity matrices shown in Figure 1. In NYT, the estimated affinity decays smoothly as the time difference between two slices increases. This implies that the a priori given diachronic structure is good enough to enhance the word embedding quality (by W2VConstr and DW2V), and estimating the affinity matrix (by W2VPred) slightly degrades the performance due to the increased number of unknown parameters to

Data	Method	general analogy tests		
		n=1	n=5	n=10
NYT	GloVe	9.40	26.41	33.58
	Skip-Gram	3.62	16.20	25.61
	CBOW	5.58	19.92	27.60
	DW2V	11.27	32.88	42.97
	W2VConstr (our)	10.90	33.01	43.12
	W2VPred (our)	10.28	31.66	41.88
WikiFoS	GloVe	6.33	23.74	32.58
	Skip-Gram	3.54	12.09	15.77
	CBOW	4.25	17.47	26.21
	W2VConstr (our)	11.91	45.96	56.88
	W2VPred (our)	11.82	45.73	56.40
	W2VDen (our)	11.61	46.50*	57.08*
WikiPhil	GloVe	2.59	17.45	24.19
	Skip-Gram	2.76	10.18	17.48
	CBOW	3.11	6.61	9.47
	W2VConstr (our)	0.42	10.37	15.02
	W2VPred (our)	4.37	31.99	41.75
	W2VDen (our)	5.96*	36.21*	46.15*

Table 3: General analogy test performance for our methods, W2VConstr and W2VPred, and baseline methods, GloVe, Skip-Gram, CBOW, and DW2V averaged across ten runs with different random seeds. The best method and the methods that are not significantly outperformed by the best is marked with a gray background, according to the Wilcoxon signed rank test for $\alpha = 0.05$. W2VDen is compared against the best method from the same data set and if it is significantly better, it is marked with an asterisk (*).

be estimated. In WikiFoS, although the estimated affinity matrix shows somewhat similar structure to the given one a priori, it is not as smooth as the one in NYT and we can recognize two instead of four clusters in the estimated affinity matrix consisting of the first two main categories (*Natural Sciences* and *Engineering & Technology*), and the last two (*Social Sciences* and *Humanities*), which we find reasonable according to Table 1. In summary, W2VConstr and W2VPred outperform baseline methods when a suitable prior structure is given. Results on the WikiPhil dataset show a different tendency: The estimated affinity by W2VPred is very different from the prior structure, which implies that the corpus structure defined by Wikipedia is not suitable for learning word embeddings. As a result, W2VConstr performs even poorer than GloVe. Overall, Table 3 shows that our proposed W2VPred robustly performs well on all datasets. In Section 4.5.3, we will further improve the performance by *denoising* the estimated

325

	n=1	n=5	n=10
GloVe	7.72	14.39	17.87
Skip-Gram	10.49	19.89	24.78
CBOW	6.35	11.36	14.59
Dynamic Word2Vec abbr	39.47	61.94	67.35
W2VConstr (our)	38.23	57.73	64.54
W2VPred (our)	41.87	64.60	69.67

Table 4: Accuracies for temporal analogies (NYT).

structure by W2VPred for the case where a prior structure is not given or is unreliable.

4.4 Ex2: Domain-specific Embeddings

4.4.1 Quantitative Evaluation

Yao et al. (2018) introduced temporal analogy tests that allow us to assess the quality of word embeddings with respect to their temporal information. Unfortunately, domain-specific tests are only available for the NYT dataset. Table 4 shows temporal analogy test accuracies on the NYT dataset. As expected, GloVe, Skip-Gram, and CBOW perform poorly. We assume this is because the individual slices are too small to train reliable embeddings. The embeddings trained with DW2V and W2VConstr are learned collaboratively between slices due to the diachronic and structure terms and significantly improve the performance. Notably, W2VPred further improves the performance by learning a more suitable structure from the data. Indeed, the learned affinity matrix by W2VPred (see Figure 1a) suggests that not the diachronic structure used by DW2V but a smoother structure is optimal.

4.4.2 Qualitative Evaluation

Since no domain-specific analogy test is available for WikiFoS and WikiPhil, we qualitatively analyzed the domain-specific embeddings by checking nearest neighboring words. Table 5 shows the 5 nearest neighbors of the word ''power'' in the embedded spaces for the 4 main categories of WikiFoS trained by W2VPred, GloVe, and Skip-Gram. We averaged the embeddings obtained by W2VPred over the subcategories in each main category. The distance between words are measured by the cosine similarity.

