Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2431–2439
Marseille, 11–16 May 2020

Embedding Space Correlation as a Measure of Domain Similarity

Anne Beyer[1], Göran Kauermann[2], Hinrich Schütze[1]
[1]Center for Information and Language Processing (CIS), [2]Department of Statistics
Ludwig-Maximilians-Universität Munich, Germany
anne.beyer@campus.lmu.de, goeran.kauermann@stat.lmu.de, inquiries@cislmu.org

Abstract

Prior work has determined domain similarity using text-based features of a corpus. However, when using pre-trained word embeddings, the underlying text corpus might not be accessible anymore. Therefore, we propose the CCA measure, a new measure of domain similarity based directly on the dimension-wise correlations between corresponding embedding spaces. Our results suggest that an inherent notion of domain can be captured this way, as we are able to reproduce our findings for different domain comparisons for English, German, Spanish and Czech as well as in cross-lingual comparisons. We further find a threshold at which the CCA measure indicates that two corpora come from the same domain in a monolingual setting by applying permutation tests. By evaluating the usability of the CCA measure in a domain adaptation application, we also show that it can be used to determine which corpora are more similar to each other in a cross-domain sentiment detection task.

Keywords: domain similarity, word embeddings, domain adaptation

1. Introduction

The application of neural network approaches to solve Natural Language Processing (NLP) tasks has led to the development of language representations that capture intrinsic linguistic information without the need for task specific, manual feature extraction by experts. This is achieved by so called *word embeddings*, which are high dimensional vector representations that can be induced from unlabeled data.

Despite the success of embeddings, there are still open research questions about what exactly these representations capture. In this paper, we address the question as how differences in domains and languages are reflected by embedding spaces. Previous work has shown that it is possible to align embedding spaces across domains and languages (e.g. Mikolov et al. (2013a), Artetxe et al. (2016), Barnes et al. (2018)), but to the best of our knowledge, no task independent, multi-lingual analysis of the underlying structural similarities has been carried out so far.

A systematic notion of similarity is also interesting from an application point of view. A widely used technique is to augment a model's capabilities by using pre-trained word embeddings, based on larger or more diverse corpora. This can be applied across different text domains or languages. In both cases, it can be useful to know how similar available pre-trained embeddings are to the target corpus in order to pick the most similar or to estimate the amount of fine tuning or adaptation necessary. While there exist approaches to determine corpus similarity based on text features (usually word frequency distributions or language model perplexity), these resources are often not available when using pre-trained word embeddings. Our contribution is therefore to propose the CCA measure, a similarity measure which is calculated directly on the embedding spaces.

This work is structured as follows: Section 2. gives an overview of the related work. Our approach is described in detail in Section 3. and further evaluated in Section 4. The predictions of our approach are compared with the results of related applications in Section 5. in order to gain insights into its usefulness for data selection in domain adaptation NLP tasks. In Section 6. we present our conclusions and propose directions for future work. Our code is available at `https://github.com/AnneBeyer/emb_sim/`. There, we also provide a link to download the embedding spaces used in this study.

2. Related Work

In neural network applications, words are mapped to multi-dimensional vector representations that encode semantic information. These can be learned task-specifically by the first layer in a neural network, but it has been shown to be more effective (in terms of computational resources and overall quality) to use word embeddings that have been pre-trained on large datasets in advance. To this end, there exist several approaches, either based on co-occurrence counts (Schütze, 1993; Bullinaria and Levy, 2007; Turney and Pantel, 2010) or learned implicitly by neural network architectures solving a language modeling or word prediction task (Mikolov et al., 2013b; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019). In order to avoid any source of additional uncertainty in the embedding spaces, we focus on the former approach in this work.

Different approaches to represent these underlying co-occurrences have been proposed (see Turney and Pantel (2010) for an overview), the most adopted of which is the Positive Pointwise Mutual Information (PPMI, cf. Levy et al. (2015)), which reflects the strength of association between word pairs (i.e., if they co-occur more often with each other than with other words, they will receive a higher score).

$$PPMI(w, c) = \max\left(\frac{\hat{P}(w, c)}{\hat{P}(w)\hat{P}(c)}, 0\right)$$

where $\hat{P}(x)$ is the relative frequency of event x in the corpus. The resulting matrices are usually very large and sparse. To make computations more feasible and the representations more general, truncated Singular Value Decomposition (SVD) can be applied. A co-occurrence matrix M

is decomposed into its left and right orthonormal eigenvectors U and V and the diagonal matrix Σ, containing the corresponding eigenvalues. As these are sorted in decreasing order according to the captured variance, a lower dimensional representation of M can be derived by selecting only the top d components of $U\Sigma$. Levy et al. (2015) study the effect of several hyperparameters on the performance and find that, with careful hyperparameter optimization, count-based models perform competitively on various word similarity and analogy tasks as compared to word2vec (Mikolov et al., 2013b) or GloVe (Pennington et al., 2014). Specifically, they show that applying context distribution smoothing consistently improves the performance. Another hyperparameter that is shown to impact the performance is to use eigenvalue weighting, i.e., using the top d components of $U\Sigma^p$, which yields better results for $p = 0.5$ and $p = 0$ (i.e., ignoring Σ) than the original approach. They also report that applying L_2 vector normalization leads to better performance.

