The Best of Both Worlds: Context-Powered Word Embedding Combinations for Longitudinal Text Analysis

Laura Rettig*, Regula Hänggli†, Philippe Cudré-Mauroux‡
* eXascale Infolab, Department of Computer Science
† Department of Mass Media and Communication Research
University of Fribourg
Fribourg — Switzerland
Email: {firstname.lastname}@unifr.ch

***Abstract*—We propose a series of methods for combining word embeddings trained on small corpora with stable embeddings from a large reference corpus, with the express purpose of preserving certain properties present in the smaller text corpora. Our methods aim to leverage i) the specificity of the representation of certain terms in a word embedding based on a small corpus pertaining to a specific subject matter while ii) transferring the stability provided by vector representations learned from a larger corpus (e.g., Wikipedia). We achieve this aim by quantifying the relevance of a word to the corpus as well as evaluating the stability of its context within the original corpus. From the scores obtained for each word, we obtain a stable—yet specific—vector that is based on a weighted combination of the aligned base vector (obtained from a large corpus) and the more specific vector (learned from a smaller corpus). We evaluate our method on state-of-the-art semantic tasks and show that we are able to retain the stability of the original corpus. In order to evaluate that our methods further retain the specificity of certain terms in the original subject matter, we demonstrate that our method aids interdisciplinary analytical tools leveraging and comparing word embeddings for researchers in other domains, such as in the understanding of the evolution of public debates on economic issues over time.**

***Index Terms*—word embeddings, natural language processing, corpus analysis, diachronic word embeddings**

I. Introduction

Word embeddings learned from text corpora that are limited in size, but which pertain to a specific topic of interest, struggle to properly represent the distributional semantics of many terms due to their low occurrence in the corpus. On the other hand, analyzing small text corpora pertaining to a particular topic or created at a particular point in time is key to many linguistic endeavors, for instance when conducting longitudinal or diachronic studies.

In fact, there are multiple reasons one might wish to work with word embeddings trained on smaller corpora. These small corpora can be constructed from languages where little data is available, on text pertaining to a niche or vertical domain, such as biomedical studies, or, as the example we will be investigating in this paper, may consist of news articles that are taken at specific points in time, showcasing public debates on political and economical issues. Being able to construct resilient embeddings from such news articles allows us to track changes in meanings, but also in a society's attitude towards an issue, over time. By aligning and comparing between news models from different time spans, researchers from other domains, such as communication or political sciences, can study these changes in an automated manner, whereas classical methods involve manually reading and summarizing large collections of text.

When leveraging embeddings trained on smaller corpora, one can of course directly observe how word neighborhoods change over time. However, the problem there is that these changes may simply be due to an instability in the resulting vectors—when little training data is given, word vectors do not necessarily best represent the semantics of all terms and are known to be unstable [1].

By figuring out which terms are adequately represented, and are to be preserved, and which terms are inadequately represented or underrepresented in a given corpus, however, one can potentially get the best of both worlds by aligning and combining different embeddings. In this paper, we provide such a solution; we analyze smaller text corpora (and their embeddings) leveraging stable embeddings obtained from a larger text corpus.

Specifically, we propose to align the embeddings obtained from a series of small text corpora to those trained on a larger, more stable encyclopedic corpus. Subsequently, we put forward different techniques to combine the embeddings from the small and large corpora on a per-term basis, in order to leverage the stability of the embeddings obtained from the large corpus while capturing the peculiarities and idiosyncrasies from the smaller (but more specific) text corpora.

In summary, our main contributions are as follows:

- we suggest that performing a weighted, per-term combination of word embeddings improves the transfer of

978-1-7281-6251-5/20/$31.00

linguistic properties from a novel corpus to a reference corpus over several baselines;
- we introduce three different methods to determine the weights of the particular terms, taking into account their respective frequencies, their relevance w.r.t. the overall corpus, and their stability in terms of textual context (Section III);
- finally, we demonstrate experimentally that our methods outperform baselines on word semantic tasks while also retaining the properties of smaller text corpora for analytical purposes (Section IV).

II. Related Work

Our work lies at the crossroad of two lines of research: combining word vectors on one hand, and capturing the semantic shift in diachronic word embeddings on the other. We briefly present the state of the art in both domains below, as well as some of their applications.

a) Word vector combinations: Rettig et al. [2] perform embedding combinations with the aim of transferring domain knowledge. They eschew the alignment problem by concatenating the corresponding vectors and performing principal component analysis (PCA) over the concatenated vectors to preserve a lower dimensionality. They also propose a metric to compare entire corpora in terms of their similarity by introducing a modified tf-idf value as word weights and treating the corpora as a distribution over terms. Similarly to that paper, we also use PCA for visualization in this work but perform alignment of the vector spaces by introducing new techniques to fine-tune the combination of vectors on a term-per-term basis.

Muromägi et al. [3] propose to combine several word embedding models into an ensemble by using least squares regression or matrix approximation techniques iteratively. They evaluate their approach on Estonian, a language for which large training corpora are not readily available, and observe that the ensemble models based on matrix approximation yield better results on synonym and analogy tasks. We perform the same alignment procedure to solve the Orthogonal Procrustes problem and show that a weighted combination performs better than the proposed unweighted alignment.

A simple yet efficient method to adapt reference models to novel data was recently put forward by Bojanowski et al. [4]. The authors suggest performing monolingual word vector alignment followed by averaging the resulting vectors. Compared against baselines—retraining on joint corpora and fine-tuning the training of the embeddings for the novel corpus—they show that their approach improves the performance on word analogy tasks while also being simpler. We empirically show in Section IV that the method we introduce in this paper, leveraging a more expressive weighted combination of word vectors instead of simple averaging by deriving weights from the immediate surroundings of a term in its original corpus, outperforms [4] in our context, especially when used to combine embeddings trained on small corpora.

b) Capturing semantic shift in embeddings: Yao et al. [5] study word evolution by developing a statistical model to learn time-aware word vector representations. Taking New York Times articles between 1990 and 2016, they do not combine or align different models but suggest instead to learn embeddings in all time slices jointly and to apply regularization terms to smooth embedding changes across time. While computationally expensive, their method yields high-fidelity embeddings, and the authors demonstrate in particular that their embeddings are robust against sudden undersampling of specific time slices.