We see that W2VPred correctly captured the domain-specifc meaning of ''power'': In *Natural Sciences* and *Engineering & Technology* the word

is used in a physical context, for example, in combination with generators, which is the closest word in both categories. In *Social Sciences* and *Humanities* on the other hand, the nearest words are ''powerful'' and ''control'', which, in combination, indicates that it refers to ''the ability to control something or someone''.[5] The embedding trained by GloVe shows a very general meaning of power with no clear tendency towards a physical or political context, whereas Skip-Gram shows a tendency towards the physical meaning. We observed many similar examples, for example, charge:electrical-legal, performance:quality-acting, resistance:physical-social, race:championship-ethnicity.

As another example in the NYT corpus, Figure 2 shows the evolution of the word *blackberry*, which can either mean the fruit or the tech company. We selected two slices (2000 & 2012) with the largest pairwise distance for the *blackberry*, and chose the top-5 neighboring words from each year. The figure plots the cosine similarities between *blackberry* and the neighboring words. The time series shows how the word *blackberry* evolved from being mostly associated with the fruit towards associated with the company, and back to the fruit. This can be connected to the release of their smartphone in 2002 and the decrease in sales number after 2011.[6,7] Interestingly, the word *apple* stays relatively close during the entire time period as its word vector also (as *blackberry*) reflects both meanings, a fruit and a tech company.

4.5 Ex3: Structure Prediction

This subsection discusses the structure prediction performance by W2VPred. We first evaluate the prediction performance by using the a priori affinity structure as the *ground-truth* structure. The results of this experiment should be interpreted with care, because we have already seen in Section 4.3 that the given a priori affinity does not necessarily reflect the similarity structure of the slices in the corpus, in particular for WikiPhil. We then analyze the correlation between the embedding quality and the structure

[5] https://www.oxfordlearnersdictionaries .com/definition/english/power_1.

[6] https://www.businessinsider.com /blackberry-smartphone-rise-fall-mobile –failure-innovate-2019-11.

[7] https://www.businessinsider.com /blackberry-phone-sales-decline-chart –2016-9.

Nat. Sci	Eng&Tech	Soc. Sci	Hum	GloVe	Skip-Gram
generator	generator	powerful	powerful	control	Power
PV	inverter	control	control	supply	inverter
thermoelectric	alternator	wield	counterbalance	capacity	mover
inverter	converter	drive	drive	system	electricity
converter	electric	generator	supreme	internal	thermoelectric

Table 5: Five nearest neighbors to the word ''power'' in the domain-specific embedding space, learned by W2VPred, of four main categories of WikiFoS (left four columns), and in the general embedding space learned by GloVe and Skip-Gram on the entire dataset (right-most columns, respectively).
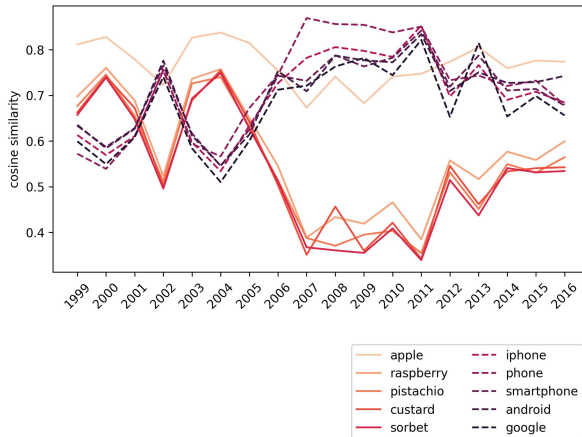

Figure 2: Evolution of the word *blackberry* in NYT. Nearest neighbors of the word *blackberry* have been selected in 2000 (blueish) and 2011 (reddish), and the embeddings have been computed with W2VPred. Cosine similarity between each neighboring word and *blackberry* is plotted over time, showing the shift in dominance between fruit and smartphone brand. The word *apple* also relates to both fruit and company, and therefore stays close during the entire time period.

Dataset Method	NYT	WikiFos	WikiPhil
GloVe	67.22	51.66	36.67
Skip-Gram	71.11	54.59	26.67
CBOW	65.28	45.00	23.33
W2VPred (our)	81.67	62.50	23.33
Burrows'delta	55.56	22.92	6.67

Table 6: Recall@k for structure prediction performance evaluation with the prior structure (Figure 1 left) used as the ground-truth.

prediction performance by W2VPred, in order to evaluate the a priori affinity as the ground-truth in each dataset. Finally, we apply W2VDen which combines the benefits of both W2VConstr and W2VPred for the case where the prior structure is not suitable.

