Proximity-Based Entity Ranking

Gustavo Gonçalves[§†], João Magalhães[§], Jamie Callan[†]
ggoncalves@cmu.edu,jm.magalhaes@fct.unl.pt,callan@cmu.edu

[§]NOVA LINCS, Universidade NOVA de Lisboa, Caparica, Portugal
[†]Language and Technology Institute, Carnegie Mellon University, Pittsburgh, USA

ABSTRACT

This work explores the value of a combination of features, including entity proximity, for improving a learning-to-rank entity ranking system.

1 INTRODUCTION

Entities are informative units of text that represent real world objects, or concepts [1]. They can provide a shallow, yet more informative textual representation by aggregating multiple words, and different word representations in a single concept. E.g. *President Trump*, *Donald Trump*, and *Trump*. Entities can be linked to knowledge bases such as Wikipedia [5], DBPedia [8], Freebase[2], and YAGO [11]. With entity links, information systems are able to relate entities and expand the vocabulary related to a given entity. Moreover, these links can also be used to point users to relevant sources of related information. The latter scenario is the main scope of the TREC News, Entity Ranking task. The task is defined as the re-ranking of a given set of entities in a news article. The entity ranking should be ordered by decreasing order of utility, for the news article reader to better comprehend the article. With our participation in the 2019 edition of the TREC News track, we explored one main hypothesis. The inverted pyramid writing scheme [6] is commonly used in journalism, where the most important content of the news article is concentrated at the beginning of the document. We would like to verify if the entities that are mentioned earlier in the document, are central and thus should be ranked higher in the scope of the entity ranking task.

2 RELATED WORK

Entities have been established as useful units of information that can be used in information retrieval tasks. Recently, an increased interest has been given to identifying the salience and centrality [4] of entities in both queries and documents. We define centrality as the entity attribute that explicitly, or implicitly, defines the main topic of a document, or query. Salience refers to the importance of a query that might be related to the central entity, but not directly the main topic of the document or query.

Recently, various state-of-the-art methods [10, 12, 13] have leveraged on entity salience estimation to determine the importance of entities and decide on how to incorporate them in re-raking pipelines.

Machine learning approaches can be used to re-rank document ranking. These Learning To Rank (LTR) [9] systems, feed signals such as, term frequency, inverse document frequency and various rankings to learn to combine features and use them to learn a better ranking that optimizes a certain metric. Since LTR can combine various retrival models in one single ranking, they are considered a strong baseline for a search engine. In this work we adapt an LTR pipeline to consider features of entities contained in a given document, thus re-ranking entities and not documents, as in a tradition LTR system.

Table 1: Features used for LTR re-ranking.

Feature ID	Retrieval Model	Field
1	TF	Title
2	TF	First 5 Paragraphs
3	TF	Body
4	TF-IDF	Title
5	TF-IDF	First 5 Paragraphs
6	TF-IDF	Body
7	BM25	Title
8	BM25	First 5 Paragraphs
9	BM25	Body
10	PMI Window/20	Title
11	PMI Window/20	First 5 Paragraphs
12	PMI Window/20	Body
13	Nr. Surface Forms	Title
14	Nr. Surface Forms	First 5 Paragraphs
15	Nr. Surface Forms	Body

3 PROPOSED APPROACH

We have indexed the Washington Post dataset using separate, non-overlapping fields for the Title, First 5 Paragraphs, and Body fields. We employed an LTR architecture using various combinations of the features in Table 1, with a Coordinate Ascent [3] algorithm as optimization method.

3.1 Feature Selection and Re-ranking

We have submitted 3 runs for evaluation. The evaluated runs consisted of LTR models trained on the 2018 queries of the Washington Post News Track [7], entity ranking task. For training, we split the data in subsets of 40 queries for training and validation. Training models were generated by rotating these splits across the data, so that every query was used for training at least once. Each training roun was done with a 10-fold cross validation to reduce data bias and overfitting. Finally, the results obtained from each trained model were averaged to be comparable with the 2018 results. With our runs, we wanted to evaluate the retrieval impact of utilizing different document representations, such as the title, first 5 paragraphs, and body, under the hypothesis that central entities would

Table 2: Results obtained while training on 2018 and re-ranking on 2019 data

Run	MAP		nDCG@5		P@5	
TREC-Median	–	–	0.5727	–	–	–
CMU_NS-1-tpb	0.5736	–	0.5051	-11.1%	0.4000	–
CMU_NS-2-tp	0.5451	–	0.4567	-20.3%	0.3846	–
CMU_NS-3-t	0.5778	–	0.4782	-16.5%	0.3692	–

appear earlier in the total document span. We used a maximum of 15 features to train our models. Table 1 presents the utilized features, where each feature is a either a retrieval model, such as TF-IDF and BM25, or a count, as is the case for TF, PMI Window/20, and Nr. of Surface Forms. TF-IDF and BM25, use corpus statistics of the Washington Post corpus to calculate the entities IDF. PMI Window/20 stands for the count of each entity pair in the document, within an unordered window of 20 words. Whereas the Nr. of Surface Forms corresponds to the number of unique surface forms that each entity displays in the document.

