Retrieving Compositional Documents Using Position-Sensitive Word Mover's Distance

Martin Trapp[1,2], Marcin Skowron[1,3], Dietmar Schabus[1]
[1]Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
[2]Signal Processing and Speech Communication Lab., Graz University of Technology, Graz, Austria
[3]Dept. of Computational Perception, Johannes Kepler University Linz, Linz, Austria
firstname.lastname@ofai.at

ABSTRACT

Retrieving similar compositional documents which consist of ranked sub-documents, such as threads of healthcare web fora containing community voted comments, has become increasingly important. However, approaches for this task have not exploited the semantic relationships between words so far and therefore do not use the effective generalization property present in semantic word embeddings. In this work, we propose an extension of the Word Mover's Distance for compositional documents consisting of ranked sub-documents. In particular, we derive a Position-sensitive Word Mover's Distance, which allows to retrieve compositional documents based on the semantic properties of their sub-documents. Additionally, we introduce a novel benchmark dataset for this task, to facilitate other researchers to work on this relevant problem. The results obtained on the novel dataset and on the well-known MovieLense dataset indicate that our approach is well suited for retrieving compositional documents. We conclude that incorporating semantic relations between words and sensitivity to the position and presentation bias is crucial for effective retrieval of such documents.

1 INTRODUCTION

Recent work in machine learning and Natural Language Processing (NLP) has collectively developed effective methods for semantic analysis of words [13, 15] and textual documents [3, 8], e.g. books. In particular, leveraging the semantic relationships between words, using word embeddings [13, 15], has led to the Word Mover's Distance (WMD) [7, 9] which has shown to be an effective approach for semantic-aware document similarity. Using the WMD, Kusner et al. [9] were able to show impressive results on various document retrieval tasks. On the other hand, comparing compositional documents which consist of ranked sub-documents, e.g. rank based on the number of community votes, has gained increasing imporance in the field of information retrieval, e.g. [4, 19]. However, despite the recent advances in semantic analysis of documents, to the best of our knowledge, there has not been a transition of such methods to the task of comparing compositional documents as of yet. Moreover, even though comparing compositional documents is a relevant problem with multiple applications, only a very limited amount of benchmark data is available.

In this paper, we tackle the problem of *finding similar compositional documents based on the similarities of their ranked sub-documents*. We consider ranked sub-documents to be textual documents, e.g. comments in a forum thread, with the order of sub-documents given by a rank which reflects the amount of community votes. Specifically, we show how to leverage recent advances in machine learning and NLP to formulate a Position-sensitive Word Mover's Distance (P-WMD) which allows to compare compositional documents based on the semantic properties of their ranked sub-documents. In addition, we introduce the twin films dataset[1], a new openly accessible benchmark dataset for this task. The twin films dataset contains community-voted short descriptions in the form of plot keywords for each film, and it is well suited as a benchmark dataset for further research in this direction.

2 BACKGROUND

In the following we briefly review relevant background material and introduce the mathematical notation used in this paper.

2.1 Word Embeddings

Word embeddings aim to represent semantic relationships of words in vector spaces which consist of fewer dimensions than the dictionary size. Recent advances in this field [13, 15] allow for efficient computation and have also gained increasing importance in the field of information retrieval, e.g. [5]. Specifically, Mikolov et al. [13] proposed an efficient architecture in which each word vector is trained by maximising the conditional log probability of neighbouring words given the current word.

2.2 Word Mover's Distance

Based on the work on semantic word embeddings, Kusner et al. [9] recently proposed the Word Mover's Distance (WMD) as an effective approach for document similarity computation. By leveraging the semantic relationships of words, captured in word embeddings, the WMD measures similarity between documents on a semantic level. At a high level explanation, the WMD computes the minimal cost required to "transport" words from one document to another, where the cost is influenced by the distance of the words in the semantic space. Therefore, the WMD can be seen as a special case of the Earth Mover's Distance (EMD) [16], which is also known as the Wasserstein distance [12], for document similarity tasks. More formally, let us assume a D-dimensional word embedding $\boldsymbol{X} \in \mathcal{R}^{D \times N}$

[1]The twin films data set and the code are available on https://github.com/trappmartin/PWMD_ICTIR2017.