4.5.1 Structure Prediction Performance

Here, we evaluate the structure prediction accuracy by W2VPred with the a priori given affinity matrix $D \in \mathbb{R}^{T \times T}$ (shown in the upper row of Figure 1) as the ground-truth. We report on recall@k averaged over all domains.

We compare our W2VPred with Burrows' Delta (Burrows, 2002) and other baseline methods based on the GloVe, Skip-Gram, and CBOW embeddings. Burrows' Delta is a commonly used method in stylometrics to analyze the similarity between

corpora, for example, for identifying the authors of anonymously published documents. The baseline methods based on GloVe, Skip-Gram, and CBOW simply learn the domain-specific embeddings separately, and the distances between the slices are evaluated by Eq. 4.

Table 6 shows recall@k (averaged over ten trials). As in the analogy tests, the best methods are in gray cells according to the Wilcoxon test. We see that W2VPred significantly outperforms the baseline methods for NYT and WikiFoS. For WikiPhil, we will further analyze the affinity structure in the following section.

4.5.2 Assessment of Prior Structure

In the following, we reevaluate the aforementioned prior affinity matrix for WikiPhil (see Figure 1). Therefore, we analyze the correlation between embedding quality and structure performance and find that a suitable ground truth affinity matrix is necessary to train good word embeddings with W2VConstr. We trained W2VPred with different parameter setting for (λ, τ) on the train set, and applied the global analogy tests and the structure prediction performance evaluation (with the prior structure as the ground-truth). For λ and τ, we considered log-scaled parameters in the ranges

Dataset	ρ
NYT	0.58
WikiFoS	0.65
WikiPhil	-0.19
WikiPhil (denoised)	-0.14

Table 7: Pearson correlation coefficients for performance on analogy tests ($n = 10$) and structure prediction evaluation (recall@k) by W2VPred for the parameters applied in the grid search for hyperparameter tuning. Linear correlation indicates that a good word embedding quality also leads to an accurate structure prediction (and vice versa). Significant correlation coefficients ($p < 0.05$) are marked in gray.

$[2^{-2} - 2^{12}]$ and $[2^4 - 2^{12}]$, respectively, and display correlation values on NYT, WikiFoS, and WikiPhil in Table 7.

In NYT and WikiFoS, we observe clear positive correlations between the embedding quality and the structure prediction performance, which implies that the estimated structure closer to the ground truth enhances the embedding quality. The Pearson correlation coefficients are 0.58 and 0.65, respectively (both with $p < 0.05$).

Whereas Table 7 for WikiPhil does not show a clear positive correlation. Indeed, the Pearson correlation coefficient is even negative with -0.19, which implies that the prior structure for WikiPhil is not suitable and even harmful for the word embedding performance. This result is consistent with the bad performance of W2VConstr on WikiPhil in Section 4.3.

4.5.3 Structure Discovery by W2VDen

The good performance of W2VPred on WikiPhil in Section 4.3 suggests that W2VPred has captured a suitable structure of WikiPhil. Here, we analyze the learned structure, and polish it with additional side information.

Figure 3 (left) shows the dendrogram of categories in WikiPhil obtained from the affinity matrix W learned by W2VPred. We see that the two pairs *Ethics-Social Philosophy* and *Cognition-Epistemology* are grouped together, and both pairs also belong to the same cluster in the original structure. We also see the grouping of *Epistemologists*, *Moral Philosophers*, *History of Logic*, and *Philosophers of Art*. This was at first glance surprising because they belong to four different
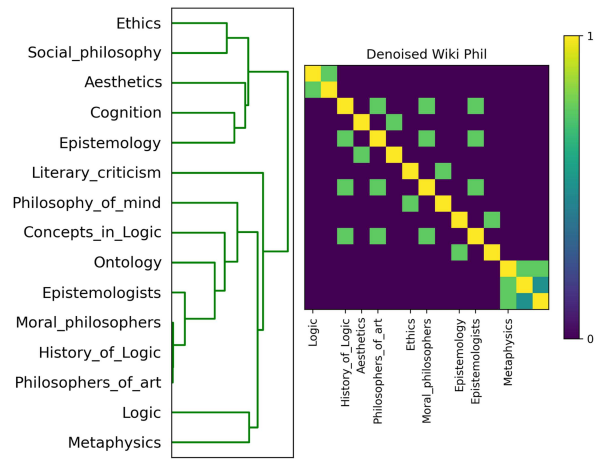

Figure 3: Left: Dendrogram for categories in WikiPhil learned by W2VPred based on the affinity matrix W. Right: Denoised Affinity matrix built from the learned structure by W2VPred. Newly formed Cluster includes *History of Logic*, *Moral Philosophers*, *Epistemologists*, and *Philosophers of Art*.

clusters in the prior structure. However, looking into the articles revealed that this is a logical consequence from the fact that the articles in those categories are almost exclusively about biographies of philosophers, and are therefore written in a distinctive style compared to all other slices.

