NLP-IISERB@Simpletext2022: To Explore the Performance of BM25 and Transformer Based Frameworks for Automatic Simplification of Scientific Texts

Sourav Saha[1], Dwaipayan Roy[2], B Yuvaraj Goud[3], Chethan S Reddy[3] and Tanmay Basu[3]

[1]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute Kolkata, India

[2]Department of Computational and Data Sciences, Indian Institute of Science Education and Research Kolkata, India

[3]Department of Data Science and Engineering, Indian Institute of Science Education and Research Bhopal, India

Abstract

CLEF SimpleText 2022 lab focuses on developing effective systems to identify relevant passages from a given set of scientific articles. The lab has organized three tasks this year. Task 1 is focused on passage retrieval from the given data for a query text. These passages can be complex and hence require further simplification to be carried out in tasks 2 and 3. The BioNLP research group at the Indian Institute of Science Education and Research Bhopal (IISERB) in collaboration with two different information retrieval research groups at IISER Kolkata and ISI Kolkata participated only in Task 1 of this challenge and submitted three runs using three different retrieval models. The paper explores the performance of these retrieval models for the given task. We used a standard BM25 model as our first run to identify 1000 relevant passages for each query. Moreover, the passages for each query were ranked based on their similarity scores generated by the BM25 model. For our second run, we used a BERT (Bidirectional Encoder Representations from Transformers) based re-ranking method, called as Mono-BERT to further rank the 1000 passages retrieved by our first run for each query. A pre-trained sequence to sequence model based re-ranking method, called MonoT5 was used as our third run to reorder the 1000 passages retrieved by the Mono-BERT model for each query. As the official results of this task are not yet announced, we cannot explore the performance of our submissions. However, we have manually checked the retrieved results of many queries for each run, which indicate that the performance improved from run 1 to run 2 and further to run 3.

Keywords

information retrieval, text simplification

1. Introduction

Scientific articles are often hard to understand as they require significant background knowledge of certain areas and use tricky terminology. Thus automatic simplification of scientific

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 05–08, 2022, Bologna, Italy

souravsaha.juit@gmail.com (S. Saha); dwaipayan.roy@iiserkol.ac.in (D. Roy); byuvaraj19@iiserb.ac.in (B. Y. Goud); chethan19@iiserb.ac.in (C. S. Reddy); welcometanmay@gmail.com (T. Basu)
https://dwaipayanroy.github.io/ (D. Roy); https://sites.google.com/view/tanmaybasu/ (T. Basu)
0000-0001-9536-8075 (T. Basu)

text is a challenge although there have been some recent efforts in this direction [1]. The SimpleText shared-task at CLEF 2022 is an initiative in this spirit [1]. The goal is to create a concise summary of several scientific documents for a given query which provides a synopsis regarding a specific topic. The classical text retrieval methods are typically designed to identify whole documents that are relevant to a given query [2]. In these techniques, the relevance between a document and the given query is determined based on the presence of query words and phrases in the documents regardless of the location or proximity of them within the document [2, 3]. As the classical information retrieval methods identify whole documents for a query instead of a concise summary of those documents, the same may not be useful for text simplification. An alternative approach is to consider each document as a set of passages and to compute the similarity of each passage to the query [2, 4, 5], where a passage is a contiguous block of text in the original document. The passages retrieved in this way from different documents can then be ranked based on the order of relevance with the query.

Classical document similarity methods such as BM25 or language model can be employed for passage retrieval methods as well where the similarity between a query and a passage can be computed based on lexical matching and frequency of common terms in query and passages following a bag of words (BOW) approach [3, 6, 7, 8]. However, this type of BOW approaches may suffer from the traditional vocabulary mismatch problem due to the use of different vocabularies by the query creator and the passage writer. Further, these models cannot capture the semantic similarity between a passage and a query, e.g., those passages that have few common terms but have a set of terms with the same meaning as the query will get a low similarity score and hence, they will be ranked lower in the ranked list [6].

As an advancement of the model, embedding based models have become popular in the past few years where significant improvements are observed by embedding based models over the BOW models [9]. In the recent years, the transformer based methods leverage the contextualized representation produced by deep language models to identify semantic relevance between a query and a passage such as BERT (Bidirectional Encoder Representations) [10, 11]. A recent model developed in this spirit is the Mono-BERT model [12], which takes query-passage pairs as the input of BERT and the similarity scores are computed based on the contextualized token representation [6]. The Mono-BERT and its variants [12, 11, 6] have demonstrated better performance than BOW based models on passage ranking tasks.

