Spectral Text Similarity Measures

Tim vor der Brück and Marc Pouly

School of Information Technology
Lucerne University of Applied Sciences and Arts
Switzerland
{tim.vorderbrueck,marc.pouly}@hslu.ch

Abstract. Estimating semantic similarity between texts is of vital importance in many areas of natural language processing like information retrieval, question answering, text reuse or plagiarism detection.
Prevalent semantic similarity estimates that are based on word embeddings are noise sensitive. Thus, small individual term similarities can have in aggregate a considerable influence on the total estimation value. In contrast, the methods proposed here exploit the spectrum of the product of embedding matrices, which leads to increased robustness when compared with conventional methods.
We apply these estimate on two tasks, which are the assignment of people to the best matching marketing target group and finding the correct match between sentences belonging to two independent translations of the same novel. The evaluation revealed that our proposed methods could increase accuracy in both scenarios.

1 Introduction

Estimating semantic document similarity is of vital importance in a lot of different areas, like plagiarism detection, information retrieval, or text summarization. One drawback of current state-of-the-art similarity estimates based on word embeddings is that small term similarities can sum up to a considerable amount and make these estimates vulnerable to noise in the data. Therefore, we propose two estimates that are based on the spectrum of the product $\mathbf{F}$ of embedding matrices belonging to the two documents to compare. In particular, we propose the spectral radius and the spectral norm of $\mathbf{F}$, where the first denotes $\mathbf{F}$'s largest absolute eigenvalue and the second its largest singular value. Eigenvalue and singular value oriented methods for dimensionality reduction aiming to reduce noise in the data have a long tradition in natural language processing. For instance, principal component analysis is based on eigenvalues and can be used to increase the quality of word embeddings [8]. In contrast, Latent Semantic Analysis [11], a technique known from information retrieval to improve search results in term-document matrices, focuses on largest singular values.

Furthermore, we investigate several properties of our proposed measures that are crucial for qualifying as proper similarity estimates, while considering both unsupervised and supervised learning.

Finally, we applied both estimates to two natural language processing scenarios. In the first scenario, we distribute participants of an online contest into several target

groups by exploiting short text snippets they were asked to provide. In the second scenario, we aim to find the correct matching between sentences originating from two independent translations of a novel from Edgar Allen Poe.

The evaluation revealed that our novel estimators performed superior to several baseline methods for both scenarios.

The remainder of the paper is organized as follows. In the next section, we look into several state-of-the-art methods for estimating semantic similarity. Sect. 3 reviews several concepts that are vital for the remainder of the paper and also for building the foundation of our theoretical results. In Sect. 4 we describe in detail, how the spectral radius can be employed for estimating semantic similarity. Some drawbacks and shortcomings of such an approach as well as an alternative method that very elegantly solves all of these issues exploiting the spectral norm are discussed in Sect. 5. The two application scenario for our proposed semantic similarity estimates are given in Sect. 6. Sect. 7 describes the conducted evaluation, in which we compare our approach with several baseline methods. The results of the evaluation are discussed in Sect. 8. So far, we covered only unsupervised learning. In Sect. 9 we investigate, how our proposed estimates can be employed in a supervised setting. Finally, this paper concludes with Sect. 10, which summarizes the obtained results.

2 Related Work

Until recently, similarity estimates were predominantly based either on ontologies [4] or on typical information retrieval techniques like Latent Semantic Analysis. In the last couple of years, however, so-called word and sentence embeddings became state-of-the-art.

The prevalent approach to document similarity estimation based on word embeddings consists of measuring similarity between vector representations of the two documents derived as follows:

1. The word embeddings (often weighted by the tf-idf coefficients of the associated words [3]) are looked up in a hashtable for all the words in the two documents to compare. These embeddings are determined beforehand on a very large corpus typically using either the skip gram or the continuous bag of words variant of the Word2Vec model[15]. The skip gram method aims to predict the textual surroundings of a given word by means of an artificial neural network. The influential weights of the one-hot-encoded input word to the nodes of the hidden layer constitute the embedding vector. For the so-called *continuous bag of words* method, it is just the opposite, i.e., the center word is predicted by the words in its surrounding.
2. The centroid over all word embeddings belonging to the same document is calculated to obtain its vector representation.

Alternatives to Word2Vec are GloVe [17], which is based on aggregated global word co-occurrence statistics and the Explicit Semantic Analysis (or shortly ESA) [6], in which each word is represented by the column vector in the tf-idf matrix over Wikipedia.

The idea of Word2Vec can be transferred to the level of sentences as well. In particular, the so-called Skip Thought Vector (STV) model [10] derives a vector representation of the current sentence by predicting the surrounding sentences.

