Improving Ad Hoc Retrieval With Bag Of Entities

Gustavo Gonçalves
João Magalhães
NOVA LINCS
Universidade NOVA de Lisboa
Caparica, Portugal
gs.goncalves@campus.fct.unl.pt
jm.magalhaes@fct.unl.pt

Chenyan Xiong
Jamie Callan
Language and Technology Institute
Carnegie Mellon University
Pittsburgh, United States of America
cx@cmu.edu
callan@cmu.edu

ABSTRACT

This work explores the value of entity information for improving a feature-based learning-to-rank search engine.

1 INTRODUCTION

Recent research shows that combining a traditional Bag-of-Words (BoW) representation with an entity-oriented representation improves retrieval accuracy in web, medical, and academic search [1, 7–9]. News are very informative documents that convey a story by investigating the participating elements of the news, its location, and if possible why it occurred. To achieve this goal news tends to focus strongly on elements such as people, companies, and countries; thus our TREC Common Core submissions investigate whether entity-oriented features also improves retrieval accuracy in the news domain.

The BoW model on top of various retrieval models is a pragmatic approach that despite its excellent results fails to grasp textual context, and vocabulary nuances such as the "*44th President of the United States*" and "*Barack Hussein Obama*". These expressions both point to the same entity and should contribute identically for the detection of relevant documents, even though they do not contribute for the same collection statistics counters. With our work we explore the utilization of entities to bring semantically similar concepts closer in an entity space aiming to improve ad-hoc retrieval. Through our participation in the 2018 edition of the TREC Core track, we intended to test two main hypothesis. First, that the utilization of entities, when used in cooperation with the traditional BoW representation, would improve the retrieval results. Second, how does the entity representation deal with short fields of text, such as titles, in comparison with BoW model.

2 RELATED WORK

Entities have been used to improve the retrieval results by exploring the added semantic value and context provided by these elements. Different approaches have been explored to take advantage of the added value of the entities. These approaches consist in performing query expansion [5] with resort to the added entity terms; building a latent entity space [4], where the relevance between a query and a document is calculated through the projections in this latent space; and finally to represent queries and documents in an entity space [7]. The Bag-of-Entities (BoE) model introduced by Xiong et al. [7] focus on the idea of transposing the documents to an entity space. The entity space is defined, for both queries and documents, by entity linking systems. The combined utilization of both term-based and entity-based representations [8] have shown excellent improvements in retrieval results, where all possible combinations of query and document in a term-based and entity-based representation, were considered and combined.

Table 1: Document and relevance distribution, by both type of document and period.

Source	Years	# Docs	% Docs	# Rel. Docs	% Rel. Docs
Article	2012	44,367	7.46%	206	5.31%
	2013	48,759	8.19%	359	9.25%
	2014	44,648	7.50%	342	8.81%
	2015	38,709	6.50%	390	10.05%
	2016	38,266	6.43%	263	6.78%
	2017	21,898	3.68%	91	2.35%
Blog	2012	15,255	2.56%	40	1.03%
	2013	48,918	8.22%	226	5.82%
	2014	67,083	11.17%	358	9.23%
	2015	71,905	12.08%	570	14.69%
	2016	122,488	20.59%	865	22.29%
	2017	32,730	5.50%	170	4.38%
All		595,026	100.00%	3,380	100.00%

The study made by Xiong et al. [7] compared three different entity tagging systems, with different knowledge bases, such as Freebase and Wikipedia. From these entity tagging systems the TagMe [2] linking system provides a balance between precision and recall for entity identification, and was selected by us to perform the entity linking of both our queries and documents.

Learning To Rank (LTR) [3] systems use a machine learning approach, exploring textual features such as document length, term frequency, inverse document frequency and various rankings to learn the features and use them to learn a ranking that optimizes a certain metric such as Mean Average Precision (MAP). These systems are becoming a new standard for what is expected from an information retrieval system, providing more relevant and robust rankings. Therefore, LTR systems can be considered a strong baseline of what is expected from a recent search system, where new systems can build upon and use as an evaluation reference.

