"Alignment is All You Need": Analyzing Cross-Lingual Text Similarity for Domain-Specific Applications

Sourav Dutta[0000−0002−8934−9166]

Huawei Research Centre, Dublin, Ireland
sourav.dutta2@huawei.com

Abstract. Cross-lingual text similarity provides an important measure to adjudge the contextual and semantic similarity between documents across different languages. Extraction of similar or aligned multi-lingual texts would enable efficient approaches for information retrieval and natural language processing applications. However, diversity of linguistic constructs coupled with domain specificity and low resources pose a significant challenge. In this paper, we present a study analyzing the performance of different existing approaches, and show that *Word Mover's Distance* on aligned language embedding provides a reliable and cost-effective cross-lingual text similarity measure to tackle evolving domain information, even when compared to advanced machine learning models.

1 Introduction

Motivation. The explosion of openly available data on the World Wide Web provides a wealth of mono-lingual text information in terms of documents, news articles, and blogs to name a few. However, with the growth of application domains catering to diverse geographic regions and target user groups, understanding of cross-lingual information has become an interesting area of research. As such, estimating the *semantic similarity* between documents, transcending the language barrier, plays an important role in information retrieval for applications like news aggregation [26], document summarization, and question answering [4]. Further, extraction of *parallel or comparable corpora* assumes a crucial role in a variety of natural language processing (NLP) tasks like machine translation [20], word-level lexicon similarities [23], and in learning large multi-lingual language models like BERT [9] and XLM [13]. Generalization of the above can lead to enhancements in diverse downstream language-agnostic NLP applications like document clustering, information fusion, translation [32], or parallel corpus generation.

Challenges. Consider an organization to expand its offerings to different global markets, naturally encompassing a multi-lingual target customer base. Scaling of operations for its automated services like Chatbots [1] and retrieval systems [15],

would require "multi-lingual capabilities" to cater to the new or emerging markets. In such scenarios, semantic relationships among such cross-lingual diverse data need to be efficiently computed for ease of information aggregation, and for business analytics based on geographical trends. Further, in terms of *event-centric* information, like election reports or natural calamities, the rapid evolution of data over time, across different sources, as well as in different languages, posing a challenge for obtaining a global understanding of the event evolution, possible interlinkings, and information propagation. Current approaches tend to rely on large pre-trained language models for computing textual semantic similarity, and also for "zero-shot learning" capabilities [31] to automatically transfer knowledge between languages – enabling decent multi-lingual document understanding and retrieval performance.

However, in both the above scenarios, existing techniques suffer from the presence of (i) domain-specificity, (ii) limited training data, and (iii) evolving information. Specifically, pre-trained language models might fail to identify contextual relationships in domain-specific applications (in terms of the operational domain of the enterprise), and would be difficult to train on evolving information (having limited training data). Further, language models like T5 [21] and GPT-3 [6] are extremely expensive and resource-intensive to train, maintain, and use in practice – potentially limiting its application to large organizations only. Other learning techniques might also be ineffective due to the presence of *limited training resources* (in terms of annotated data), and in general fail to sufficiently generalize to morphologically rich and low-resourced languages.

Contributions. In this paper, we study the problem of *multi-lingual text alignment* and explore the efficacy of unsupervised strategies for accurately capturing semantic similarity in cross-lingual domain-specific contents. To this end, we show that the *Word Mover's Distance* [12] measure applied on *aligned vector space vocabulary embedding* across languages is quite effective in terms of accuracy, comparable to state-of-the-art large language model architectures. Experiments on various domains and languages showcase the above strategy to be reliable in handling not only domain-specific data, but also morphologically rich and low-resourced languages – providing a generalizable, cost-effective, multi-lingual text similarity measure for information retrieval, aggregation and fusion.

2 Understanding Cross-Lingual Text Similarity

Initial approaches for cross-lingual text similarity relied on differences across documents based on descriptor terms from multi-lingual thesaurus [30]. However, to reduce the dependency on expensive manually created parallel dictionaries, translation-based text similarity approaches using lexical and syntactic features along with overlap of synonymous words from resources like WordNet [19] were proposed [27]. With automated translation techniques and growth of parallel data availability, machine learning models were used to detect semantically similar documents [26], which were further extended to zero-shot environments via

transfer learning from pre-trained language models [14]. We next briefly discuss possible state-of-the-art strategies to measure text similarity.