If vector representations of the two documents to compare were successfully established, a similarity estimate can be obtained by applying the cosine measure to the two vectors. [18] propose an alternative approach for ESA word embeddings that establishes a bipartite graph consisting of the best matching vector components by solving a linear optimization problem. The similarity estimate for the documents is then given by the global optimum of the objective function. However, this method is only useful for sparse vector representations. In case of dense vectors, [14] suggested to apply the Frobenius kernel to the embedding matrices, which contain the embedding vectors for all document components (usually either sentences or words, cf. also [9]). However, crucial limitations are that the Frobenius kernel is only applicable if the number of words (sentences respectively) in the compared documents coincide and that a word from the first document is only compared with its counterpart from the second document. Thus, an optimal matching has to be established already beforehand. In contrast, the approach as presented here applies to arbitrary embedding matrices. Since it compares all words of the two documents with each other, there is also no need for any matching method.

Before going more into detail, we want to review some concepts that are crucial for the remainder of this paper.

3 Similarity Measure / Matrix norms

According to [2], a similarity measure on some set $\mathbf{X}$ is an upper bounded, exhaustive and total function $s : X \times X \to I \subset \mathbb{R}$ with $|I| > 1$ (therefore I is upper bounded and $\sup I$ exists). Additionally, a similarity measure should fulfill the properties of reflexivity (the supremum is reached if an item is compared to itself) and symmetry. We call such a measure normalized if the supremum equals 1 [1]. Note that an asymmetric similarity measure can easily be converted into a symmetric by taking the geometric or arithmetic mean of the asymmetric measure applied twice to the same arguments in switched order.

A norm is a function $f : V \to \mathbb{R}$ over some vector space V that is absolutely homogeneous, positive definite and fulfills the triangle inequality. It is called matrix norm if its domain is a set of matrices and if it is sub-multiplicative, i.e., $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$. An example of a matrix norm is the spectral norm, which denotes the largest singular value of a matrix. Alternatively, one can define this norm as: $\|\mathbf{A}\|_2 := \sqrt{\rho(\mathbf{A}^\top \mathbf{A})}$, where the function ρ returns the largest absolute eigenvalue of the argument matrix.

4 Document Similarity Measure based on the Spectral Radius

For an arbitrary document t we define the embeddings matrix $E(t)$ as follows: $E(t)_{ij}$ is the i-th component of the normalized embeddings vector belonging to the j-th word of the document t. Let t, u be two arbitrary documents, then the entry (i, j) of a product $\mathbf{F} := E(t)^\top E(u)$ specifies the result of the cosine measure estimating the semantic similarity between word i of document t and word j of document u.

The larger the matrix entries of $\mathbf{F}$ are, the higher is usually the semantic similarity of the associated texts. A straight-forward way to measure the magnitude of the matrix is just to summate all absolute matrix elements, which is called the L_{11}-norm. However, this approach has the disadvantage that also small cosine measure values are included in the sum, which can have in aggregate a considerable impact on the total similarity estimate making such an approach vulnerable to noise in the data. Therefore we propose instead to apply an operator, which is more robust than the $L_{1,1}$ norm instead and which is called the spectral radius.

This radius denotes the largest absolute eigenvalue of the input matrix and constitutes a lower bound of all matrix norms. It also insinuates the convergence of the matrix power series $\lim_{n\to\infty} \mathbf{F}^n$. The series converges if and only if the spectral radius does not exceed the value of one.

Since the vector components obtained by Word2Vec can be negative, the cosine measure between two word vectors can also assume negative values (rather rarely in practice though). Akin to zeros, negative cosine values indicate unrelated words as well. Because the spectral radius usually treats negative and positive matrix entries alike (the spectral radius of a matrix A and of its negation coincide), we replace all negative values in the matrix by zero. Finally, since our measure should be restricted to values from zero to one, we have to normalize it. Formally, we define our similarity measure as follows:

$$sn(t,u) := \frac{\rho(R(E(t)^\top E(u))}{\sqrt{\rho(R(E(t)^\top E(t)) \cdot \rho(R(E(u)^\top E(u)))}}$$

where $E(t)$ is the embeddings matrix belonging to document t, where all embedding column vectors are normalized. $R(\mathbf{M})$ is the matrix, where all non-zero entries are replaced by zero, i.e. $R(\mathbf{M})_{ij} = \max\{0, \mathbf{M}_{ij}\}$.