3 DATASET AND RELEVANCE ANALYSIS

The Washington Post news dataset was the elected collection for this edition of the core track. The recency of the adoption of this document collection required an analysis to determine what kind

Table 2: Collection Analysis

News Field Types	Frequency	Dataset Coverage	Avg. Length
kicker	594,527	99.91%	11
title	595,037	100.00%	61
image	409,421	68.81%	116
byline	545,659	91.71%	17
date	595,026	99.99%	N/A
sanitized html	592,341	99.55%	241
author info	449,711	75.58%	135
video	108,941	27.54%	77
gallery	44,814	7.53%	114
list	13,811	2.32%	N/A
deck	9,443	1.59%	73
tweet	56,412	9.48%	13
instagram	5,091	0.86%	114
pull quote	149	0.03%	128
inline story	22	0.00%	70
top deck	5	0.00%	60
animated video	1	0.00%	0
ar-wikitude	1	0.00%	228

of textual content was included in the documents, and what fields provided the most useful content for indexation.

A preliminary overview of the dataset was provided regarding the distribution of documents from each source and respective time period. With the data in Table 1 we can conclude that approximately 65% of the documents are between 2014 and 2016, with the majority of the documents in the collection being blog posts. There are 11 blog news that do not contain the *published date* field, thus the number of total documents is slightly different from the published one of 595037. After manually inspecting some documents we did not notice any significant difference between the article and blog documents. In general, blog documents are mostly opinion articles and tend to be shorter than article documents. However, this is not a rule since there are also documents classified as blog documents that were updated multiple times regarding an early lead news, matching the length of an article document.

With the released relevance judgments we analyzed the relevant documents distribution considering the type of document and time period. In Table 1 we can observe that 71.85% of the relevant documents are from 2014 to 2016, which approximately matches the overall document distribution. We can conclude that the participant systems contributed accordingly to the initial data distribution, which could indicate that few or even no temporal approaches were used. Moreover, we can also conclude that the assessors did not consider temporal recency as a major criteria for the assessments.

In Table 2 we can observe the elements available across every document of the collection, and their respective proportion in the dataset. Upon manual inspection we decided to index the fields that appear in bold, due their textual content and relative contribution towards the dataset constitution. The sanitized html field constitutes the majority of the documents, being sanitized html paragraphs from the news. Multimedia elements such as videos, images and galleries, contained small captions or blurbs that we decided to index, due to the small yet very informative descriptions. Finally, the text from the social media posts that was embedded in the news, and also scraped by the Washington Post, was also included since they were highly related to the news.

The data indexation was made by separating each document in its title and body. We have used the virtual document model [6], by treating the pieces of text contained in the multimedia captions and embedded social-media posts as part of the news; in opposition to considering only the body of the news, this is the sanitized html field shown in Table 2. Hence, the text contained in the highlighted fields in Table 2 were considered as part of the body of the news.

4 PROPOSED APPROACH

We started by creating a entity representation of both the corpus and the queries. The entities were indexed in analogous fields, one for the title entities and another for the extended body entities. Secondly, we employed an LTR architecture by re-ranking the results provided by a BM25 rank of 1000 documents, using various combinations of the features in Table 3.

4.1 Entity Linking

The entity linking was done with TagMe [2]. The entity linking was made with a Wikipedia dump from January 1st of 2016. The rationale behind the entity space [7] is to link similar concepts to the same entity representation, thus reducing the dispersion between semantic synonyms. In practice this will result in a exact match between queries and document terms by the entity representation, or ID. Considering the following news title from the Washington Post collection, "*John Wall, Wizards seek to pick up their spirits*" TagMe will identify the entities "*John Wall (basketball); Washington Wizards; Item (gaming); and Spirit*". TagMe aims for a balance between precision and recall, thus incorrectly linking some entities. This linking is not a major issue due to TagMe's consistency, thus still approaching the terms in the entity space. However, following the previous example it is possible that the entity "*Washington Wizards*" can be misidentified as it happened for the following news title "*Wizards vs. Magic: Washington snaps eight-game losing streak*", where the identified entities were "*Magician (fantasy); Vs. (Pearl Jam album); Magic (paranormal); Washington Huskies football; Snap (gridiron football); and Losing streak*", in this case the Wizards keyword links to a different entity despite the news still being related to the basketball team. This situation will harm our results and we acknowledge it as one of the limitations of our work.

