HOTTER: Hierarchical Optimal Topic Transport with Explanatory Context Representations

Sabine Wehnert
Christian Scheel
Georg Eckert Institute
Leibniz Institute for International
Textbook Research, Germany
Otto von Guericke University
Magdeburg, Germany

Simona Szakács-Behling
Maret Nieländer
Patrick Mielke
Georg Eckert Institute
Leibniz Institute
for International
Textbook Research, Germany

Ernesto William De Luca
Georg Eckert Institute
Leibniz Institute
for International
Textbook Research, Germany
Otto von Guericke University
Magdeburg, Germany

Abstract

Natural language processing (NLP) is often the backbone of today's systems for user interactions, information retrieval and others. Many of such NLP applications rely on specialized learned representations (e.g. neural word embeddings, topic models) that improve the ability to reason about the relationships between documents of a corpus. Paired with the progress in learned representations, the similarity metrics used to compare representations of documents are also evolving, with numerous proposals differing in computation time or interpretability. In this paper we propose an extension to a specific emerging hybrid document distance metric which combines topic models and word embeddings: the Hierarchical Optimal Topic Transport (HOTT). In specific, we extend HOTT by using context-enhanced word representations. We provide a validation of our approach on public datasets, using the language model BERT for a document categorization task. Results indicate competitive performance of the extended HOTT metric. We furthermore apply the HOTT metric and its extension to support educational media research, with a retrieval task of matching topics in German curricula to educational textbooks passages, along with offering an auxiliary explanatory document representing the dominant topic of the retrieved document. In a user study, our explanation method is preferred over regular topic keywords.

1 Introduction

Topic models have been employed for more than a decade to capture latent semantics that help to organize documents in a dataset. Latent Dirichlet Allocation (LDA) by Blei et al. (2001) illustrates this approach. On the other hand, word embedding models have been proposed in recent times, for uses in many natural language processsing (NLP) tasks. Word embeddings are able to map phrases and documents to a dense, high-dimensional space, where (according to how the embedding model was trained) semantic similarity or analogy relationships between expressions are facilitated. In recent years, both approaches - topic models and word embeddings - are used in combination. An emerging metric for such a scenario is the Hierarchical Optimal Topic Transport (HOTT) by Yurochkin et al. (2019), which computes the distance between top *n* topic words using their word embedding representation, weighted by the document-topic distribution. As approaches for creating word embeddings evolve, newer contextual variants emerge. Hence, a reasonable **first research question** could be to consider the impact on HOTT; *Does HOTT benefit from contextual word embeddings?* In this paper we perform targeted experiments regarding this first research question: We select the BERT language model (Devlin et al., 2019) and try different variants of applying its contextual word embeddings to extend the HOTT metric.

A particular advantage of using HOTT over current solutions relying solely on the Word Mover's Distance (WMD) is a better interpretability of the document distances, since the metric relies on the known top *n* words of each topic for the respective document. Supporting a better interpretability is a demand for practical applications in many domains (Neitmann and Scheel, 2020). In this context, it is natural to ask how truly interpretable are those top *n* topic words to a domain expert? While the answer to this question depends highly on the quality of the topic model and parameter choices, we instead develop an alternative, simpler method for explaining the results by HOTT and then compare this method to the interpretability of the top *n* topic words. To this end, we employ the HOTT metric in a retrieval setting. Particularly, we use the dominant topic keywords of the retrieved document for extracting a representative auxiliary document from the corpus, containing the respective topic word and scoring highest on the particular topic

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4856–4866
November 7–11, 2021. ©2021 Association for Computational Linguistics

that the word belongs to. This in turn suggests for our study a **second research question**: *Does an auxiliary document that is close to the dominant topic keyword for retrieving a document offer a better explanation to a humanist than the top n topic keywords?* We carry out experiments for the second research question on educational media research data, where the aim is to match themes in a teaching curriculum to the parts of a textbook corpus covering those themes. To summarize, the overall aim of this work is to adapt the **H**ierarchical **O**ptimal **T**opic **T**ransport to **E**xplanatory Context **R**epresentations (HOTTER). First, we investigate whether using contextual word embeddings offers a benefit for HOTT performance, and second, we examine the interpretability of HOTTER results. Our core contributions can be stated as follows:

- We extend the HOTT method by contextual word embeddings from the BERT model.
- We gain insights about the *interpretability* of the top *n* topic words, compared to selecting one of these topic words and offering an auxiliary document from the corpus for the keyword, which is chosen to both represent and explain why the retrieved document is close to the topic keyword in the vector space.

The remainder of this work is structured as follows: Section 2 collects related work about combining word embeddings with topic models, Section 3 contains foundations required to understand our contribution, covering document distances in the word embedding space, Hierarchical Optimal Topic Transport and contextual word embeddings. Section 4 describes our proposed HOTTER approach in detail. Section 5 includes our experimental results and the corresponding discussion. Section 6 concludes our findings.