In contrast to matrix norms that can be applied to arbitrary matrices, eigenvalues only exist for quadratic matrices. However, the matrix $\mathbf{F}^* := R(E(t)^\top E(u))$ that we use as basis for our similarity measures is usually non-quadratic. In particular, this matrix would be quadratic, if and only if the number of terms in the two documents t and u coincide. Thus, we have to fill up the embedding matrix of the smaller one of the two texts with additional embedding vectors. A quite straightforward choice, which we followed here, is to just use the centroid vector for this. An alternative approach would be to sample the missing vectors.

A further issue is that eigenvalues are not invariant concerning row and column permutations. The columns of the embedding matrices just represent the words appearing in the texts. However, the word order can be arbitrarily for the text representing the marketing target groups (see Sect. 6.1 for details). Since a similarity measure should not be dependent on some random ordering, we need to bring the similarity matrix $\mathbf{F}^*$ in some normalized format. A quite natural choice would be to enforce the ordering that maximizes the absolute value of the largest eigenvalue (which is actually our target value). Let us formalize this. We denote with $\mathbf{F}^*_{P,Q}$ the matrix obtained from $\mathbf{F}^*$ by applying the permutation P on the rows and the permutation Q on the columns. Thus, we can define our similarity measure as follows:

$$sim_{sr}(t,u) = \max_{P,Q} \rho(\mathbf{F}^*_{P,Q}) \tag{1}$$

However, solving this optimization problem is quite time-consuming. Let us assume the matrix $\mathbf{F}^*$ has m rows and columns. Then we would have to iterate over $m!\cdot m!$ different possibilities. Hence, such an approach would be infeasible already for medium sized texts. Therefore, we instead select the permutations that optimize the absolute value of the arithmetic mean over all eigenvalues, which is a lower bound of the maximum absolute eigenvalue.

Let $\lambda_i(\mathbf{M})$ be the i-th eigenvalue of a matrix $\mathbf{M}$. With this, we can formalize our optimization problem as follows:

$$\begin{aligned} \tilde{sim}_{sr}(t,u) =& \rho(\mathbf{F}^*_{\tilde{P},\tilde{Q}}) \\ \tilde{P},\tilde{Q} =& \arg\max_{P,Q} |\sum_{i=1}^{m} \lambda_i(\mathbf{F}^*_{P,Q})| \end{aligned} \tag{2}$$

The sum over all eigenvalues is just the trace of the matrix. Thus,

$$\tilde{P},\tilde{Q} = \arg\max_{P,Q} |\operatorname{tr}(\mathbf{F}^*_{P,Q})| \tag{3}$$

which is just the sum over all diagonal elements. Since we constructed our matrix $\mathbf{F}^*$ in such a way that it contains no negative entries, we can get rid of the absolute value operator.

$$\tilde{P},\tilde{Q} = \arg\max_{P,Q} \{\operatorname{tr}(\mathbf{F}^*_{P,Q})\} \tag{4}$$

Because the sum is commutative, the sequence of the individual summands is irrelevant. Therefore, we can leave either the row or column ordering constant and only permutate the other one.

$$\begin{aligned} sim_{sr}(t,u) =& \rho(\mathbf{F}^*_{\tilde{P},id}) \\ \tilde{P} =& \arg\max_{P} \{\operatorname{tr}(\mathbf{F}^*_{P,id})\} \end{aligned} \tag{5}$$

P can be found by solving a binary linear programming problem in the following way. Let X be the set of decision variables and let further $X_{ij} \in X$ be one if and only if row i is changed to row j in the reordered matrix and zero otherwise. Then the objective function is given by $\max_X \sum_{i=1}^{m}\sum_{j=1}^{m} X_{ji}F^*_{ji}$. A permutation denotes an 1:1 mapping, i.e.

$$\begin{aligned} &\sum_{i=1}^{m} X_{ij} = 1 \; \forall j = 1,\ldots,m \\ &\sum_{j=1}^{m} X_{ij} = 1 \; \forall i = 1,\ldots,m \\ &X_{ij} \in \{0,1\} \; \forall i,j = 1,\ldots,m \end{aligned} \tag{6}$$

5 Spectral Norm

The similarity estimate as described above has several drawbacks.

- The boundedness condition is violated in some cases. Therefore, this similarity does not qualify as a normalized similarity estimate according to the definition in Sect. 3.
- The largest eigenvalues of a matrix depends on the row and column ordering. However, this ordering is arbitrary for our proposed description of target groups by keywords (cf. Sect. 6.1 for the details). To ensure a unique eigenvalue, we apply linear optimization, which is an expensive approach in terms of runtime.
- Eigenvalues are only defined for square matrices. Therefore, we need to fill up the smaller of the embedding matrices to meet this requirement.