To confirm that the discovered structure captures the semantic sub-corpora structure, we defined a new structure for WikiPhil, which is shown in Figure 3 (right), based on our findings above and also define a new structure for WikiFoS: A minor characteristic that we found in the structure of the prediction of W2VPred in comparison with the assumed structure is that the two sub-corpora Humanities and Social Sciences and the two sub-corpora Natural Sciences and Engineering are a bit closer than other combinations of sub-corpora, which also intuitively makes sense. We connected the two sub-corpora by connecting their root node respectively and then apply W2VDen. The general analogy tests performance by W2VDen is given in Table 3. In WikiFoS, the improvement is only slightly significant for $n = 5$ and $n = 10$ and not significant for $n = 1$. This implies that the structure that we previously assumed for WikiFoS already works well. This shows that applying W2VDen is in fact a general purpose method that can be applied on any of the data sets but it is especially useful when there is a mismatch between the assumed structure and the structure predicted by W2VPred. In WikiPhil,

	NYT						WikiFos					WikiPhil				
	GloVe	Skip-G.	CBOW	DW2V	W2VC	W2VP	GloVe	Skip-G.	CBOW	W2VC	W2VP	GloVe	Skip-G.	CBOW	W2VD	W2VP
RW-STANFORD	.36	.50	.55	.51	.52	.53	.38	.43	.42	.42	.42	.34	.43	.38	.38	.34
MTurk-771	.41	.54	.52	.58	.59	.59	.58	.60	.59	.62	.62	.49	.55	.49	.59	.58
RG-65	.46	.51	.40	.55	.53	.54	.60	.67	.60	.74	.72	.42	.50	.38	.63	.55
WS-353-ALL	.44	.56	.56	.57	.56	.57	.57	.65	.63	.62	.62	.53	.59	.57	.59	.58
MTR-3k	.50	.63	.60	.68	.68	.68	.66	.60	.59	.65	.65	.59	.56	.48	.62	.60
WS-353-REL	.34	.48	.47	.50	.49	.51	.52	.59	.57	.55	.55	.51	.54	.51	.55	.54
MC-30	.50	.56	.40	.58	.54	.54	.71	.68	.59	.77	.75	.47	.62	.54	.66	.58
YP-13	.38	.37	.31	.39	.40	.40	.33	.41	.43	.44	.43	.31	.40	.39	.43	.40
WS-353-SIM	.53	.64	.62	.64	.64	.64	.63	.67	.66	.70	.70	.59	.64	.61	.67	.64
MTurk-287	.55	.64	.64	.66	.67	.66	.44	.62	.64	.68	.69	.44	.63	.61	.66	.65
SimVerb-350	.15	.22	.25	.23	.23	.23	.17	.17	.20	.21	.21	.11	.19	.17	.19	.18
SIMLEX-999	.26	.32	.33	.30	.30	.30	.28	.28	.29	.29	.29	.23	.27	.25	.27	.26
Count	–	2	3	7	5	7	1	3	–	8	7	–	4	–	9	1

Table 8: Correlation values from word similarity tests on different datasets (one per row). The best method and the methods that are not significantly outperformed by the best is marked with gray background, according to the Wilcoxon signed rank test for $\alpha = 0.05$. In this table, we use a shorter version of the method names (W2VC for W2VConstr, etc.)

	NYT	WFos	WPhil
GloVe	0.26	0.29	0.27
Skip-Gram	0.28	0.30	0.29
CBOW	0.29	0.31	0.29
Dynamic Word2Vecabbr	0.28	—	—
W2VConstr (our)	0.29	0.32	—
W2VDen (our)	—	—	0.30
W2VPred (our)	0.28	0.32	0.29

Table 9: QVEC results: Correlation values of the aligned dimension between word embeddings and linguistic word vectors.

we see that W2VDen further improves the performance by W2VPred, which already outperforms all other methods with a large margin. The correlation between the embedding quality and the structure prediction performance—with the denoised estimated affinity matrix as the ground truth—is shown in Table 7. The Pearson correlation is still negative, -0.14, but no longer statistically significant ($p = 0.11$).

