Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop, pages 32–37
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

A Comparison of Unsupervised Methods for Ad hoc Cross-Lingual Document Retrieval

Elaine Zosa, Mark Granroth-Wilding, Lidia Pivovarova
University of Helsinki
Helsinki, Finland
firstname.lastname@helsinki.fi

Abstract

We address the problem of linking related documents across languages in a multilingual collection. We evaluate three diverse unsupervised methods to represent and compare documents: (1) multilingual topic model; (2) cross-lingual document embeddings; and (3) Wasserstein distance. We test the performance of these methods in retrieving news articles in Swedish that are known to be related to a given Finnish article. The results show that ensembles of the methods outperform the stand-alone methods, suggesting that they capture complementary characteristics of the documents.

1. Introduction

We address the problem of retrieving related documents across languages through unsupervised cross-lingual methods that do not use translations or other lexical resources, such as dictionaries. There is a multitude of multilingual resources on the Internet such as Wikipedia, multilingual news sites, and historical archives. Many users may speak multiple languages or work in a context where discovering related documents in different languages is valuable, such as historical enquiry. This calls for tools that relate resources across language boundaries.

We choose to focus on methods that do not use translations because lexical resources and translation models vary across languages and time periods. Our goal is to find methods that are applicable across these contexts without extensive fine-tuning or manual annotation. Much work on cross-lingual document retrieval (CLDR) has focused on *cross-lingual word embeddings* but topic-based methods have also been used (Wang et al., 2016). Previous work has applied such cross-lingual learning methods to *known item search* where the task is to retrieve one relevant document given a query document (Balikas et al., 2018; Josifoski et al., 2019; Litschko et al., 2019). We are interested in *ad hoc retrieval* where there could be any number of relevant documents and the task is to rank the documents in the target collection according to their relevance to the query document (Voorhees, 2003).

Here we evaluate three existing unsupervised or weakly supervised methods previously used in CLDR for slightly different tasks: (1) multilingual topic model (MLTM); (2) document embeddings derived from cross-lingual reduced rank ridge regression or Cr5 (Josifoski et al., 2019) and; (3) Wasserstein distance for CLDR (Balikas et al., 2018). These methods link documents across languages in fundamentally different ways. MLTM induces a shared cross-lingual topic space and represents documents as a language-independent distribution over these topics; Cr5 obtains cross-lingual document embeddings; and the Wasserstein distance as used by (Balikas et al., 2018) computes distances between documents as sets of cross-lingual word embeddings (Speer et al., 2016). The methods broadly cover the landscape of recent CLDR methods. To our knowledge, this is the first comparison of Cr5 and Wasserstein for ad hoc retrieval.

This paper adds to the literature on CLDR in three ways: (1) evaluating unsupervised methods for retrieving related documents across languages (ad hoc retrieval), in contrast to retrieval of a single corresponding document; (2) evaluating different ensembling methods; and (3) demonstrating the effectiveness of relating documents across languages through complementary methods.

2. Related Work

Previous work on linking documents across languages has used translation-based features, where the query is translated into the target language and the retrieval task proceeds in the target language (Hull and Grefenstette, 1996; Litschko et al., 2018; Utiyama and Isahara, 2003). Other methods used term-frequency correlation (Tao and Zhai, 2005; Vu et al., 2009), sentence alignment (Utiyama and Isahara, 2003), and named entities (Montalvo et al., 2006). In this paper, we are interested in language-independent models with minimal reliance on lexical resources and other metadata or annotations.

2.1. Multilingual topic model

The multilingual topic model (MLTM) is an extension of LDA topic modelling (Blei et al., 2003) for comparable multilingual corpora (De Smet and Moens, 2009; Mimno et al., 2009). In contrast to LDA, which learns topics by treating each document as independent, MLTM relies on a topically aligned corpus, which consists of tuples of documents in different languages discussing the same themes. MLTM learns separate but aligned topic distributions over the vocabularies of the languages represented in the corpus. One of the main advantages of MLTM is that it can extend across any number of languages, not just two, as long as there is a topically aligned corpus covering these languages. This can be difficult because aligning corpora is not a trivial task, especially as the number of languages gets larger. For this reason, Wikipedia, currently in more than 200 languages, is a popular source of training data for MLTM.

