Core IR Project Proposal

Aditya Kunar(5074274), Nikhil Saldanha(4998707), Sharwin Bobde()

20th March 2020

1 Problem description

We are interested in examining the problem of non-factoid answer passage retrieval on the WikiPassageQA collection as specified in the paper titled: "WikiPassageQA: A Benchmark Collection for Research on Non-factoid Answer Passage Retrieval" [1].

Our main goal is to reproduce the paper to the best of our ability and in doing so understand the most fundamental aspects of information retrieval with a substantial emphasis on the use of different relevance modelling strategies. We chose this paper because the collection is of the ultimate quality with a wide range of queries and long passages annotated with excellent workers with high inter-annotator agreement scores. Moreover, the task at hand is of sophisticated complexity and is extremely fascinating as present information needs require an information retrieval system to possess the ability to find relevant parts of a document and not solely the document itself.

2 Resources

As mentioned previously, the dataset we will be working with is the WikiPassageQA collection [1]. In addition, we will be using information retrieval toolkits such as Indri and Lemur for their built-in implementations when it comes to creating our own information retrieval models. Finally we will use the lecture videos from the course and the recommended reading material based on our needs and requirements for the project.

[1] https://ciir.cs.umass.edu/downloads/wikipassageqa

3 Methodology

Before starting with the implementation, we would like to get a deeper understanding of the data that we are dealing with. We also believe that the results from this analysis might aid us in creating a baseline model. The following are some of the studies we aim to conduct:

- Frequency counts of starting words of queries
- Frequency distribution of answer passage lengths

We may do more as we progress with the implementation.

From Cohen et.al. [1], we would like to reproduce the Benchmark results of different methods on WikiPassageQA dataset which the authors have presented in table 3. Although, we will limit ourselves to base and traditional IR rows of the table. We omit the neural baselines from our reproduction study as it may be too time consuming and we fear that we may not have the expertise to build complex Convolutional and Recurrent Neural Network Models in the short time frame available to us. To recompense, we implement two additional baselines:

- **Language Model** with a prior incorporating features of passages(maybe length, words common with query, etc; informed from the data analysis)
- **Learning to rank model** by using already extracted features, i.e. tf-idf, vector space model, BM25 and language model scores

4 Background readings

- Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval [6]
- A Language Modeling Approach to Information Retrieval [5]
- Learning to Rank for Information Retrieval [3]
- Benchmark for Complex Answer Retrieval. [4]
- The Importance of Prior Probabilities for Entry Page Search. [2]

5 Evaluation

The authors split the full dataset into training, validation and testing in the ratio: 0.8 : 0.1 : 0.1, partition on queries rather than articles as the authors suggest. We would like to keep this the same so our results can be comparable. Additionally, for the learning to rank models, we would like to create a cross-validation set from training + validation so as to be able to tune hyper-parameters.

As far as metrics are concerned, we stick with the ones that the authors have used, viz. MAP, MRR, P5, P10, nDCG, Recall5, Recall10, Recall20. The main reason for this is to allow for straightforward comparison between the authors' and our results.

Additionally, we would like to try and justify why some models perform better than others on specific metrics.

References

[1] Daniel Cohen, Liu Yang, and W. Croft. Wikipassageqa: A benchmark collection for research on non-factoid answer passage retrieval. pages 1165–1168, 06 2018.

[2] Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. The importance of prior probabilities for entry page search. 10 2002.

[3] Tie-Yan Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, March 2009.

[4] Federico Nanni, Bhaskar Mitra, Matt Magnusson, and Laura Dietz. Benchmark for complex answer retrieval, 2017.

[5] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 275–281, New York, NY, USA, 1998. Association for Computing Machinery.

[6] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Bruce W.

Croft and C. J. van Rijsbergen, editors, *SIGIR '94*, pages 232–241, London, 1994. Springer London.