2 Related Work

In this section, we briefly describe the most important related work about combining the LDA topic model with word embeddings. Topic Models such as LDA are popular for clustering a document collection. They learn a topic distribution for each document in a corpus and infer a word distribution for each topic. Viewing the top *n* words of a topic can lead to insights about the themes the topic captures. However, this is not the only application of a topic model. The probabilistic nature makes it easy to interpret since the topic distribution of a document and the word distribution of each topic, respectively, sum up to one. This property makes them suited for featurizing and tagging documents for both, end user applications and further processing. A drawback in standard LDA implementations is connected to the text representation they use. Often a simple bag-of-words approach is chosen, leading to positional information being lost and the resulting topics carrying a notion of "relatedness" between words instead of semantic similarity (Bunk and Krestel, 2018). While there are n-gram topic model implementations such as the work by Tam and Schultz (2008), this is not employed frequently due to model sparsity. Therefore, the traditional unigram LDA approach is still applied and combined with other approaches, such that they can complement each other.

Word embeddings by Mikolov et al. (2013) are nowadays a common choice for experimentation with other text representations due to their disruptive performance on many Natural Language Processing tasks, and their capability of capturing term analogies and semantic similarity. In this approach word vectors are trained by maximizing the average log probability of the next word, which is different from topic model probabilities that are normalized to one. For this reason, standard word embedding values cannot be interpreted as probabilities, but we can assume that words with similar vectors have a similar meaning. Another difference is that word embeddings are usually pre-trained on substantially larger corpora than topic models and then optionally fine-tuned on domain-specific text. Taking those aspects into account, word embeddings and topic models have the potential to enrich each other. The inclined reader may refer to a more in-depth discussion on differences between topic models and word embeddings by Bunk and Krestel (2018) or Li et al. (2016b) which we omit due to space restrictions. There are numerous works which combine topic models and word embeddings, resulting in two main groups of approaches (Bunk and Krestel, 2018): those using a topic model architecture with features from word embeddings, as opposed to those using the neural network architecture to obtain word embeddings and topic representation during training, such as LDA2Vec (Moody, 2016) or TWE (Liu et al., 2015). The basis for our work - the HOTT meta distance - belongs to the former group, hence we focus on related research in this regard. The Vec2Topic approach extracts word embeddings and combines them with a topic model that has been trained with K-Means Clustering, while using agglomerative clustering on the word vectors to score the topic based on the keyword similarity and importance (depth and degree) (Randhawa et al., 2016). Bunk and Krestel (2018) use word embeddings to improve topic models with Gibbs sampling to exchange top n topic words from the topic model with more salient terms. The GPU-DMM method is specifically developed for short text, since short texts rarely contain co-occurrences of semantically similar words (Li et al., 2016a). Another approach incorporating topic correlation has been proposed by Xun et al. (2017). Overall, the numerous works in this field suggest a complementary nature of word embeddings capturing local semantic similarity, compared to topic models which generate a notion of semantic relatedness. To the best of our knowledge, there is no work on enhancing topic models with contextual word embeddings yet, which is one key contribution.

3 Background

We proceed with the fundamental concepts required for understanding our approach. First, we cover the Word Mover's Distance (WMD), second, we explain the HOTT meta distance and afterwards, we describe contextualized word embeddings.

3.1 Word Mover's Distance

The WMD allows us to quantify the minimal transportation cost between multidimensional word embedding vectors by minimizing the Euclidean distance, which we denote as $||\cdot||_2$. The basis for computing the distance is the sparse nBOW document representation which captures counts for all unique words in the vocabulary. Then, the travelling costs c between two words i and j are defined as $c(i,j) = ||x_i - x_j||$, where x refers to the embedding vector of the respective word (Kusner et al., 2015). The optimization problem for the transportation costs for two documents is depicted below, with the vocabulary size v and T the transportation flow matrix (Kusner et al., 2015)

$$\min_{T \geq 0} \sum_{i,j=1}^{v} T_{ij} c(i,j). \tag{1}$$

3.2 Hierarchical Optimal Topic Transport

The HOTT metric combines word embeddings and LDA. While topic models characterize a document according to its topic distribution, the distance between documents using word embeddings is computed by the pairwise transportation costs between all individual words in a document. Since the WMD is an accurate, but expensive operation, Yurochkin et al. (2019) define HOTT for a set of topics $T = \{t_1, t_2, \ldots, t_{|T|}\} \in \Delta^{|V|}$ distributed over our vocabulary V and document-topic distributions $\bar{d^i} \in \Delta^{|T|}$:

$$HOTT(d^1, d^2) = W_1\left(\sum_{k=1}^{|T|} \bar{d}_k^1 \delta_{t_k}, \sum_{k=1}^{|T|} \bar{d}_k^2 \delta_{t_k}\right), \tag{2}$$