A method for tracking and detecting statistically significant linguistic shifts in the meaning and usage of words using distributional characteristics inferred from word co-occurrences was introduced by Kulkani et al. [6]. The authors leverage linear transformations to align all word representations from the different time snapshots to the initial embedding space, and then capture linguistic shift by constructing time series capturing distances in the embedding space across time. They demonstrate the scalability of their approach by tracking linguistic change across years of Twitter messages, a decade of product reviews and a century of written books using the Google Book-ngrams. Our proposed method serves rather exploratory purposes and as such leverages specific, small-corpus embeddings.

Along similar lines, Zhang et al. [7] tackle temporal counterpart search, which given an input term returns semantically similar terms from the past in longitudinal document collections. Their approach is based on embeddings and vector space transformations mapping the word representation of present text to those from the past. The authors propose a technique for automatically constructing seed pairs of terms that are used to identify the proper transformation, and test their approach over time frames as long as 100 years.

c) Applications of diachronic word embeddings: Mining text to obtain diachronic word embeddings are of interest in many interdisciplinary settings. In this paper, we focus on their potential application to political science. As reported by Hänggli et al. [8], understanding the link between policies and discourse is key to our knowledge of political communications. In that sense, the authors analyze the degree of diversity and the communication style of debates in the media. As most studies in that field, the authors resort to experts and manual analyses to evaluate the debates, a time-consuming task that could be somewhat alleviated by automated techniques, as we show in Section IV.

Along similar lines, Boydstun et al. [9] point out the importance of tracking political attention by analyzing how attention is distributed across several issues, instead of focusing on a single issue such as front-page coverage of a specific term. The authors study how to measure attention diversity, and point out

that despite its importance, the community lacks a standard for how best to measure it.

Baumgartner et al. [10] study the media coverage of the death penalty in the United States over time. They suggest the use of Evolving Factor Analysis to identify sets of terms that cluster over time. Much of their analysis is hand-coded, however. The techniques we develop in this work could be used to automate part of that process, by automatically tracking the use and neighborhoods of terms over time.

III. Method

In this work, we aim to obtain combined word embedding vectors across different textual corpora in order to conduct diachronic studies on word semantics.

Given a reference embedding $\mathcal{E}_{ref}$ that is general-purpose and trained on a large corpus $\mathcal{T}_{ref}$, as well as a potentially small novel corpus $\mathcal{T}_{new}$ and its respective word vectors $\mathcal{E}_{new}$, we aim to obtain a new set of word vectors $\mathcal{E}'$ by updating the vectors in the reference model with new properties from the novel embedding.

A. Vector Space Alignment

In order to combine embedding spaces, $\mathcal{E}_{ref}$ and $\mathcal{E}_{new}$ must be aligned to bring vectors representing the same terms as close together as possible in a target vector space. We perform the alignment by solving the orthogonal Procrustes problem [11] on the vector spaces, the state-of-the-art method for aligning mono- and multilingual word embeddings [3], [4], [12]–[14]. We solve the problem by performing singular value decomposition (SVD) on ${\mathcal{E}_{new}}^T\mathcal{E}_{ref}$ (as such fixing $\mathcal{E}_{ref}$), yielding USV^T, from which we can obtain the orthogonal transformation matrix

$$P = UV^T \tag{1}$$

such that the difference between $\mathcal{E}_{ref}(w)$ and $P \cdot \mathcal{E}_{new}(w)$ is minimized for all $w \in \mathcal{T}_{ref} \cap \mathcal{T}_{new}$, that is, the intersection of the vocabularies between the corpora.

There will, however, be differences in the aligned vector spaces. In the case of terms that change in meaning between the corpora, these differences are desirable; in other cases, where the term is underrepresented in the corpus to be analyzed and therefore the embedding is of low quality, this difference is to be minimized. We tackle those issues in the following.

B. Term Relevance and Quality Measures

In order to conduct diachronic studies and measure the quality of a word vector $\mathcal{E}(w)$ and a term's relevance in the original corpus $\mathcal{T}$, we first need to evaluate the term in the context of the original corpus it appears in. The context of the corpus—whether it be the immediate term neighborhood or the overall collection of texts—provides a plethora of information for a preliminary estimation of the quality of a resulting embedding for a given term. A term that occurs rarely within a corpus is unlikely to yield a stable vector space representation due to lack of training data. Similarly, as embedding algorithms such as *word2vec* utilize the context words in learning a term's vector representation, the immediate surroundings of a term also have an impact on its fidelity—if context words exhibit low consistency, i.e., if a term appears in a variety of contexts, its semantic representation is negatively affected.

In the following, we present three methods for measuring the relevance and estimated quality of a term and its vector, respectively, in the context of its original corpus. These measures yield weights α that are then used in weighted combinations. The different measures are then compared empirically in Section IV.

1) Relative term frequency: Intuitively, words that are relevant to the subject of a new corpus should have a higher relative frequency compared to the general-purpose reference corpus. On the other hand, terms that have a lower frequency than in the reference corpus are likely to yield vectors of lower quality.

Using this intuition, for every unique word $w \in \mathcal{T}_i$, we assign a weight

$$\alpha_i(w) = \frac{\#\mathcal{T}_i(w)}{|\mathcal{T}_i|} \tag{2}$$

which quantifies how often w appears relative to the total number of words in $\mathcal{T}_i$.

2) Term-corpus relevance: We can refine the relative frequency to our problem by taking into consideration a term's frequency in one corpus $\mathcal{T}_i$ compared to its frequency in another corpus $\mathcal{T}_j$, giving more information as to the specificity of w for corpus $\mathcal{T}_i$, as suggested in [2]. We obtain a score

$$\alpha_i(w) = \frac{\#\mathcal{T}_i(w)}{\#\mathcal{T}_i(w) + \#\mathcal{T}_j(w)} \frac{|\mathcal{T}_i| + |\mathcal{T}_j|}{|\mathcal{T}_i|}, \tag{3}$$

where $\alpha_i(w)$ represents the ratio between the frequency of w in $\mathcal{T}_i$ and its frequency over both corpora. A word that appears frequently in the novel corpus, but relatively rarely in the reference corpus, will have higher weight, assuming that it is more representative of the topic at hand.