4.6 Ex4: Evaluation in Word Similarity Tasks

We further evaluate word embeddings on various word similarity tasks where human-annotated similarity between words is compared with the cosine similarity in the embedding space, as proposed in Faruqui and Dyer (2014). Table 8 shows the correlation coefficients between the human-annotated similarity and the embedding cosine similarity, where, again, the best method and the runner-ups (if not significantly outperformed) are highlighted.[8] We observe that W2VPred outperforms the other methods in 7 out of 12 datasets for NYT, and W2VConstr in 8 out of 12 for WikiFoS. For WikiPhil, since we already know that W2VConstr with the given affinity matrix does not improve the embedding performance, we instead evaluated W2VDen, which outperforms 9 out of 12 datasets in WikiPhil. In addition, W2VPred gives comparable performance to the best method over all experiments.

We also apply QVEC, which measures component-wise correlation between distributed word embeddings, as we use them throughout the paper, and linguistic word vectors based on WordNet (Fellbaum, 1998). High correlation values indicate high saliency of linguistic properties and thus serve as an intrinsic evaluation method that has been shown to highly correlate with downstream task performance (Tsvetkov et al., 2015). Results are shown in Table 9, where we observe that W2VConstr (as well as W2VDen for WikiPhil) outperforms all baseline methods, except CBOW in NYT, on all datasets, and W2VPred performs comparably with the best method.

---

[8]We removed the dataset *VERB-143* since we are using lemmatized tokens and therefore catch only a very small part of this corpus. We acknowledge that the human annotated similarity is not domain-specific and therefore not optimal for evaluating the domain-specific embeddings. However, we expect that this experiment provides another aspect of the embedding quality.

329

4.7 Summarizing Discussion

In this section, we have shown a good performance of W2VConstr and W2VPred in terms of global and domain-specific embedding quality on news articles (NYT) and articles from Wikipedia (WikiFoS, WikiPhil). We have also shown that W2VPred is able to extract the underlying sub-corpora structure from NYT and WikiFoS.

On the WikiPhil dataset, the following observations implied that the prior sub-corpora structure, based on the Wikipedia's definition, was not suitable for analyzing semantic relations:

- Poor general analogy test performance by W2VConstr (Table 3),
- Low structure prediction performance by all methods (Table 6)
- Negative correlation between embedding accuracy and structure score (Table 7).

Accordingly, we analyzed the learned structure by W2VPred, and further refined it by *denoising* with human intervention. Specifically, we analyzed the dendrogram from Figure 3, and found that 4 categories are grouped together that we originally assumed to belong to 4 different clusters. We further validated our reasoning by applying W2VDen with the structure shown in Figure 3 resulting in the best embedding performance (see Table 3).

This procedure poses an opportunity to obtain good global and domain-specific embeddings and extract, or validate if given a priori, the underlying sub-corpora structure by using W2VConstr and W2VPred. Namely, we first train W2VPred, and also W2VConstr if prior structure information is available. If both methods similarly improve the embeddings in comparison with the methods without using any structure information, we acknowledge that the prior structure is at least useful for word embedding performance. If W2VPred performs well, while W2VConstr performs poorly, we doubt that the given prior structure would be suitable, and update the learned structure by W2VPred. When no prior strucutre is given, we simply apply W2VPred to learn the structure.

We can furthermore refine the learned structure with side information, which results in a clean and human interpretable structure. Here W2VDen is used to validate the new structure, and to provide

enhanced word embeddings. In our experiment on the WikiPhil dataset, the embeddings obtained this way significantly outperformed all other methods. The improved performance from W2VPred is probably due to the fewer degrees of freedom of W2VConstr, that is, once we know a reasonable structure, the embeddings can be more accurately trained with the fixed affinity matrix.

5 Application on Digital Humanities

We propose an application of W2VPred to the field of Digital Humanities, and develop an example more specifically related to Computational Literary Studies. In the renewal of literary studies brought by the development and implementation of computational methods, questions of authorship attribution and genre attribution are key to formulating a structured critique of the classical design of literary history, and of Cultural Heritage approaches at large. In particular, the investigation of historical person networks, knowledge distribution, and intellectual circles has been shown to benefit significantly from computational methods (Baillot, 2018; Moretti, 2005). Hence, our method and its capability to reveal connections between sub-corpora (such as authors' works), can be applied with success to these types of research questions. Here, the use of quantitative and statistical models can lead to new, hitherto unfathomed insights. A corpus-based statistical approach to literature also entails a form of emancipation from literary history in that it makes it possible to shift perspectives, e.g., to reconsider established author-based or genre-based approaches.