where the 1-Wasserstein distance is denoted as W_1. Each topic t_k supports a probability distribution Dirac delta δ_{t_k}. For practical use and to increase the stability of the approach, the topic-word distribution can be truncated to the top n topic words, i.e., the words which have the highest probability for a given topic, without significant performance losses (Yurochkin et al., 2019). The main considerations for using topic models along with word embeddings are computation time and interpretability, according to Yurochkin et al. (2019). Using LDA leads to more interpretable distances between documents because we obtain a notion of the top n topic words whose distances are computed in the embedded space. In contrast, measuring the distance between all word embeddings of a document is computationally expensive. Thus, this process generally benefits from a weighting mechanism of words, such as Term Frequency Inverse Document Frequency (TF-IDF) or selecting the top n words of a topic in the manner of HOTT. As an addition in this work, the distances are refined with contextual word embeddings which we present henceforth.

3.3 Contextualized Word Embeddings

Recently, word embeddings come from contextual models, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). These approaches convert a sequence of text into word embeddings at run-time, such that each word's vector depends on the other vectors. In other words, there is no static representation of the same word in different sentences. Furthermore, instead of entire words, BERT and ELMo process subwords, which makes them more robust against the out-of-vocabulary problem. For instance, a rare word such as *"unceremoniously"*, may not be part of the pre-trained vocabulary of the language model BERT, however, it can generate a representation for the prefix *"un"* and the other subwords *"##cer"*, *"##emon"*,

"##iously". For those reasons the performance of contextual word embeddings is at this time state-of-the-art for many applications. Considering the previously introduced distance metrics HOTT and WMD, contextual word embeddings may need a different treatment than traditional word embeddings because they are not intended to model the words without a context. However, there is emerging research, for example the BERTScore, which computes pairwise cosine distances between the words of two sentences (Zhang et al., 2020). The cosine distance is a common choice for contextual word embeddings (Bao et al., 2020). Ethayarajh (2019) proposed metrics for measuring the contextuality of those embeddings, among them are the so-called Self-Similarity (SelfSim) - measuring the cosine similarity between all contextual representations for the same word - and Maximum Explainable Variance (MEV), quantifying to which extent the first principal component can explain the variance in a contextual word embedding vector. Since there is no consensus in the research community yet, we test different methods for the aggregation of contextual word embeddings to one representation for each word in the vocabulary.

4 HOTTER

In this section, we present the HOTTER approach by first incorporating contextual word embeddings into the HOTT document distance metric. Furthermore, we explore retrieving representative sentences based on the obtained contextual embeddings for the top n topic words.

4.1 Incorporating Contextual Embeddings into Hierarchical Optimal Topic Transport

Figure 1 provides an overview of our proposed process. In the first step, all documents are preprocessed. For this, the documents are tokenized and (optionally) stemmed or filtered by part-of-speech tags (POS) for the topic modeling part. The choice of preprocessing techniques depends on language and data characteristics. Meanwhile, the BERT language model operates with its own tokenizer which is creating subwords. Since BERT subwords would not be readable in the top n words of each topic, we refrain from using the BERT tokenizer for the pipeline leading to the LDA topic model input, but rather consolidate after obtaining the contextual word embeddings and the topic model individually. Also, keeping the tokenizer for LDA on a word basis makes HOTTER more comparable to HOTT. After preprocessing, the second step begins where the LDA model is generated as in the original HOTT implementation by using Gibbs sampling. The pre-trained BERT model processes a sequence of 512 tokens at a time and we extract the embeddings from the last layer. As a result, we have for each token a 768-dimensional context vector. In step three, we first have to create a mapping between the subwords of the BERT model and the vocabulary used by the topic model, so that we can use the contextual embeddings for the top n topic words. To achieve a guaranteed mapping for each word, we apply exact matching. If that fails, we resolve the existing subwords (indicated by ## in the BERT model). As a last option we employ the longest matches of left-bound substring comparison. For the fourth step of finding a common ground between the topic model and the contextual embeddings, we recall that there is no scientific consensus regarding the use of the context vector on a word basis for distance computations. Since it is not thoroughly studied how to apply the context-dependent word vectors in the same fashion as regular word embeddings, we test multiple approaches. The first option (S-HOTTER) is the naïve method of taking the contextual embedding vector for each word in the vocabulary and then averaging the vectors for the same word, regardless of context. Consider w as a word that appears in documents $\{d_1, d_2, \ldots, d_p\}$ of a corpus at indices $\{i_1^{d_j}, i_2^{d_j}, \ldots, i_{m_j}^{d_j}\}$ in each document d_j, so that $w = d_1[i_1^{d_1}] = \ldots = d_p[i_{m_p}^{d_p}]$. Then $e_\ell(d,i)$ is the contextual embedding vector obtained from the language model's layer ℓ for the token at index i in document d. The S-HOTTER aggregation of layer ℓ for the word w is

$$\mathcal{S}_\ell(w) = \frac{1}{\sum_{j=1}^{p} m_j} \left(\sum_{j=1}^{p} \sum_{k=1}^{m_j} e_\ell(d_j, i_k^{d_j}) \right) \quad (3)$$