3) Context stability: Intuitively, for a term to yield a stable embedding representation it needs to appear within relevant and stable contexts in the training corpus, that is, it needs to frequently co-occur with similar terms while not appearing too frequently with noisy terms. Pointwise Mutual Information (PMI) [15] captures this precise association—words that appear jointly with a higher probability than individually receive higher PMI scores and are thus deemed more stable within the chosen corpus.

We quantify the stability of the context of a term by evaluating the relationship between a given word w_i and its context words v_i appearing in a n-word window S_w before and after the occurrence of the term within the corpus $\mathcal{T}_i$, using

PMI to reflect context relevance. For every word w in the vocabulary, we gather its context terms into a set S_w. For every context word v, we compute PMI, which is then normalized to yield a score as

$$\alpha_i(w) = \frac{\sum_{v \in S_w} PMI(w, v)}{|S_w|}. \quad (4)$$

C. Weighted Vector Combinations

Equations 2, 3 and 4 assign a score $\alpha_i(w)$ to each word w in $\mathcal{T}_i$ and $\alpha_j(w)$ for words w_j in $\mathcal{T}_j$. In order to combine two aligned vectors from the corresponding embeddings $\mathcal{E}_i$ and $\mathcal{E}_j$, these scores need to be normalized to sum up to 1:

$$\alpha_i(w)' = \frac{\alpha_i(w)}{\alpha_i(w) + \alpha_j(w)}. \quad (5)$$

In order to obtain the vector for a word w in $\mathcal{E}'$, we thus compute

$$\mathcal{E}'(w) = \alpha_i(w)' * \mathcal{E}_i(w) + \alpha_j(w)' * (\mathcal{E}_j(w) \cdot P). \quad (6)$$

In total, the combined embedding $\mathcal{E}'(w)$ is obtained as follows:

1) Take a pair of $\mathcal{E}_{ref}$ and $\mathcal{E}_{new}$ and compute their alignment matrix P as in 1.
2) Process $\mathcal{T}_{ref}$ and $\mathcal{T}_{new}$ in order to get the weights $\alpha_{ref}(w)$ and $\alpha_{new}(w)$, according to 2, 3 or 4, for each word in each corpus.
3) Normalize the scores as in 5 and obtain a set of vectors by applying 6 to each word in the union of all words in the vocabularies of the two corpora; if a word is missing in one corpus, the vector from the other corpus is taken with a weight of 1.

IV. Evaluation

This section analyzes in detail the performance of our proposed combined word embedding models, showing that our method is capable of retaining the stability and quality of a reference embedding (subsection IV-A) while also transferring information such as word evolution over time from smaller text corpora (subsection IV-B).

TABLE I
Dataset specifications.

	English	**German**
Time period	1995-2016	1997-2018
Articles in total	ca. 2,000,000	85,424
Articles per 5-year time slice	ca. 430,000	ca. 17,000

a) Data: We evaluate our approach on two languages, English and German. We pick the collection of all Wikipedia articles in a respective language as reference corpus $\mathcal{T}_{ref}$, as this gives us a general-purpose collection of training data of sufficient size. As novel corpora $\mathcal{T}_{new}$, we choose collections of newspaper articles. In English, we take a corpus of 2 million New York Times articles published between 1995 and 2016 [5]. In German, we curated a collection of 85,424 articles from several Swiss newspapers published between 1997 and 2018. From these collections of articles, we build smaller corpora of articles published within shorter time frames, in order to allow us to analyze the evolution of language over time.

b) Setup: We tokenize sentences [16] and words [17] and lowercase all text. We then utilize the skip-gram variant of the well-known word2vec algorithm [18] with a window size of 5 and 300 dimensions to learn embeddings from the text corpora[1]. Words appearing fewer than 5 times are removed. For the *context stability* method, we remove any word pairs appearing only once from PMI. The window size for context words in this method is also configured to 5.

A. Semantic Task Performance

First, we evaluate the performance of our combined embeddings on state-of-the-art word similarity tasks, that is, we measure how well distances between pairs of words in the embedding reflect human evaluations of their lexical semantic relationships. We evaluate against the ground-truth datasets *SimLex-999* [19] and *WordSim353* [20], both of which are also available in German [21]. For this evaluation, we report the Spearman correlation between the ground-truth similarity scores of the words and the cosine distance of their corresponding embeddings in the vector space.

We compare the performance of our weighted combination methods against two baselines:

- **Unaligned**: We average the corresponding word vectors per dimension without aligning their respective vector spaces [22].
- **Unweighted**: We align the vector spaces as described in section III-A and then average the vectors for corresponding words on each dimension [4].

We also compare against the performance of the reference embeddings *wiki* used individually. It should be noted that the general-purpose reference embeddings, i.e., *wiki*, are expected to perform better on the given tasks. However, they fail to address our proposed analytical tasks over time or on the specific corpora and as such are of no use beyond enhancing the smaller, more specific corpora. Our aim in comparing against this general baseline lies thus more in showing a range of improvement compared to the evaluation of the novel corpora individually. We also evaluate the performance of our novel *news* corpora individually, which due to their small training data sets are expected to perform significantly worse than when combined with a reference corpus.

We split the news corpora into 5-year time slices and report their average performances individually. For all methods

[1]Our aim here is not to innovate in terms of training method, but rather to show that our alignment techniques enable us to conduct diachronic studies taking advantage of standard embedding algorithms.

TABLE II
COMPARISON OF THE PERFORMANCE OF BASELINE AND PROPOSED METHODS ON WORD SIMILARITY TASKS (SPEARMAN CORRELATION).

