Text Similarity Estimation Based on Word Embeddings and Matrix Norms for Targeted Marketing

Tim vor der Brück
School of Information Technology
Lucerne University of
Applied Sciences and Arts
Switzerland
tim.vorderbrueck@hslu.ch

Marc Pouly
School of Information Technology
Lucerne University of
Applied Sciences and Arts
Switzerland
marc.pouly@hslu.ch

Abstract

The prevalent way to estimate the similarity of two documents based on word embeddings is to apply the cosine similarity measure to the two centroids obtained from the embedding vectors associated with the words in each document. Motivated by an industrial application from the domain of youth marketing, where this approach produced only mediocre results, we propose an alternative way of combining the word vectors using matrix norms. The evaluation shows superior results for most of the investigated matrix norms in comparison to both the classical cosine measure and several other document similarity estimates.

1 Introduction

Estimating semantic document similarity is of utmost importance in a lot of different areas, like plagiarism detection, information retrieval, or text summarization. We will focus here on an NLP application that has been less researched, i.e., the assignment of people to the best matching target group to allow for running precise and customer-oriented marketing campaigns.

Until recently, similarity estimates were predominantly based either on deep semantic approaches 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 in measuring the similarity between the vector representations of the two documents derived as follows:

1. The word embeddings (often weighted by the tf-idf coefficients of the associated words (Brokos et al., 2016)) 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 (Mikolov et al., 2013).
2. The centroid over all word embeddings belonging to the same document is calculated to obtain its vector representation.

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.

Let $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ be the word vectors of two documents. The cosine similarity value between the two document centroids C_1 und C_2 is given by:

$$\begin{aligned}&\cos(\angle(\frac{1}{m}\sum_{i=1}^{m} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i))\\ =&\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\langle x_i, y_j\rangle}{mn\|C_1\|\|C_2\|}\end{aligned} \quad (1)$$

Hence, potentially small values of $\langle x_i, y_j\rangle$ can have in aggregate a considerable influence on the total similarity estimate, which makes this estimate vulnerable to noise in the data. We propose an alternative approach that is based on matrix norms and which proved to be more noise-robust by focusing primarily on high word similarities.

Finally, we conducted an evaluation where we achieved with our method superior accuracy in target group assignments than several traditional word embedding based methods.

2 Related Work

The most popular method to come up with word vectors is Word2Vec, which is based on a 3 layer neural network architecture in which the word vectors are obtained as the weights of the hidden layer. Alternatives to Word2Vec are GloVe

Proceedings of NAACL-HLT 2019, pages 1827–1836
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

(Pennington et al., 2014), which is based on aggregated global word co-occurrence statistics and the Explicit Semantic Analysis (or shortly ESA) (Gabrilovic and Markovitch, 2009), 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 model (STV) (Kiros et al., 2015) derives a vector representation of the current sentence by predicting the surrounding sentences.

(Song and Roth, 2015) propose an alternative approach to applying the cosine measure to the two word vector centroids for ESA word embeddings. In particular, they establish 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, (Mijangos et al., 2017) 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 (Hong et al., 2015)). 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 matrix norm approach as presented here applies to arbitrary embedding matrices. Since it conducts a pairwise comparison of all words contained in the two documents, there is also no need for any matching method.

Another simlarity estimate that employs the entire embedding matrix is the word mover‘s distance (Kusner et al., 2015), which is a special case of the earth mover‘s distance, a well studied transportation problem. Basically, this approach determines the minimum effort (with respect to embedding vector changes) to transform the words of one text into the words of another text. The word mover‘s distance requires a linear optimization problem to be solved. Linear optimization is usually tackled by the simplex method, which has in the worst case, which rarely occurs however, exponential runtime complexity.