The second option (A-HOTTER) averages all vectors from a document containing a given word in the vocabulary, and in turn computes the mean of all these average embeddings for all occurrences of that word within the corpus.

$$\mathcal{A}_\ell(w) = \frac{1}{p} \left(\sum_{j=1}^{p} \frac{1}{|d_j|} \sum_{k=1}^{|d_j|} e_\ell(d_j, k) \right) \quad (4)$$

These two methods may have the drawback of neglecting homonyms, thus we introduce two

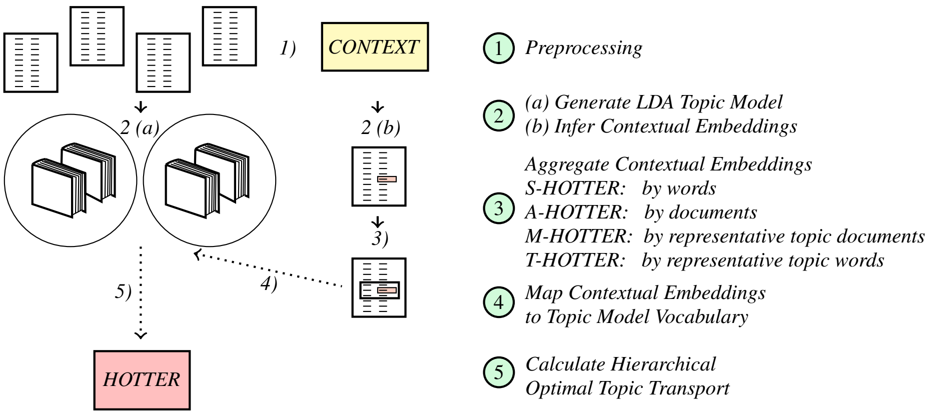
1)
CONTEXT
2 (a)
2 (b)
3)
4)
5)
HOTTER
1 Preprocessing
2 (a) Generate LDA Topic Model
(b) Infer Contextual Embeddings
3 Aggregate Contextual Embeddings
S-HOTTER: by words
A-HOTTER: by documents
M-HOTTER: by representative topic documents
T-HOTTER: by representative topic words
4 Map Contextual Embeddings to Topic Model Vocabulary
5 Calculate Hierarchical Optimal Topic Transport

Figure 1: Overview of the HOTTER workflow for measuring contextual document distance.

more methods (document-averaged M-HOTTER and token-based T-HOTTER) which only take into account the most probable documents containing the respective word according to our topic model. We compute this as shown in the pseudocode of Algorithm 1 by using the topic-word distribution to obtain top n topic words. We continue by tracking for each document in the corpus the most probable topics t. Then this information is used to find out for each topic the most representative documents until all topics obtain at least one probable document (see also lines 15-20 in Algorithm 1). Finally, for top n topic words, only the embedding vectors of the most probable documents are averaged.

Hence, if a word belongs to the top n topic words, the aggregation of the embedding vectors changes in the M-HOTTER variant to a subset of documents $\mathcal{R}$ which are representative of a topic:

$$\mathcal{M}_\ell(w) = \frac{1}{p^{\mathcal{R}}}\left(\sum_{j=1}^{p^{\mathcal{R}}}\frac{1}{|d_j^{\mathcal{R}}|}\sum_{k=1}^{|d_j^{\mathcal{R}}|} e_\ell(d_j^{\mathcal{R}},k)\right) \qquad (5)$$

For other words than the top n topic words, the M-HOTTER variant falls back to the equation 4 of A-HOTTER. Analogously, in case of a top n topic word, the T-HOTTER variant averages the contextual word embeddings appearing in the representative set of documents $\mathcal{R}$ (see equation 6), and reverts to the S-HOTTER equation 3, otherwise:

$$\mathcal{T}_\ell(w) = \frac{1}{\sum_{j=1}^{p^{\mathcal{R}}} m_j}\left(\sum_{j=1}^{p^{\mathcal{R}}}\sum_{k=1}^{m_j} e_\ell(d_j^{\mathcal{R}}, i_k^{d_j^{\mathcal{R}}})\right) \qquad (6)$$

Overall, all aforementioned methods offer one aggregated contextual embedding for each word in the vocabulary. In the fifth step, we calculate the