An alternative to the spectral radius is the spectral norm, which is defined by the largest singular value of a matrix. Formally, the spectral norm based estimate is given as:

$$sn_2(t,u) := \frac{\|(R(E(t)^\top E(u))\|_2}{\sqrt{\|R(E(t)^\top E(t))\|_2 \cdot \|R(E(u)^\top E(u))\|_2}}$$

where $\|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^\top \mathbf{A})}$.

By using the spectral norm instead of the spectral radius, all of the issues mentioned above are solved. The spectral norm is not only invariant to column or row permutations, it can also be applied to arbitrary rectangular matrices. Furthermore, the boundedness is guaranteed as long as no negative cosine values occur as it is stated in the following proposition.

Proposition 1. *If the cosine similarity values between all embedding vectors of words occurring in any of the documents are non-negative, i.e., if* $R(E(t)^\top E(u)) = E(t)^\top E(u)$ *for all document pairs* (t,u)*, then* sn_2 *is a normalized similarity measure.*

Symmetry

Proof. At first, we focus on the symmetry condition.

Let $\mathbf{A} := E(t)$, $\mathbf{B} := E(u)$, where t and u are arbitrary documents. Symmetry directly follows, if we can show that

$$\|\mathbf{Z}\|_2 = \|\mathbf{Z}^\top\|_2$$

for arbitrary matrices $\mathbf{Z}$, since with this property we have

$$\begin{aligned} sn_2(t,u) =& \frac{\|\mathbf{A}^\top \mathbf{B}\|_2}{\sqrt{\|\mathbf{A}^\top \mathbf{A}\|_2 \cdot \|\mathbf{B}^\top \mathbf{B}\|_2}} \\ =& \frac{\|(\mathbf{B}^\top \mathbf{A})^\top\|_2}{\sqrt{\|\mathbf{B}^\top \mathbf{B}\|_2 \cdot \|\mathbf{A}^\top \mathbf{A}\|_2}} \\ =& \frac{\|\mathbf{B}^\top \mathbf{A}\|_2}{\sqrt{\|\mathbf{B}^\top \mathbf{B}\|_2 \cdot \|\mathbf{A}^\top \mathbf{A}\|_2}} \\ =& sn_2(u,t) \end{aligned} \tag{7}$$

Let $\mathbf{M}$ and $\mathbf{N}$ be arbitrary matrices such that $\mathbf{MN}$ and $\mathbf{NM}$ are both defined and quadratic, then (see [5])

$$\rho(\mathbf{MN}) = \rho(\mathbf{NM}) \tag{8}$$

where $\rho(\mathbf{X})$ denotes the largest absolute eigenvalue of a squared matrix $\mathbf{X}$.

Using identity 8 one can easily infer that:

$$\|\mathbf{Z}\|_2 = \sqrt{\rho(\mathbf{Z}^\top \mathbf{Z})} \;=\; \sqrt{\rho(\mathbf{Z}\mathbf{Z}^\top)} \;=\; \|\mathbf{Z}^\top\|_2 \tag{9}$$

Boundedness

Proof. The following property needs to be shown:

$$\frac{\|\mathbf{A}^\top \mathbf{B}\|_2}{\sqrt{\|\mathbf{A}^\top \mathbf{A}\|_2 \cdot \|\mathbf{B}^\top \mathbf{B}\|_2}} \leq 1 \tag{10}$$

In the proof, we exploit the fact that for every positive-semidefinite matrix $\mathbf{X}$, the following equation holds

$$\rho(\mathbf{X}^2) = \rho(\mathbf{X})^2 \tag{11}$$

We observe that for the denominator

$$\begin{aligned}
&\|\mathbf{A}^\top \mathbf{A}\|_2 \cdot \|\mathbf{B}^\top \mathbf{B}\|_2 \\
=&\sqrt{\rho((\mathbf{A}^\top \mathbf{A})^\top \mathbf{A}^\top \mathbf{A})}\sqrt{\rho((\mathbf{B}^\top \mathbf{B})^\top \mathbf{B}^\top \mathbf{B})} \\
=&\sqrt{\rho((\mathbf{A}^\top \mathbf{A})^\top (\mathbf{A}^\top \mathbf{A})^\top)}\sqrt{\rho((\mathbf{B}^\top \mathbf{B})^\top (\mathbf{B}^\top \mathbf{B})^\top)} \\
=&\sqrt{\rho([(\mathbf{A}^\top \mathbf{A})^\top]^2)}\sqrt{\rho([(\mathbf{B}^\top \mathbf{B})^\top]^2)} \\
\overset{(11)}{=}&\sqrt{\rho((\mathbf{A}^\top \mathbf{A})^\top)^2}\sqrt{\rho((\mathbf{B}^\top \mathbf{B})^\top)^2} \\
=&\rho((\mathbf{A}^\top \mathbf{A})^\top)\rho((\mathbf{B}^\top \mathbf{B})^\top) \\
\overset{(9)}{=}&\|\mathbf{A}\|_2^2 \cdot \|\mathbf{B}\|_2^2
\end{aligned} \tag{12}$$