Name	Definition
Frob. norm	$\|\mathbf{A}\|_F := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \lvert\mathbf{A}_{ij}\rvert^2}$
2-norm	$\|\mathbf{A}\|_2 := \sqrt{\rho(\mathbf{A}^\top \mathbf{A})}$
$L_{1,1}$-norm	$\|\mathbf{A}\|_{L_{1,1}} := \sum_{i=1}^{m} \sum_{j=1}^{n} \lvert\mathbf{A}_{ij}\rvert$
1-norm	$\|\mathbf{A}\|_1 := \max_{1 \leq j \leq n} \sum_{i=1}^{m} \lvert\mathbf{A}_{ij}\rvert$
∞-norm	$\|\mathbf{A}\|_\infty := \max_{1 \leq i \leq m} \sum_{j=1}^{n} \lvert\mathbf{A}_{ij}\rvert$

Table 1: Examples of matrix norms; $\mathbf{A}$ is an $m \times n$ matrix; $\rho(\mathbf{X})$ denotes the largest absolute eigenvalue of a squared matrix $\mathbf{X}$.

A drawback of conventional similarity estimates as described above is that slightly related word pairs can have in aggregate a considerable influence on their values, i.e., these estimates are sensitive to noise in the data. In contrast, several of our matrix norm based similarity estimates focus primarily on strongly related word pairs and are therefore less vulnerable to noise.

3 Similarity Measure / Matrix Norm

Before going more into detail, we want to review some concepts that are crucial for the remainder of this paper. According to (Belanche and Orozco, 2011), 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, it 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 (Attig and Perner, 2011). 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., $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$. Several popular matrix norms are given in Table 1. Note that the Frobenius norm can also be represented by $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^\top)}$.

4 Document Similarity Measure based on Matrix Norms

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 $E(t)^\top E(u)$ specifies the result of the cosine measure estimating the semantic similarity between word i of document t and word j of document u. The value of a matrix norm $\|E(t)^\top E(u)\|$ is then a measure for the similarity of the two documents. 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). Negative cosine values indicate negatively correlated words and should be handled akin to the uncorrelated case. Because a matrix norm usually treats negative and positive matrix entries alike, we replace all negative values in the matrix by zeros. Finally, since our measure should be restricted to values from zero to one, we have to normalize it.

Formally, we define our similarity measure $sn(t, u)$ as follows :

$$\frac{\|K(E(t)^\top E(u))\|}{\sqrt{\|K(E(t)^\top E(t))\| \cdot \|K(E(u)^\top E(u))\|}}$$

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

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 $K(E(t)^\top E(u)) = E(t)^\top E(u)$ for all document pairs (t, u), then sn is a normalized similarity measure for the 2-norm, the Frobenius norm and the $L_{1,1}$-norm.*

Proof. We give the proof for the 2-norm here and for the other two norms in the appendix.

Symmetry 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

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

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

$$\begin{aligned} sn(t,u) =& \frac{\|\mathbf{A}^\top\mathbf{B}\|}{\sqrt{\|\mathbf{A}^\top\mathbf{A}\| \cdot \|\mathbf{B}^\top\mathbf{B}\|}} \\ =& \frac{\|(\mathbf{B}^\top\mathbf{A})^\top\|}{\sqrt{\|\mathbf{B}^\top\mathbf{B}\| \cdot \|\mathbf{A}^\top\mathbf{A}\|}} \\ =& \frac{\|\mathbf{B}^\top\mathbf{A}\|}{\sqrt{\|\mathbf{B}^\top\mathbf{B}\| \cdot \|\mathbf{A}^\top\mathbf{A}\|}} = sn(u,t) \end{aligned} \tag{2}$$

Let $\mathbf{M}$ and $\mathbf{N}$ be arbitrary matrices such that $\mathbf{MN}$ and $\mathbf{NM}$ are both defined and quadratic, then (see (Chatelin, 1993))

$$\rho(\mathbf{MN}) = \rho(\mathbf{NM}) \tag{3}$$

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

Using identity 3 one can easily infer that:

$$\|\mathbf{Z}\|_2 = \sqrt{\rho(\mathbf{Z}^\top\mathbf{Z})} = \sqrt{\rho(\mathbf{Z}\mathbf{Z}^\top)} = \|\mathbf{Z}^\top\|_2 \tag{4}$$

□

Boundedness

Proof. The following property needs to be verified:

$$\frac{\|\mathbf{A}^\top\mathbf{B}\|_2}{\sqrt{\|\mathbf{A}^\top\mathbf{A}\|_2 \cdot \|\mathbf{B}^\top\mathbf{B}\|_2}} \leq 1 \tag{5}$$