To compare embedding spaces, they have to be mapped into a shared space. Different supervised, semi-supervised and unsupervised approaches for mapping embedding spaces have been proposed (Mikolov et al., 2013a; Artetxe et al., 2016; Lample et al., 2018; Faruqui and Dyer, 2014; Rastogi et al., 2015; Lu et al., 2015). In this work, we apply Canonical Correlation Analysis (CCA, Hotelling (1936)), which maps two multi-dimensional spaces into a shared space in which they are maximally correlated. It was first proposed for mapping embedding spaces by Faruqui and Dyer (2014). Given two count-based embedding spaces $\Sigma \in \mathbb{R}^{n_1 \times d_1}$ and $\Omega \in \mathbb{R}^{n_2 \times d_2}$ where n_1 and n_2 denote the vocabulary sizes and d_1 and d_2 are the embedding dimensions, they first extract the sub-spaces Σ' and Ω' with shared vocabulary n. For every corresponding vector pair $x \in \Sigma'$ and $y \in \Omega'$ CCA then finds two transformations v and w such that xv and yw are maximally correlated, yielding the transformation matrices $V \in \mathbb{R}^{d_1 \times d}$ and $W \in \mathbb{R}^{d_2 \times d}$ with $d = min\{rank(V), rank(W)\}$. These are then used to project the original embedding spaces into the maximally correlated spaces Σ^* and Ω^*.

Prior work on measuring domain similarity often relies on the frequency distribution of words in different corpora. Barnes et al. (2018), for example, use the Jensen-Shannon Divergence, which is a symmetric adaptation of the Kullback-Leibler Divergence, to compute a similarity score between word frequency distributions. We will compare our approach to their results in Section 5. Another approach is to use an approximation of the A-distance (Ben-David et al., 2007), based on the generalization error of a separately trained domain classifier. While this has also been used to align representations across domains prior to the emergence of pre-trained embedding spaces, we will focus on more recently proposed mapping approaches and leave a comparison with this method for future work.

3. The CCA measure

This section describes our approach to measuring domain similarity in terms of embedding space correlation.

3.1. Data

In the following, we compare several corpora from different text domains. The corpora were selected following an intuitive notion of domain to cover a wide range of different text types. All corpora are available in four languages: English, German, Spanish and Czech. The corpora marked with * are retrieved from the OPUS project page (Tiedemann, 2012).[1] The names in parentheses will be used to refer to the corpora throughout this study.

Wikipedia (wiki) One of the largest corpora available in many languages is the Wikipedia corpus.[2] We use the dumps from February 22, 2019. The raw article texts are extracted using the wikiextractor tool[3] and split into sentences using the NLTK Tokenizer Package.[4] Due to its encyclopedic nature it is treated as a general-domain corpus here, even though it covers a wide range of topics in itself. As pre-trained embeddings are generally available based on the whole corpus,[5] a more thorough examination of semantic sub-parts will be left for future work.

Europarl* (euro) The Europarl corpus is a parallel multilingual collection of the transcriptions of the proceedings in the European Parliament, originally compiled and aligned by Koehn (2005) as a resource for machine translation. As our approach does not necessarily require parallel corpora, we downloaded the monolingual untokenized raw texts to make use of the largest amount of data available per language. For the comparison in Section 3.4. we also use the parallel corpora for English and our selected languages.

OpenSubtitles* (sub) The OpenSubtitles corpus (Lison and Tiedemann, 2016) consists of pre-processed movie subtitles.[6] We use the 2018 version and again downloaded the monolingual untokenized raw texts.

Acquis Communautaire* (dgt) This corpus consists of the publicly accessible Translation Memory provided by the Directorate-General for Translation of the European Commission, which contains translations of the European Unions legislative documents (Acquis Communautaire).[7] It is similar to the Europarl corpus, but differs in the form of language (written documents vs. transcribed speeches).