Algorithm 1 Topic-Representative Documents

```
 1: procedure GET_DOCUMENTS $\mathcal{R}$
 2:     $sort\_de$ ← sort in descending order
 3:     $Topics$ ← get $topic\_word$ distribution
 4:     $top$ ← select top words from a list
 5:     $n$ ← set top topic words
 6: Obtain top_n words:
 7:     for $t$, $topic\_word$ in $Topics$ do
 8:         $top\_n \leftarrow top(n, sort\_de(topic\_word[t]))$
 9: Select representative documents:
10:     $d\_t$ ← document-topic distribution
11:     $top\_d$ ← map for a topic's top documents
12:     $Documents$ ← document $d$ corpus
13:     $i$ ← index for the topic's probability in $d$
14:     $i \leftarrow 0$
15:     for $t$ in $Topics$ do
16:         while $t$ has an empty $top\_d$ do
17:             for $d$ in $Documents$ do
18:                 $top\_d \leftarrow sort\_de(d\_t[d])[i]$
19:             if $top\_d[t]$ empty then
20:                 $i \leftarrow +1$
21: Mapping of top_n words to top_d:
22:     $word\_d$← map word to $d$ in Documents
23:     $\mathcal{R}$ ← map word to representative $d$
24:     for $t$ in Topics do
25:         for $w$ in $top\_n[t]$ do
26:             for $d$ in $word\_d[w]$ do
27:                 if $d$ in $top\_d[t]$ then
28:                     $\mathcal{R}[w] \leftarrow d$
```

cost matrices as inputs for the HOTT computation. Since we brought the contextual embeddings on a word-level, there is no difference to the original HOTT process with regular word embeddings. Hence, adjusting preprocessing steps and adding aggregation and a mapping between the contextual word embeddings and the vocabulary used by the LDA mode enable us to employ the BERT model for the HOTTER meta distance.

4.2 Interpretability of Representative Documents

Blei and Lafferty (2009) already stated a decade ago that the bag-of-words top *n* topic words may not be enough for successfully interpreting them, and suggest to use a different representation for visualizing salient topic features. We propose to select representative documents for each of the top *n* topic words to improve topic interpretability. For a given topic we select the documents which have the highest probability assigned to this topic. From those documents we retrieve a set of closest documents for each top *n* topic word according to the aggregated embedding value of the respective word. Then, for all top *n* topic words, we have at least one representative document. From those documents, we select the one which is closest to the retrieved document that shall be explained. This auxiliary text should offer an insight into which nearest neighbor represents a keyword from the top *n* topic words that are also prominent in the retrieved document. That way, the user may develop an understanding of the context in which the keyword is prevalent within the given topic, as opposed to the list of top *n* topic words without further context information. In the following, we present a use case for this approach in educational media research.

5 Evaluation

5.1 Educational Media Research Data

In education, curricula are one or more documents describing the wanted knowledge or skill set for a students of a specific school subject, level of education and geographical region. In curricula, topics are referred to as learning units. A learning unit contains a description of wanted knowledge or skills, for instance for the “French Revolution” or “World War I”. Often these descriptions are separated in levels of difficulty, so that there is a description of the lowest expectation and a description containing the maximal knowledge or skill set. If the textbooks’ topics match these learning units, they will be approved. Since the curricula are localized, there are often customized versions of a textbook for each region. In educational media research, the learning units are of particular interest, because this socially, pedagogically and scientifically sanctioned knowledge forms the young generations. Students often depend on the school as the only source of knowledge and are thus especially vulnerable, if the knowledge imparted is altered or omitted. Additionally, popular knowledge in historical textbooks helps to understand worldviews and thought flows of specific periods and regions. The ability to match textbook content to learning units is important for educational media research. Textbooks often contain additional content to the required learning units for the following reasons:
Regional: The book may be used in more regions.
Temporal: When the textbook is older (but still valid) and this knowledge was either required in earlier curricula or for future curricula.
Propaganda: If the topic has a regional, political, religious or ideological reference.
In the following, we focus on evaluating the linking of learning units in curricula to topics in textbooks. In our first experiment, we rely on pretrained GloVe embeddings[1] and the *bert-base-cased*[2] model for German language (embedding size: 768 dimensions, 12 layers, maximum sequence length 512 tokens). We collected 87 German curricula of the year 2016. These curricula only target history and society-related school subjects. In general, each German federal state publishes its own curriculum, along with a list of approved textbooks. The digitization of these textbooks resulted in 36,018 pages containing sentences or similar structural parts like table cells or bullet points. The evaluation of this work is based on a corpus with 127 of these approved textbooks. We consider each page of those textbooks as a document. The learning units in the curricula have been manually extracted. Extracted descriptions for each learning unit always include the best achievable knowledge descriptions, without any duplicate text or skill set descriptions. The evaluation of this work is based on 5 learning units, for history lessons. On average, there are 175 words describing each learning unit. The language used in

[1] https://deepset.ai/german-word-embeddings

[2] https://deepset.ai/german-bert

curricula and textbooks is fundamentally different. While curricula often only state the topic names, e.g. "Students should know about the end of World War II.", textbooks offer a deeper insight. Because these learning units originate from different curricula, some may cover similar or equal topics. The results on this task are presented in the following.