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

$$\rho(\mathbf{X}^2) = \rho(\mathbf{X})^2 \tag{6}$$

We observe that for the denominator

$$\begin{aligned} &\|\mathbf{A}^\top\mathbf{A}\|_2 \cdot \|\mathbf{B}^\top\mathbf{B}\|_2 \\ =& \sqrt{\rho((\mathbf{A}^\top\mathbf{A})^\top\mathbf{A}^\top\mathbf{A})}\sqrt{\rho((\mathbf{B}^\top\mathbf{B})^\top\mathbf{B}^\top\mathbf{B})} \\ =& \sqrt{\rho((\mathbf{A}^\top\mathbf{A})^\top(\mathbf{A}^\top\mathbf{A})^\top)}\sqrt{\rho((\mathbf{B}^\top\mathbf{B})^\top(\mathbf{B}^\top\mathbf{B})^\top)} \\ =& \sqrt{\rho([(\mathbf{A}^\top\mathbf{A})^\top]^2)}\sqrt{\rho([(\mathbf{B}^\top\mathbf{B})^\top]^2)} \\ \overset{(6)}{=}& \sqrt{\rho((\mathbf{A}^\top\mathbf{A})^\top)^2}\sqrt{\rho((\mathbf{B}^\top\mathbf{B})^\top)^2} \\ =& \rho((\mathbf{A}^\top\mathbf{A})^\top)\rho((\mathbf{B}^\top\mathbf{B})^\top) \\ \overset{(4)}{=}& \|\mathbf{A}\|_2^2 \cdot \|\mathbf{B}\|_2^2 \end{aligned} \tag{7}$$

Putting things together we finally obtain

$$\begin{aligned} \frac{\|\mathbf{A}^\top\mathbf{B}\|_2}{\sqrt{\|\mathbf{A}^\top\mathbf{A}\|_2\|\mathbf{B}^\top\mathbf{B}\|_2}} \overset{\text{sub-mult.}}{\leq}& \frac{\|\mathbf{A}^\top\|_2 \cdot \|\mathbf{B}\|_2}{\sqrt{\|\mathbf{A}^\top\mathbf{A}\|_2\|\mathbf{B}^\top\mathbf{B}\|_2}} \\ \overset{(4)}{=}& \frac{\|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2}{\sqrt{\|\mathbf{A}^\top\mathbf{A}\|_2\|\mathbf{B}^\top\mathbf{B}\|_2}} \\ \overset{(7)}{=}& \frac{\|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2}{\sqrt{\|\mathbf{A}\|_2^2 \cdot \|\mathbf{B}\|_2^2}} = 1 \end{aligned} \tag{8}$$

□

However, proposition 1 is not sufficient in all cases, since negative cosine similarity values can occur in practice. Therefore, we also prove a stronger claim stated in the following proposition.

Proposition 2. *If the cosine measure values between embedding vectors belonging to words of the same document are all non-negative, then sn is a normalized similarity measure for the Frobenius and the $L_{1,1}$-norm.*

Proof. The proof of symmetry and reflexivity is analogous to proposition 1. So we only prove boundedness of sn. Since the cosine measure for two embedding vectors emb belonging to words of the same document cannot be negative, we have $\langle emb(w_i), emb(w_k)\rangle \geq 0$ for i,k with $1 \leq i \leq k \leq |t|$ and therefore $K(E(t)^\top E(t)) = E(t)^\top E(t)$. We furthermore have $\|K(E(t)^\top E(u))\| \leq \|E(t)^\top E(u)\|$ for the Frobenius and $L_{1,1}$-norm, since replacing a zero entry with another value can never decrease the value of the norm. Thus,

$$\begin{aligned}&\frac{\|K(E(t)^\top E(u))\|}{\sqrt{\|K(E(t)^\top E(t))\| \cdot \|K(E(u)^\top E(u))\|}} \leq \\ &\frac{\|E(t)^\top E(u)\|}{\sqrt{\|E(t)^\top E(t)\| \cdot \|E(u)^\top E(u)\|}} \leq 1.\end{aligned} \tag{9}$$

The last inequality follows from proposition 1. □

However, the proposed normalization factor $\sqrt{\|K(E(t)^\top E(t))\| \cdot \|K(E(u)^\top E(u))\|}$ is not eligible for all types of matrix norms, which is an immediate consequence of the following proposition.