To this end, we applied W2VPred to high literature texts (*Belletristik*) from the lemmatized versions of DTA (German Text Archive), a corpus selection that contains the 20 most represented authors of the DTA text collection for the period 1770-1900. We applied W2VPred in order to predict the connections between those authors with $\lambda = 512, \tau = 1024$ (same as WikiFoS).

As a measure of comparison, we extracted the year of publication as established by DTA, and identified the place of work for each author[9] and categorized each publication into one of three genre categories (ego document, verse, and fiction). Ego documents are texts written in the first person that document personal experience in their

---

[9] via the German Integrated Authority Files Service (GND) where available, adding missing data points manually.
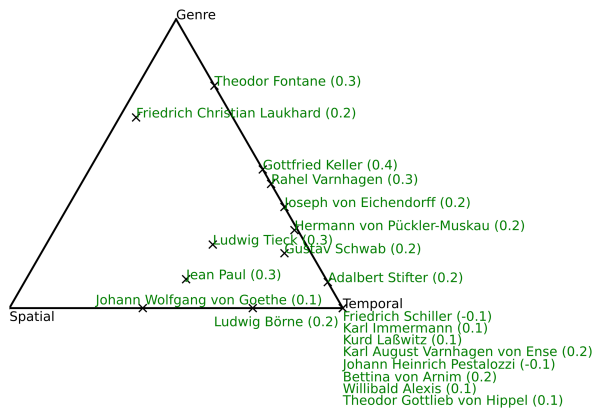
Figure 4: Author's points in a barycentric coordinates triangle denote the mixture of the prior knowledge that has the highest correlation (in parentheses) with the predicted structure of W2VPred. The correlation excludes the diagonal, meaning the correlation between the author itself.

historical context. They include letters, diaries, and memoirs and have gained momentum as a primary source in historical research and literary studies over the past decades. We created pairwise distance matrices for all authors based on the spatial, temporal, and genre information. Temporal distance was defined as the absolute distance between the average publication year, the spatial distance as the geodesic distance between the average coordinates of the work places for each author and the genre difference as cosine distance between the genre proportions for each author. For each author, we correlated linear combinations of this (normalized) spatio-temporal-genre prior knowledge with the structure found by our method, which we show in Figure 4.

Reference Dimensions In this visualization we want to compare the pairwise distance matrix that our method predicted with the distance matrices that can be obtained by meta data available in the DTA corpus—the reference dimensions:

1. Temporal difference between authors. We collect the publication year for each title in the corpus and compute the average publication year for each author. The temporal distance between one author A_{t1} and another author A_{t2} is computed by $|A_{t1} - A_{t2}|$, the absolute difference of the average publication year.

2. Spatial difference between authors. We query the German Integrated Authority File for the authors' different work places and extract

them as longitude and latitude coordinates on the earths surface. We compute the average coordinates for each author by converting the coordinates into the Cartesian system and take the average on each dimension. Then, we convert the averages back into the latitude, longitude system. The spatial distance between two authors is computed by the geodesic distance as implemented in GeoPy.[10]

3. Genre difference between authors. We manually categorized each title in the corpus into one of the three categories ego document, verse, and fiction. A genre representation for an author $A_g = (A_{g_{ego}}, A_{g_{verse}}, A_{g_{fiction}})$ is the relative frequency of the respective genre for that author. The distance between one author A_{g1} and another author A_{g2} is computed by $1 - \frac{A_{g1} \cdot A_{g2}}{||A_{g1}|| \cdot ||A_{g2}||}$, the cosine distance.

Calculating the Correlations For each author t, we denote the predicted distance to all other authors as $X_t \in \mathbb{R}^{T-1}$ where T is the number of all authors. $Y_t \in \mathbb{R}^{(T-1) \times 3}$ denotes the distances from the author t to all other authors in the three meta data dimensions: space, time, and genre. For the visualization, we seek for the coefficients of the linear combination of Y that has the highest correlation with X. For this, Non-Negative Canonical Correlation Analysis with one component is applied. The MIFSR algorithm is used as described by Sigg et al. (2007).[11] The coefficients are normalized to comply with the sum-to-one constraint for projection on the 2d simplex.

For many authors, the strongest correlation occurs with a mostly temporal structure and fewer correlate strongest with the spatial or the genre model. Börne and Laukhard, who have a similar spatial weight and thereby form a spatial cluster, both resided in France at that time. The impact of French literature and culture on Laukhard's and Börne's writing deserves attention, as suggested by our findings.