(A) Aligned Semantic Distance. The success of distributional word representations like Word2Vec [18] and FastText [5] in capturing word meanings has been established in a wide range of NLP tasks. Document embedding techniques like *doc2vec* [18] with cosine distance have been traditionally used to effectively capture semantic similarity between texts. The *Word Mover's Distance* (WMD) [12] provides a far more effective *unsupervised distance metric* by formulation as an optimal transport problem based on the Earth Mover's or Wasserstein distance. Mathematically, given two distributions $\mathcal{X}$ and $\mathcal{Y}$ (set of word embeddings), WMD computes the minimum effort of transforming one distribution to the other by solving $min_{\Gamma} \sum_{ij} \Gamma_{ij} \mathcal{C}_{ij}$, where Γ_{ij} is the amount of transformation required and $\mathcal{C}_{ij}$ is the associated distance between points i and j across the two distributions (or documents). However, in multi-lingual settings, the distributed word vector representations for the different languages are created in potentially different embedding spaces, as the monolingual embeddings are learnt in a relative fashion, and thus might have different orientations and degrees of freedom [2]. As such, direct application of the original WMD formulation in this setting would not be appropriate.

Cross-lingual word embedding alignment entails mapping the vocabularies of the different languages onto a single vector space to capture syntactic and semantic similarity of words across language boundaries [8]. Leveraging the similarity of geometric properties between different monolingual word embedding spaces [17], several supervised and unsupervised frameworks employing adversarial learning coupled with refinement strategies were proposed such as MUSE [8], VecMap [3], and RCSLS [11]. In general, to align the different independently learnt monolingual word embeddings (for the different languages) onto a common vector representation space, the above approaches tend to learn a transformation $\mathcal{T} : \mathcal{X} \rightarrow \mathcal{Y}$ between the two language embeddings. Mathematically this involves the optimization of $min_{\mathcal{T}} ||\mathcal{X} - \mathcal{T}(\mathcal{Y})||_F^2$, where $||\cdot||_F$ is the Frobenius norm. This, when constrained to orthonormal matrices, results in the closed-form *orthogonal Procrustes problem* [28], used in literature as a refinement strategy.

Once the vector spaces are aligned onto a common representation domain, the WMD formulation can now be directly applied on this new shared space, as presented in [4]. We refer to this approach as *Word Mover's Distance - Aligned* (WMD-A), while the use of the original non-aligned embeddings for WMD is denoted as WMD-NA. Alternatively, a naïve approach would be to obtain text representation via averaged aligned word embeddings, and using cosine distance between the contents as a measure of similarity. However, use of such weighting strategies tend to depict lower accuracy [25], and are hence not considered as baselines in our analysis.

(B) Multi-Domain Wasserstein Distance. Since Wasserstein distance is not compatible for comparing distances across different geometric domain spaces (as mentioned above), *Gromov-Wasserstein* (GW) distance [16] was proposed. To generalize between different domains, the GW distance takes into account

the metric spaces and compares the distance between points in one domain space to the distances of points in the other space, i.e., finds the optimal using $\min_{\Gamma} \mathcal{C}_{ij,kl}\Gamma_{ik}\Gamma_{jl}$, where Γ specifies the transformation between the pairs of points from the two spaces (i, j and k, l) and $\mathcal{C}$ is the difference in the distances between the point pairs within the individual domains. In this scenario, WMD coupled with GW distance (henceforth referred to as *WMD-GW*), provides a viable option for computing multi-lingual document semantic similarity based on (non-aligned) word embeddings across different languages.

(C) Pre-trained Language Models. *Contextual language models* (CLM) like BERT [9] take into account the context of a word occurrence to provide "dynamic" word embeddings, where the same word in different contexts is represented by different vector embeddings, capturing possibly different meanings based on the usage. A natural extension to multi-lingual settings were explored, and CLMs like M-BERT [9] and XLM-R [13] were developed using the *transformer* architecture with multi-lingual parallel corpus and shared vocabulary. The presence of open-source data enabled the creation of huge pre-trained language models from large repositories of Wikipedia pages. These language models were shown to be adept at several NLP tasks like question answering [22], text summarization, document similarity, text generation [6], and zero-shot transfer learning, with "near human-level" language "understanding" in certain scenarios [10]. Thus, text embeddings from multi-lingual CLMs like M-BERT coupled with cosine similarity measure are usually used for computing cross-lingual document similarity.

Larger language models were shown to be better for downstream NLP tasks, leading to enormous models like T5 [21] and GPT-3 [6] with 11B and 175B parameters respectively. Although, their performance in text similarity, comprehension, translation, and zero-shot transfer learning were astounding, these models require high-end compute resources for training (re-training to capture relationships in domain specific and evolving data is not practical), and are susceptible to low-resourced languages.