We measure the performance of the retrieval task in terms of precision@20. Due to the corpus size, we omit evaluating recall. For 5 distinct curricula topics, we evaluate the top 20 results obtained by T-HOTTER trained with 70 topics. As our baseline, we choose a standard BM25 scoring. Previous results by Yurochkin et al. (Yurochkin et al., 2019) allowed a comparison of HOTT to many metrics, including TF-IDF. Since TF-IDF has been a competitive scoring method we chose to experiment with BM25 on this corpus. First, we let the three experts judge the relevance of each retrieved document with a binary label. We instructed the experts to view relevance in terms of their perceived contribution of the document content to the learning unit. Then the experts compare the retrieved document with the provided explanations and judge each of them also with a binary label. We included three explanations: the context explanation obtained by German BERT embeddings (Ex-Contextual), the context explanation resulting from static German GloVe embeddings (Ex-Static) and the top *n* topic words (n=20) of the dominant topic assigned to the curriculum (Ex-Keywords). The explanations were assessed by the experts with respect to the question whether the keywords / auxiliary documents were understandably related to the retrieved document. We list the results separated by curriculum theme in Table 1. The retrieval precision is measured for documents retrieved by the T-HOTTER metric, given curriculum text as query. We choose T-HOTTER among all other aggregation methods because it offers the highest amount of context-sensitivity due to its preference on aggregating only representative topic tokens. Most of the top 20 retrieved documents by T-HOTTER for each curriculum topic were relevant (74%), with the highest average precision score (85%) achieved on the "French Revolution" theme and the lowest score (50%) in "Decolonization". The BM25 baseline is outperformed significantly in all cases by T-HOTTER. Interestingly, the themes "Western Modernity" and "Decolonization" in particular are difficult to retrieve using BM25 (with scores of 20% and 3%, respectively) because the curriculum descriptions are formulated in a more abstract manner. The T-HOTTER metric also yielded comparatively low scores on those two themes, however at least half of the results (scoring 50% and 68%) were relevant. We measured inter-annotator agreement using Fleiss' Kappa (Fleiss, 1971) and Krippendorff's Alpha (Krippendorff, 1970). Both metrics have a strong correlation. For BM25, the experts have the strongest agreement, with scores of 64% for Fleiss' Kappa and 63% for Krippendorff's Alpha. For the documents retrieved by T-HOTTER, the agreement is lower with 44% and 43%. After close inspection we found that T-HOTTER has retrieved several documents which are related to the given theme, but the relationship may not always be as evident as with the BM25 results. This is an advantage for the HOTTER approach when there are term mismatches with the curriculum theme, which a keyword-based approach such as BM25 does not overcome and therefore also not retrieve. Given the results on precision@20, T-HOTTER has the potential to be used as for textbook (page) retrieval in educational textbook research.

Further, we let the experts also evaluate explanations which we provide in addition to our retrieved documents. We obtain the auxiliary document serving as an explanation of the nearest neighborhood within the prominent topic of the originally retrieved document. This document is either obtained using T-HOTTER (Ex-Contextual) or HOTT (Ex-Static). The latter is called static because it is based on the cost between the German static GloVe vectors. It achieves the best scores overall, outperforming the contextual explanations drawn with T-HOTTER. Possible reasons for this behavior are corrupt tokens from the optical character recognition process within the used corpus, or that we did not further pre-train the German BERT model. The corrupt tokens are discarded by the regular HOTT approch if stemming does not return any valid term within the respective static word embeddings. On the other hand, the BERT model may use its subword mechanism and incorporate corrupt tokens which could affect the resulting contextual word embeddings. Nevertheless, both explanations provided with the HOTT variants have been assessed with average precision scores of 55% and 59%, respectively, whereas the explanation using only the top *n* topic keywords obtained very poor scores. The major criticism by the experts was that the

Learning Unit	BM25	T-HOTTER	Ex-Contextual	Ex-Static	Ex-Keywords
Cold War	0.67±0.18	**0.83**±0.08	0.68±0.20	**0.70**±0.20	0.30±0.50
Egypt	0.67±0.08	**0.83**±0.10	0.51±0.31	**0.55**±0.35	0.00±0.00
French Revolution	0.55±0.05	**0.85**±0.00	0.47±0.37	**0.51**±0.41	0.00±0.00
Western Modernity	0.20±0.13	**0.68**±0.16	0.61±0.14	**0.65**±0.13	0.00±0.00
Decolonization	0.03±0.06	**0.50**±0.26	0.47±0.08	**0.55**±0.10	0.03±0.06
Average	0.42±0.28	**0.74**±0.19	0.55±0.23	**0.59**±0.24	0.06±0.23
Fleiss' Kappa	**0.64**	0.44	0.34	0.35	0.00
Krippendorff's Alpha	**0.63**	0.43	0.30	0.30	-0.07