Medical documents (med) The UFAL Medical Corpus is a pairwise bilingual (one of which is always English) collection of sentences from several medical and other domains.[8] It is made available upon registration on the website. We extracted all segments from the medical domains (which we will treat as one domain) for our languages of

[1] http://opus.nlpl.eu/index.php
[2] http://dumps.wikimedia.your.org/backup-index.html
[3] https://github.com/attardi/wikiextractor
[4] https://www.nltk.org/_modules/nltk/tokenize.html
[5] e.g. https://wikipedia2vec.github.io/wikipedia2vec/pretrained/
[6] http://www.opensubtitles.org/
[7] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory
[8] http://ufal.mff.cuni.cz/ufal_medical_corpus

interest and joined the respective English parts into one English corpus.

We extract the raw text and tokenize it using the gensim tokenizer.[9] Duplicates are removed, only lines with at least two words are kept and all lines are shuffled. In order to avoid effects based on corpus size, we randomly select subsets from all corpora. The smallest amount of data available for English is the dgt corpus (260 MB). We selected this as our sample size. For the other languages, some corpora were smaller. In these cases we used the available amount of data. Table 1 reports the number of tokens in each corpus after pre-processing. We also report results on different corpus sizes by comparing the full sized corpora for English in Section 3.4. In the following analysis, we use Wikipedia as our reference corpus and create two distinct samples from this corpus as an instance of two corpora that come from the same domain.

3.2. Embedding Spaces

To create the PPMI + SVD embeddings, we use the implementation by Levy et al. (2015).[10] We set the window size to 5, restrict the vocabulary to words that appear at least 50 times and reduce the dimensionality of the resulting word vectors to 100. As our work is not concerned with the application of word embeddings to a specific task, we do not fine tune the hyperparameters but restrict ourselves to one setting that follows their general suggestions by normalizing the vectors to unit length, applying context distribution smoothing with $\alpha = 0.75$ and not using the eigenvalues ($eig = 0$). All other parameters use the default values from their implementation.

For mapping the embedding spaces, we use CCA in an implementation based on Rastogi et al. (2015). In addition to being more efficient in computation time as compared to other CCA implementations, generalized CCA has the advantage of allowing a simultaneous alignment of several embedding spaces. In this work, however, we will only focus on binary comparisons and leave a multi-domain/multi-language comparisons for future work. In the monolingual versions, we use the shared vocabulary as supervision lexicon, in the bilingual comparisons we incorporate the ground-truth bilingual lexicons from the MUSE project (Lample et al., 2018).[11] As we don't want to find a mapping for unseen data but are rather interested in the best possible mapping for the embedding spaces at hand, we use all the available shared vocabulary for calculating the mapping. Section 3.4. investigates how results vary depending on the shared vocabulary size.

3.3. Correlation Measure

To measure the relatedness of two corpora from different domains or languages, we first map their corresponding embeddings as described above. As CCA is mapping the embedding spaces by maximizing their dimension-wise correlations, we further use these correlations to determine their overall similarity. We first compute the correlation matrix of the two mapped embedding spaces based on their shared vocabulary. This gives us a $2d \times 2d$ matrix of the form

$$R = \left[\begin{array}{c|c} R_x & R_{xy} \\ \hline R_{yx} & R_y \end{array}\right]$$

where R_x contains the pair-wise correlations for all dimensions within the first embedding space, R_y those within the second embedding space and R_{xy} and R_{yx} contain the correlations between the dimensions of the two embedding spaces, respectively. As we are only interested in the cross-embedding correlations, which are symmetric in this case, we only consider R_{xy} further. To gain a first insight into the correlation patterns, we plot the correlation scores on the diagonal as shown in Figure 1. To convert these correlation scores into the CCA measure score, we calculate the mean of these dimension-wise correlations along the diagonal of R_{xy}.

3.4. Results

In this section, we report the results of different domain comparisons. We first focus on the monolingual approach and motivate some further studies based on the findings. The second part reports the results of the cross-lingual comparisons and the follow-up questions and results that arose. The results will be discussed in Section 3.5.

3.4.1. Monolingual Domain Comparison

Figure 1 illustrates the dimension-wise correlations between the embedding space of the first Wikipedia sample and all other English embedding spaces.

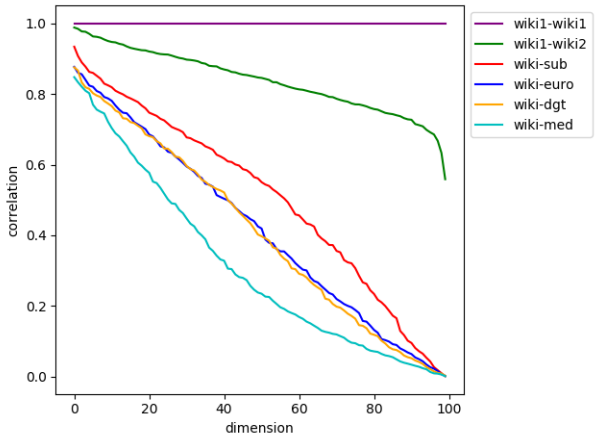


Figure 1: Correlation per dimension between embedding spaces (i) trained on samples from wiki and (ii) (a) the same embedding space (wiki1), (b) different wiki sample (wiki2), (c) sub, (d) euro, (e) dgt, and (f) med

