Master Thesis Proposal

Faculty of Science
Radboud University Nijmegen

Date: 26-03-2020

Author:	**Tom Janssen Groesbeek**	Supervisor:	Prof.dr.ir. A.P. de Vries
E-mail:	tomjg@hotmail.nl	E-mail:	A.deVries@cs.ru.nl
Phone:	+31 6 55626118	Phone:	+31 24 3652354
Student nr.:	S4229738		
Study:	Computer Science		
Specialization:	Data Science		

Proposed Topic:

Re-ranking BERT; A Thorough Analysis on MS MARCO Passage Re-ranking with BERT

Topic Characteristics:
My thesis will focus on re-evaluating BERT on its performance on passage re-ranking using the MS MARCO dataset. Currently, the MS MARCO dataset contains only a few relevant passages per query, mostly one relevant passage per query. My hypothesis is that this is inaccurate and that in fact the dataset contains more relevant passages per query. Therefore, as part of my research, I will gather more query-passage pair relevancy assessments for the MS MARCO passage ranking dataset. With these additional labels I then plan to re-evaluate BERT on the task of passage re-ranking. Furthermore, I plan to perform a similar evaluation using graded relevance labels, to see if performance is affected when dealing with non-binary relevance labels. Finally, recent research has tested BERT's behaviour on specially constructed axiomatic datasets to test if BERT adheres to the information retrieval heuristics. For this thesis I would like to explore if similar results can be obtained when using the MS MARCO dataset with the newly acquired labels.

Research Questions:

1. Does the MS MARCO dataset contain more relevant passages per query than currently labelled?
2. Given that the MS MARCO dataset is updated with more relevant passages per query, how does this affect BERT performance on the MS MARCO passage re-ranking task?
3. What is BERT's performance on the MS MARCO passage re-ranking task when graded relevance labels are used?
4. Does BERT adhere to the information retrieval heuristics when utilizing the MS MARCO passage re-ranking dataset.
    a. How are the results affected when the updated MS MARCO dataset is utilized?

Methodology:
In order to gather more query-passage relevancy assessments, I set out to create an online assessment webpage. On this webpage the subjects who consented to help me with my research, will be able to read different queries with each a set of corresponding passages. They will then be able to assess those passages on their level of relevancy compared to the query. It will be possible to select a level of relevancy between 1 and 5, where 1 is totally irrelevant and 5 is perfectly relevant. Thus making it possible to gather graded relevancy labels for the MS MARCO dataset. Each query will need to be assessed by at least 3 assessors. In case of discrepancies in the answers by the 3 assessors, a majority voting will help to decide on the resulting label. The BM25 algorithm will be used to obtain an initial

passage ranking. After which BERT will be used to re-rank the passages in either the current MS MARCO dataset or an updated version including the newly collected relevancy labels. In order to evaluate BERT's performance on the passage re-ranking task, I will make use of the MRR@10 metric as this was used by the Microsoft organizers. Aside from this metric, I will also make use of the MAP and nDCG metric in order to deal with multiple (graded) relevancy labels. To be able to check if BERT adheres to the information retrieval heuristic, I will follow the paper by Camara and Hauff and construct diagnostic datasets from the MSMARCO data.

References:

Aslam, J. A., & Yilmaz, E. (2007). *Inferring Document Relevance from Incomplete Information*. 633–642.

Bailey, P., Craswell, N., De Vries, A. P., & Soboroff, I. (2007). Overview of the TREC 2007 enterprise track. *NIST Special Publication*.

Bailey, P., Thomas, P., Craswell, N., De Vries, A. P., Soboroff, I., & Yilmaz, E. (2008). Relevance assessment: Are judges exchangeable and does it matter? *ACM SIGIR 2008 - 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Proceedings*, 667–674. https://doi.org/10.1145/1390334.1390447

Borlund, P. (2003). The concept of relevance in information retrieval. *Journal of the American Society for Information Science and Technology*, *54*(10), 913–925. https://doi.org/http://dx.doi.org/10.1002/asi.10286

Camara, A. & Hauff, C. (2017). *Diagnosing BERT with Retrieval Heuristics*. 605–618. https://doi.org/10.1007/978-3-030-45439-5

Crijns, T. & de Vries, A. (2019). *Have a chat with BERT ; passage re-ranking using conversational context*.