Putting things together we finally obtain

$$\begin{aligned}
\frac{\|\mathbf{A}^\top \mathbf{B}\|_2}{\sqrt{\|\mathbf{A}^\top \mathbf{A}\|_2 \|\mathbf{B}^\top \mathbf{B}\|_2}} &\overset{\text{sub-mult.}}{\leq} \frac{\|\mathbf{A}^\top\|_2 \cdot \|\mathbf{B}\|_2}{\sqrt{\|\mathbf{A}^\top \mathbf{A}\|_2 \|\mathbf{B}^\top \mathbf{B}\|_2}} \\
&\overset{(9)}{=} \frac{\|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2}{\sqrt{\|\mathbf{A}^\top \mathbf{A}\|_2 \|\mathbf{B}^\top \mathbf{B}\|_2}} \\
&\overset{(12)}{=} \frac{\|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2}{\sqrt{\|\mathbf{A}\|_2^2 \cdot \|\mathbf{B}\|_2^2}} = 1
\end{aligned} \tag{13}$$

The question remains, how the similarity measure value induced by matrix norms performs in comparison with the usual centroid method. General statements about the

spectral-norm based similarity measure are difficult, but we can draw some conclusions, if we restrict to the case, where $\mathbf{A}^\top\mathbf{B}$ is a square diagonal matrix. Hereby, one word of the first text is very similar to exactly one word of the second text and very dissimilar to all remaining words. The similarity estimate is then given by the largest eigenvalue (the spectral radius) of $\mathbf{A}^\top\mathbf{B}$, which equals the largest cosine measure value. Noise in form of small matrix entries is completely ignored.

6 Application Scenarios

We applied our semantic similarity estimates to the following two scenarios:

6.1 Market Segmentation

Market segmentation is one of the key tasks of a marketer. Usually, it is accomplished by clustering over behaviors as well as demographic, geographic and psychographic variables [12]. In this paper, we will describe an alternative approach based on unsupervised natural language processing. In particular, our business partner operates a commercial youth platform for the Swiss market, where registered members get access to third-party offers such as discounts and special events like concerts or castings. Actually, several hundred online contests per year are launched over this platform sponsored by other firms, an increasing number of them require the members to write short free-text snippets, e.g. to elaborate on a perfect holiday at a destination of their choice in case of a contest sponsored by a travel agency. Based on the results of a broad survey, the platform provider's marketers assume five different target groups (called *milieus*) being present among the platform members: *progressive postmodern youth* (people primarily interested in culture and arts), *young performers* (people striving for a high salary with a strong affinity to luxury goods), *freestyle action sportsmen*, *hedonists* (rather poorly educated people who enjoy partying and disco music) and *conservative youth* (traditional people with a strong concern for security). A sixth milieu called *special groups* comprises all those who cannot be assigned to one of the upper five milieus. For each milieu (with the exception of *special groups*) a keyword list was manually created by describing its main characteristics. For triggering marketing campaigns, an algorithm shall be developed that automatically assigns each contest answer to the most likely target group: we propose the youth milieu as best match for a contest answer, for which the estimated semantic similarity between the associated keyword list and user answer is maximal. In case the highest similarity estimate falls below the 10 percent quantile for the distribution of highest estimates, the special groups milieu is selected.

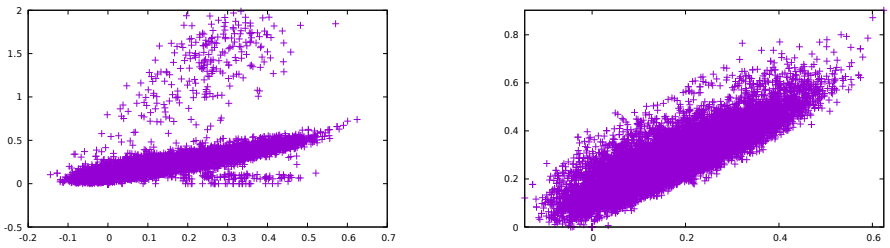
Since the keyword list typically consists of nouns (in the German language capitalized) and the user contest answers might contain a lot of adjectives and verbs as well, which do not match very well to nouns in the Word2Vec vector representation, we actually conduct two comparisons for our Word2Vec based measures, one with the unchanged user contest answers and one by capitalizing every word beforehand. The final similarity estimate is then given as the maximum value of both individual estimates.