	SimLex-999	WordSim-353
German		
wiki	0.398	0.606
news, 5-year	0.129	0.253
unaligned	0.251	0.382
unweighted	0.256	0.401
relative frequency	0.269	0.444
term-corpus relevance	0.209	0.365
context stability	0.295	0.415
English		
wiki	0.308	0.657
news, 5-year	0.320	0.565
unaligned	0.313	0.616
unweighted	0.325	0.641
relative frequency	0.326	0.626
term-corpus relevance	0.326	0.626
context stability	0.311	0.542

(baseline and new), the combination is performed based on the 5-year news corpora aligned to the Wikipedia reference embedding. This results in five 5-year models for each language (English: 1995-2016, German: 1997-2018). Each 5-year corpus contains on average 17k (German) and 430k (English) news articles. The results reported in Table II list the average Spearman correlation for the results obtained from combining each 5-year slice of news with the reference corpus according to the method in the first column, and averaging their correlation scores.

We focus our analyses on the German setting, where the news corpora we use are relatively small, hence representing a good use-case for the applications of our methods. As expected, we observe that the smaller text corpora taken alone (*news, 5-year*) yield subpar results. Also, we see that we are able to significantly improve the performance with our combination methods. For the English corpora, the new corpora perform well (as they are comparatively big); however, most aligned methods still yield substantial performance improvement.

While the baseline methods *unaligned* and *unweighted* both use weights of 0.5 (as is the nature of averaging vectors), the average weight assigned to the evaluated words using all of our methods lies in the range of $0.6-0.7$, that is, our methods include the novel vectors to a higher degree.

We further evaluate our methods on the English news article corpus on two word analogy tasks. Table III lists the results of the same corpora and methods on two state-of-the-art test sets, the Google analogy test set [18] as well as the Bigger analogy test set (BATS) [23]. Unfortunately, no such word analogy corpus is available for the German language. Similarly to the previous word similarity tasks, we can once again demonstrate that a combination of small corpus temporal slices with a general-purpose corpus improves performance on the word analogy task over the news article corpora on their own.

TABLE III
COMPARISON OF THE PERFORMANCE OF BASELINE AND PROPOSED METHODS ON WORD ANALOGY TASKS.

	Google Analogy Test Set	BATS
English		
wiki	0.715	0.304
news, 5-year	0.5538	0.2348
unaligned	0.6882	0.2768
unweighted	0.716	0.308
relative frequency	0.7074	0.2982
term-corpus relevance	0.7074	0.2982
context stability	0.4994	0.2112

B. Qualitative Analysis of Temporal Models

Since we do in fact expect certain terms to evolve over time, those terms should exhibit different correlations than in the general, time-agnostic baseline. To that end, we leverage the weights introduced in Section III, specifically, due to its simplicity and good performance in table II, we further analyze the results produced by the *relative frequency* method. We identify in table IV among the words used by *SimLex-999* i) the words with high and low changes over time (see below) and ii) the words with high and low weights α as defined in Section III-B. Changes over time are computed by summing up the cosine distance scores between the vectors for the same word over time, again, on combined embeddings with 5-year slices of the news corpora. The higher the distance, the further a term has moved in the vector space over time.

As expected, there is an overlap between the terms that display high change and those assigned high weights. This follows from the method, as terms with low weights given to the vectors in the novel corpora will be closer to their vector in the reference embedding and thus exhibit little change over time in the combined embedding.

We can leverage this knowledge to split the *SimLex* dataset into two halves: one half containing words with low changes only — where we expect to achieve better performance on the word similarity task, specifically *SimLex-999*, given that these terms do not change their relationships over time — and the other half that we assume to have a negative impact on the overall score as reported in table II due to their change in meaning and thus also in relationships between the words. Table V reports these scores. For every word pair in the *SimLex* ground-truth dataset, it is considered in the *high change* evaluation if one or both words are in the top half of the sum of cosine distances over time, and it is considered *low change* only if both words appear in the bottom half of this list, making the evaluation dataset for *low change* significantly smaller.

TABLE IV
TOP 6 TERMS WITH HIGHEST AND LOWEST CHANGE OVER TIME AS WELL AS HIGHEST AND LOWEST WEIGHTS α AMONG WORDS IN SIMLEX-999.

	English	**German**
High change	self, suds, brow, dreary, bush, contemplate	frust, täten, neulich, trostlos, raten, debattieren
Low change	population, politician, august, football, boundary, saint	erzbischof, stumpf, neffe, samen, infektion, gleichung
High weights	say, think, salad, drizzle, sure, self	zuversichtlich, neulich, gewiss, froh, arbeitnehmer, sparen
Low weights	boundary, conquest, august, saint, song, population	kapelle, erzbischof, turm, samen, schlacht, glocke

