Sentence Similarity Measures Revisited: Ranking Sentences in Pubmed Documents

Qingyu Chen*, Sun Kim*, W. John Wilbur and Zhiyong Lu†
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health
8600 Rockville Pike, Bethesda, MD, USA
{qingyu.chen, sun.kim, john.wilbur, zhiyong.lu}@nih.gov

ABSTRACT

While various measures are available for computing sentence similarity, few studies have examined their performance in the biomedical domain. Motivated by BIOSSES, an earlier study for biomedical sentence similarity, we here explore the effectiveness of multiple similarity measures via sentence ranking in PubMed abstracts. Ranking sentences is a crucial component for text summarization and biocuration evidence attribution. Applied to the "natural language processing" and "computational biology" datasets, our experimental results show that the off-the-shelf measures for sentence similarity may not be effective for ranking sentences. Neither lexical nor semantic measures provided more than 0.60 NDCG scores at the top 1 ranked document. It necessitates the development of a large-scale benchmark set and more effective measures.

CCS CONCEPTS

• Computing methodologies → Natural language processing • Applied computing → Health informatics • Information systems → Retrieval effectiveness

KEYWORDS

Natural language processing; biomedicine; textual similarity

ACM Reference format:
Q. Chen, S. Kim, J. Wilbur, Z. Lu. Sentence similarity measures revisited: ranking sentences in PubMed documents. In ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, *August 29-September 1, 2018, Washington, DC, USA.*, ACM, NY, NY, USA 2 pages. https://doi.org/10.1145/3233547.3233640

1 MOTIVATION

Information retrieval (IR) plays a crucial role given the rapid growth of scientific literature. Sentences, the intermediate blocks in the word-sentence-paragraph-document hierarchy, are vital to represent the semantics of an article [1]. Accordingly, computing the similarity between sentences is a common practice in IR applications [2]. The applications using sentence similarity include question-answer retrieval (e.g., find best candidate sentences for a question) [3], text summarization (e.g., summarize a document based on its key sentences) [4] and citation-reference linkage (e.g., find the most relevant sentences in a reference given a sentence that cites the reference) [5].

Sentence similarity is also important in the biomedical domain. For instance, evidence sentence retrieval, a task of searching and ranking sentences in biomedical literature based on user-defined biological expressions is essential in database curation [6]. Other cases that use sentence-level features include sentence/document classification [7, 8], question-answer retrieval [9] and text summarization [10].

Unlike a more general genre, biomedical sentences are often lengthy and grammatically complex. For example, bio-entities tend to be ambiguous and may have many variant names; sentences may also contain complex relations such as protein-protein interactions and characterization of gene products. Moreover, core ideas may be expressed in multiple sentences, which is challenging for biocurators to produce evidence attribution (link the most relevant sentence to a biological annotation) [11]. To date, few studies have tackled sentence similarity in the biomedical domain. One pilot study, proposed a model titled BIOSSES [12] and examined a range of string similarity measures for biomedical sentences, such as q-gram, Jaccard, and paragraph vector-based approaches. It evaluated the performance of different measures on a 100-pair dataset where sentence pairs are manually annotated in one of five different relevance levels from entirely different (score 0) to semantically equivalent (score 4). The best results showed over 0.80 correlation to manual annotations. However, the BIOSSES study treats the five different relevance levels equally, i.e. ignoring the fact that high scoring relevance is more important. As mentioned in [6, 11], a good sentence similarity measure should model human relevance judgement such that it ranks the most similar sentences at a higher rank given a query sentence. The top-ranked sentences, in turn, can benefit biological users: for example, biocurators could search annotation statements (represented as sentences) to find most similar sentences, and select the appropriate ones for evidence attribution without manually investigating every sentence.

*Contributed equally
†Corresponding author

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA
© 2018 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-5794-4/18/08.
https://doi.org/10.1145/3233547.3233640

2 METHOD AND RESULTS

In this work, we measure the effectiveness of sentence similarity metrics by ranking sentences from PubMed documents. We examine whether the similarity measures can find relevant sentences and give them a high rank in two domains, “Natural Language Processing” (NLP) and “Computational Biology/Education” (CB). We chose these topics for variety since the BIOSSES study focuses on the biological domain.