6.2 Translation Matching

The novel *The purloined letter* authored by Edgar Allen Poe was independently translated by two translators into German[1]. We aim to match a sentence from the first translation to the associated sentence of the second by looking for the assignment with the highest semantic relatedness disregarding the sentence order. To guarantee an 1:1 sentence mapping, periods were partly replaced by semicolons.

7 Evaluation

For evaluation we selected three online contests (language: German), where people elaborated on their favorite travel destination (contest 1, see Appendix A for an example), speculated about potential experiences with a pair of fancy sneakers (contest 2) and explained why they emotionally prefer a certain product out of four available candidates. In order to provide a gold standard, three professional marketers from different youth marketing companies annotated independently the best matching youth milieus for every contest answer. We determined for each annotator individually his/her average inter-annotator agreement with the others (Cohen's kappa). The minimum and maximum of these average agreement values are given in Table 2. Since for contest 2 and contest 3, some of the annotators annotated only the first 50 entries (last 50 entries respectively), we specified min/max average kappa values for both parts. We further compared the youth milieus proposed by our unsupervised matching algorithm with the majority votes over the human experts' answers (see Table 3) and computed its average inter-annotator agreement with the human annotators (see again Table 2). The obtained accuracy values for the second scenario (Matching translated sentences) are given in Table 4.



(a) W2VC / Spectral radius.

(b) W2VC / Spectral norm.

Fig. 1: Scatter Plots of Cosine between Centroids of Word2Vec Embeddings (W2VC) vs similarity estimates induced by different spectral measures.

The Word2Vec word embeddings were trained on the German Wikipedia (dump originating from 20 February 2017) merged with a Frankfurter Rundschau newspaper

[1] This corpus can be obtained under the URL `https://www.researchgate.net/publication/332072718_alignmentPurloinedLettertar`

Table 1: Corpus sizes measured by number of words.

Corpus	# Words
German Wikipedia	651 880 623
Frankfurter Rundschau	34 325 073
News journal *20 Minutes*	8 629 955

Table 2: Minimum and maximum average inter-annotator agreements (Cohen's kappa) / average inter-annotator agreement values for our automated matching method.

Method	Contest		
	1	2	3
Min kappa	0.123	0.295/0.030	0.110/0.101
Max. kappa	0.178	0.345/0.149	0.114/0.209
Kap. (spectral norm)	0.128	0.049/0.065	0.060/0.064
# Entries	1544	100	100

Table 3: Obtained accuracy values for similarity measures induced by different matrix norms and for five baseline methods. (W)W2VC=Cosine between (weighed by tf-idf) Word2Vec Embeddings Centroids.

Method	Contest			
	1	2	3	all
Random	0.167	0.167	0.167	0.167
ESA	0.357	0.254	**0.288**	0.335
ESA2	0.355	0.284	0.227	0.330
W2VC	0.347	**0.328**	0.227	0.330
WW2VC	0.347	0.299	0.197	0.322
Skip-Thought-Vectors	0.162	0.284	0.273	0.189
Spectral Norm	**0.370**	0.299	**0.288**	**0.350**
Spectral Radius	0.353	0.313	0.182	0.326
Spectral Radius+W2VC	0.357	0.299	0.212	0.334

Corpus and 34 249 articles of the news journal *20 minutes*[2], where the latter is targeted to the Swiss market and freely available at various Swiss train stations (see Table 1 for a comparison of corpus sizes). By employing articles from *20 minutes*, we want to ensure the reliability of word vectors for certain Switzerland specific expressions like *Velo* or *Glace*, which are underrepresented in the German Wikipedia and the Frankfurter Rundschau corpus.

ESA is usually trained on Wikipedia, since the authors of the original ESA paper suggest that the articles of the training corpus should represent disjoint concepts, which

[2] http://www.20min.ch

is only guaranteed for encyclopedias. However, Stein and Anerka [7] challenged this hypothesis and demonstrated that promising results can be obtained by applying ESA on other types of corpora like the popular Reuters newspaper corpus as well. Unfortunately, the implementation we use (Wikiprep-ESA[3]) expects its training data to be a Wikipedia Dump. Furthermore, Wikiprep-ESA only indexes words that are connected by hyperlinks, which are usually lacking in ordinary newspaper articles. So we could train ESA on Wikipedia only but we have developed meanwhile a reimplementation of ESA that can be applied to arbitrary corpora and which was trained on the full corpus (Wikipedia+Frankfurter Rundschau+20 minutes). In the following we refer to this implementation as ESA2.