(D) Sentence Embedding. Recent approaches like dual-encoder based *Universal Sentence Encoder* [7] or Siamese network based SBERT [24] for generating contextualized sentence embeddings involve a layer of deep learning architecture atop the pre-trained contextualized language models, providing multi-lingual sentence embeddings. Specifically, the sentence transformer architectures of multi-lingual SBERT uses the teacher-student knowledge distillation framework coupled with fine-tuned language model to generate effective multi-lingual text vector representations. Such models have been shown to outperform sentence embeddings, obtained directly from the CLMs, for text similarity tasks. Multi-lingual SBERT with cosine similarity measure is considered as a strong state-of-the-art technique for capturing semantic similarity across multi-lingual short texts.

3 Experimental Analysis

We analyze the performance of the above techniques for computing *multi-lingual text semantic similarity*, in presence of domain-specificity and linguistic diversity.

Baselines. We analyze the following existing approaches in cross-lingual settings:
(i) WMD-NA – Word Mover's Distance on the independently learnt monolingual embeddings (without any alignment) obtained from FastText (fasttext.cc);
(ii) WMD-GW – Word Mover's Distance coupled with Gromov-Wasserstein distance, implemented using Python Optimal Transport library (pythonot.github.io);
(iii) WMD-A – Word Mover's Distance on aligned FastText word embeddings (alignment for Xhosa to English obtained from VecMap (github.com/artetxem/vecmap);
(iv) M-BERT – token embeddings from pre-trained multi-lingual BERT language model (using github.com/hanxiao/bert-as-service) is used to compute cosine similarity between texts; and,
(v) SBERT – text embeddings obtained from multi-lingual sentence transformer based teacher-student architecture (github.com/UKPLab/sentence-transformers) is used with cosine similarity for content similarity.

Dataset. We use document across different languages and diverse domains from *OPUS*, the open parallel corpus, obtained from opus.nlpl.eu. Specifically, we use sentence translation pairs from the following six domain-specific collections as:
(i) EMEA (Medical) – a parallel corpus from the European Medicines Agency;
(ii) JRC-Acquis (Judicial) – a collection of legislative text of the European Union;
(iii) Bible-uedin (Religious) – created from the translations of the Holy Bible;
(iv) MultiUN (Legislative) – collection of documents from the United Nations;
(v) TedTalks (Generic) – corpus of transcribed and translated TED talks; and,
(vi) XhosaNavy (Maritime) – contains maritime texts from South African Navy.

We also considered six different languages (including morphologically rich and low-resourced), namely German (de), Finnish (fi), Romanian (ro), Russian (ru), Croatian (hr), and Xhosa (xh) – containing a diverse combination of *isolating, fusional and agglutinative language with dependent and mixed marking* [29]. For each of the above datasets, $5K$ parallel sentences across each language pair were randomly sampled to form the dataset.

Task. Given a language pair (X, Y), for each input sentence in X the corresponding translation in Y is extracted by using cosine similarity computation. For example, given a sentence x_i in language X, its similarity is computed to sentences y_j (for $\forall j$) of language Y, and the one with the maximum cosine similarity score is reported as the translation of x_i by the algorithm.

Evaluation Measure. We evaluate the accuracy of the competing algorithms using the *Precision-at-Rank-1* (P@1) and *Precision-at-Rank-5* (P@5) measures. This reports the percentage of times the ground-truth translations between the input sentences (across language pairs) are present in the extracted top-1 and top-5 results, respectively. Note, equivalent sentence translations should represent the highest semantic similarity, and should be reported as top-1.

4 Experimental Observations

Table 1 tabulates the multi-lingual sentence semantic similarity accuracy scores as obtained by the different approaches across the languages. We observe that

Table 1. Cross-Lingual domain-specific sentence similarity accuracy across languages.