The results show that there exist quite substantial differences among the embedding spaces. As expected, the two Wikipedia corpora have the highest dimension-wise correlations among the comparisons of different corpora. Two spaces trained on the exact same corpus result in the same embedding space with dimension-wise correlation scores of 1. The two European legal domains show very similar correlation patterns with Wikipedia while the subtitles

[9] https://radimrehurek.com/gensim/corpora/wikicorpus.html

[10] https://bitbucket.org/omerlevy/hyperwords/src/f5a01ea3e44c/

[11] https://github.com/facebookresearch/MUSE

	en	de	es	cs
wiki1	43,954,946	37,303,212	42,024,088	35,718,709
wiki2	43,956,083	37,315,139	42005,477	35,728,850
sub	53,214,198	44,536,422	46,676,844	42,376,509
euro	44,373,711	37,645,317	42,103,150	12,936,864
dgt	41,806,529	31,204,671	41,285,490	31,066,266
med	42,896,554	36,958,007	9,912,670	13,402,521

Table 1: Number of tokens in each corpus after sub-sampling similar sized splits and pre-processing

domain exhibits slightly higher values and the medical domain is the most different in terms of dimension-wise correlations.

As described in the previous section, we convert these graphs into a score by computing the mean of their values. Table 2 displays results for English as well as the other languages.

	en	de	es	cs
wiki1-wiki2	0.84	0.79	0.81	0.87
wiki-sub	0.51	0.36	0.41	0.40
wiki-euro	0.42	0.34	0.36	0.36*
wiki-dgt	0.41	0.35*	0.38	0.38*
wiki-med	0.31	0.25	0.31*	0.29*

Table 2: CCA measure scores for domain similarity. Based on the shared vocabulary for Wikipedia compared with different domains for English (en), German (de), Spanish (es) and Czech (cs). In cells marked with *, the second corpus was smaller than the first due to the limited availability of resources.

The comparison with the other languages shows that a very similar ranking of corpus similarities can be obtained for all languages. The only exception is the comparison with euro and dgt, where the ranking is reversed – but the CCA measure scores for these two corpora are in general quite similar.

Based on these results, we investigate additional factors as reported below.

Vocabulary Size While we mostly controlled the corpus size in the previous setting, the different corpus combinations differ in terms of vocabulary overlap. Table 3 shows the size of shared vocabulary for the comparisons in Table 2. In order to test whether the CCA measure can simply be explained by the size of shared vocabulary, as the numbers in Table 3 might suggest, we also compute the CCA measure on a controlled vocabulary size. Table 4 contains the CCA measure scores computed by selecting a random sample of 5000 from the shared vocabulary for each corpus in a comparison.

While we see some variation in the CCA measure score, the same rankings in terms of similarity can be observed as when using the whole available vocabulary for the mapping and correlation computation. We will therefore continue using the original settings (i.e., using all of the shared vocabulary) for the mapping and correlation analysis in the remainder.

	en	de	es	cs
wiki1-wiki2	30895	38898	35783	50442
wiki-sub	18785	17157	19474	22865
wiki-euro	11668	12930	13362	12303
wiki-dgt	11891	12412	12872	17813
wiki-med	9460	8753	7095	8160

Table 3: Amount of shared vocabulary in domain comparisons

	en	de	es	cs
wiki1-wiki2	0.84	0.80	0.81	0.87
wiki-sub	0.51	0.38	0.42	0.41
wiki-euro	0.43	0.35	0.38	0.37
wiki-dgt	0.42	0.36	0.39	0.39
wiki-med	0.32	0.27	0.32	0.30

Table 4: CCA measure score for domain similarity. Based on a random sample of 5000 words from the shared vocabulary for the same comparisons as in Table 2

Corpus Size The varying amounts of data available in the languages other than English (see Table 1) do not seem to have a big influence on the overall ranking in Table 2. To further investigate the effect of corpus size, we now compute the CCA measure for English when using the entire dataset for each condition, i.e., without sampling. Table 5 shows results.

	CCA measure	corpus size
wiki	–	2,681,657,224
sub	0.41	717,628,087
euro	0.29	54,824,832
dgt	0.32	41,889,595
med	0.24	87,142,718

Table 5: Comparison of the English Wikipedia with all other domains using all available data, i.e., without subsampling. The second column shows the number of tokens in each corpus. As we use all data, no comparison of two wiki samples is performed.