The STVs (Skip Thought Vectors) were trained on the same corpus as our estimates and Word2Vec embedding centroids (W2VC). The actual document similarity estimation is accomplished by the usual centroid approach. An issue we are faced with for the first evaluation scenario of market segmentation (see Sect. 6.1) is that STVs are not bag of word models but actually take the sequence of the words into account and therefore the obtained similarity estimate between milieu keyword list and contest answer would be dependent on the keyword ordering. However, this order could have arbitrarily been chosen by the marketers and might be completely random. A possible solution is to compare the contest answers with all possible permutation of keywords and determine the maximum value over all those comparisons. However, such an approach would be infeasible already for medium keyword list sizes. Therefore, we apply for this scenario a beam search to extends the keyword list iteratively while keeping only the n-best performing permutations.

Table 4: Accuracy value obtained for matching a sentence of the first to the associated sentence of the second translation (based on the first 200 sentences of both translations).

Method	Accuracy
ESA	0.672
STV	0.716
Spectral Radius	0.721
W2VC	0.726
Spectral Norm	**0.731**

8 Discussion

The evaluation showed that the inter-annotator agreement values vary strongly for contest 2 part 2 (minimum average annotator agreement according to Cohen's kappa of 0.03 while the maximum is 0.149, see Table 2). On this contest part, our spectral norm based matching obtains a considerably higher average agreement than one of the annotators. Regarding baseline systems, the most relevant comparison is naturally the one

[3] https://github.com/faraday/wikiprep-esa

with W2VC, since it employs the same type of data. The similarity estimate induced by the spectral norm performs quite stable over both scenarios and clearly outperforms the W2VC approach. In contrast however, the performance of the spectral radius based estimate is rather mixed. While it performs well on the first contest, the performance on the third contest is quite poor and lags there behind the Word2Vec centroids. Only the average of both measures (W2VC+Spectral Radius) performs reasonable well on all three contests. One major issue of this measure is its unboundedness. The typical normalization with the geometric mean of comparing the documents with itself results in values exceeding the desired upper limit of one in 1.8% of the cases (determined on the largest contest 1). So still some research is needed to come up with a better normalization.

Finally, we conducted a scatter plot (see Fig. 1), plotting the values of the spectral similarity estimates against W2VC. While the spectral norm is quite strongly correlated to W2VC, the spectral radius behaves much more irregular and non-linear. In addition, its values exceed several times the desired upper limit of 1, which is a result of its non-boundedness. Furthermore, both of the spectral similarity estimates tend to assume larger values than W2VC, which is a result of its higher robustness against noise in the data.

Note that a downside of both approaches in relation to the usual Word2Vec centroids method is the increased runtime, since it requires the pair-wise comparison of all words contained in the input documents. In our scenario with rather short text snippets and keyword lists, this was not much of an issue. However, for large documents, such a comprehensive comparison could become soon infeasible. This issue can be mitigated for example by constructing the embedding matrices not on basis of individual words but on entire sentences, for instance by employing the skip-thought-vector representation.

9 Supervised Learning

So far, our two proposed similarity measures were only applied in an unsupervised setting. However, supervised learning methods usually obtain superior accuracy. For that, we could use our two similarity estimates as kernels for a suppport vector machine [19] (not yet evaluated however), potentially combined with an RBF kernel applied to an ordinary feature representation consisting of tf-idf-weights of word forms or lemmas. One issue here is to investigate, whether our proposed similarity estimates are positive semidefinite and qualify as regular kernels. In case of non positive-semidefiniteness, the SVM (short for support vector machine) training process can stuck in a local minimum resulting in failing to reach the global minimum for the hinge loss.

The estimate induced by the spectral radius and also the spectral norm in case of negative cosine measure values between word embedding vectors can possibly violate the boundedness constraint and therefore, it cannot constitute a positive-semidefinite kernel. To see this, let us regard the kernel matrix $\mathbf{K}$. According to Mercer's theorem [13, 16], an SVM kernel is exactly then positive-semidefinite, if for any possible set of inputs, the associated kernel matrices are positive-semidefinite. So we must show that there is at least one kernel matrix that is not positive-semidefinite. Let us select one

kernel matrix $\mathbf{K}$ with at least one violation of boundedness. We can assume that $\mathbf{K}$ is symmetric, since symmetry is a prerequisite for positive-semidefiniteness.