Proposition 3. *There exist no mean value function $mean : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that*

$$\begin{aligned}sn_1 :=& \frac{\|K(E(t_0)^\top E(u_0))\|_1}{m}\\ m :=& mean(\|K(E(t_0)^\top E(t_0))\|_1,\\ &\|K(E(u_0)^\top E(u_0))\|_1)\end{aligned} \tag{10}$$

is bounded by one. Hence, sn_1 cannot be a normalized similarity measure.

Proof. We give a counter-example for the maximum mean, for which we show that sn_1 exceeds the value of 1:

$$E(t_0) = \begin{bmatrix}0.1644 & 0.5025\\ 0 & 0.5025\\ 0.9864 & 0.7035\end{bmatrix}$$

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

Table 2: Corpus sizes measured by number of words.

$$E(u_0) = \begin{bmatrix}0.1204 & 0.9759 & 0.2722\\ 0.2408 & 0.0976 & 0.9526\\ 0.9631 & 0.1952 & 0.1361\end{bmatrix}$$

Since the maximum mean $mean_{max}(a, b) = \max\{a, b\}$ is greater or equal to all other means (including the geometric mean), we have that:

$$\begin{aligned}&\frac{\|K(E(t_0)^\top E(u_0))\|_1}{mean(\|K(E(t_0)^\top E(t_0))\|_1, \|K(E(u_0)^\top E(u_0))\|_1)}\\ &\geq \frac{\|K(E(t_0)^\top E(u_0))\|_1}{\max\{\|K(E(t_0)^\top E(t_0))\|_1, \|K(E(u_0)^\top E(u_0))\|_1\}}\\ &= 1.0284 > 1\end{aligned} \tag{11}$$

for arbitrary type of means $mean$. □

Note that the matrices used in the counter-example can be extended to any number of embedding dimensions by adding additional zeros.

A further issue is, whether the similarity measure is invariant to word permutations. Actually, this is the case for our matrix norm similarity estimates, which is stated in the following proposition.

Proposition 4. *The obtained similarity estimate for all of the considered matrix norms is independent of the word sequence of the input texts.*

This property is quite beneficial in our scenario since one of the texts to compare constitutes of an unordered key word list (see more details in the next section).

Proof. We focus in this proof on the 2-norm, for which this property is not directly obvious like for the other regarded norms. For simplicity, we first concentrate on the special case that all cosine values between word embeddings are non-negative. This proof can easily be extended to the general case, too. In particular, we show that the similarity estimate does not change, if two columns of the first matrix are exchanged, which can be expressed by postmultiplying this matrix with a permutation matrix $\mathbf{P}$. By employing symmetry and induction this proof can be applied to arbitrary sequence permutations and to the second argument matrix as well. With this, the similarity estimate is given as:

$$
\begin{aligned}
sn_2(t,u) =& \|((\mathbf{AP})^\top \mathbf{B})\|_2 \\
=& \sqrt{\rho(((\mathbf{AP})^\top \mathbf{B})^\top ((\mathbf{AP})^\top \mathbf{B}))} \\
=& \sqrt{\rho(\mathbf{B}^\top \mathbf{APP}^\top \mathbf{A}^\top \mathbf{B})} \\
& (\mathbf{P} \text{ is an orthogonal matrix}) \qquad (12) \\
=& \sqrt{\rho(\mathbf{B}^\top \mathbf{AIA}^\top \mathbf{B})} \\
=& \sqrt{\rho(\mathbf{A}^\top \mathbf{B})^\top (\mathbf{A}^\top \mathbf{B})} \\
=& \|\mathbf{A}^\top \mathbf{B}\|_2
\end{aligned}
$$

By exploiting that $K(\mathbf{MP}) = K(\mathbf{M})\mathbf{P}$ for arbitary matrices $\mathbf{M}$, this proof can be generalized to negative cosine measure values as well. □