Table 1: Precision@20 scores with their standard deviation and inter-annotator agreement.

topic keywords did not appear to be related, neither to each other, nor to the respective retrieved document. For this experiment we did not perform any hyperparameter optimization on the topic model, which could have had a positive effect. However, the HOTT retrieval and explanations were based on the same topic model and performed reasonable. Considering the bad precision score for the topic keyword-based explanation which has been outperformed by the HOTTER approach, the impact of the topic model quality on the HOTT metric performance is yet to be studied. We found that the topics for the German educational textbook corpus were not crisp and did not seem coherent in many cases. However, the results retrieved by HOTT were nevertheless relevant and most explanations using the HOTTER method were more useful than the keyword-based ones. Our corpus can contain multiple issues of the same book title, thus we find language artifacts reflected by the topic model due to common text passages among a few books in the corpus. Those artifacts also impact the top *n* topic words, so that they may be another reason for bad interpretability. Given that we provide auxiliary documents as explanations in the other approaches, we point out that interpretability is assessed by humans who want to understand the context of the retrieved documents better, while at the same time having a limited tolerance for too much information. The size of the auxiliary documents which we chose as an explanation was a page within a textbook. Depending on the user it may be favorable to improve upon that simplified segmentation and dissect a textbook into paragraphs. Further work on the keyword-based explanations should consider the problem of overlapping topic keywords. This can be possibly mitigated by post-filtering the top keywords using the Term Frequency - Inverse Topic Frequency (TF-ITF) (Usui et al., 2006; Xie et al., 2008). As shown, our approach is very useful in the context of educational media research, when applied on a retrieval task. With the additional documents, the researcher can follow a reasoning, investigate additional concepts found on the textbook pages and start to compare different approaches to knowledge dissemination, for instance in the search for missing, altered or omitted knowledge.

5.2 Experiments on Public Datasets

Since HOTTER did not perform better on the educational media dataset than the original HOTT approach, we validate HOTTER on the same public datasets as Kusner et al. (2015). In our experiments[3], we use a pytorch implementation of the pre-trained BERT model (embedding size: 768 dimensions, 12 layers, maximum sequence length 512 tokens)[4] and then continue pre-training one model for each dataset individually for one epoch on batch size 32 to adapt to the domain. We set 70 topics for the LDA model. The performance of the alternative HOTTER aggregation methods is compared against HOTT and further baseline metrics, also used by Yurochkin et al. (2019), with the test error from a *k*-NN classifier on the seven multi-class datasets. We see from the results in Figure 2 that the performance of HOTTER is generally competitive, if not better. Although the differences in performance are rather small, we did not perform any hyperparameter tuning on the topic model or vary the random seed selection. T-HOTTER has the best scores among all baselines for 20NEWS and BBCSPORT. Therefore, we investigate the benefit of applying contextual embeddings by measuring the degree of contextualization within the top 20 topic words within all 70 topics for all datasets. For the challenging OHSUMED dataset we computed

[3] Code: https://github.com/anybass/HOTTER

[4] https://huggingface.co/bert-base-uncased

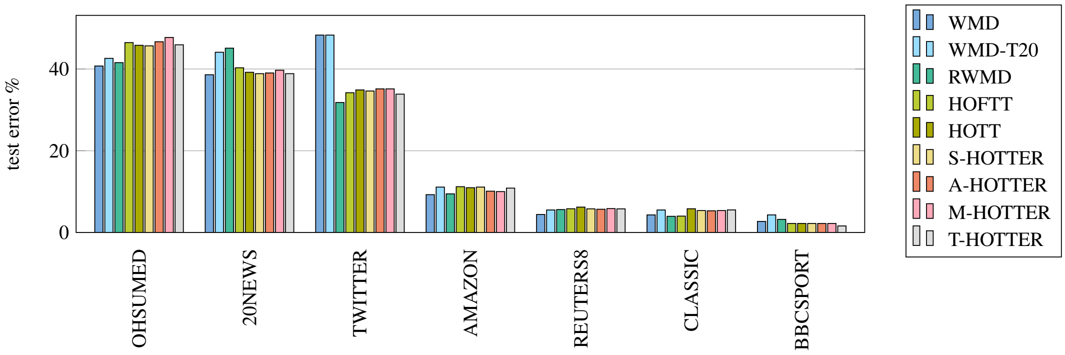

Figure 2: Test error in percent across seven datasets.