ICTIR '17, October 01–04, 2017, Amsterdam, Netherlands.
 978-1-4503-4490-6/17/10.
DOI: https://doi.org/10.1145/3121050.3121084

for a set of N words and let $\boldsymbol{x}_i \in \mathcal{R}^D$ be the embedding vector of the ith word in the vocabulary. Let $c_i \geq 0$ define the frequency count of the ith word and let $z = \sum_{i=1}^{N} c_i$ be a normalisation constant. We define $\boldsymbol{d} = \{d_i\}_{i=1}^{N}$ and $\boldsymbol{d'} = \{d'_i\}_{i=1}^{N}$ to be the normalised Bag Of Words (BOW) representations of two documents. Each column $d_i = \frac{1}{z} c_i$ represents the relative frequency count of a word in the respective document. As for the EMD, the WMD uses a transportation matrix $\boldsymbol{T} \in \mathbb{R}_{\geq 0}^{N \times N}$, where T_{ij} describes how much of word i in document $\boldsymbol{d}$ travels to word j in $\boldsymbol{d'}$. Formally, the WMD solves the following linear program:

$$\begin{aligned} &\underset{\boldsymbol{T} \in \mathbb{R}_{\geq 0}^{N \times N}}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{N} T_{ij} ||x_i - x_j||_2 \\ &\text{subject to} \sum_{j=1}^{N} T_{ij} = d_i, \ \sum_{i=1}^{N} T_{ij} = d'_j \ \forall i, j \end{aligned} \tag{1}$$

Informally, the WMD assigns a smaller distance to documents that share many semantically similar words than to documents with many semantically different words. Intuitively, the WMD is therefore well suited for various retrieval tasks in the natural language domain. In the following, we will show how to utilize the WMD when the aim is to compute the distance between compositional documents consisting of ranked sub-documents.

3 POSITION-SENSITIVE WORD MOVER'S DISTANCE

As discussed in prior work [4], compositional documents with ranked sub-documents are susceptible to a position bias [17]. In particular, displaying sub-documents in an order affects the perception, resulting in top-ranked items being more popular than low-ranked items. In addition, presenting summary information of the ranked sub-documents to the user can lead to a presentation bias [21]. Recent work by Lee et al. [11] proposed to include the position and presentation bias into the modelling process, using a Bayesian nonparametric model and allowing the model to be sensitive to the bias. We refer to [14, 20] for more details on Bayesian nonparametric models. Lee at al. [11] also showed that the sensitivity to the position and presentation bias depends on the community. In particular, the community *stackoverflow* turned out to have a higher sensitivity than the related community *mathoverflow*. Therefore, it is crucial for effective retrieval of such compositional documents to integrate the bias, but also to control the sensitivity. In the case of compositional documents, the linear program described in Equation 1 can be extended as follows:

$$\begin{aligned} &\underset{\boldsymbol{T} \in \mathbb{R}_{\geq 0}^{N \times N}}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{N} T_{ij} ||x_i - x_j||_2 \\ &\text{subject to} \sum_{j=1}^{N} T_{ij} = \frac{1}{z} \sum_{p=1}^{P} c_{pi} \ \forall i \\ &\sum_{i=1}^{N} T_{ij} = \frac{1}{z'} \sum_{p=1}^{P} c'_{pj} \ \forall j \end{aligned} \tag{2}$$

where c_{pi} and c'_{pi} are frequency counts of the ith and jth word in the pth sub-document of a document. Note that we are using P to indicate the number of ranked sub-documents for both documents. Without loss of correctness, the number of ranked sub-documents can vary for the two documents. By integrating the bias directly into the normalised BOW representation, we can allow the WMD to be sensitive to the position and presentation bias. Formally, we define r_p to be the rank of the pth sub-document. Further, we borrow the bias term by Lee at al. [11] and define the position and presentation bias as:

$$b_p = \left(\frac{1}{1 + r_p} \right)^{\gamma} \tag{3}$$

where $\gamma \geq 0$ is a sensitivity parameter which allows us to control the effect of the position and presentation bias. We refer to Lee at al. [11] for more details on the choice of the bias term. We can now define the normalised rank-weighted BOW representation $\hat{\boldsymbol{d}} = \{\hat{d}_i\}_{i=1}^{N}$ as follows:

$$\hat{d}_i = \frac{1}{z} \sum_{p=1}^{P} b_p c_{pi} \tag{4}$$

The Position-sensitive Word Mover's Distance (P-WMD) is therefore defined by solving the following linear program:

$$\underset{\boldsymbol{T} \in \mathbb{R}_{\geq 0}^{N \times N}}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{N} T_{ij} ||x_i - x_j||_2 \tag{5}$$

$$\text{subject to} \sum_{j=1}^{N} T_{ij} = \hat{d}_i, \ \sum_{i=1}^{N} T_{ij} = \hat{d}'_j \ \forall i, j \tag{6}$$

Note that this formulation allows us to use the word centroid distance [9], providing efficient computation of the transportation problem. As described in previous work [1, 10], word embeddings tend to suffer from the hubness problem which is critical for retrieval tasks. We propose to reduce the hubness of the P-WMD using mutual proximity [18], which transforms the P-WMD to a statistical distance.