Another issue facing topic models is that the choice of hyperparameters can significantly affect the quality and nature of topics extracted from the corpus and, consequently,

its performance in the downstream task we want use it for. There are three main hyperparameters in LDA-based models: the number of topics to extract, K; the document concentration parameter, α, that controls the sparsity of the topics associated with each document; and the topic concentration parameter, β, which controls the sparsity of the topic-specific distribution over the vocabulary.

2.2. Cross-lingual document embeddings

Cross-lingual reduced-rank ridge regression (Cr5) was recently introduced as a novel method of obtaining cross-lingual document embeddings (Josifoski et al., 2019). The authors formulate the problem of inducing a shared document embedding space as a linear classification problem. Documents in a multilingual corpus are assigned language-independent concepts. The linear classifier is trained to assign the concepts to documents, learning a matrix of weights W that embeds documents in a concept space close to other documents labelled with the same concept and far from documents expressing different concepts.
They train on a multilingual Wikipedia corpus, where articles are assigned labels based on language-independent Wikipedia concepts. They show that the method outperforms the state-of-the-art cross-lingual document embedding method from previous literature (Litschko et al., 2018). Cr5 is trained to produce document embeddings, but can also be used to obtain embeddings for smaller units, such as sentences and words. One disadvantage is that it requires labelled documents for training. However, the induced cross-lingual vectors can then be used for any tasks in which the input document is made up of words in the vocabulary of the corresponding language in the training set.

2.3. Wasserstein distances for documents

Wasserstein distance is a distance metric between probability distributions and has been previously used to compute distances between text documents in the same language (*Word Mover's Distance* (Kusner et al., 2015)). In (Balikas et al., 2018) the authors propose the Wasserstein distance to compute distances between documents from different languages. Each document is a set of cross-lingual word embeddings (Speer et al., 2016) and each word is associated with some weight, such as its term frequency inverse document frequency (tf.idf). The Wasserstein distance is then the minimum cost of transforming all the words in a query document to the words in a target document. They then demonstrate that using a regularized version of the Wasserstein distance makes the optimization problem faster to solve and, more importantly, allows multiple associations between words in the query and target documents.

3. Experimental setup

3.1. Task and dataset

We evaluate using a dataset of Finnish and Swedish news articles published by the Finnish broadcaster YLE and freely available for download from the Finnish Language Bank[1]. The articles are from 2012-18 and are written separately in the two languages (not translations and not parallel). This dataset contains 604,297 articles in Finnish and 228,473 articles in Swedish. Each article is tagged with a set of keywords describing the subject of the article. These keywords were assigned to the articles by a combination of automated methods and manual curation. The keywords vary in specificity, from named entities, such as *Sauli Niinisto* (the Finnish president), to general subjects, such as *talous* (sv: *ekonomi*, en: economy). On average, Swedish articles are tagged with five keywords and 15 keywords for Finnish articles. Keywords are provided in Finnish and Swedish regardless of the article language so no additional mapping is required.
To build a corpus of related news articles for testing, we associate one Finnish article with one or more Swedish articles if they share three or more keywords and if the articles are published in the same month. From this we create three separate test sets: 2013, 2014, and 2015. For each month, we take 100 Finnish articles to use as queries, providing all of the related Swedish articles as a candidate set visible to the models.
To build a topically aligned corpus for training MLTM, we match a Finnish article with a Swedish article if they were published within two days of each other and share three or more keywords. As a result no Finnish article is matched with more than one Swedish article and vice-versa so that we have a set of aligned unique article pairs. To train MLTM we use a year which is preceding the testing year: e.g., we train a model using articles from 2012 and test it on articles from 2013. Unaligned articles are not used for either training or testing. The script for article alignment will be provided in the Github repository for this work.
Table 1 shows the statistics of the training and test sets. As can be seen in the last column of the table, one Finnish article corresonds to almost twenty Swedish articles for the 2013 dataset and more than thirty for the other two datasets. This is typical for large news collections, since one article may have an arbitrary number of related articles. Thus, our corpus is more suitable for ad-hoc search evaluation than Wikipedia or Europarl corpus, since they contain only one-to-one relation[2].