3.4.2. Cross-lingual Domain Comparison

To test whether our CCA measure scores also hold across languages, we also report the results for a cross-lingual

comparison in Figure 2. We compare each of the languages with English and use the same corpus samples as above. As described in Section 3.2., we first extract the shared vocabulary between two embedding spaces using a dictionary, and then apply the mapping and correlation calculation on the corresponding vector space as before.
Figure 2 also contains the results of the monolingual comparison for English as a reference in purple. In this setting, the English wiki corpus is compared with copora in four different languages in five different domains for a total of 20 comparisons.

Parallel Corpora To compare this to actual parallel corpora, we also compute the CCA measure score for the parallel portions of the Europarl corpus for English and our languages of interest. The results can be found in the first column of Table 6. It should be noted that the Europarl corpora for en–es and en–de are roughly three times as large as the en–cs corpus.

	parallel	cleaned
en–es	0.57	0.71
en–de	0.53	0.71
en–cs	0.50	0.64

Table 6: CCA measure scores for cross-lingual comparison. For parallel portion of the Europarl corpus and additionally using a cleaned dictionary.

Cleaned Dictionary In the cross-lingual setting, our approach has another influencing factor: As we apply a supervised mapping algorithm, we rely on the use of a dictionary. The MUSE dictionaries are created using a translation tool and can therefore also contain erroneous translations (a short inspection of the English-German dictionary revealed also English-English pairs). Lubin et al. (2019) propose a noise-aware mapping approach based on the EM algorithm that combines the Orthogonal Procrustes mapping with a dictionary cleaning task and thereby determines the "useful" portion of the dictionary.[12] We exploit their mapping approach to create cleaned dictionaries from the overlap of the MUSE dictionaries with the shared vocabulary for the parallel Europarl corpora. We then use these cleaned vocabularies for our CCA mapping on the original embeddings. The CCA measure scores with these cleaned vocabularies are displayed in the second column Table 6.

3.5. Discussion

In general, we can conclude from the results that the CCA measure allows us to determine a ranking among corpora that is consistent within and across multiple languages.
The results of our monolingual variation studies show that corpus size has the biggest impact on the CCA measure scores. Here we can see that the order for dgt versus euro is flipped, now matching the predictions for the other languages. The amount of shared vocabulary considered only slightly increases the CCA measure scores across all comparisons and did not influence the order of the rankings. If we abstract away from the magnitude of the CCA measure scores and only consider the ranking, the results are stable across all variations, except for the comparison with euro and dgt. But their scores are quite similar to begin with, so that a changed ranking can easily occur due to noise.
It should be noted that, since we focus on the comparison of domains in a broader sense and not restrict ourselves to the availability of aligned parallel data, our cross-lingual results do not compare the same text resources per language (except for the results reported in Table 6), but instead random samples from each domain. However, Figure 2 shows that ranking patterns similar to the monolingual results can also be observed in the cross-lingual setting. This can be taken as evidence that the CCA measure captures some language independent, intrinsic notion of domain similarity when used as a ranking score.
The cross-lingual results in Table 2 suggest that languages and domains seem to have a similar impact on the change of embedding spaces. Comparing the monolingual result for wiki-sub with the cross-lingual wiki-wiki scores shows a similar decrease in CCA measure scores.
Comparing the results of the wiki1-wiki2 comparison with the similarity of parallel Europarl corpora yields comparable results, suggesting that the notion of domain is captured in the same way, independent of the explicit content overlap. However, it is important to note that the quality of the dictionary used plays a very important role. While the CCA measure scores in Table 6 are still smaller than for the monolingual equivalent of comparing two Wikipedia samples, the difference can be reduced quite substantially by relying on a high-quality dictionary.
In the current setting, the CCA measure does not distinguish between domain and language differences but rather "adds" them up to some extent. The question whether a more sophisticated and/or non-linear mapping algorithm could reduce the impact of language and allow an independent domain comparison across languages can not be answered in the scope of this work, but should be examined in future work. Here, it could also be useful to apply the mapping for multiple languages at the same time, which is possible with the Generalized CCA method, but will require a multilingual dictionary.
The question of the "true" ranking of similarities and if it can be modeled by the CCA measure cannot be answered by an intrinsic examination but has to be addressed with an extrinsic evaluation of the performance in downstream tasks. We will look into this more closely in Section 5.

4. Analysis of the CCA measure

Section 3. has shown that the CCA measure is able to give a consistent ranking of domains in and across different languages. We now further investigate the results with respect to the interpretability of the CCA measure, namely, we want to determine if the CCA measure can reliably answer the question whether two corpora come from the same domain or not.