Table 1: Examples from twin films dataset.

First Film	Second Film
Oscar Wilde (1960)	The Trials of Oscar Wilde (1960)
Prefontaine (1997)	Without Limits (1998)
Kundun (1997)	Seven Years in Tibet (1997)
A Hijacking (2012)	Captain Phillips (2013)

4 TWIN FILMS DATASET

Twin films can be characterized as films which have the same or very similar plot but were produced by two different studios around the same time. Explanations for the phenomenon are diverse, ranging from industrial espionage to topical issues [2], e.g. the refugee crisis. We acquired a dataset based on an list of twin film examples listed on Wikipedia[2] to tackle the problem of finding the twin film. The dataset consists of 111 twin film pairs (first film, second film) and is composed of 221 unique films which have been carefully

[2] https://en.wikipedia.org/wiki/Twin_films, last accessed: 16.02.2016

Table 3: R-precision scores on twin films dataset. The best result is highlighted in bold.

Method	Random	BOW	TFIDF	WMD	P-WMD
R-precision	0.005	0.171	0.072	0.459	**0.523**

revised. The films present in the dataset range over several genres and sub-genres and cover a wide range of production dates (1938–2016). Table 1 lists a few samples from our twin films dataset, where we treat the first film to be considered as the query object. To summarize the synopsis of the films, we additionally extracted the plot keywords listed at IMDB[3] for each of the 221 films. The position of each plot keyword on IMDB depends on the number of up-votes given by the community members resulting in a position bias of the keywords. Additionally, the plot keywords listed on IMDB are biased by the presentation used on the film description page, as only the five highest ranked keywords are shown.

Table 2 lists the top five plot keywords for a few samples of our twin films dataset. Note that the IMDB plot keywords do not necessarily contain only a single token but rather describe properties of the plot using concatenated tokens, e.g. *long-distance-runner* in Prefontaine (1997). We have therefore pre-processed all plot keywords as follows: (1) each keyword has been split into its individual words using hyphen as the delimiter, (2) all occurrences of stop words have been removed and (3) all tokens have been transformed to lowercase. Note that it was not possible to retrieve plot keywords for all 221 films, resulting in a sub-set of 108 twin film pairs which contain plot keywords for both films.

5 EXPERIMENTAL RESULTS

To asses the effectiveness of our approach, we performed a set of evaluations on the introduced twin films dataset and on the MovieLense dataset [6]. We used a sensitivity of $\gamma = 0.75$ for all experiments, as this value roughly represents the sensitivity of the majority of online communities discussed in [11]. Note that an optimal sensitivity value can be found using grid search if a validation set is available.

5.1 Twin Films Benchmark

We compared the performance of our P-WMD against three other methods: Bag Of Words (BOW), Term Frequency–Inverse Document Frequency (TFIDF) and the original Word's Mover Distance (WMD). As for each query object (first twin film) exactly one film has to be retrieved, we used the R-precision metric. In particular we define the R-precision as the average precision with the number of relevant documents equal to one ($R = 1$). The R-precision of random guessing can therefore be computed by $\frac{R}{\#Films}$, where $\#Films$ is the number of films in the dataset. All R-precision scores on the twin films dataset, including random guessing, are shown in Table 3. As indicated in the results table, our P-WMD approach outperforms all other approaches by a clear margin. In addition to the evaluation using R-precision scores, we assessed the recall@k for the range of $k = \{1, \ldots, 10\}$. The resulting recall@k scores are shown in Figure 1. The recall@k scores indicate that our P-WMD achieves an improvement upon the WMD for all values of k considered in the evaluation. Recall@k of the random guessing has been estimated using 200 random trials. The stable recall@k results indicated that the P-WMD is a robust approach for retrieving compositional documents based on the semantic properties of their ranked sub-documents. These results and the R-precision results indicate that the P-WMD is well suited for such retrieval tasks.