4.2 Feature Selection and Re-ranking

We have submitted 10 runs for evaluation. The evaluated runs consisted of LTR models trained on 150 queries of the Wall Street Journal collection between the years 1987 to 1992. The training was made with a 10-fold cross validation to reduce data bias and reduce overfitting. To generate our evaluation runs we reranked a BM25 run using the default parameters (k=1.2, b=0.75) using the trained models, on the Washington Post collection. With our runs, we wanted to evaluate the retrieval impact of the utilization of different document and query representations, namely the advantages of using a BoE representation versus a traditional BoW one. Moreover, we were also interested in evaluating the impact of utilizing these

Table 3: Features used for LTR re-ranking.

Feature ID	Retrieval Model	Target	Representation
1	BM25	Body	BoW
2	BM25	Body	BoE
3	BM25	Title	BoW
4	BM25	Title	BoE
5	Coord Match	Body	BoW
6	Coord Match	Body	BoE
7	Coord Match	Title	BoW
8	Coord Match	Title	BoE
9	IDF	Body	BoW
10	IDF	Body	BoE
11	IDF	Title	BoW
12	IDF	Title	BoE
13	LM 2 Stage	Body	BoW
14	LM 2 Stage	Body	BoE
15	LM 2 Stage	Title	BoW
16	LM 2 Stage	Title	BoE
17	LM Dirichlet	Body	BoW
18	LM Dirichlet	Body	BoE
19	LM Dirichlet	Title	BoW
20	LM Dirichlet	Title	BoE
21	TF	Body	BoW
22	TF	Body	BoE
23	TF	Title	BoW
24	TF	Title	BoE
25	TF-IDF	Body	BoW
26	TF-IDF	Body	BoE
27	TF-IDF	Title	BoW
28	TF-IDF	Title	BoE

different representations over short and long textual elements, such as the title and body fields of the documents. We used a maximum of 28 features to train our models. Table 3 presents the features, where each feature corresponds to a ranking using a given retrieval model, such as BM25 or query likelihood with two-staged smoothing, over a target textual representation, such as the body field in a BoE representation.

We chose the following combinations of the features in Table 3 as our submitted runs:

1. BM25-b-BoW: A BM25 baseline obtained from the documents' body, using the default parameters k=1.2, b=0.75.
2. b-BoW: A LTR run combining all considered BoW features over the documents' body.
3. bt-BoW: A LTR run combining all considered BoW features over the documents' body and title.
4. b-BoE: A LTR run combining all considered BoE features over the documents' body.
5. t-BoE: A LTR run combining all considered BoE features over the documents' title.
6. bt-BoE: A LTR run combining all considered BoE features over the documents' body and title.
7. eb-boost: A LTR run that combines all BoW and BoE features over the documents' body, and the BoE features over the documents' title. Here the entities features were manually boosted, doubling their original weight given by the Coordinate Ascent algorithm.
8. bt-BoWBoE: A LTR run combining all features over the documents' body and title, also denoted as *FullModel*.
9. b-BoWBoE-t-BoE: Equivalent to the *FullModel* run, without the BoE features over the documents' body.
10. b-BoW-t-BoWBoE: Equivalent to the *FullModel* run, without the BoW features over the documents' title.

5 EXPERIMENTAL RESULTS

Our systems' performance can be observed in Tables 4 and 5. The runs in both tables are clustered by our research questions. The first line of the Table 4 correspond to the BM25 baseline that we proposed ourselves to beat. The t-BoE, b-BoE and bt-BoE runs were submitted to determine whether the utilization of the entity representations alone, would be able to beat the BoW representation. Despite these entity oriented runs were not able to beat the BM25 baseline, we can observe that when used in combination with the BoW representation, namely the bt-BoWBoE run, it improves substantially the BoW LTR runs, bt-BoW and b-BoW respectively. Regarding the impact of the length of the available text, we can observe some curious results where the entities alone used on a very small text field, as the title, provides siginificant results of approximately 0.4 in nDCG@10. In general when combining both the title and body textual representations the system's performance improves for both BoW and BoE representations; Despite also verifying that the models performance is improved when we remove the title field from the models bt-BoW and bt-BoWBoE, these losses are not significant to indicate the presence of a trend.