For Fontane, we do not observe a notable spatial proportion, which is surprising because his sub-corpus mostly consists of ego documents describing the history and geography of the area surrounding Berlin, his workplace. However, in

---

[10] https://geopy.readthedocs.io/en/stable/.

[11] We use $\epsilon = .00001$.

contrast to the other authors residing in Berlin, the style is much more similar to a travel story. In W2VPred 's predicted structure, the closest neighbor of Fontane is, in fact, Pückler (with a distance of .052), who also wrote travel stories.

In the case of Goethe, the maximum correlation at the (solely spatio-temporal) resulting point is relatively low and, interestingly, the highest disagreement between W2VPred and the prior knowledge is between Schiller and Goethe. The spatio-temporal model represents a close proximity; however, in W2VPred's found structure, the two authors are much more distant. In this case, the spatio-temporal properties are not sufficient to fully characterize an author's writing and the genre distribution may be skewed due to the incomplete selection of works in the DTA and due to the limitations of the labeling scheme, as in the context of the 19th century, it is often difficult to distinguish between ego documents and fiction.

Nonetheless, we want to stress the importance of the analysis where linguistic representation and structure, captured in W2VPred, is in line with these properties and, also, where they disagree. Both agreement and disagreement between the prior knowledge and the linguistic representation found by W2VPred can help identifying the appropriate ansatz for a literary analysis of an author.

6 Conclusion

We proposed novel methods to capture domain-specific semantics, which is essential in many NLP tasks: Word2Vec with Structure Constraint (W2VConstr) trains domain-specific word embeddings based on prior information on the affinity structure between sub-corpora; Word2Vec with Structure Prediction (W2VPred) goes one step further and predicts the structure while learning domain-specific embeddings simultaneously. Both methods outperform baseline methods in benchmark experiments with respect to embedding quality and the structure prediction performance. Specifically, we showed that embeddings provided by our methods are superior in terms of global and domain-specific analogy tests, word similarity tasks, and the QVEC evaluation, which is known to highly correlate with downstream performance. The predicted structure is more accurate than the baseline methods including Burrows' Delta. We also proposed and successfully demonstrated a procedure, Word2Vec with

Denoised Structure Constraint (W2VDen), to cope with the case where the prior structure information is not suitable for enhancing embeddings, by using both W2VConstr and W2VPred. Overall, we showed the benefits of our methods, regardless of whether (reliable) structure information is given or not. Finally, we were able to demonstrate how to use W2VPred to gain insight into the relation between 19th century authors from the German Text Archive and also how to raise further research questions for high literature.

A Implementation Details

A.1 Ex1

All word embeddings were trained with $d = 50$.

GloVe We run GloVe experiments with $\alpha = 100$ and minimum occurrence $= 25$.

Skip-Gram, CBOW We use the Gensim (Řehůřek and Sojka, 2010) implementation of Skip-Gram and CBOW with min_alpha $= 0.0001$, sample $= 0.001$ to reduce frequent words and for Skip-Gram, we use 5 negative words and ns_component $= 0.75$.

Parameter Selection The parameters λ and τ for DW2V, W2VConstr and W2VPred were selected based on the performance in the analogy tests on the train set. In order to flatten the contributions from the n nearest neighbors (for $n = 1, 5, 10$), we rescaled the accuracies: For

each n, accuracies are scaled so that the best and the worst method is 1 and 0, respectively. Then, we computed their average and maximum.

Analogies Each analogy consists of two word pairs (e.g., countryA - capitalA; countryB - capitalB). We estimate the vector for the last word by $\widehat{v}$ = capitalA - countryA + countryB, and check if capitalB is contained in the n nearest neighbors of the resulting vector $\widehat{v}$.

A.2 Ex2

Temporal Analogies Each of two word pairs consists of a year and a corresponding term, as for example, 2000 - Bush; 2008 - Obama, and the inference accuracy of the last word by vector operations on the former three tokens in the embedded space is evaluated. To apply these analogies, GloVe, Skip-Gram, and CBOW are trained individually on each year on the same vocabulary as W2VPred (same parameters for GloVe as before, with minimum occurrence=10). For the other methods, DW2V, W2VConstr, and W2VPred, we can simply use the embedding obtained in Section 4.3. Note that the parameters τ and λ were optimized based on the general analogy tests.

A.3 Ex3

Burrows It compares normalized bag-of-words features of documents and sub-corpora, and provides a distance measure between them. Its parameters specify which word frequencies are taken into account. We found that considering the 100th to the 300th most frequent words gives the best structure prediction performance on the train set.