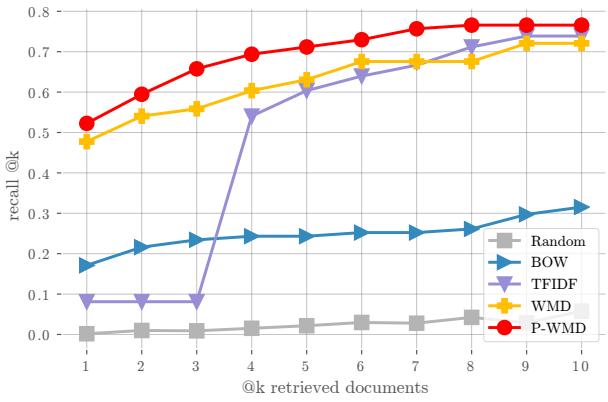

Figure 1: Comparison of recall at different values of k on the twin films dataset.

5.2 MovieLense Benchmark

In addition to the evaluation on the twin films dataset, we assessed the performance of our proposed P-WMD on the MovieLense dataset [6] by measuring the relative improvement over a baseline. The MovieLense dataset consists of 100k film ratings from 1000 users on 1700 films. The data was collected through the MovieLense web site and is an established dataset for research on recommender systems. In order to allow for an evaluation consistent to those on the twin films dataset, we considered the problem of retrieving the genre of a film based on a similarity space constructed from plot keywords. Note that the plot keywords do not necessarily reflect information about the genres, resulting in weak performance ($F1$ score) of all approaches.

We automatically retrieved IMDB IDs and IMDB plot keywords for all 1700 films. As the MovieLense dataset provides information on the film genre, we used the provided data as ground truth. We assessed the performance of this multi-label task using the macro-averaged F_1 score and used leave-one-out to estimate the generalization error. To obtain predictions for the film genres given the plot keywords of a film, we computed the majority vote using the k-Nearest Neighbour classifier. We used Bag Of Words (BOW) as the baseline approach and measured the relative improvements of the leave-one-out macro-averaged F_1 scores over the baseline score. The relative improvement of a method is computed using $\frac{F1_{method}}{F1_{baseline}} - 1$. Figure 2 shows the improvements obtained by all approaches when $k = 2$ neighbours or $k = 5$ neighbours are considered in the estimation of the genres for a movie. Considering the large margin to TFIDF and WMD with $k = 2$, our P-WMD seems to

[3] http://www.imdb.com, last accessed: 16.02.2016

Table 2: Top five plot keywords of examples from twin films dataset.

Film	Plot Keywords (Top 5)
Oscar Wilde (1960)	homosexual-history, grapes, playwright, grape, london-fog
The Trials of Oscar Wilde (1960)	gay-husband, gay-interest, homosexuality, homosexual, gay
Prefontaine (1997)	oregon, long-distance-runner, runner, olympics, watching-television
Without Limits (1998)	oregon, car-crash, death, university-of-oregon, coach
Kundun (1997)	tibet, chinese, dalai-lama, lama, tibetan
Seven Years in Tibet (1997)	dalai-lama, tibet, austria, mountain, himalaya
A Hijacking (2012)	somali-pirate, pirate, cargo-ship, ransom, ceo
Captain Phillips (2013)	ship, hostage, lifeboat, somalian-pirate, leader

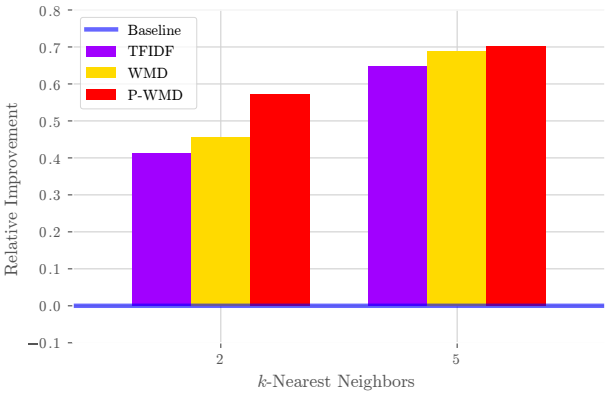

Figure 2: Relative improvements, estimated using leave-one-out, of the macro-averaged F_1-score over BOW using k-Nearest Neighbour classification.

be more robust for compositional documents resulting in less noise in the distance space and therefore resulting in a higher relative improvement than the other methods. As expected, with increased k the relative improvement over the baseline of the P-WMD is approximately the same as for the original Word Mover's Distance.