	MLTM Train set	Test set	
	articles per lang	**#candidates**	**#related**
2012	7.2K	-	-
2013	7.2K	1.3K	19.5
2014	7.2K	1.4K	31.8
2015	-	1.5K	35.9

Table 1: Statistics of the training set for training MLTMs and test sets for each year. #candidates is the average size of the candidate articles set and #related is the average number of Swedish articles related to each Finnish article.

3.2. Models

We use our in-house implementation of MLTM training using Gibbs sampling[3]. The training corpus was tokenized, lemmatized and stopwords were removed. We limited the

[1] https://www.kielipankki.fi/corpora/

[2] CLEF 2000-2003 ad-hoc retrieval Test Suite, which also contains many-to-many relations, is not freely available

[3] https://github.com/ezosa/cross-lingual-linking.git

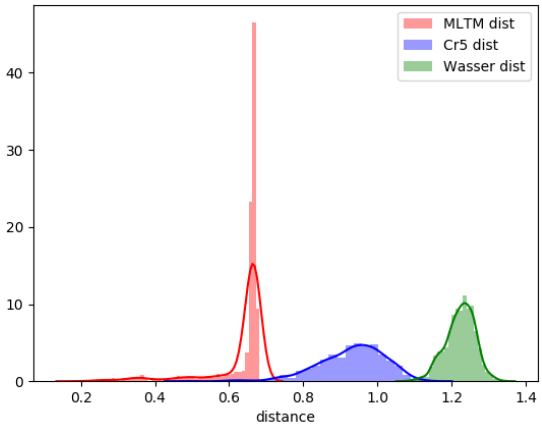

Figure 1: Density plots of the distances between one query document and the candidate documents.

vocabulary to the 9,000 most frequent terms for each language. We train three separate models for 2012, 2013, and 2014 (for the 2013, 2014, and 2015 test sets, respectively). We train all three models with $K = 100$ topics, $\alpha = 1/K$ and $\beta = 0.08$. We use 1,000 iterations for burn-in and then infer vectors for unseen documents by sampling every 25th iteration for 200 iterations. To obtain distances between documents, we compute the Jensen-Shannon (JS) divergence between the document-topic distributions of the query document and each of the candidate documents.

For Cr5, we use pretrained word embeddings for Finnish and Swedish provided by the authors[4]. We construct document embeddings according to the original method – by summing up the embeddings of the words in the document weighted by their frequency. We compute the distance between documents as the cosine distance of the document embeddings.

For Wasserstein distance, we use code provided by the authors for computing distances between documents and use the same cross-lingual embeddings they did in their experiments[5] (Speer et al., 2016). Wasserstein distance has a regularization parameter λ that controls how the model matches words in the query and candidate documents. The authors suggested using $\lambda = 0.1$ because it encourages more relaxed associations between words. Higher values of λ create stronger associations while too low values fail to associate words that are direct translations of each other. In this task, it might make more sense to use lower λ values, though an experiment with $\lambda = 0.01$ brought no noticeable improvement in performance (see Section 3.3.).

We created ensemble models by averaging the document distances from the stand-alone models and ranking candidate documents according to this score. We construct four ensemble models by combining each pair of models, as well as all three: **MLTM_Wass**; **Cr5_Wass**; **MLTM_Cr5**; and **MLTM_Cr5_Wass**.

