ExSiM: Explainable Methodology to Upgrade Sentence Similarity Metrics to Document-Level

Matthew "Hugh" C. Williams Jr., Shubhra Kanti Karmaker ("Santu")
Department of Computer Science & Software Engineering
Auburn University, Alabama, USA
{mcw0097,sks0086}@auburn.edu

Abstract

Document similarity metrics tend to be black boxes. This poses a challenge for determining how to improve these metrics alongside what the correct use cases are for them. To open a new direction that will alleviate these issues, this paper introduces a methodology that can take sentence similarity metrics and expand them into document similarity metrics. This is also done in a general way to allow other such transformations, like paragraph to document. This is achieved through an analytic intuition-based methodology for constructing document similarity metrics inspired by how humans read texts. Thus, we hope to allow greater explainability for document similarity metrics while also paving the way for further improvements in the domain.

1 Introduction

Finding the similarity of two documents is an important task. The primary use that will be discussed here is in unsupervised grading of AI generated documents, but there are other use cases like searching over large datasets (Gutiérrez-Soto et al., 2019).

Typically, such metrics rely on encoders with some simple analytic methodology tacked on, see contemporary practices and (Gahman and Elangovan, 2023). These merely output a normalized score, but in this paper we propose a metric that can dot this while offering more explainability . Furthermore, our proposed metric is non-commutative which is critical in cases when comparing documents while knowing one of them to be correct, i.e., unsupervised grading.

Typical metrics also possess other shortcomings that are solved with this metric. For one, popular encoders like BERT possess token limits (Devlin et al., 2019a) which make them unusable on long documents out of the box. Furthermore, BERT was not trained to generate embeddings whose cosine similiarities correlate well with human intuitions (Devlin et al., 2019b), and so many methods have been proposed to solve this (Reimers and Gurevych, 2019) (Zhang et al., 2020).

Our metric responds to both these flaws by, one, greatly fleshing out the analytic portion of the metric to allow arbitrarily long documents in an explainable manner and, two, using encoders where they function well, on lower-level documents like sentences.

2 Proposed Metric

ExSiM (Explainable Similarity Metric) is an analytic methodology that upgrades a similarity metric for a given level (sentence, paragraph, etcetera) and upgrades it into the next higher level, whatever that is determined to be. This paper will focus on ExSiM being used to upgrade highly optimized sentence-level similarity metrics into document-level metrics. The documents within the used datasets are a few paragraphs at most, so in essence this method is being used to upgrade *sentence similarity metrics* to paragraph-level.

In our fleshed out analytic approach, we suggest matching sentences and using this information in a way that mimics the general human reader. We have driven to construct a methodology that accomplishes this as intuitively as possible by imagining our similarity metric as measuring the ease of narratively transforming one document to another. This makes our metric non-commutative.

This process of transformation is done by segmenting the documents into semantic pieces, matching these pieces, then trying to recreate the flow between corresponding pieces from one document in another.

2.1 Explainability

The methodology of ExSiM is such that machine learning neophytes can understand, which in itself poses great benefits. But another surprising outcome is the ability to derive new metrics besides

	Wikipedia	
	Synthetic	Handpicked
Deberta E_D	73.0%	80.3%
Roberta E_D	76.0%	84.2%
Avg. SBERT E_S	76.2%	92.1%
Avg. MiniLM E_S	77.1%	**94.0%**
Avg. GloVe E_W	70.7%	88.7%
ExSiM(Mini LM E_S)	**77.7%**	91.4%

Table 1: Wikipedia Dataset Results

	Correlation with Human Annotations	
	Similarity	Ordering
Deberta E_D	0.632	0.589
Avg. SBERT E_S	0.526	0.547
Avg. MiniLM E_S	0.537	0.558
Avg. GloVe E_W	0.537	0.537
ExSiM(Mini LM E_S) (Commutative)	0.621	0.579
ExSiM(Mini LM E_S) (Non-Commutative)	**0.768**	**0.716**

Table 2: Kendeall Tau Human Annotation Results

mere similarity. We were able to come up with these: seeing how similarity in between two sets of documents varies along the length of individual documents, the tendency for sentences to fuse or split between documents, and use of ExSiM as an ordering metric.

3 Experimentation

We seek to show that our ExSiM correlates well with human intuitions on the similarity between documents. Furthermore, *our metric uses completely unoptimized hyper-parameters* so as to verify we were comparing apples to apples.

As for the terminology used in our results, we use E_D, E_S, and E_W to mean document, sentence, and word encoders respectively. For example, "ExSiM(MiniLM E_S)" means the MiniLM sentence encoder upgraded by ExSiM into a document similarity metric (DSM).

3.1 Results

3.1.1 Wikipedia Triplets

Wikipedia Triplets is a dataset created to test DSMs (Dai et al., 2015). The rows are triplets of Wikipedia links such that the first two articles are more similar than the second is to the third. Looking at table 1, it lists the accuracy of each respective models' generated similarity scores. A model's similarity scores are computed between the first two documents and the last two in each triplet, and the model is correct if the score between the first two documents is higher than between the last two.

One can see that ExSiM achieves comparable performance with state-of-the-art DSMs. It edges out all the other metrics on the synthetic while coming short of two averaged sentence encoders on the handpicked. We argue these results show great promise because ExSiM has superior explainability which allows other metrics to be derived.

3.1.2 Human Annotations

For *Human Annotations*, we tasked five Auburn University undergraduate students with annotating 20 article pairs as sourced and modified from CNN Daily Mail data set as provided by (See et al., 2017) and (Hermann et al., 2015). Each of five documents were modified by four models, then the students were asked to compare the generated documents to their correct originals, both in terms of general similarity and similarity of the ordering of sentences.

Looking at table 2, one can see the correlation between any given models' similarity scores and the human annotated ground truths. These correlations were computed using Kendall Tau. For a table that uses the Spearman ranking correlation metric instead, see table 3 in appendix A.

Here, our ExSiM boasts better performance and this is because the task is a non-commutative one wherein the correct document is treated differently than the generated document its being compared to. This pans out through the high variance in correlation between our commuative and non-commutative models. Clearly the non-commutative models correlate better with human intuitions than any other model, especially on the question of how well documents were ordered.

4 Conclusion

We believe ExSiM shows great promise as a timeless methodology that will allow the upgrading of contemporary and future semantic similarity metrics. Explainability is critically important, especially in an age rife with black boxes. Understanding what we use not only allows methodical improvement, but is superior when in actual use because of the trust and intuition it engenders.

References

Andrew M. Dai, Christopher Olah, and Quoc V. Le. 2015. Document embedding with paragraph vectors.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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.

Nicholas Gahman and Vinayak Elangovan. 2023. A comparison of document similarity algorithms.

Claudio Gutiérrez-Soto, Arturo Curiel Díaz, and Gilles Hubert. 2019. Comparing the effectiveness of query-document clusterings using the qdsm and cosine similarity. In *2019 38th International Conference of the Chilean Computer Science Society (SCCC)*, pages 1–8.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A Human Annotations Spearman Scores

See table 3, the results are similar and warrant the same analysis.

	Correlation with Human Annotations	
	Similarity	Ordering
Deberta E_D	0.700	0.677
Avg. SBERT E_S	0.593	0.586
Avg. MiniLM E_S	0.622	0.623
Avg. GloVe E_W	0.641	0.606
ExSiM(Mini LM E_S) (Commutative)	0.691	0.671
ExSiM(Mini LM E_S) (Non-Commutative)	**0.864**	**0.830**

Table 3: Spearman Human Annotation Results
* matches between concatenations disallowed