4.1. Method

To this end, we perform random permutation tests to approximate the distributions for every corpus pair from the

[12] https://github.com/NoaKel/Noise-Aware-Alignment

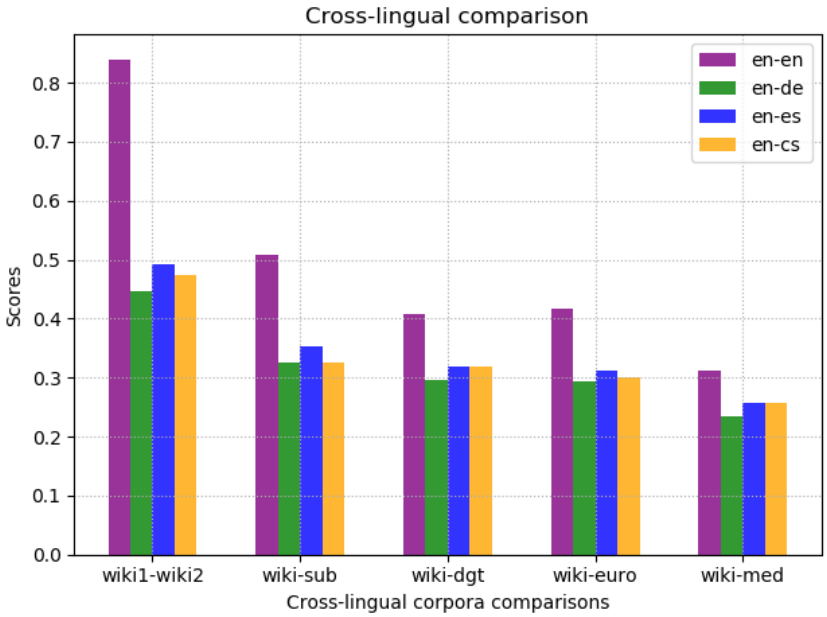


Figure 2: Cross-lingual domain comparison of the English Wikipedia with the other domains in four different languages. The purple bars are the monolingual reference.

comparison of the monolingual English corpora.

Under the assumption that both corpora come from the same distribution, we generate (for each corpus pair) 100 random sample corpora by shuffling them together and then splitting them again, thereby creating splits containing varying amounts from both corpora. For each pair of splits, we compute the CCA measure (i.e., we compute the two embedding spaces, map them using CCA, compute the dimension-wise correlation coefficients, and then compute the mean of these coefficients). We then compare the CCA score for the true data split with the distribution of CCA scores from our simulations.

In order to gain an insight into the cross-lingual comparison, we further adapt the above approach as follows: As we do not want to train mixed language embeddings, we first convert one language into the other by word-to-word translation. For this study, we again rely on the parallel Europarl corpus and inspect our approach for the English-German pair. We re-use the cleaned vocabulary from Section 3.4.2. and use this to pre-translate the German part of the parallel corpus. Words not present in the dictionary are omitted in the translation.[13] We then perform the same simulation as described above for 5 randomly permutated corpus pairs.

4.2. Results

Figure 3 illustrates the results of our monolingual simulation study for 100 sampled corpora pairs for each domain comparison. In the case of the same domain comparison (a), the original data point lies within the simulated points, whereas in the case of less similar spaces it lies far to the left of the simulated distribution. A ranking of these distances corresponds to the ranking based on the CCA measure as reported in Table 2. Across all corpus combinations, the means of the sampled distribution lie in the interval between 0.78 and 0.87.

Even though we had to restrict the amount of simulations due to limited time resources in the cross-lingual comparison, the results presented in Figure 4 seem to indicate a similar pattern as the monolingual cross-domain simulations in Figure 3 (d-e).

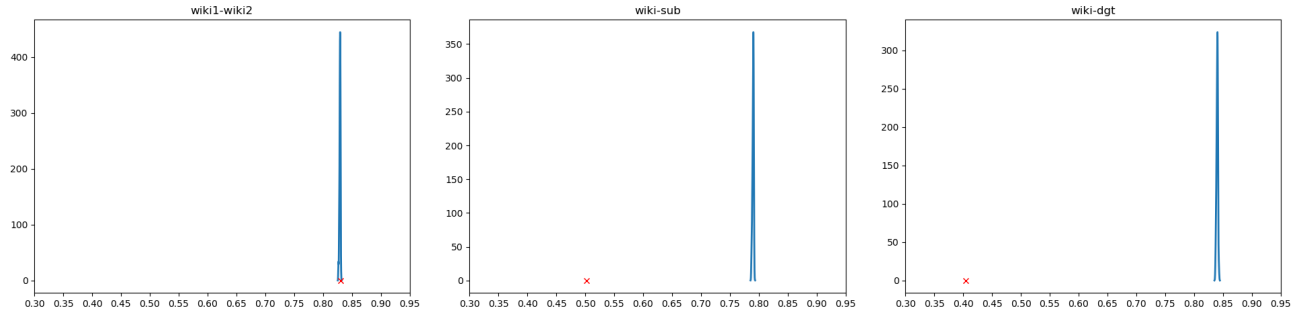
4.3. Discussion

Assuming that both embedding spaces come from the same distribution, simulated by shuffling the corpora together, the resulting distributions are centered close to 0.8 for all corpus combinations, as shown in Figure 3. This confirms our findings from Table 2, where the comparison of two distinct Wikipedia samples served as a reference for two corpora coming from the same domain, yielding CCA measure scores in the same range. We can therefore state that values ≥ 0.8 according to the CCA measure are a strong indicator for two corpora actually coming from the same domain in a monolingual setting.