The SimpleText lab at CLEF 2022 has organized three shared tasks this year regarding scientific text simplification. The first task is focused on passage retrieval from the given corpus for a query text. These passages can be complex and hence require further simplification to be carried out in the scond and third tasks. However, we, the BioNLP research group at the Indian Institute of Science Education and Research Bhopal (IISERB) in collaboration with two different information retrieval research groups at IISER Kolkata and ISI Kolkata participated only in Task 1 of this challenge and submitted three runs using three different retrieval models. The organizers released the corpus and a set of queries for task 1. We have submitted three runs for task 1. The standard BM25 model [7] was used as our first run to identify 1000 relevant passages for each query. The passages retrieved for each query were ranked based on their similarity scores generated by the BM25 model. The Mono-BERT model [12] was used as the second run, which further re-ranks the 1000 passages retrieved by our first run for each query. As a third run, we used the MonoT5 model [11], which is a pre-trained sequence to sequence model based re-ranking method. This model reorder the 1000 passages retrieved by the Mono-BERT model for each query and returned the 100 best passages based on the similarity score. As the official results of this task are not yet announced, we cannot explore or compare the performance of our submissions. However, we have manually checked the retrieved results for various queries for each run which indicate the performance improvement over run 1 by run 2 and further by the run 3 where MonoT5 based model was used.

The paper is organized as follows. The runs submitted by our team for the first task is described in section 2. Section 3 presents the analysis of experimental results. Eventually we conclude with scopes of further works in section 4.

2. Proposed Frameworks

We employ a multistage [11] ranking pipeline for this task. As our first run, BM25 model is used to rank a set of top k documents from the collection. The Mono-BERT model is utilized to re-rank the k top passages returned by the BM25 model, which is submitted as our second run. The Mono-BERT model [12] encodes query and the documents with BERT [13] based language model. In Mono-BERT model, BERT is used as a binary classification model, where the output of the CLS token is passed over to a feed-forward neural network to obtain a probability score for each query-passage pair. Furthermore, the MonoT5 model is implemented as our third run over the ranked passages returned by the Mono-BERT model for each query. MonoT5 is based on the architecture of T5 [14]which is a sequence-to-sequence model that uses a similar masked language modeling objective as BERT to pre-train its encoder–decoder architecture [15]. As we do not have explicit relevance information for query-passage pairs, the pre-trained version of both the models and used here.

3. Experimental Evaluation

3.1. Dataset

The organizers released a DBLP corpus in JSON file format of size 3 GB. The JSON file of the corpus contains 4894063 entries for different publications. Each entry contain the detail information of a publication i.e., title, abstract, authors name, year of publication, publisher name, citation etc. The abstracts of individual publications were considered as individuals documents. The queries were released in a CSV file, where there are 114 unique queries. The query id and link from where the query was collected were also given in that file. The objective was to identify relevant passages for each query from the documents of the corpus.

3.2. Experimental Setup

The collection was indexed with Apache Lucene[1]. The standard analyzer[2] with the default stopword list was employed. Queries were formulated by appending both the topic and query text. For implementing the BM25 model in the first run, k i.e., the number of passages retrieved per query was set to 1000. BM25 parameters, specifically k_1 and b, respectively controlling the term frequency (tf) scaling and the document length normalization, were set to 1.2 and 0.75. All the other parameters were set to their default values.

Table 1
Performance of Different Teams for Task 1

Team	#Topics	Avg #Doc.	NDCG		
			5	10	20
CYUT	114	4.9	0.5866	0.5636	0.5536
UAMS	114	95.5	0.3531	0.3776	0.4073
UAMS-MF⋆	69	2.7	0.3494	0.3328	0.3270
NLP@IISERB 1	30	92.5	0.0605	0.0680	0.0819
NLP@IISERB 2	114	100	0.0503	0.0640	0.0815
NLP@IISERB 3	114	100	0.0467	0.0522	0.0722

⋆ *Manual run.*

3.3. Analysis of Results

The performance of all the teams participated in task 1 are reported in Table 1 in terms of normalized discounted cumulative gain (NDCG) [16]. It can be seen from Table 1 that none of our runs achieve a place among the top three runs of task 1. None of our runs implemented the text preprocesing techniques like stemming or lemmatization before generating the inverted index. This may be one of the reasons of poor performance. We used the default parameters of the Mono-BERT and Mono-T5 models and did not tune different relevant parameters of these models, which may lead to poor performance. In future, we will focus on addressing these limitations of the proposed approaches.