De Beer, J., & Moens, M. F. (2006). Rpref - A generalization of bpref towards graded relevance judgments. *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *2006*, 637–638.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *Mlm*. http://arxiv.org/abs/1810.04805

Hazen, T. J., Dhuliawala, S., & Boies, D. (2019). *Towards Domain Adaptation from Limited Data for Question Answering Using Deep Neural Networks*. http://arxiv.org/abs/1911.02655

Ingale, V. & Singh, P. (2019). *Datasets for Machine Reading Comprehension: A Literature Review*.

Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, *53*(13), 1120–1129. https://doi.org/10.1002/asi.10137

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. *1*. http://arxiv.org/abs/1907.11692

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). MS MARCO: A human generated MAchine reading COmprehension dataset. *CEUR Workshop Proceedings*, *1773*(Nips 2016), 1–11.

Nogueira, R., & Cho, K. (2019). *Passage Re-ranking with BERT*. 1–5. http://arxiv.org/abs/1901.04085

Padigela, H., Zamani, H., & Croft, W. B. (2019). *Investigating the Successes and Failures of BERT for Passage Re-Ranking*. http://arxiv.org/abs/1905.01758

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). *Language Models as Knowledge Bases?* 2463–2473. https://doi.org/10.18653/v1/d19-1250

Robertson, S. E., Kanoulas, E., & Yilmaz, E. (2010). Extending average precision to graded relevance judgments. *SIGIR 2010 Proceedings - 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 603–610. https://doi.org/10.1145/1835449.1835550

Sakai, T. (2007). Alternatives to Bpref. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07*, 71–78. https://doi.org/10.1145/1277741.1277756

Sakai, T. (2007). On the reliability of information retrieval metrics based on graded relevance. *Information Processing and Management*, *43*(2), 531–548. https://doi.org/10.1016/j.ipm.2006.07.020

Sheetrit, E., Shtok, A., Kurland, O., & Shprincis, I. (2018). Testing the cluster hypothesis with focused and graded relevance judgments. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, *June 2018*, 1173–1176. https://doi.org/10.1145/3209978.3210120

Soboroff, I. (2014). Computing Confidence Intervals for Common IR Measures. *Proceedings of the 6th EVIA Workshop*, 6–9.

Sormunen, E. (2002). Liberal relevance criteria of TREC - Counting on negligible documents? *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 324–330.

Tenney, I., Das, D., & Pavlick, E. (2019). *BERT Rediscovers the Classical NLP Pipeline*. 4593–4601. https://doi.org/10.18653/v1/p19-1452

Turc, I., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Well-Read Students Learn Better: On the Importance of Pre-training Compact Models*. *Mlm*, 1–13. http://arxiv.org/abs/1908.08962

Van Aken, B., Löser, A., Winter, B., & Gers, F. A. (2019). How does BERT answer questions? A layer-wise analysis of transformer representations. *International Conference on Information and Knowledge Management, Proceedings*, 1823–1832. https://doi.org/10.1145/3357384.3358028

Weissenborn, D., Wiese, G., & Seiffe, L. (2017). Making neural QA as simple as possible but not simpler. *CoNLL 2017 - 21st Conference on Computational Natural Language Learning, Proceedings*, 271–280. https://doi.org/10.18653/v1/k17-1028

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2019). *BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis*. http://arxiv.org/abs/1904.02232

Yang, P., Fang, H., & Lin, J. (2018). Anserini: Reproducible ranking baselines using lucene. *Journal of Data and Information Quality*, *10*(4). https://doi.org/10.1145/3239571

Yang, W., Xie, Y., Tan, L., Xiong, K., Li, M., & Lin, J. (2019). *Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering*. http://arxiv.org/abs/1904.06652

Yilmaz, E., & Aslam, J. A. (2008). Estimating average precision when judgments are incomplete. In *Knowledge and Information Systems* (Vol. 16, Issue 2). https://doi.org/10.1007/s10115-007-0101-7

Yilmaz, E., Kanoulas, E., & Aslam, J. A. (2008). A simple and efficient sampling method for estimating AP and NDCG. *ACM SIGIR 2008 - 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Proceedings*, 603–610. https://doi.org/10.1145/1390334.1390437