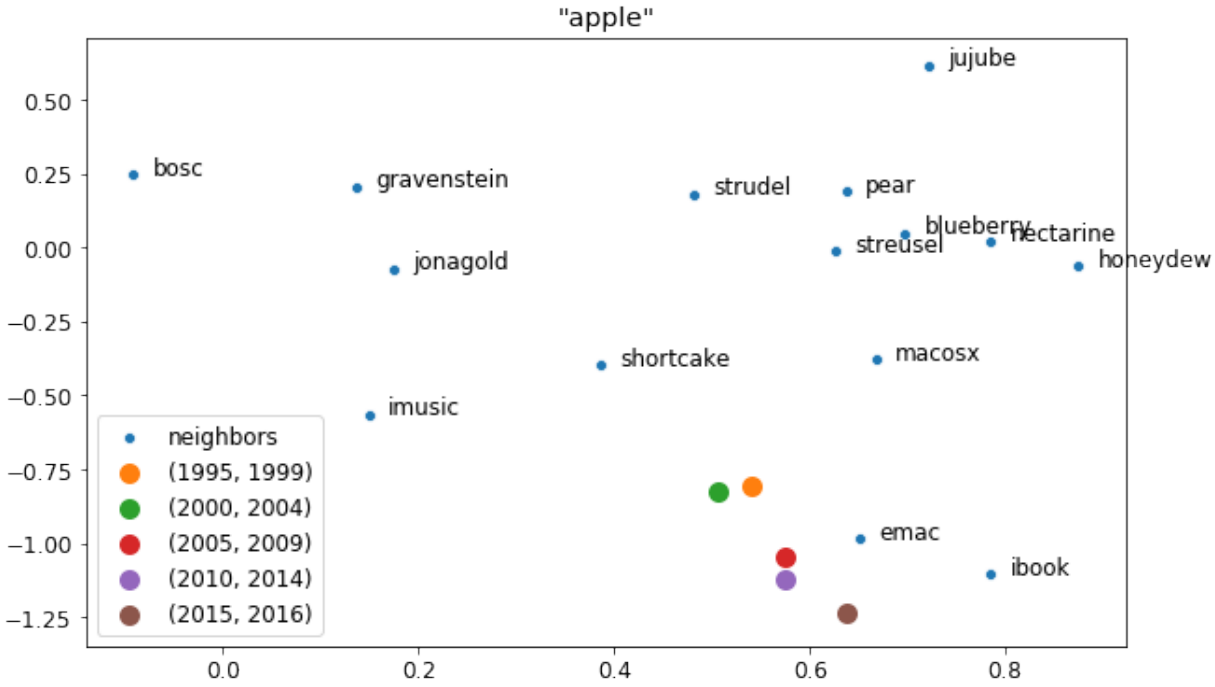

Fig. 1. Change over time of the word "apple" in the English corpora.

TABLE V
SCORES FOR *relative frequency* WHEN SPLITTING THE GROUND TRUTH DATASET INTO TWO HALVES: ONE WITH WORDS EXHIBITING HIGH CHANGE, AND THE OTHER WITH WORDS EXHIBITING LOW CHANGE.

	English	**German**
SimLex-999 overall	0.3264	0.269
SimLex-999: high change	0.2588	0.2468
SimLex-999: low change	0.5252	0.4456

1) Visual analysis of diachronic word embeddings: One of the main motivations behind this work lies in the ability to analyze changes in the vector space over time. To that end, we apply Primary Component Analysis (PCA) in order to reduce the 300-dimensional word vectors to two dimensions, such that we can plot the words in their neighborhoods. We take neighborhoods of stable contextual terms and analyze the location of the term of interest in different time slices. In blue, we give the vector locations of neighboring terms. The other points locate the word that is being analyzed over time.

For the English term "apple", we can see in Figure 1 that there is some observable change from the fruit to the technology company; however, it is a term that is used with multiple senses to this day and thus a clear delineation is difficult, as already between 1995 and 1999 the company apple as well as the fruit have both been relevant. Choosing the term "salad", on the other hand, a word that has been identified in table IV as a word with high change, we observe that it trends with various other food terms over time in Figure 2.

Taking the German corpus, we observe how the focus of the word "Arbeitnehmer" (employee) changes over time in Figure 3, from discussing income to companies to unemployment and social security contexts. On the other hand, Figure 4 showcases a term ("Erzbischof" i.e., archbishop, a low-change word) that is of lower relevance to the German news over time and exhibits little change.

Based on our proposed method, we are able to track the

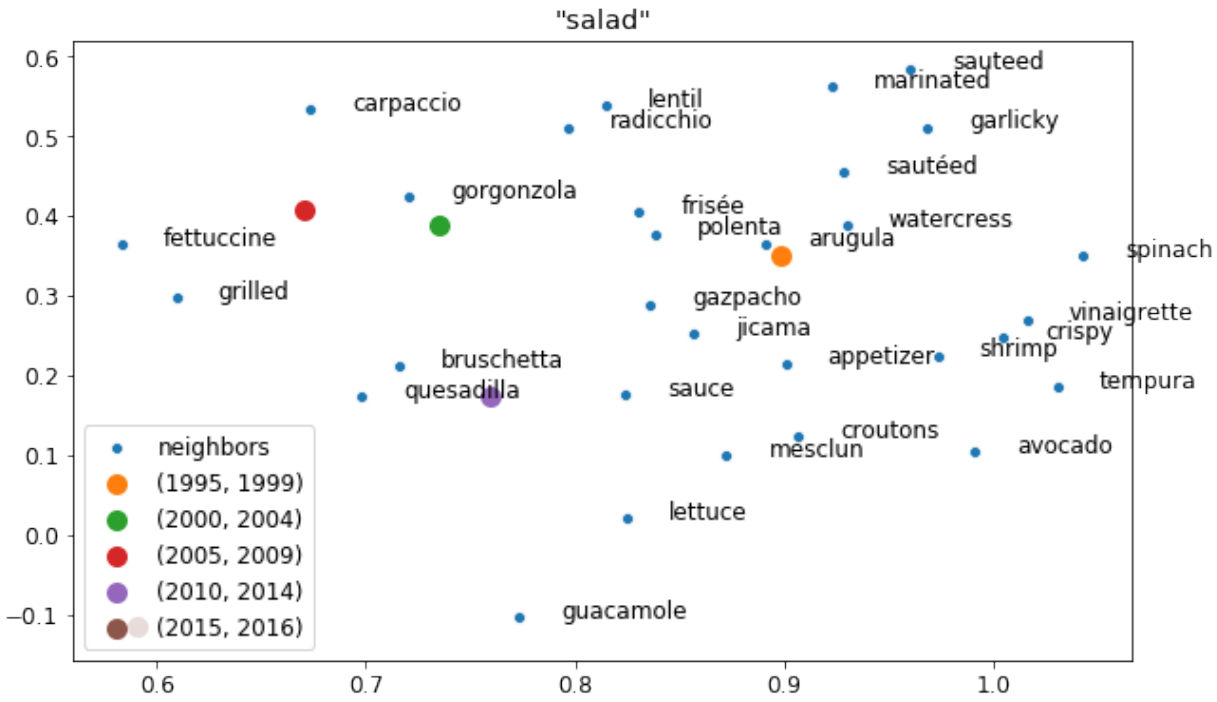

Fig. 2. Change over time of the word "salad" in the English corpora.

movement of terms over time, and as such, we can view shifts in topics, for example in political debates. Consider the acronym "SVP" (Swiss People's Party) in Figure 5 and how it has shifted over time away from semantic similarity to the acronyms of other parties or politicians, supposedly in developing discourse around unique characteristics. Being able to track such developments is intriguing and can be done easily for any term, for example for individual politicians or other public actors as well, as well as on the basis of other data and information sources.