The question remains, how the similarity measure value induced by matrix norms performs in comparison to the usual centroid method. Let us first focus on L_{11} and the Frobenius norm. Actually, both are special cases of a norm that raises the absolute values of the matrix components to a certain power e. If this exponent e becomes large, then:

$$
\begin{aligned}
& sn_{L_{e,1}}(t,u) \\
=& \frac{\|E(u)^\top E(t)\|_{L_{e,1}}}{\sqrt{\|E(t)^\top E(t)\|_{L_{e,1}} \cdot \|E(u)^\top E(u)\|_{L_{e,1}}}} \\
\approx& \frac{(\#p)^{1/e}}{\sqrt{\left\| \begin{bmatrix} 1 & 0 & \dots \\ & \ddots & \\ 0 & \dots & 1 \end{bmatrix} \right\|_{L_{e,1}} \cdot \left\| \begin{bmatrix} 1 & 0 & \dots \\ & \ddots & \\ 0 & \dots & 1 \end{bmatrix} \right\|_{L_{e,1}}}} \\
& (\text{mainly the diagonal elements of the} \\
& \text{matrices in the denominator assume } 1) \\
\approx& \left(\frac{\#p}{\sqrt{nm}}\right)^{1/e}
\end{aligned}
\qquad (13)
$$

where $\#p$ denotes the number of perfect matches (similarity value of 1.0) between words of the two documents and n (m) is the number of words in text t (u). Thus, with an increasing exponent, $sn_{L_{e,1}}$ tends to focus on very good matches and disregards the others. This property is quite beneficial in our scenario, where often only one or two words of the contest answers (cf. next section) indicate the right target group.

General statements about the 2-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 (also called the spectral radius) of $\mathbf{A}^\top \mathbf{B}$, which equals the largest cosine measure value. Thus, the 2-norm based similarity estimate is able to filter out noise (low word similarity values) akin to the Frobenius norm.

Let us now take a look at the similarity measure sn_1, which is induced by the 1-norm. sn_1 assumes high values, if there is one word of the second document that matches very well with all words of the first document. All other less matching words of the second document do not contribute to the assumed similarity estimate at all.

5 Application to Targeted Marketing

Market segmentation is one of the key tasks of a marketer. Usually, it is accomplished by clustering over demographic variables, geographic variables, psychographic variables and behaviors (Lynn, 2011). 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 to describe its main characteristics.

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
GloVe	0.350	0.269	0.258	0.328
STV	0.157	0.313	0.258	0.189
$sn_{L_{1,1}}$	0.337	**0.328**	0.197	0.318
GM	**0.377**	0.313	0.227	**0.350**
sn_1	0.372	0.299	0.212	0.343
sn_∞	0.243	0.254	0.273	0.248
sn_2	0.370	0.299	**0.288**	**0.350**
sn_F	0.367	0.254	0.242	0.337

Table 3: Obtained accuracy values for similarity measures induced by different matrix norms and for four baseline methods. GM = Geometric Mean of sn_1 and sn_∞. (W)W2VC=(tf-idf-weighted) Word2Vec Embedding Centroids.

Method	Contest		
	1	2	3
Min kap.	0.123	0.295/0.030	0.110/0.101
Max. kap.	0.178	0.345/0.149	0.114/0.209
sn_2	0.128	0.049/0.065	0.060/0.064
sn_1	0.124	-0.032/0.033	0.024/0.017
sn_F	0.129	0.041/0.042	0.039/0.045
# Entr.	1544	100	100

Table 4: Minimum and maximum average inter-annotator agreements (Cohen's kappa) / average inter-annotator agreement values for our automated matching method, FN=Frobenius norm.

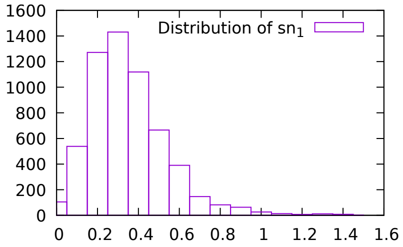


Figure 1: Distribution of sn_1 determined on contest 1.