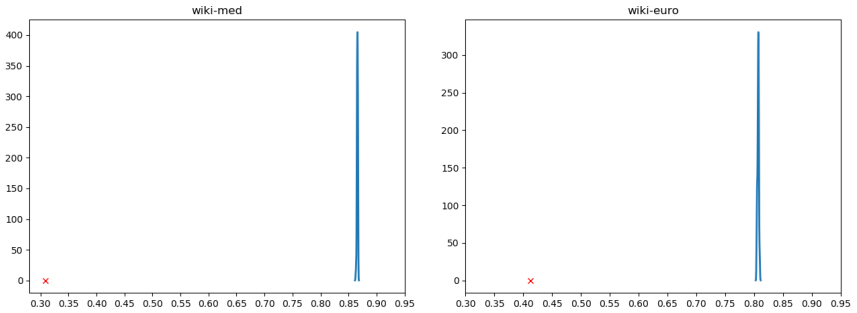
Figure 4 displays the cross-lingual results from comparing the same corpus in two different languages, but shows a similar pattern as the different domain comparisons in Figure 3 (b-e), i.e., the actual CCA score lying to the left of the simulated distribution. This suggests that the CCA measure does not distinguish language differences from domain differences.

Future work should address whether a more sophisticated mapping approach can close the gap between the original and the simulated datapoints in the cross-lingual comparison, or whether this is due to inherent language differences (words without exact 1-1 correspondences, ambiguity pat-

[13] When computing the word embeddings, this can have the effect of increasing the window size for some word pairs similar to sub-sampling and the deletion of rare words described in (Levy et al., 2015). However, they note that this was shown to only have a small effect in preliminary experiments.



(a) Two Wikipedia corpora (b) Wikipedia and subtitle corpus (c) Wikipedia and legislative corpus



(d) Wikipedia and medical corpus (e) Wikipedia and Europarl corpus

Figure 3: The blue curve shows the distribution of CCA measures for 100 random corpus splits (after permutation). The red x is the CCA measure for the original (unpermuted) corpus pair.

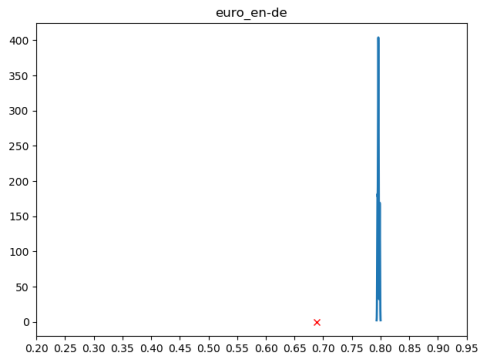


Figure 4: Resampling results for comparison of English and German parallel Europarl corpora for 5 samples

terns, syntax etc.) that cannot be mapped into a shared space.

The main result of this section is that this experiment confirms the utility of the CCA measure. We can define a threshold for the CCA measure that distinguishes corpus pairs that come from the same distribution from those that come from different distributions.

5. Comparison with Domain Adaptation Results and Other Similarity Measures

In this section, we investigate the utility of the CCA measure for predicting domain similarity. We use a dataset created for domain adaptation research. Specifically, we focus on the Amazon dataset often used for domain adaptation in sentiment prediction (Blitzer et al., 2007) and compare the CCA measure to (Blitzer et al., 2007)'s results as well as to the Jensen-Shannon Divergence results reported by Barnes et al. (2018). The Amazon dataset contains product reviews for four different categories, namely books, DVDs, electronics and kitchen, annotated for sentiment. As Barnes et al. (2018), we use the unlabeled parts of the dataset to extract word embeddings and compare the domains using the CCA measure. Table 7 compares CCA measure and Jensen-Shannon Divergence, as reported by Barnes et al. (2018).[14]

	book	DVD	electronics	kitchen
book	1.000	0.940	0.870	0.864
DVD	0.505	1.000	0.873	0.866
electronics	0.365	0.365	1.000	0.908
kitchen	0.365	0.354	0.457	1.000

Table 7: Scores of the CCA measure (gray) and Jensen-Shannon Divergence (orange) on the Amazon datasets. Both approaches yield scores of 1 when a corpus is compared to itself.