V. Discussion

We evaluated our embedding combination methods against semantic tasks, as reported in tables II and III. While the improvements achieved through various combination methods on the English news corpora were comparatively minor (while indeed representing a small improvement over the use of the NYT news corpora on their own), the evaluation of the German news corpora—which are relatively small and hence represent good use-cases for the application of our methods—highlights the benefits of our method and the impact small corpora have on the performance of the embeddings on similarity tasks. The English news corpora, being comparatively larger (430k articles on average per time slice vs 17k articles on average in German), can be assumed to represent the terms of interest adequately, given the reported results. For the German corpus, we hence observe some significant improvement compared to the individual corpus embeddings, indicating that our combinations improve the quality of the resulting vectors for downstream tasks, be it machine learning tasks or manual analyses. We thus conclude that our methods are most valuable in contexts where the initial novel corpora are quite small and perform poorly individually. We do note however that, given the desired outcome of change in the vectors over time, we can expect to see some deviation in the correlation to ground truth, as the experiments on which we report in table V demonstrate, where we achieved very high scores, even well above the reference corpus, on stable terms.

Comparing with the methods introduced in section III, it becomes apparent that *relative frequency*, while the simplest, performs best of all methods on SimLex-999 in English, and falls second behind *context stability* in German; on WordSim-353, it performs best for both German and English. *Context stability* displays the greatest improvement in SimLex-999 for the German corpus. We assume this is due to being most closely linked to the final word vectors due to its evaluation of the immediate context words. However, it is in comparison by far the most costly method of the ones we introduced, as PMI scores have to be computed for all words and their 5-word context windows, which is why all further analyses were performed using the simple yet efficient *relative frequency* method. *Term-corpus relevance*, being essentially a slight modification of the *relative frequency* method, does not appear to significantly improve the performance of the vectors. We hypothesize that it may be more useful in a context where more than two text corpora are compared, as in the setting

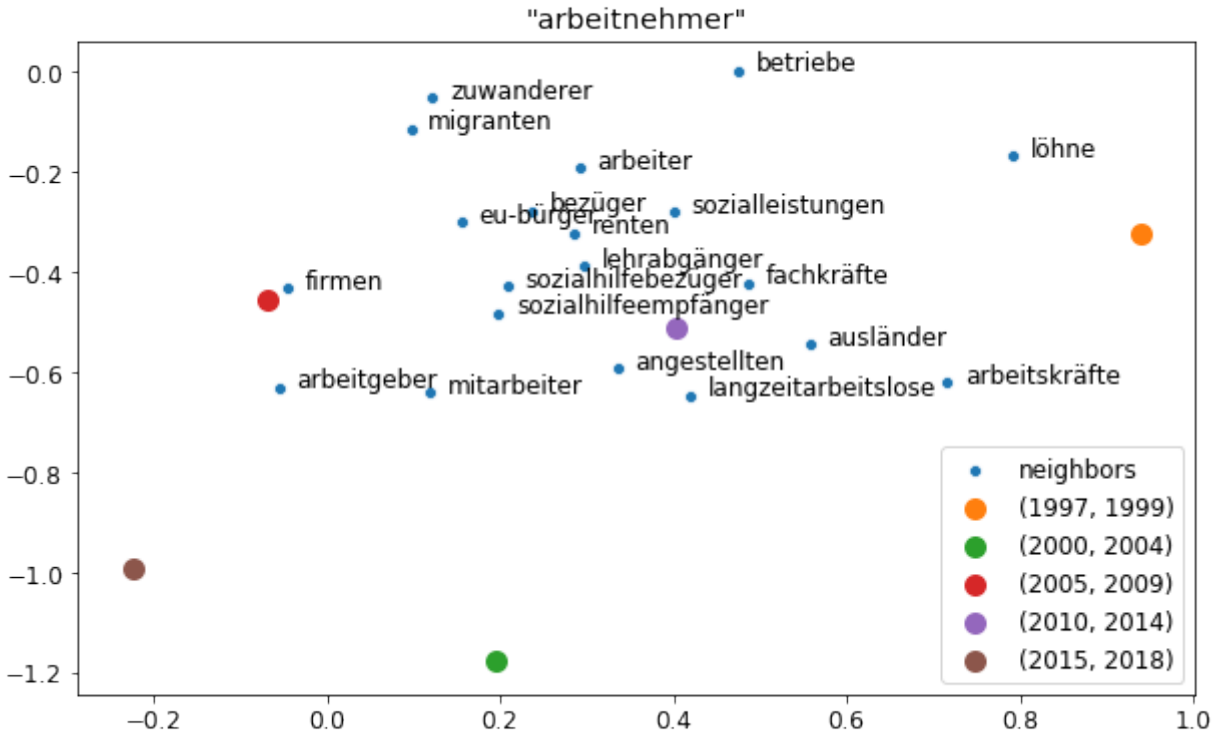

Fig. 3. Changes over time for the word “Arbeitnehmer” in the German corpora.

used by Rettig et al. [2].

In a more qualitative analysis, we showed that we are successfully able to visualize the stability of a term’s meaning and usage by projecting the temporal vectors associated with the term and its stable neighbors onto a two-dimensional space. In doing so, we were able to show in Figure 1 how a homonym such as “apple”, capturing multiple meanings in one vector, exhibits only a slight shift, whereas a term such as “salad”, that follows different trends over time, can be seen located with different terms over time, in changing semantic neighborhoods (Figure 2).