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 Evaluation

For evaluation, we selected three online contests (language: German), where people elaborated on their favorite travel destination 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. We experimented with different keyword list sizes but obtained the best results with rather few and therefore precise keywords. In particular, we used the following number of keywords for the individual milieus:

- Action Sportsman: 3
- Young Performer: 4
- Hedonist: 7
- Conservative Youth: 4
- Progressive Postmodern Youth: 6

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 4. 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)[1]. Moreover, we computed its average inter-annotator agreement with the human annotators (see again Table 4), quasi treating the predictions like additional annotations.

The Word2Vec word embeddings were trained on the German Wikipedia (dump originating from 20 February 2017) merged with a Frankfurter Rundschau newspaper 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 2 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 is only guaranteed for encyclopedias. However, Stein and Anerka (Gottron et al., 2011) 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 Wikiprep-ESA on Wikipedia only but additionally have developed a version of ESA that can be applied on arbitrary corpora (in the following referred to as ESA2) and which was trained on the full corpus (Wikipedia+Frankfurter Rundschau+20 minutes). The STVs were also trained on the same corpus as our matrix norms based estimates and Word2Vec embedding centroids. The actual document similarity estimation is accomplished by the usual centroid approach (we did not evaluate matrix norms here). An issue we were faced with 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 use a beam search approach instead, which extends the keyword list iteratively and keeps only the n-best performing permutations.

Method	Accuracy
ESA	0.672
STV	0.716
W2VC	0.726
sn_2	0.731
$sn_{L_{1,1}}$	0.741
sn_F	0.781

Table 5: Accuracy value otbained for matching a sentence of the first to the associated sentence of the second translation.

Finally, to verify the general applicability of our approach, we conducted a second experiment, where a novel from Edgar Allen Poe (The purloined letter) was independently translated by two translators into German. 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. The obtained accuracy values based on the first 200 sentences of both translations are given in Table 5. To guarantee an 1:1 sentence mapping, periods were partly replaced by semicolons.

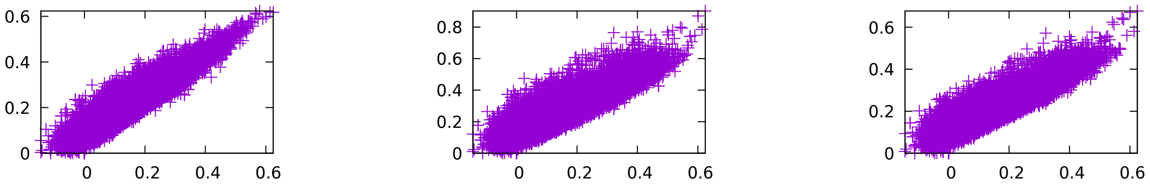
7 Discussion

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 4). On this contest part, our matrix norm-based matching (2-norm and Frobenius-norm) obtains a considerably higher average agreement than one of the annotators. Regarding baseline systems, the most relevant comparison is naturally the one with Word2Vec cen-

[1] Note that the geometric mean of the 1- and ∞-norm as specified in Table 3 is not a matrix norm itself, since it lacks submultipicativity.

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

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



(a) W2VC vs $sn_{L_{11}}$

(b) W2VC vs sn_2

(c) W2VC vs sn_F

Figure 2: Scatter plots of cosine between centroids of Word2Vec embeddings (W2VC) vs sn.

troids, since it employs the same type of data. Hereby we reached higher accuracy values for the best performing matrix norms on two of the three contests including the largest contest 1. Note that the elimination of negative values from the embedding matrix product proved to be important. If we omit this step, the obtained accuracy of sn_f for instance will drop by around 0.023 determined over all three contests (column: *all*).

It is quite striking that, although sn_1 lacks two properties of a normalized similarity measure (boundedness by 1 and symmetry), it reaches quite good results on contest 1. As you can see in Figure 1, which shows the distribution of sn_1 in contest 1, the value of 1 is indeed exceeded several times (the maximum value is 1.5), but this occurs rather rarely in our experiment. Actually, 99% of its values fall into the interval [0,1]. Thus, the non-boundedness is much less a problem in practice than the theoretical results indicate.