3.3. Results and Discussion

Table 2 shows the results for each model and ensemble on each of the three test sets, reporting the precision of the top-ranked k results and mean reciprocal rank (MRR). Cr5 is the best-performing stand-alone model by a large margin. Cr5 was originally designed for creating cross-lingual document embeddings by classifying Wikipedia documents according to concepts. We did not retrain it for our particular task. Nevertheless, using these pre-trained word embeddings we were able to retrieve articles that discuss similar subjects in this different domain. However, it is worth noting that Cr5 can only be trained on languages for which labels are available for *some* similarly transferable training domain.

MLTM, being a topic-based model, would seem like the obvious choice for a task like this because we want to find articles that share some broad characteristics with the query document, even if they do not discuss the same named entities or use similar words. However, Cr5 outperforms MLTM on its own. One reason may be that 100 topics are too few. We chose this number because it seemed to give topics that are specific enough for short articles but still broad enough that they could reasonably be used to describe similar articles. Another drawback of this model is that it does not handle out-of-vocabulary words and the choice of using a vocabulary of 9,000 terms might be too low.

Wasserstein distance is the worst-performing of the stand-alone models especially for the 2014 and 2015 test sets where it offers little improvement when ensembled with Cr5 (Cr5_Wass). A possible reason is that it attempts to transform one document to another and therefore favors documents that share a similar vocabulary to the query document. The technique might be suitable for matching Wikipedia articles, as shown in (Balikas et al., 2018) because they talk about the same subject at a fine-grained level and use similar words, whilst in our task the goal is to make broader connections between documents.

In Figure 1, the density plots of the distances of one query document and the candidate documents. We see that MLTM and Wasserstein tend to have sharper peaks while Cr5 distances are flatter. MLTM has minimum and maximum distances of 0.2 and 0.68, respectively, while Cr5 has 0.49 and 1.14, and Wasserstein has 1.08 and 1.34. Topic modelling tends to predict that most of the target documents are far from the query document (peaks at the right side). This is not only true for this particular query document but for other query documents in our test set as well. We also see that Wasserstein has larger distances which is potentially problematic. We tried normalizing the distances produced by the models such that they are centered at zero and using these distances for the ensembled model however it produces the same document rankings as the unnormalized distances. This might be because we are only concerned with the documents with the smallest distances where Wasserstein does not contribute much.

For the ensemble models, combining all three models per-

[4]https://github.com/epfl-dlab/Cr5

[5]https://github.com/balikasg/WassersteinRetrieval

Test set:	2013				2014				2015			
Measure:	**P@1**	**P@5**	**P@10**	**MRR**	**P@1**	**P@5**	**P@10**	**MRR**	**P@1**	**P@5**	**P@10**	**MRR**
MLTM	21.8	18.2	16.3	31.6	24.1	22.4	20.6	34.8	30.8	29.0	27.1	41.6
Wass	21.1	13.7	11.3	30.8	21.0	16.9	14.7	31.9	25.1	20.6	17.9	37.2
Wass $\lambda = 0.01$	20.3	13.5	11.1	30.0	21.3	16.8	14.6	32.0	25.1	20.1	17.3	36.6
Cr5	32.5	24.5	21.2	41.7	38.3	30.2	26.0	48.0	43.1	37.1	33.5	53.8
MLTM_Wass	24.6	21.3	19.1	35.2	27.3	25.5	23.4	38.2	30.4	31.4	30.1	42.9
Cr5_Wass	35.4	27.4	23.2	45.2	38.1	32.2	28.2	49.2	41.2	37.7	34.9	52.9
MLTM_Cr5	36.4	28.2	24.4	46.6	**44.8**	34.3	30.1	53.6	42.7	40.1	36.9	54.5
MLTM_Cr5_Wass	**40.7**	**30.7**	**26.3**	**50.3**	43.0	**36.1**	**31.9**	**53.8**	**44.5**	**41.3**	**38.5**	**55.9**

Table 2: Precision at k and MRR of cross-lingual linking of related news articles obtained by three stand-alone models and four ensemble models.