We collected the PubMed documents that belong to two MeSH terms, *natural language processing* and *computational biology/education.* Twenty PubMed abstracts were then randomly chosen from each topic. The selected articles were restricted that they must contain at least ten sentences in the abstracts for evaluation purposes. For each article, we measured the similarity between its title and the abstract sentences using six high performance similarity measures [12]: cosine similarity, block (or Manhattan) distance, q-gram (or n-gram) similarity, Jaccard similarity, Levenshtein distance and paragraph vector similarity. The first five metrics are string-based that measure the sentence similarity at the character or word level. The paragraph vector metric extends the string-based metrics by converting sentences into vectors, and in turn the sentence similarity is computed based on the similarity between the vectors. We used existing implementations provided by BIOSSES [12]. For each abstract the sentences were annotated manually in terms of their relevance to the title on a five-point scale (1 to 5), from an IR perspective, i.e. whether the title and the sentence are closely related (for example, the sentence has similar meanings to the title or it provides extended information which users may be interested in). The sentences in the abstract were ranked based on their similarity to the title in descending order using each BIOSSES metric. The ranking performance was measured using NDCG (normalized discounted cumulative gain) [13] at top positions 1, 3, and 5.

The average NDCG scores are presented in Table 1. The scores in the table are distinctive compared to the results from BIOSSES [12], which indicates that high-performing measures in BIOSSES do not apply in our task. In particular, our results show that no similarity measures perform well on finding the most relevant sentences; e.g., the highest NDCG score at position 1 for the NLP set is only 0.60. The performance for the CB set is worse: the highest NDCG score is only 0.52. This suggests that simply comparing word overlaps at the lexical level or semantic relatedness at the sentence level is not an effective means for the similar sentence search in our experiments.

Interestingly, the paragraph vector score is lower than most other measures in this task. It may be due to our task being intra-document search, i.e. the title and the abstract are from the same article. Authors tend to use the same words/terminologies in this case, hence string-based metrics show better performance, especially for the metrics that favor shared words in two sentences. Another factor may be that paragraph vectors were trained and optimized on PubMed Central full-text articles.

Our study has a few limitations. We assume that titles are descriptive enough and abstracts are an extended version of titles. While this might be true for most of the articles, it is difficult to judge title-abstract sentence pairs from documents with short titles. Future work includes exploring extra-document search for the same task and developing more effective similarity measures for sentence search.

	NLP			CB		
	NDCG@1	@3	@5	@1	@3	@5
Cosine	0.59	0.67	0.72	0.48	0.65	0.72
Block	0.59	0.67	0.72	0.43	0.60	0.69
q-gram	0.46	0.61	0.68	0.51	**0.70**	0.76
Jaccard	**0.60**	0.68	0.73	0.47	0.64	0.71
Leven	0.25	0.40	0.52	0.45	0.50	0.55
Overlap	0.59	**0.70**	**0.75**	**0.52**	0.68	**0.78**
Paragraph	0.43	0.59	0.63	0.37	0.56	0.66

Table 1. Average NDCG scores at different ranks.

ACKNOWLEDGMENTS

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

REFERENCES

[1] Le Q, Mikolov T. Distributed representations of sentences and documents. International Conference on Machine Learning 2014:1188-96.

[2] Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 Task 1: Semantic Textual Similarity-Multilingual and Cross-lingual Focused Evaluation. arXiv preprint arXiv:1708.00055 2017.

[3] Severyn A, Moschitti A. Learning to rank short text pairs with convolutional deep neural networks. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval 2015:373-82.

[4] Tan J, Kotov A, Pir Mohammadiani R, Huo Y. Sentence Retrieval with Sentiment-specific Topical Anchoring for Review Summarization. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management 2017:2323-6.

[5] Nomoto T. NEAL: A neurally enhanced approach to linking citation and reference. Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) 2016:168-74.

[6] Rastegar-Mojarad M, Komandur Elayavilli R, Liu H. BELTracker: evidence sentence retrieval for BEL statements. Database 2016;2016.

[7] Kim S, Kim W, Comeau D, Wilbur WJ. Classifying gene sentences in biomedical literature by combining high-precision gene identifiers. Proceedings of the 2012 Workshop on Biomedical Natural Language Processing 2012:185-92.

[8] Chen Q, Panyam NC, Elangovan A, Davis M, Verspoor K. Document Triage and Relation Extraction for Protein-Protein Interactions affected by Mutations. Proceedings of the BioCreative VI Workshop 2017;6:52.1.

[9] Sarrouti M, El Alaoui SO. A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering. Journal of biomedical informatics 2017;68:96-103.

[10] Chandu K, Naik A, Chandrasekar A, Yang Z, Gupta N, Nyberg E. Tackling Biomedical Text Summarization: OAQA at BioASQ 5B. BioNLP 2017 2017:58-66.

[11] Hirschman L, Burns GA, Krallinger M, Arighi C, Cohen KB, Valencia A, et al. Text mining for the biocuration workflow. Database 2012;2012.

[12] Soğancıoğlu G, Öztürk H, Özgür A. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics 2017;33:i49-i58.

[13] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) 2002;20:422-46.