Finally, we determined the scatter plots (see Figure 2) showing cosine of Word2Vec embeddings (W2VC) vs several matrix norm based similarity estimates. These scatter plots exhibits that the score distributions of sn_f and sn_2 are quite similar and their values often exceed the cosine measure value due to the fact that a few very strong word matches can already result in a high similarity estimate. The scatter plot for sn_{L11} reveals that this measure is much closer to W2VC than the other two matrix norm based similarity estimates.

Note that a downside of our approach in relation to the usual Word2Vec centroids method is the increased runtime, since it requires the pairwise 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. One possible solution for this performance issue is to apply our proposed estimates to sentence embeddings instead of word embeddings, which on the one hand would reduce the dimensionality of the embedding matrices and on the other hand would take word order into account.

8 Conclusion

We proposed a novel similarity measure to compare word embeddings from different documents, which makes use of matrix norms. This measure was evaluated on the task to assign users to the best matching marketing target groups. We obtained superior results compared to the usual centroid / cosine measure similarity estimation for most of the investigated matrix norm especially for the largest contest 1. Furthermore, we proved elementary properties for our proposed similarity measure regarding its well-definedness and its performance in comparison to the usual centroid-based approach.

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

Anja Attig and Petra Perner. 2011. The problem of normalization and a normalized similarity measure by online data. *Transactions on Case-Based Reasoning*, 4(1).

Lluís A. Belanche and Jorge Orozco. 2011. 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.

Georgios-Ioannis Brokos, Prodromos, and Ion Androutsopoulos. 2016. 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*, pages 114–118, Berlin, Germany.

Françoise Chatelin. 1993. *Eigenvalues of Matrices - Revised Edition*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

Evgeniy Gabrilovic and Shaul Markovitch. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34.

Thomas Gottron, Maik Anderka, and Benno Stein. 2011. Insights into explicit semantic analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1961–1964, Glasgow, UK.

Ki-Joo Hong, Gai-Hui Lee, and Han-Joon Kom. 2015. Enhanced document clustering using wikipedia-based document representation. In *Proceedings of the 2015 International Conference on Applied System Innovation (ICASI)*.

Ryand Kiros, Yukun Zhu, Rusland Salakhudinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fiedler. 2015. Skip-thought vectors. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, Montréal, Canada.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilias Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 957–966.

Michael Lynn. 2011. Segmenting and targeting your market: Strategies and limitations. Technical report, Cornell University. Online: http://scholorship.sha.cornell.edu/articles/243.

Victor Mijangos, Gerardo Sierra, and Azuncena Montes. 2017. Sentence level matrix representation for document spectral clustering. *Pattern Recognition Letters*, 85.

Tomas Mikolov, Ilya Sutskever, Chen Ilya, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119, Lake Tahoe, Nevada.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Katar.

Yangqiu Song and Dan Roth. 2015. 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.

A Proof of Proposition 1

In this section we give the proof of proposition 1 for Frobenius and $L_{1,1}$-norm.

Reflexivity is trivially fulfilled for Frobenius, 2- and $L_{1,1}$-norm. Therefore, we only prove that the induced similarity measures are symmetric and upper-bounded by 1.

Let $\mathbf{A} := E(t)$, $\mathbf{B} := E(u)$, where t and u are arbitrary documents.

A.1 Frobenius norm

Symmetry The trace of a matrix is known to be invariant under cyclic permutations. With this property, the symmetry of sn_F can easily be deduced.

Boundedness

Proof. We need to show that:

$$\frac{\|\mathbf{A}^\top\mathbf{B}\|_F}{\sqrt{\|\mathbf{A}^\top\mathbf{A}\|_F \cdot \|\mathbf{B}^\top\mathbf{B}\|_F}} \leq 1 \tag{14}$$

For that we leverage the Cauchy-Schwarz inequality that states that for an inner product space $\mathcal{H}$ we have

$$|\langle x, y\rangle| \leq \sqrt{\langle x, x\rangle \cdot \langle y, y\rangle} \tag{15}$$

for all $x, y \in \mathcal{H}$, where equality holds if, and only if, x is a scalar multiple of y. In particular, the function $f(\mathbf{Y}, \mathbf{X}) = \mathrm{tr}(\mathbf{Y}^\top\mathbf{X})$ is such an inner product.