In analyzing debates in communication and political sciences, one can identify relevant issues in the context of political science like unemployment in a visual manner, such as investigating the labor market around the term “Arbeitnehmer” (employee, Figure 3). Unemployment is a topic of great concern to the general public that is widely debated by the public at large. Public debates are driven either by institutional-driven policy reforms or by events [24]. These news articles are curated from Swiss sources and are thus relevant to the political scene during the given time frames. Our method can identify conflictual elements or relevant problems in a country. Switzerland has had comparatively low levels of unemployment and related structural problems. By contrast, the most important problems related to the labor market and unemployment are comparatively high long-term unemployment (appears in the discussion as “Langzeitarbeitslose” (long-term unemployed) or “Sozialhilfeempfänger” (social security recipients) because they do not receive unemployment insurance money after a certain period), and that Switzerland does not have enough workers with qualification (appears visually as “Fachkräfte”, or as “Zuwanderer”, “Migranten” or “EU-Bürger” because Switzerland solves this also by getting specialists from abroad). Tracking the temporal evolution of the debate around employees (“Arbeitnehmer”), we consider the important policy reforms that took place during the time being analyzed: In the late 1990s there were votes about EU membership and bilateral agreements (reflecting a relationship to the terms “Zuwanderer” (immigrants) and “EU-Bürger” (EU citizen)). In 2004, the Swiss population voted about the old-age and survivors’ insurance (thus term context relationships to words as “Sozialhilfe” and “Arbeitgeber”), and later there were votes on mass immigration (linked to terms such as “Ausländer”, and discussed alongside the lack of skilled workers (“Fachkräfte”), clearly visible in the 2010-2014 corpus). Having analyzed those graphs jointly with political scientists, we find their results striking and are confident that our methods indeed pave the way to a new generation of political communication studies based on diachronic word embeddings.

VI. Conclusion

In this paper, we introduced the novel idea of improving pre-trained small-corpus embeddings by aligning and combining them with a bigger and more stable reference corpus. We

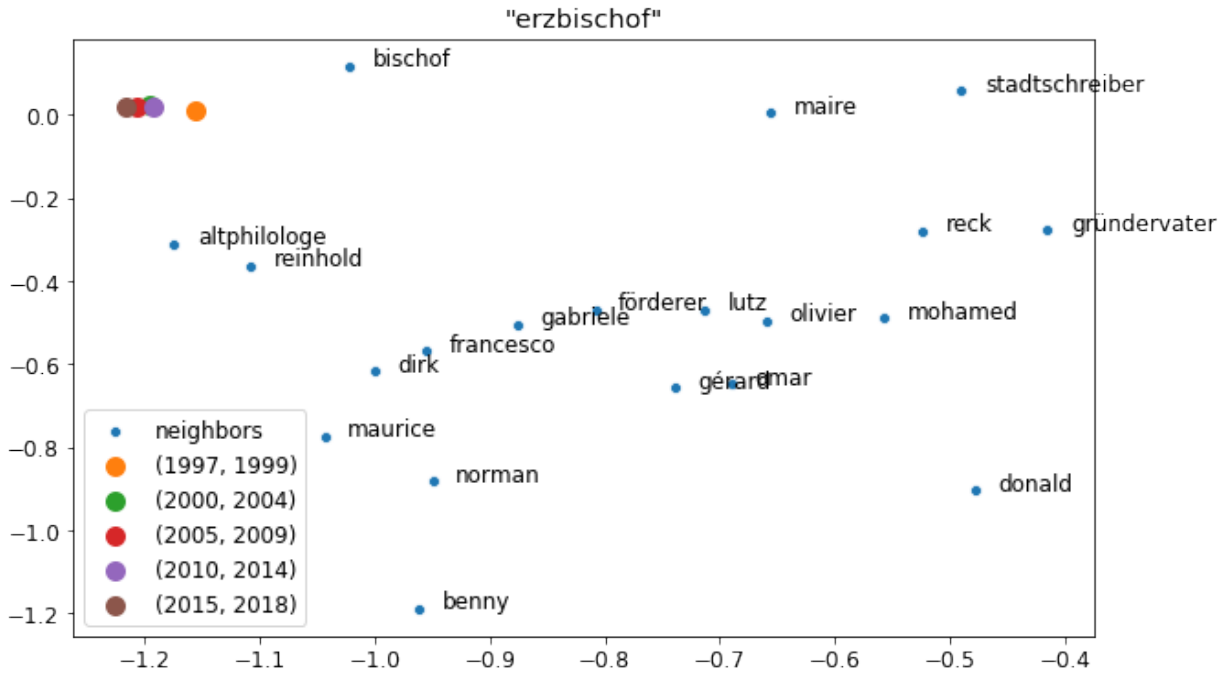

Fig. 4. Changes over time for the word "Erzbischof" (archbishop) in the German corpora.

proposed a set of methods that evaluate the original corpora and assign scores (weights) to each word in the vocabularies. From these scores, we update the vectors in the reference model with new properties from the novel embedding on a word-by-word basis. Our method is hence capable of retaining the stability and quality of the reference embedding while also transferring information such as word evolution over time from the novel corpora. Our empirical analyses showed that we are able to achieve significant improvement over state-of-the-art unweighted combination methods, especially on very small corpora. In addition—and most importantly—we showed how our methods can be used to obtain embeddings that are suitable for longitudinal or diachronic studies analyzing the evolution of words over time, without requiring fastidious manual efforts or computationally intensive temporal word embedding models.

We regard these results as very promising for future research on embeddings from small corpora. In future work, we intend to further investigate the relationships between terms over time by quantifying the impact of one word on the evolution of another word, in particular in the context of political communications studies. We are also considering approaches that view word embeddings as distributions rather than point vectors, allowing to better capture the various elements that jointly make up the semantics of a given term. Such approaches could prove particularly useful when handling polysemous words, and would allow us to conduct more detailed temporal analyses by modeling fine-grained changes in the distributions attached to each word or topic of interest.

Acknowledgment

This work was funded by the *Hasler Foundation* in the context of the *City-Stories* project. Thanks to Florin Zai for his help with data collection.

References

[1] M. Antoniak and D. Mimno, "Evaluating the stability of embedding-based word similarities," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 107–119, 2018. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/1202

[2] L. Rettig, J. Audiffren, and P. Cudré-Mauroux, "Fusing vector space models for domain-specific applications," in *IEEE 31st International Conference on Tools with Artificial Intelligence, ICTAI*, 2019.

[3] A. Muromägi, K. Sirts, and S. Laur, "Linear ensembles of word embedding models," in *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, Gothenburg, Sweden*, 2017.