Dataset /	**EMEA** (Medical)						**JRC** (Judicial)						**Bible** (Religious)					
	en-de		*en-fi*		*en-ro*		*en-de*		*en-fi*		*en-ro*		*en-de*		*en-fi*		*en-ro*	
Method	$P_{@1}$	$P_{@5}$	$P_{@1}$	$P_{@5}$	$P_{@1}$	$P_{@5}$	$P_{@1}$	$P_{@5}$	$P_{@1}$	$P_{@5}$	$P_{@1}$	$P_{@5}$	$P_{@1}$	$P_{@5}$	$P_{@1}$	$P_{@5}$	$P_{@1}$	$P_{@5}$
WMD-NA	0.011	0.066	0.011	0.054	0.022	0.077	0.011	0.042	0.012	0.071	0.012	0.06	0.0	0.05	0.01	0.06	0.01	0.04
WMD-GW	0.143	0.33	0.086	0.28	0.10	0.297	0.063	0.263	0.083	0.345	0.084	0.349	0.13	0.31	0.10	0.26	0.04	0.17
M-BERT	0.55	0.79	0.33	0.57	0.50	0.735	0.375	0.677	0.32	0.546	0.283	0.778	0.28	0.51	0.24	0.58	0.17	0.33
SBERT	**0.901**	**0.956**	*0.54*	*0.80*	*0.847*	**0.98**	**0.842**	**1.00**	*0.417*	*0.802*	**0.545**	**1.00**	**0.98**	**1.00**	*0.16*	*0.38*	*0.29*	*0.57*
WMD-A	*0.791*	*0.901*	**0.806**	**0.935**	**0.868**	*0.967*	*0.726*	*0.905*	**0.809**	**0.893**	*0.373*	*0.831*	*0.91*	**1.00**	**0.80**	**0.96**	**0.80**	**0.97**

Table 2. Text similarity for (a) domain-specificity with distant languages and (b) different language base on the JRC dataset.

(a)

Dataset	**MultiUN** (en-ru)		**TedTalks** (en-hr)		**XhosaNavy** (en-xh)	
/ Method	$P_{@1}$	$P_{@5}$	$P_{@1}$	$P_{@5}$	$P_{@1}$	$P_{@5}$
WMD-NA	0.011	0.069	0.01	0.06	0.012	0.047
WMD-GW	0.08	0.264	0.13	0.36	0.023	0.08
M-BERT	0.523	0.727	0.71	*0.89*	0.04	0.10
SBERT	**0.875**	**0.943**	*0.76*	0.88	*0.28*	*0.45*
WMD-A	*0.828*	**0.943**	**0.95**	**0.97**	**0.43**	**0.59**

(b)

Language	*de-ro*	de-fi	fi-ro
/ Method	$P_{@1}$	$P_{@1}$	$P_{@1}$
SBERT	0.487	0.782	0.427
WMD-A	0.467	0.766	0.375

for high-resource languages like German, *SBERT* performs the best, showcasing robustness to domain-specificity, while *WMD-A* provides comparable results. However, on morphologically rich and low-resource languages like Finnish and Romanian, *WMD-A* is seen to outperform the other methods, in almost all the domains. Intuitively, the presence of sufficient training resource for well-documented languages provides enhanced performance for supervised methods like SBERT. But, for morphologically rich and low-resourced languages training of language models is challenged by linguistic diversity and scarcity of resources.

To further explore the robustness of the algorithms, Table 2(a) considers a more challenging setting of domain-specificity coupled with distant languages. We observe that in this case, unsupervised *WMD-A* based on aligned vector embeddings consistently outperforms the other techniques in nearly all the scenarios. As expected, WMD on non-aligned embedding space (WMD-NA) performs the worst, as the optimal transportation is not theoretically geared for comparison across different domains. The use of Gromov-Wasserstein distance brings a healthy improvement in WMD, however falls significantly short compared to the other algorithms; possibly due to the presence of limited context. However, it is interesting to note that *M-BERT* fails to perform well, possibly due to its lack of robustness to domain-specificity for generating embeddings.

The above observations are applicable when the language pairs are coupled with English, however in certain scenarios semantic similarity in texts across other language pairs might be necessary. For completeness, we compare the performance of the methodologies with different language pair bases as shown in Table 2(b). We observe that in these scenarios, SBERT performs the best while WMD-A showcases comparable results. As most cross-lingual vocabulary alignment techniques consider English as the base shared embedded space, the

dependency of WMD-A on English is portrayed in this analysis. In general, we find that *WMD-A* provides an effective method for textual semantic similarity, across diverse domains and languages.

5 Conclusion

In this paper, we compare and analyze the performance of state-of-the-art approaches for text similarity in the face of domain-specificity and diverse linguistic variations. We observed that *Word Mover's Distance* based on *aligned vector space embedding* provides an efficient and unsupervised technique for computing cross-lingual textual similarity. It is robust to domain-specific data even on morphologically rich and distant languages, and might be easily applicable to evolving event-centric information (without the need of any training process). Overall *WMD-A* provides an effective method for textual semantic similarity, comparable to state-of-the-art advanced machine learning methods and language models – depicting *"alignment is all you need"*.