Let $\mathbf{X} = \mathbf{A}\mathbf{A}^\top$ and $\mathbf{Y} = \mathbf{B}\mathbf{B}^\top$ such that $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times m}$. Then we can infer

$$\begin{aligned}
\mathrm{tr}(\mathbf{B}\mathbf{B}^\top\mathbf{A}\mathbf{A}^\top) &= \mathrm{tr}((\mathbf{B}\mathbf{B}^\top)^\top\mathbf{A}\mathbf{A}^\top)\\
&= \mathrm{tr}(\mathbf{Y}^\top\mathbf{X})\\
&\overset{(15)}{\leq} \sqrt{\mathrm{tr}(\mathbf{X}^\top\mathbf{X}) \cdot \mathrm{tr}(\mathbf{Y}^\top\mathbf{X})}\\
&= \sqrt{\mathrm{tr}((\mathbf{A}\mathbf{A}^\top)^\top\mathbf{A}\mathbf{A}^\top) \cdot \mathrm{tr}((\mathbf{B}\mathbf{B}^\top)^\top\mathbf{B}\mathbf{B}^\top)}\\
&= \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^\top\mathbf{A}\mathbf{A}^\top) \cdot \mathrm{tr}(\mathbf{B}\mathbf{B}^\top\mathbf{B}\mathbf{B}^\top)}
\end{aligned} \tag{16}$$

Next, we observe that for the numerator

$$\begin{aligned}
\|\mathbf{A}^\top\mathbf{B}\|_F &= \sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{B}(\mathbf{A}^\top\mathbf{B})^\top)}\\
&= \sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{B}\mathbf{B}^\top\mathbf{A})}\\
&= \sqrt{\mathrm{tr}(\mathbf{B}\mathbf{B}^\top\mathbf{A}\mathbf{A}^\top)}\\
&\overset{(16)}{\leq} \sqrt[4]{\mathrm{tr}(\mathbf{A}\mathbf{A}^\top\mathbf{A}\mathbf{A}^\top) \cdot \mathrm{tr}(\mathbf{B}\mathbf{B}^\top\mathbf{B}\mathbf{B}^\top)}\\
&= \sqrt[4]{\mathrm{tr}(\mathbf{A}^\top\mathbf{A}^\top\mathbf{A}\top\mathbf{A}) \cdot \mathrm{tr}(\mathbf{B}\top\mathbf{B}\mathbf{B}^\top\mathbf{B})}\\
&= \sqrt{\|\mathbf{A}^\top\mathbf{A}\|_F \cdot \|\mathbf{B}^\top\mathbf{B}\|_F}
\end{aligned} \tag{17}$$

Hence, we finally obtain

$$\begin{aligned} \frac{\|\mathbf{A}^\top\mathbf{B}\|_F}{\sqrt{\|\mathbf{A}^\top\mathbf{A}\|_F \cdot \|\mathbf{B}^\top\mathbf{B}\|_F}} &\overset{(17)}{\leq} \frac{\sqrt{\|\mathbf{A}^\top\mathbf{A}\|_F \cdot \|\mathbf{B}^\top\mathbf{B}\|_F}}{\sqrt{\|\mathbf{A}^\top\mathbf{A}\|_F \cdot \|\mathbf{B}^\top\mathbf{B}\|_F}} \\ &= 1. \end{aligned}$$

□

A.2 $L_{1,1}$-norm

Proof for symmetry. For the $L_{1,1}$-norm, we have that

$$\begin{aligned} \|A\|_{L_{1,1}} &= \sum_{i=1}^{m}\sum_{j=1}^{n}|A_{ij}| = \sum_{j=1}^{n}\sum_{i=1}^{m}|A^\top_{ji}| \\ &= \|A^\top\|_{L_{1,1}} \end{aligned} \tag{18}$$

□

The boundedness follows directly from the Cauchy-Schwarz inequality, since the induced similarity estimate $sn_{L_{1,1}}$ is an inner product.