6 CONCLUSION

We presented an effective approach for retrieving compositional documents consisting of ranked sub-documents by incorporating the position and the presentation bias into the Word Mover's Distance. As datasets for such retrieval tasks are rare and difficult to obtain, we additionally introduced a new benchmark dataset on twin films. While the formulation of our Position-sensitive Word Mover's Distance allows for efficient computation, integrating the position and the presentation bias has shown to lead to an improvement over state-of-the-art approaches on both the twin films and the established MovieLense dataset. We could further identify a larger improvement if a small number of neighbours is used in the k-Nearest Neighbour classification task, indicating that our approach produces less noise and is suitable for retrieving compositional documents based on their ranked sub-documents. We further conclude that exploiting semantic properties of words and integrating the position and presentation bias is important to achieve convincing results. In further work, we will investigate the integration of the position and presentation bias into the supervised Word Mover's Distance.

ACKNOWLEDGMENTS

This research is partially funded by the Austrian Science Fund (FWF): P 27530.

REFERENCES

[1] M. Artetxe, G. Labaka, and E. Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of EMNLP*.
[2] Henrik Arvidsson. 2016. Först till kvarn i Drömfabriken. (16 02 2016). http://www.dn.se/kultur-noje/film-tv/forst-till-kvarn-i-dromfabriken
[3] J. A Aslam and M. Frost. 2003. An Information-theoretic Measure for Document Similarity. In *Proceedings of ACM SIGIR*. 449–450.
[4] J. HD Cho, P. Sondhi, C. Zhai, and B. R Schatz. 2014. Resolving healthcare forum posts via similar thread retrieval. In *Proceedings of ACM BCB*. 33–42.
[5] D. Ganguly, D. Roy, M. Mitra, and G. JF Jones. 2015. Word embedding based generalized language model for information retrieval. In *Proceedings of ACM SIGIR*. 795–798.
[6] F Maxwell Harper and Joseph A Konstan. 2016. The movielens datasets: History and context. *ACM TiiS* 5, 4 (2016), 19.
[7] G. Huang, C. Guo, M. J Kusner, Y. Sun, F. Sha, and K. Q Weinberger. 2016. Supervised Word Mover's Distance. In *Proceedings of NIPS*. 4862–4870.
[8] R. Kiros, Y. Zhu, R. R Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. 2015. Skip-Thought Vectors. In *Proceedings of NIPS*. 3294–3302.
[9] M. J Kusner, Y. Sun, N. I Kolkin, and K. Q Weinberger. 2015. From Word Embeddings To Document Distances.. In *Proceedings of ICML*, Vol. 15. 957–966.
[10] A. Lazaridou, G. Dinu, and M. Baroni. 2015. Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning. In *Proceedings of ACL*.
[11] M. Lee, S. H Jin, and D. Mimno. 2016. Beyond Exchangeability: The Chinese Voting Process. In *Proceedings of NIPS*. 4934–4942.
[12] E. Levina and P. Bickel. 2001. The earth mover's distance is the mallows distance: Some insights from statistics. In *Proceedings of ICCV*, Vol. 2. 251–256.
[13] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*. 3111–3119.
[14] P. Orbanz and Y. W Teh. 2011. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*. Springer, 81–89.
[15] J. Pennington, R. Socher, and C. D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP*. 1532–1543.
[16] Y. Rubner, C. Tomasi, and L. J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *IJCV* 40, 2 (2000), 99–121.
[17] M. J Salganik, P. S Dodds, and D. J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311, 5762 (2006), 854–856.
[18] D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer. 2012. Local and global scaling reduce hubs in space. *JMLR* 13 (2012), 2871–2902.
[19] A. Singh, P. Deepak, and D. Raghu. 2012. Retrieving Similar Discussion Forum Threads: A Structure Based Approach. In *Proceedings of ACM SIGIR*. 135–144.
[20] M. Trapp. 2015. BNP.jl: Bayesian nonparametrics in Julia. In *Bayesian Nonparametrics: The Next Generation Workshop at NIPS*.
[21] Y. Yue, R. Patel, and H. Roehrig. 2010. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of WWW*. 1011–1018.