Test set:	2013	2014	2015	AVG
MLTM, Wass	-0.039	-0.016	-0.022	-0.026
Cr5, Wass	0.128	0.027	0.026	0.060
MLTM, Cr5	0.156	0.164	0.178	0.166

Table 3: Mean Spearman correlation of the ranks of candidate documents for each pair of models.

forms best overall for all three test sets and all but one precision level—the only exception is P1 for 2014 where MLTM_Cr5 achieves roughly the same performance. This tells us that each model sometimes finds relevant documents not found by the other models. The correlation of candidate document rankings between the different methods is quite low (Table 3). We compute the correlation between the ranks for each of the 1200 query documents (100 queries for each month) for each year of our test set and average them. As can be seen in the table the correlations are rather low, which means that they retrieve documents based on different principles. The highest correlation is between **MLTM** has the **Cr5** while correlation between **MLTM** and **Wass** is the lowest.

This suggests that there are different ways of retrieving related documents across languages and that the three methods of cross-lingual embeddings, cross-lingual topic spaces and cross-lingual distance measures capture complementary notions of similarity. A simple combination of their decisions is thus able to make better judgements than any can make on its own.

As an example, in Table 4 we show excerpts from a query article in Finnish and some of the related Swedish articles correctly predicted by the different models. For this article, Cr5 gave 10 correct predictions in its top 10 (perfect precision), MLTM gave 8 correct predictions and Wasserstein only 4. Like Cr5, the ensemble model MLTM_Cr5_Wass also achieved perfect precision. MLTM and MLTM_Cr5_Wass shared 4 correct predictions while Cr5 and MLTM_Cr5_Wass shared 7. All the articles correctly predicted by Wasserstein were also predicted by the other models. We show articles from Cr5, MLTM and MLTM_Cr5_Wass that was correctly predicted by that model only and for Wasserstein, we show the top correct article that it predicted.

4. Conclusions and Future work

In this paper we compare three different methods for cross-lingual ad hoc document retrieval by applying them to the task of retrieving Swedish news articles that are related to a given Finnish article. We show that a word-embedding based model, Cr5, performs best followed by the multilingual topic model and the distance-based Wasserstein model has the worst results of the stand-alone models. We then demonstrate that combining at least two of these methods by averaging their distances yields better results than the models used on their own. Finally we show that combining the three models yields the best results. These results tell us that relating documents based on different techniques such as embedding-based or topic-based techniques yields different results and that pooling these results make for a better model.

In the future we plan to investigate the performance of word embedding-based multilingual topic models in this task. There is already some work done on developing topic models that use word embeddings (Batmanghelich et al., 2016; Das et al., 2015). To our knowledge, they have not yet been applied to cross-lingual embeddings. Such a model could potentially combine the benefits of the multilingual topic model with word embeddings for retrieving similar documents across languages.

We also plan to further experiments with multilingual topic models for languages where the amount of linked documents is scarce. In this work, we trained the topic model with thousands of linked articles because the articles were annotated with tags however this might not always be the case, for instance with historical data sets or under-resourced languages where there are not readily available annotated data and manual annotation is time-consuming or requires expert knowledge. In such cases, we could still train a multilingual topic model with smaller amounts of aligned training data or perhaps a training set where some articles do not have a counterpart article in the other language.