the SelfSim values and MEV, which were comparatively high. On the BBCSPORT dataset with the best results we find above-average SelfSim values and lowest MEV scores. We conclude that the results may be explained by the extent of contextualization present in the datasets: If there is a low variance among the contextual word embeddings, HOTTER has the potential to outperform HOTT and other metrics which are using static word vectors. This observation has also been made by Ethayarajh (2019), who shows that contextual word embeddings can outperform static ones when there is low contextualization. We therefore suppose that the best use cases for the HOTTER approach could be datasets with a domain focus that the underlying language model has been trained on, so that the aggregation of its contextualized embeddings for a given word can pose an advantage over static embeddings that were trained on a less focused dataset and which are thereby failing to capture the right context. Although the initial pre-training and possible further fine-tuning comes with significant costs, at run-time HOTTER was computed in a comparable time to regular HOTT. In general, the results can be sensitive to the number of topics, such that increasing them could improve the retrieval performance. There are also multiple factors stemming from the contextual embeddings which can impact the retrieval results. In our experiments, we used the last layer of the BERT language model to obtain contextual word embeddings, contrary to what the results by Ethayarajh (2019) suggest. However, we also checked the effect of extracting contextual embeddings from the first layer which is supposed to have the lowest degree of contextualization and to be a viable alternative to static word embeddings. The results were similar, the last layer gave us a small increase in the score though. The BERT model is usually employed after fine-tuning on a supervised downstream task, which we did not perform in our experiments. Hence, we find many further research opportunities regarding the way the embeddings are generated and then employed with the HOTT metric. Furthermore, choosing a distance metric for BERT embeddings is still subject to ongoing research, it is yet to be empirically validated on a broader range of tasks that the contextual representation is beneficial on a token basis.

6 Conclusion

In this paper we investigated two research questions for the recently proposed meta-distance HOTT, which computes optimal transport between documents using topic models and word embeddings. We showed that enhancing HOTT by contextual word embeddings from the BERT model is competitive. Our experiments on public datasets indicate that further pre-training of the language model offers an advantage over the original static HOTT variant. Leaving out further pre-training shows static word embeddings to perform better on the explanation component which we developed for the second research question in a retrieval setting on educational media data. Therefore, adapting contextual word embeddings to their domain via further pre-training may make a difference. Overall, the explanations offered by our HOTT variants are more interpretable than dominant topic keywords. Given those findings, we may improve the existing method by enforcing crisp topics. Although real-world corpora potentially include multiple versions of the same document, it may be worthwhile to employ document consolidation to different document versions in order to obtain coherent topics. Finding subtle differences between several textbook issues can be treated as a separate task in order to reduce language artifacts in the corpus.

References

Wei Bao, Hongshu Che, and Jiandong Zhang. 2020. Will_Go at SemEval-2020 task 3: An accurate model for predicting the (graded) effect of context in word similarity based on BERT. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 301–306, Barcelona (online). International Committee for Computational Linguistics.

David M Blei and John D Lafferty. 2009. Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 601–608. MIT Press.

Stefan Bunk and Ralf Krestel. 2018. WELDA: enhancing topic models by incorporating local word context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018*, pages 293–302. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016a. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 165–174. ACM.

Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. 2016b. Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 666–675, Berlin, Germany. Association for Computational Linguistics.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical word embeddings. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2418–2424. AAAI Press.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Christopher E. Moody. 2016. Mixing dirichlet topic models and word embeddings to make lda2vec. *CoRR*, abs/1605.02019.

Sedef Neitmann and Christian Scheel. 2020. Digitalisierung von (geistes)wissenschaftlichen arbeitspraktiken im alltag: Entwicklung und einführung eines werkzeugs zur digitalen annotation. *Berliner Blätter*, 82:119–132.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Ramandeep S. Randhawa, Parag Jain, and Gagan Madan. 2016. Topic modeling using distributed word embeddings. *CoRR*, abs/1603.04747.

Yik-Cheung Tam and Tanja Schultz. 2008. Correlated bigram LSA for unsupervised language model adaptation. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 1633–1640. Curran Associates, Inc.

Shiro Usui, Paulito P. Palmes, Kazunori Nagata, Tatsuki Taniguchi, and Naonori Ueda. 2006. Extracting keywords from research abstracts for the neuroinformatics platform index tree. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2006, part of the IEEE World Congress on Computational Intelligence, WCCI 2006, Vancouver, BC, Canada, 16-21 July 2006*, pages 5045–5050. IEEE.

Shasha Xie, Yang Liu, and Hui Lin. 2008. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *2008 IEEE Spoken Language Technology Workshop, SLT 2008, Goa, India, December 15-19, 2008*, pages 157–160. IEEE.

Guangxu Xun, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. 2017. A correlated topic model using word embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4207–4213. ijcai.org.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M. Solomon. 2019. Hierarchical optimal transport for document representation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1599–1609.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.