Retrieving Scientific Abstracts using Venue- and Concept-based Approaches: CincyMedIR at TREC 2019 Precision Medicine Track

Danny T.Y. Wu, PhD, MSI[1], Wu-Chen Su, MS[1], James J. Lee, PhD[2]
[1]Deparmtnet of Biomedical Informatics, University of Cincinnati, Cincinnati, OH; [2]Digital Scholarship Center, University of Cincinnati, Cincinnati, OH

Introduction

The CincyMedIR group led by Dr. Danny T.Y. Wu at the University of Cincinnati (UC) College of Medicine participated in the Text Retrieval Conference 2019 Precision Medicine Track (TREC-PM). Dr. Wu was part of the MedIER group in TREC 2015, 2017, and 2018, and formed his own group this year. CincyMedIR only worked on the scientific abstracts but not clinical trial documents this year.

Method

All scientific abstracts were downloaded from the TREC-PM and indexed using Elasticsearch on Amazon Web Services. The retrieval pipeline was simplified based on our previous approach[1]. Specifically, we did not include ML-based re-ranking or iterative re-retrieval since these approaches did not dramatically improve the system performance. Instead, we developed a venue-based and a concept-based approach to re-rank the documents. For the venue-based approach, we obtained a list of journal titles from the released result in TREC-PM 2018. Documents with a journal title on this list were moved to the top of the retrieval results. For the concept-based approach, we extracted the medical concepts using MetaMapLite. Similarly, documents with the highest concept matches were moved to the top of retrieval results. We used a threshold (i.e., 300) based on experiments to decide whether each topic had sufficient records. If not, we used the maximum records obtained for the evaluation. We then conducted the experiment with different parameters (venue- or concept-based approach or both) and the ranking algorithm (Okapi BM25 or LM-Drichlet). Based on the results of TREC-PM 2018, our system was able to retrieve competitive results in the cancer specialty without relying on any advanced retrieval mechanisms.

Results

Table 1 shows the evaluation results. The evaluation scores were close in all runs. Specifically, MedIR3 performed the best since its scores in infNDCG and R-prec were the highest and P@10 was comparable. The overall results show that a concept-based approach can improve infNDCG and P@10. Applying the venue-based approach does not seem to help much. Merging the results of BM25 and LM improved P@10, but decreased infNDCG and R-prec slightly.

Table 1. Summary of submitted run result of TREC-PM 2019

RUN	Algorithm	Venue-based	Concept-based	infNDCG	P @ 10	R-prec
MedIR1	BM25 and LM	X	X	0.4735	**0.5675**	0.2744
MedIR2	BM25	X	X	0.4674	0.5650	0.2737
MedIR3	BM25		X	**0.4801**	0.5600	**0.3111**
MedIR4	BM25	X		0.4430	0.5250	0.2710
MedIR5	BM25			0.4534	0.5125	0.3056

Conclusion

Using the venue- and concept-based approaches and the baseline ranking algorithms, we were able to achieve competitive results . We are eager to learn the techniques of the top teams in TREC-PM 2019 to enhance the performance of our system for next year.

Acknowledgement

This project was partially funded by the Andrew W. Mellon foundation through the grant received by the Digital Scholarship Center (DSC) at the UC Libraries. Dr. James Lee is one of the principal investigators (co-PI) of the grant and the director of the DSC.

References

1. Jinghui Liu, Clair Kronk, Wu-Chen Su, Danny TY Wu, and VG Vinod Vydis- waran. Retrieving scientific abstracts iteratively: Medier at trec 2018 precision medicine track. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Text REtrieval Conference, TREC*, November 2018