There is also scope for further exploration of ensemble methods, going beyond the simple combination of distance metrics we have applied here. As well as combining models in different ways, further, potentially complementary,

Query article	Yleisradion YleX-kanavan kymmenen suosituimman kappaleen listalla,valtaosa on suomalaisartisteja tai -yhtyeitä. Radio Suomen kaikki,kymmenen eniten kuultua kappaletta ovat odotetusti kotimaisia. YleX ja Radio Suomi ovat koonneet listan eniten soittamastaan musiikista vuonna 2012.
MLTM	På min låtlista finns låtar som på olika sätt och från olika perspektiv beskriver livets grundläggande vemod eller ”life bitter-sweet”, som man brukar säga på Irland. Det säger Tom Sjöblom, som har valt musiken denna vecka i [Min musik.]
Cr5	De isländska banden tar över världen, vi träffade Sóley som nyligen varit på USA-turné med sina isländska kollegor Of Monsters And Men. **Sóley** är isländska och betyder solros. Sóley är också namnet på sångerskan som är en av de mest intressanta nya musikexporterna som kommit från Island.
Wasserstein	Både Radio Vega och Radio Extrem har börjat spela låtar som tävlar i Tävlingen för ny musik UMK. Radio Extrem har tagit in både Krista Siegfrids __Marry me__ och Diandras __Colliding into you__ på spellistan, och låtarna kommer att spelas två gånger om dagen åtminstone nu i början.
MLTM_Cr5_Wass	Smakproven på 30 sekunder av de tolv UMK låtarna kittlade fantasin så,där passligt, men nu behöver vi inte längre gissa oss till hur sångerna,låter i sin helhet. De färdigt producerade bidragen kan nu höras på,Arenan.

Table 4: Excerpt from a query Finnish article and some related Swedish articles correctly predicted by the models. The query article is about popular songs on Finnish radio.

measures of document similarity could be included: for example, explicitly taking into account overlap of named entities, or document publishing metadata if such information is available.

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye) and 825153 (EMBEDDIA).

References

Balikas, G., Laclau, C., Redko, I., and Amini, M.-R. (2018). Cross-lingual document retrieval using regularized Wasserstein distance. In *European Conference on Information Retrieval*, pages 398–410. Springer.

Batmanghelich, K., Saeedi, A., Narasimhan, K., and Gershman, S. (2016). Nonparametric spherical topic modeling with word embeddings. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2016, page 537. NIH Public Access.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Das, R., Zaheer, M., and Dyer, C. (2015). Gaussian LDA for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804.

De Smet, W. and Moens, M.-F. (2009). Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the 2nd ACM workshop on Social web search and mining*, pages 57–64. ACM.

Hull, D. A. and Grefenstette, G. (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–57. Citeseer.

Josifoski, M., Paskov, I. S., Paskov, H. S., Jaggi, M., and West, R. (2019). Crosslingual document embedding as reduced-rank ridge regression. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 744–752. ACM.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Litschko, R., Glavaš, G., Ponzetto, S. P., and Vulić, I. (2018). Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1253–1256. ACM.

Litschko, R., Glavaš, G., Vulic, I., and Dietz, L. (2019). Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1109–1112. ACM.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics.

Montalvo, S., Martinez, R., Casillas, A., and Fresno, V. (2006). Multilingual document clustering: an heuristic approach based on cognate named entities. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1145–1152. Association for Computational Linguistics.

Speer, R., Chin, J., and Havasi, C. (2016). ConceptNet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975.

Tao, T. and Zhai, C. (2005). Mining comparable bilingual text corpora for cross-language information integration. In *Proceedings of the eleventh ACM SIGKDD interna-*

tional conference on Knowledge discovery in data mining, pages 691–696. ACM.

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 72–79. Association for Computational Linguistics.

Voorhees, E. (2003). Overview of TREC 2003. pages 1–13, 01.

Vu, T., Aw, A. T., and Zhang, M. (2009). Feature-based method for document alignment in comparable news corpora. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 843–851. Association for Computational Linguistics.

Wang, Y.-C., Wu, C.-K., and Tsai, R. T.-H. (2016). Cross-language article linking with different knowledge bases using bilingual topic model and translation features. *Knowledge-Based Systems*, 111:228–236.