Recall@k Let $\hat{D} \in \mathbb{R}^{T \times T}$ be the predicted structure. We report on recall@k averaged over all domains:

$$\text{recall@k} = \frac{1}{T}\sum_{t}^{T}\text{recall@k}_t, \quad \text{where}$$

$$\text{recall@k}_t = \frac{\text{TP}_t(k)}{\text{TP}_t(k) + \text{FN}_t(k)},$$

$$\text{TP}_t(k) = \sum_{t'}^{T} b(D_t, t', k) \ \& \ b(\hat{D}_t, t', k),$$

$$\text{FN}_t(k) = \sum_{t'}^{T} b(D_t, t', k) \ \& \ \neg b(\hat{D}_t, t', k), \text{and}$$

$$b(x, i, k) = \begin{cases} 1 & x_i \text{ is one of the k smallest in } x, \\ 0 & \text{otherwise.} \end{cases}$$

For NYT, we chose $k = 2$, which means relevant nodes are the two next neighbors, that is,

the preceding and the following years. For WikiFoS and WikiPhil, we respectively chose $k = 3$ and $k = 2$, which corresponds to the number of subcategories that each main category consists of.

W2VPred Hyperparameters for W2VPred were selected on the train set where we maximized the accuracy on the global analogy test as before.

References

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518. https://doi.org/10.1145/3132847.3132878

Anne Baillot. 2018. Die Krux mit dem Netz Verknüpfung und Visualisierung bei digitalen Briefeditionen. In Toni Bernhart, Marcus Willand, Sandra Richter, and Andrea Albrecht, editors, *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*, pages 355–370. De Gruyter. https://doi.org/10.1515/9783110523300-016

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. *arXiv preprint arXiv:1702.08359*.

Erik Bleich, Hasher Nisar, and Rana Abdelhamid. 2016. The effect of terrorist events on media portrayals of Islam and Muslims: Evidence from New York Times headlines, 1985–2013. *Ethnic and Racial Studies*, 39(7):1109–1127. https://doi.org/10.1080/01419870.2015.1103886

John Burrows. 2002. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287. https://doi.org/10.1093/llc/17.3.267

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://aclanthology.org/N19-1423

Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 19–24. https://doi.org/10.3115/v1/P14-5004

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database and some of its applications. https://doi.org/10.7551/mitpress/7287.001.0001

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555. https://doi.org/10.18653/v1/2020.acl-main.51

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with Wasserstein Procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR.

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. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1141

Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze. 2020. Dynamic contextualized word embeddings. *arXiv preprint arXiv:2010.12684*. https://doi.org/10.18653/v1/2021.acl-long.542

Ganesh Jawahar and Djamé Seddah. 2019. Contextualized diachronic word representations. In

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 35–47. https://doi.org/10.18653/v1/W19-4705

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*. https://doi.org/10.18653/v1/E17-2068

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*. https://doi.org/10.3115/v1/W14-2517

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint 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*, pages 625–635. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2736277.2741627

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

Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. 2017. Content analysis of 150 years of british periodicals. *Proceedings of the National Academy of Sciences*, 114(4):E457–E465. https://doi.org/10.1073/pnas.1606380114, PubMed: 28069962

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

Jani Marjanen, Lidia Pivovarova, Elaine Zosa, and Jussi Kurunmäki. 2019. Clustering ideological terms in historical newspaper data with diachronic word embeddings. In *5th International Workshop on Computational History, HistoInformatics 2019*. CEUR-WS.

334

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Franco Moretti. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162

Stephen D. Reese and Seth C. Lewis. 2009. Framing the war on terror: The internalization of policy in the US press. *Journalism*, 10(6):777–797. https://doi.org/10.1177/1464884909344480

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1003–1011. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3185999

Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. 2016. Exponential family embeddings. In *Advances in Neural Information Processing Systems*, pages 478–486.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76. https://doi.org/10.18653/v1/D19-1007

C. Sigg, B. Fischer, B. Ommer, V. Roth, and J. Buhmann. 2007. Non-negative CCA for audio-visual source separation. In *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*. https://doi.org/10.1109/MLSP.2007.4414315

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

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2049–2054. https://doi.org/10.18653/v1/D15-1243

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 673–681. ACM. https://doi.org/10.1145/3159652.3159703, PubMed: 30409101

Ziqian Zeng, Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Socialized word embeddings. In *IJCAI*, pages 3915–3921. https://doi.org/10.24963/ijcai.2017/547

Yating Zhang, Adam Jatowt, Sourav S. Bhowmick, and Katsumi Tanaka. 2016. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807. https://doi.org/10.1109/TKDE.2016.2591008