We chose the following combinations of the features in Table 1 as our submitted runs:

1. CMU_NS-1-tpb: This run uses all features presented in Table 1. With this run we aimed to set an upper bound by combining all features across all considered fields, where tpb stands for title, paragraph, and body, respectively;
2. CMU_NS-2-tp: This run focus on using only the features that focus on the title and paragraph fields. The objective of this run was to examine the performance of adding the first 5 paragraphs of the document, to observe if the entities contained in this portion of text were central, and thus should be ranked higher;
3. CMU_NS-3-t: Finally, run 3 focus only on features based on the title field. With this run we aimed to set a lower bound of the retrieval performance that could be achieved with the chosen entities, and the smallest document field that is also expected to contain the most important entities to a news article.

4 EXPERIMENTAL RESULTS

Our systems' performance can be observed in Table 2. The first line of the table corresponds to the TREC Median result in terms of NDCG@5. Each following line of the table corresponds to one of our runs. The results are presented in terms of MAP, nDCG@5 which is the main metric of the task, and P@5. A percentual comparison is provided in terms of the nDCG@5 metric. Unfortunately, our runs were not successful in ranking the entities for each document, displaying significant losses when compare with the TREC median. This indicates that our set o features was not able to distinguish between central and important entities. However, we were able to verify with our runs that the inclusion of paragraph based features actually contributed to harming the results, thus not confirming the entity importance based on the inverted pyramid hypothesis. The other two runs based on all features, or just features based on the title field displayed an expected relative performance variation, where features combinations based on fields with less text tend to under-perform when compared to features that include more of the document's text.

5 CONCLUSION

Through our participation we were able to test one research hypothesis. We actually verified that our assumption that entities that occur in an earlier position of the document did not improve our model's performance. However, further research is necessary to determine in fact the effect of the entities position in the document providing a signal for relevance. Careful tuning of the proximity word windows in required to further assess the obtained results.

REFERENCES

[1] Krisztian Balog. 2018. *Entity-Oriented Search*. The Information Retrieval Series, Vol. 39. Springer. https://doi.org/10.1007/978-3-319-93935-3

[2] Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, Jason Tsong-Li Wang (Ed.). ACM, 1247–1250. https://doi.org/10.1145/1376616.1376746

[3] Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. 2005. Learning to Rank Using Gradient Descent. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005 (ACM International Conference Proceeding Series)*, Luc De Raedt and Stefan Wrobel (Eds.), Vol. 119. ACM, 89–96. https://doi.org/10.1145/1102351.1102363

[4] Jesse Dunietz and Daniel Gillick. 2014. A New Entity Salience Task with Millions of Training Examples. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, Gosse Bouma and Yannick Parmentier (Eds.). The Association for Computer Linguistics, 205–209.

[5] Paolo Ferragina and Ugo Scaiella. 2012. Fast and Accurate Annotation of Short Texts with Wikipedia Pages. *IEEE Software* 29, 1 (2012), 70–75. https://doi.org/10.1109/MS.2011.122

[6] Pottker Horst. 2003. News and Its Communicative Quality: The Inverted Pyramid—When and Why Did It Appear?. In *Journalism Studies*, Vol. 4. Routledge, 501–511. https://doi.org/10.1080/1461670032000136596

[7] Shudong Huang, Ian Soboroff, and Donna Harman. 2018. TREC 2018 News Track. In *Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval Co-Located with 40th European Conference on Information Retrieval (ECIR 2018), Grenoble, France, March 26, 2018 (CEUR Workshop Proceedings)*, Dyaa Albakour, David Corney, Julio Gonzalo, Miguel Martinez-Alvarez, Barbara Poblete, and Andreas Valochas (Eds.), Vol. 2079. CEUR-WS.org, 57–59.

[8] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195. https://doi.org/10.3233/SW-140134

[9] Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval* 3, 3 (2009), 225–331. https://doi.org/10.1561/1500000016

[10] Yuanyuan Qi, Jiayue Zhang, Weiran Xu, and Jun Guo. 2019. Finding Salient Context Based on Semantic Matching for Relevance Ranking. *arXiv:1909.01165 [cs]* (Sept. 2019). arXiv:cs/1909.01165

[11] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy (Eds.). ACM, 697–706. https://doi.org/10.1145/1242572.1242667

[12] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-Hoc Ranking with Kernel Pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 55–64. https://doi.org/10.1145/3077136.3080809

[13] Chenyan Xiong, Zhengzhong Liu, Jamie Callan, and Tie-Yan Liu. 2018. Towards Better Text Understanding and Retrieval through Kernel Entity Salience Modeling. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 575–584. https://doi.org/10.1145/3209978.3209982