Since our normalization procedure gurantees reflexivity, a text compared with itself always yields the estimated similarity of one. Therefore, the value of one can only be exceeded for off-diagonal elements. Let us assume the entry $K_{ij} = K_{ji}$ with $i < j$ of the kernel matrix equals $1+\epsilon$ for some $\epsilon > 0$. Consider a vector $\mathbf{v}$ with $v_i = 1$, $v_j = -1$ and all other components equal to zero. Let $\mathbf{w} := \mathbf{v}^\top \mathbf{K}$ and $q := \mathbf{v}^\top \mathbf{K}\mathbf{v} = \mathbf{w}\mathbf{v}$, then $w_i = 1 - (1 + \epsilon) = -\epsilon$ and $w_j = 1 + \epsilon - 1 = \epsilon$. With this, it follows that $q = -\epsilon - \epsilon = -2\epsilon$. And therefore $\mathbf{K}$ cannot be positive-semidefinite.

Note that sn_2 can be a proper kernel in certain situations. Consider the fact that all of the investigated texts are so dissimilar that the kernel matrices are diagonal dominant for all possible sets of inputs. Since diagonal dominant matrices with non-negative diagonal elements are positive-semidefinite, the kernel is positive-semidefinite as well. It is still an open question if this kernel can also be positive-semidefinite if not all of the kernel matrices are diagonal dominant.

10 Conclusion

We proposed two novel similarity estimates based on the spectrum of the product of embedding matrices. These estimates were evaluated on a two task, i.e., assigning users to the best matching marketing target groups and matching sentences of a novel translation with its counterpart from a different translation. Hereby, we obtained superior results compared to the usual centroid of word2vec vectors (W2VC) method. Furthermore, we investigated several properties of our estimates concerning boundness and positive-definiteness.

A Example Contest Answer

The following snippet is an example user answer for the travel contest (contest 1):

1. Jordanien: Ritt durch die Wüste und Petra im Morgengrauen bestaunen bevor die Touristenbusse kommen
2. Cook Island: Schnorcheln mit Walhaien und die Seele baumeln lassen
3. USA: Eine abgespaceste Woche am Burning Man Festival erleben

English translation:

1. Jordan: Ride through the desert and marveling Petra during sunrise before the arrival of tourist buses
2. Cook Island: Snorkeling with whale sharks and relaxing
3. USA: Experience an awesome week at the Burning Man Festival

Acknowledgement

Hereby we thank the Jaywalker GmbH as well as the Jaywalker Digital AG for their support regarding this publication and especially for annotating the contest data with the best-fitting youth milieus.

References

1. Attig, A., Perner, P.: The problem of normalization and a normalized similarity measure by online data. Transactions on Case-Based Reasoning **4**(1) (2011)
2. Belanche, L.A., Orozco, J.: Things to know about a (dis)similarity measure. In: Proceedings of the 15th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. Karlsruhe, Germany (2011)
3. Brokos, G.I., Prodromos, Androutsopoulos, I.: Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. pp. 114–118. Berlin, Germany (2016)
4. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of semantic relatedness. Computational Linguistics **32**(1) (2006)
5. Chatelin, F.: Eigenvalues of Matrices - Revised Edition. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania (1993)
6. Gabrilovic, E., Markovitch, S.: Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research **34** (2009)
7. Gottron, T., Anderka, M., Stein, B.: Insights into explicit semantic analysis. In: Proceedings of the 20th ACM international conference on Information and knowledge management. pp. 1961–1964. Glasgow, UK (2011)
8. Gupta, V.: Improving Word Embeddings Using Kernel Principal Component Analysis. Master's thesis, Bonn-Aachen International Center for Information Technology (B-IT) (2018)
9. Hong, K.J., Lee, G.H., Kom, H.J.: Enhanced document clustering using wikipedia-based document representation. In: Proceedings of the 2015 International Conference on Applied System Innovation (ICASI) (2015)
10. Kiros, R., Zhu, Y., Salakhudinov, R., Zemel, R.S., Torralba, A., Urtasun, R., Fiedler, S.: Skip-thought vectors. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS). Montréal, Canada (2015)
11. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Processes **25**, 259–284 (1998)
12. Lynn, M.: Segmenting and targeting your market: Strategies and limitations. Tech. rep., Cornell University (2011), online: `http://scholorship.sha.cornell.edu/articles/243`
13. Mercer, J.: Functions of positive and negative type and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society A **209**, 441–458 (1909)
14. Mijangos, V., Sierra, G., Montes, A.: Sentence level matrix representation for document spectral clustering. Pattern Recognition Letters **85** (2017)
15. Mikolov, T., Sutskever, I., Ilya, C., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS). pp. 3111–3119. Lake Tahoe, Nevada (2013)
16. Murphy, K.P.: Machine Learning - A Probabilistic Perspective. MIT Press (2012)
17. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Katar (2014)
18. Song, Y., Roth, D.: Unsupervised sparse vector densification for short text similarity. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). Denver, Colorado (2015)
19. Vapnik, V.: Statistical Learning Theory. John Wiley and Sons, Inc., New York (1998)