4. Conclusion

The objective of SimpleText lab at CLEF 2022 is to involve the researchers to develop models to generate simplified summary of scientific literature for a given query. We have submitted three runs for the first shared task of the lab. The runs comprised of the classical BOW based BM25 models and transformer based Mono-BERT and MonoT5 models to further improve the performance of BM25 model. The transformer based re-ranking methods have been widely used for the last few years. Hence we used these models to rearnk the passages retrieved by the BM25 model for each given query. However, we could not achieve good performance as

[1] https://lucene.apache.org/

[2] https://lucene.apache.org/core/8_8_1/core/org/apache/lucene/analysis/standard/StandardAnalyzer.html

we already mentioned and the same needs to be addressed in future. Moreover, the aim is to explore the performance of other transformer based method for the given task that have been widely used for other applications.

References

[1] L. Ermakova, P. Bellot, J. Kamps, D. Nurbakova, I. Ovchinnikova, E. SanJuan, E. Mathurin, S. Araújo, R. Hannachi, S. Huet, et al., Automatic simplification of scientific texts: Simpletext lab at clef-2022, in: European Conference on Information Retrieval, Springer, 2022, pp. 364–373.

[2] M. Kaszkiel, J. Zobel, Passage retrieval revisited, in: ACM SIGIR Forum, volume 31, ACM New York, NY, USA, 1997, pp. 178–185.

[3] C. D. Manning, P. Raghavan, H. Schutze, Introduction to Information Retrieval, Cambridge University Press, New York, 2008.

[4] X. Liu, W. B. Croft, Passage retrieval based on language models, in: Proceedings of the eleventh international conference on Information and knowledge management, 2002, pp. 375–382.

[5] A. Mallia, O. Khattab, T. Suel, N. Tonellotto, Learning passage impacts for inverted indexes, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1723–1727.

[6] S. Zhuang, G. Zuccon, Tilde: Term independent likelihood model for passage re-ranking, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1483–1492.

[7] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Found. Trends Inf. Retr. 3 (2009) 333–389. URL: https://doi.org/10.1561/1500000019. doi:10.1561/1500000019.

[8] F. Song, W. B. Croft, A general language model for information retrieval, in: Proceedings of the Eighth International Conference on Information and Knowledge Management, CIKM '99, Association for Computing Machinery, New York, NY, USA, 1999, p. 316–321. URL: https://doi.org/10.1145/319950.320022. doi:10.1145/319950.320022.

[9] K. D. Onal, Y. Zhang, I. S. Altingovde, M. M. Rahman, P. Karagoz, A. Braylan, B. Dang, H. Chang, H. Kim, Q. McNamara, A. Angert, E. Banner, V. Khetan, T. McDonnell, A. T. Nguyen, D. Xu, B. C. Wallace, M. de Rijke, M. Lease, Neural information retrieval: at the end of the early years, Inf. Retr. J. 21 (2018) 111–182. URL: https://doi.org/10.1007/s10791-017-9321-y. doi:10.1007/s10791-017-9321-y.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies„ Minnesota, USA, 2019, pp. 4171–4186.

[11] R. Nogueira, W. Yang, K. Cho, J. J. Lin, Multi-stage document ranking with bert, ArXiv abs/1910.14424 (2019).

[12] R. Nogueira, K. Cho, Passage re-ranking with BERT, CoRR abs/1901.04085 (2019). URL: http://arxiv.org/abs/1901.04085.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[15] R. Nogueira, Z. Jiang, J. Lin, Document ranking with a pretrained sequence-to-sequence model, arXiv preprint arXiv:2003.06713 (2020).

[16] Y. Wang, L. Wang, Y. Li, D. He, T.-Y. Liu, A theoretical analysis of ndcg type ranking measures, in: Conference on learning theory, PMLR, 2013, pp. 25–54.