Even though the CCA measure does not directly rely on frequency information, it still captures the same similarities as the frequency-distribution-based Jensen-Shannon Diver-

[14] We do not compare to the additional SemEval datasets used in their study because the amount of data available is very small, and as the CCA measure is based on word embeddings, it is not suited for corpora as small as 1 or 2.5 MB.

gence. The CCA measure is also consistent with the domain adaptation findings by Blitzer et al. (2007) and the ones reported in Barnes et al. (2018), who find that the book and DVD categories are closer to each other than to any of the other two and likewise kitchen and electronics are closer to each other than the other two in terms of adaptation loss and classification accuracy.

Another proposed similarity measure is the perplexity of a language model when trained on one domain and tested on another (as for example reported by Hu et al. (2019) Appendix A.2). This measure, however, is not symmetric, which makes it not straightforward to compare to. As symmetry is not necessarily required, it would be interesting to adapt the CCA measure by using different mapping algorithms, for example based on the Orthogonal Procrustes problem (Mikolov et al., 2013a; Artetxe et al., 2016; Artetxe et al., 2017; Lample et al., 2018; Lubin et al., 2019). Examining this will be left for future work.

6. Conclusion

In this work, we presented the CCA measure, a new measure to capture domain similarity based on the dimension-wise correlation of embeddings spaces. In contrast to other approaches of domain similarity, our CCA measure does not rely on

To investigate the relations between embedding spaces from different domains and languages, we map PPMI+SVD embeddings into a maximally correlated space by applying CCA to a selected set of pairwise combinations. We define the CCA measure as the mean of the dimension-wise correlations. This allows us to rank corpus pairs according to their similarity. We have shown that the same ranking of different domains can be produced for English, German, Spanish and Czech monolingual corpora as well as in cross-lingual comparisons.

In the cross-lingual setting, the CCA measure scores show a similar pattern across domains, but are lower in general. We compared parallel corpora to measure the similarity of embedding spaces across languages excluding the factor domain differences. The results suggest that the difference can be reduced by using a high quality bilingual lexicon but are still below the monolingual CCA measure scoress. Further investigation is necessary to determine whether these differences are due to "unbridgeable" differences between languages (words without exact 1-1 correspondences, ambiguity patterns, syntax etc.) or whether more complex mapping algorithms could make cross-language and cross-domain scores of the CCA measure more comparable.

To examine the interpretability of the CCA measure, we applied permutation tests. This allowed us to define a threshold for the CCA measure that distinguishes corpus pairs that come from the same distribution from those that come from different distributions.

As long as enough data is available to train reliable word embeddings, the CCA measure is a reliable measure for corpus similarity. We confirmed this on a domain adaptation dataset for sentiment analysis. The CCA measure yields the same rankings as text based measures such as the Jensen-Shannon Divergence of frequency distributions.

The extent to which the CCA measure is applicable to embeddings trained by different algorithms, based on characters or sub-word units, and of differing embedding dimensionalities will have to be studied in future work.

7. Acknowledgments

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content.

8. Bibliographical References

Artetxe, M., Labaka, G., and Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Artetxe, M., Labaka, G., and Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 451–462.

Barnes, J., Klinger, R., and Schulte im Walde, S. (2018). Projecting Embeddings for Domain Adaption: Joint Modeling of Sentiment Analysis in Diverse Domains. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 818–830.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007). Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 19*, pages 137–144.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, Bollywood, Boom-boxes and Blenders:Domain Adaptation for Sentiment Classification. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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*, pages 4171–4186.

Faruqui, M. and Dyer, C. (2014). Improving Vector Space Word Representations Using Multilingual Correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, 28(3-4):321–377.

Hu, J., Xia, M., Neubig, G., and Carbonell, J. (2019). Domain Adaptation of Neural Machine Translation by Lexicon Induction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2989–3001.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jgou, H. (2018). Word translation without parallel data. In *International Conference on Learning Representations*.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 923–929.

Lu, A., Wang, W., Bansal, M., Gimpel, K., and Livescu, K. (2015). Deep Multilingual Correlation for Improved Word Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256.

Lubin, N. Y., Goldberger, J., and Goldberg, Y. (2019). Aligning Vector-spaces with Noisy Supervised Lexicon. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 460–465.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013a). Exploiting Similarities among Languages for Machine Translation. *arXiv preprint*, arXiv:1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Rastogi, P., Van Durme, B., and Arora, R. (2015). Multiview LSA: Representation Learning via Generalized CCA. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–566.

Schütze, H. (1993). Word Space. In *Advances in Neural Information Processing Systems 5*, pages 895–902.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 2214–2218.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.