[4] P. Bojanowski, O. Celebi, T. Mikolov, E. Grave, and A. Joulin, "Updating pre-trained word vectors and text classifiers using monolingual alignment," *arXiv preprint arXiv:1910.06241*, 2019.

[5] Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong, "Dynamic word embeddings for evolving semantic discovery," in *WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining*, 2018.

[6] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena, "Statistically significant detection of linguistic change," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 625–635.

[7] Y. Zhang, A. Jatowt, S. S. Bhowmick, and K. Tanaka, "The past is not a foreign country: Detecting semantically similar terms across time," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2793–2807, 2016.

[8] R. Hänggli and R. van der Wurff, *Quality of Public Debates*. Cambridge University Press, 2019, p. 257–284.

[9] A. E. Boydstun, S. Bevan, and H. F. Thomas III, "The importance of attention diversity and how to measure it," *Policy Studies Journal*, vol. 42, no. 2, pp. 173–196, 2014.

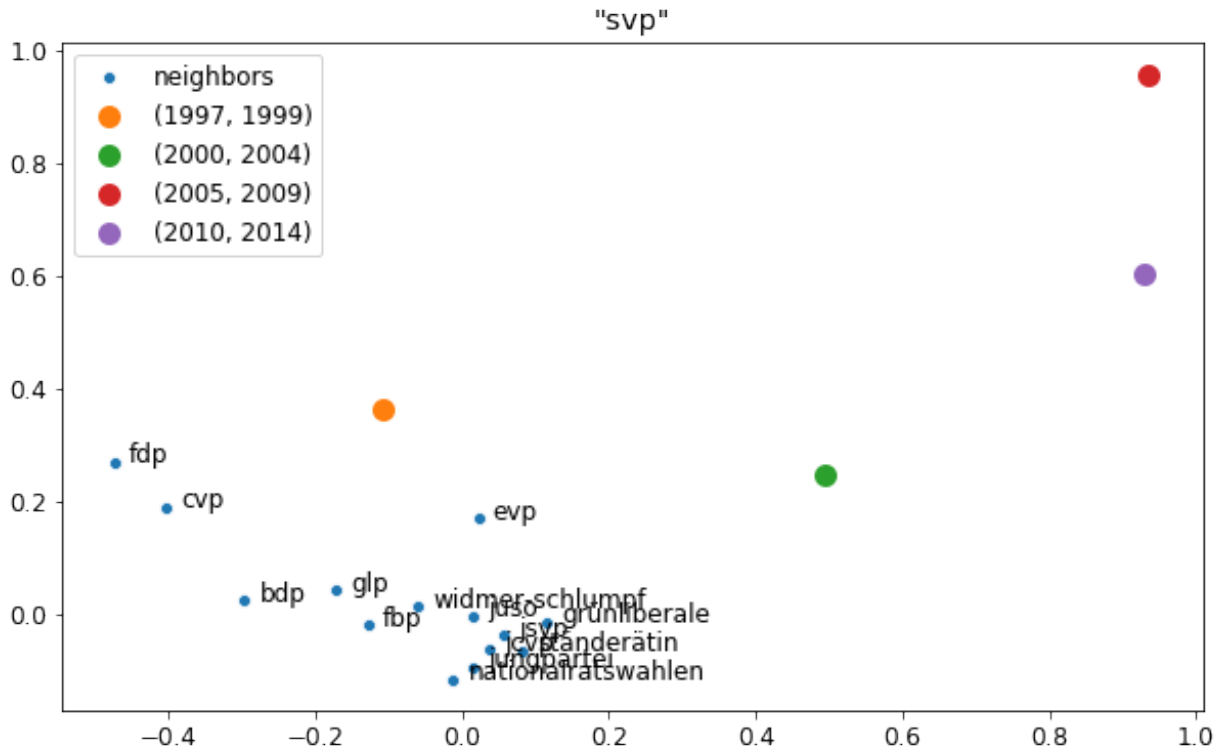

Fig. 5. Changes over time for the term “SVP” (acronym, “Schweizerische Volkspartei” or Swiss People’s Party) in the German corpora.

[10] F. R. Baumgartner, S. L. De Boef, and A. E. Boydstun, *The Decline of the Death Penalty and the Discovery of Innocence*. Cambridge University Press, 2008.

[11] P. H. Schönemann, “A generalized solution of the orthogonal procrustes problem,” *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[12] W. L. Hamilton, J. Leskovec, and D. Jurafsky, “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)*, Berlin, Germany, Aug. 2016.

[13] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” in *Sixth International Conference on Learning Representations, ICLR*, 2018.

[14] C. Xing, D. Wang, C. Liu, and Y. Lin, “Normalized word embedding and orthogonal transform for bilingual word translation,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015.

[15] K. W. Church and P. Hanks, “Word association norms, mutual information, and lexicography,” *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[16] T. Kiss and J. Strunk, “Unsupervised multilingual sentence boundary detection,” *Computational linguistics*, vol. 32, no. 4, pp. 485–525, 2006.

[17] M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger, “The penn treebank: annotating predicate argument structure,” in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 114–119.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[19] F. Hill, R. Reichart, and A. Korhonen, “Simlex-999: Evaluating semantic models with (genuine) similarity estimation,” *Computational Linguistics*, vol. 41, no. 4, pp. 665–695, 2015.

[20] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, “Placing search in context: The concept revisited,” in *Proceedings of the 10th international conference on World Wide Web*, 2001, pp. 406–414.

[21] I. Leviant and R. Reichart, “Separated by an un-common language: Towards judgment language informed vector space modeling,” *arXiv preprint arXiv:1508.00106*, 2015.

[22] J. Coates and D. Bollegala, “Frustratingly easy meta-embedding – computing meta-embeddings by averaging source word embeddings,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018.

[23] A. Gladkova, A. Drozd, and S. Matsuoka, “Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn’t.” in *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 8–15. [Online]. Available: https://www.aclweb.org/anthology/N16-2002

[24] R. Hänggli, *Framing Strategies: Important Messages in Public Debates*. Cambridge University Press, 2019, p. 191–211.