References

1. Abbet, C.e.a.: Churn Intent Detection in Multilingual Chatbot Conversations and Social Media. In: CoNLL. pp. 161–170 (2018)
2. Alvarez-Melis, D., Jaakkola, T.: Gromov-Wasserstein Alignment of Word Embedding Spaces. In: EMNLP. pp. 1881–1890 (2018)
3. Artetxe, M., Labaka, G., Agirre, E.: A Robust Self-learning Method for Fully Unsupervised Cross-lingual Mappings of Word Embeddings. In: ACL. pp. 789–798 (2018)
4. Balikas, G., Laclau, C., Redko, I., Amini, M.: Cross-Lingual Document Retrieval Using Regularized Wasserstein Distance. In: ECIR. pp. 398–410 (2018)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
6. Brown, T.B.e.a.: Language Models are Few-Shot Learners (2020), arXiv:2005.14165
7. Cer, D.e.a.: Universal Sentence Encoder (2018), arXiv preprint arXiv:1803.11175
8. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word Translation Without Parallel Data. In: ICLR. pp. 1–14 (2018)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT. pp. 4171–4186 (2019)
10. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In: ICLR. pp. 1–20 (2021)
11. Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., Grave, E.: Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In: EMNLP. pp. 2979–2984 (2018)
12. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From Word Embeddings To Document Distances. In: ICML. pp. 957–966 (2015)
13. Lample, G., Conneau, A.: Cross-Lingual Language Model Pretraining (2019), arXiv:1901.07291

14. MacAvaney, S., Soldaini, L., Goharian, N.: Teaching a New Dog Old Tricks: Resurrecting Multilingual Retrieval Using Zero-Shot Learning. In: ECIR. pp. 246–254 (2020)
15. Mass, Y., Carmeli, B., Roitman, H., Konopnicki, D.: Unsupervised FAQ Retrieval with Question Generation and BERT. In: ACL. pp. 807–812 (2020)
16. Mémoli, F.: GromovWasserstein Distances and the Metric Approach to Object Matching. Foundations of Computational Mathematics **11**, 417–487 (2011)
17. Mikolov, T. Le, Q.V., Sutskever, I.: Exploiting Similarities among Languages for Machine Translation (2013), arXiv preprint arXiv:1309.4168
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: NIPS. pp. 3111–3119 (2013)
19. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography **4**(3), 235–244 (1990)
20. Munteanu, D.S., Marcu, D.: Improving Machine Translation Performance by Exploiting Non-parallel Corpora. Computational Linguistics **31**(4), 477–504 (2005)
21. Raffel, C.e.a.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)
22. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: EMNLP. pp. 2383–2392 (2016)
23. Rapp, R.: Automatic Identification of Word Translations from Unrelated English and German Corpora. In: ACL. pp. 519–526 (1999)
24. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: EMNLP. pp. 3982–3992 (2019)
25. Rücklé, A., Eger, S., Peyrard, M., Gurevych, I.: Concatenated Power Mean Word Embeddings as Universal Cross-Lingual Sentence Representations (2018)
26. Rupnik, J., Muhič, A., Leban, G., Fortuna, B., Grobelnik, M.: News Across Languages - Cross-Lingual Document Similarity and Event Tracking. In: IJCAI. pp. 5050–5054 (2017)
27. Santosh, G.S.K., Kumar, N.K., Varma, V.: Ranking Multilingual Documents Using Minimal Language Dependent Resources. In: CICLing. pp. 212–220 (2011)
28. Schönemann, P.H.: A Gen. Sol. of the Ort. Procrustes Prob. Psychometrika **31**(1), 1–10 (1966)
29. Søgaard, A., Ruder, S., Vulić, I.: On the Limitations of Unsupervised Bilingual Dictionary Induction. In: ACL. pp. 778–788 (2018)
30. Steinberger, R., Pouliquen, B., Hagman, J.: Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. In: CICLing. pp. 415–424 (2002)
31. Xian, Y., Schiele, B., Akata, Z.: Zero-shot Learning - the Good, the Bad and the Ugly. In: CVPR. pp. 4582–4591 (2017)
32. Zhang, M., Liu, Y., Luan, H., Sun, M., Izuha, T., Hao, J.: Building Earth Mover's Distance on Bilingual Word Embeddings for Machine Translation. In: AAAI. pp. 2870–2876 (2016)