6 CONCLUSION

With our work we were able to answer our two research questions. We verified that the use of entities alone is not enough to beat the current baselines, but when used together with the BoW representation it greatly improves the system's performance. We were also able to confirm that the utilization of both document representations, title and body, together follows the expected results of improving the system's performance when used together. However, further research is required to determine the effect of short textual representations, with our experiments showing promising results when using the entity representation over the documents' titles. These results confirm that the use of an entity representation can improve the performance of search results, thus being a promising tool to be further researched.

REFERENCES

[1] Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity Query Feature Expansion Using Knowledge Base Links. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin (Eds.). ACM, 365–374. https://doi.org/10.1145/2600428.2609628

[2] 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

[3] 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

Table 4: Comparison of the different runs with a BM25 baseline over the documents' body. Following the BM25 run we can observe the median values of all 72 submitted runs. The LTR models were trained on the WSJ dataset, applied on WaPo dataset.

Run	MAP		nDCG@1000		P@10	
BM25-b-BoW	0.2219	–	0.5066	–	0.4140	–
t-BoE	0.1049	-52.73%	0.3977	-21.50%	0.2640	-36.23%
b-BoE	0.2001	-9.82%	0.4810	-5.05%	0.3800	-8.21%
bt-BoE	0.2054	-7.43%	0.4866	-3.95%	0.3860	-6.76%
eb-boost	0.2171	-2.16%	0.4993	-1.44%	0.4200	%
bt-BoW	0.2340	5.45%	0.5127	1.20%	0.4400	6.28%
b-BoW	0.2360	6.35%	0.5188	2.41%	0.4480	8.21%
bt-BoWBoE	**0.2468**	11.22%	**0.5231**	3.26%	**0.4640**	12.08%
TREC Median	0.2310	4.10%	0.5085	0.38%	0.4480	8.21%

Table 5: Comparison of the run that combines all features, also denoted as *FullModel*, with the median values of all 72 submitted runs, and runs that remove the title body, and the body of entities from the *FullModel*, respectively. The LTR models were trained on the WSJ dataset, applied on WaPo dataset.

Run	MAP		nDCG@1000		P@10	
bt-BoWBoE	**0.2468**	–	0.5231	–	0.4640	–
$FullModel_{-t-BoW}$	0.2457	-0.45%	**0.5249**	0.34%	**0.4800**	3.45%
$FullModel_{-b-BoE}$	0.2389	-3.20%	0.5179	-0.99%	0.4660	0.43%
TREC Median	0.2310	-6.40%	0.5085	-2.79%	0.4480	-3.45%

[4] Xitong Liu and Hui Fang. 2015. Latent Entity Space: A Novel Retrieval Approach for Entity-Bearing Queries. *Inf. Retr. Journal* 18, 6 (2015), 473–503. https://doi.org/10.1007/s10791-015-9267-x

[5] Edgar Meij, Dolf Trieschnigg, Maarten de Rijke, and Wessel Kraaij. 2010. Conceptual Language Models for Domain-Specific Retrieval. *Inf. Process. Manage.* 46, 4 (2010), 448–469. https://doi.org/10.1016/j.ipm.2009.09.005

[6] Tao Tao, Xuanhui Wang, Qiaozhu Mei, and ChengXiang Zhai. 2006. Language Model Information Retrieval with Document Expansion. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*, Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson (Eds.). The Association for Computational Linguistics.

[7] Chenyan Xiong, Jamie Callan, and Tie-Yan Liu. 2016. Bag-of-Entities Representation for Ranking. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, September 12-6, 2016*, Ben Carterette, Hui Fang, Mounia Lalmas, and Jian-Yun Nie (Eds.). ACM, 181–184. https://doi.org/10.1145/2970398.2970423

[8] Chenyan Xiong, Jamie Callan, and Tie-Yan Liu. 2017. Word-Entity Duet Representations for Document Ranking. 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, 763–772. https://doi.org/10.1145/3077136.3080768

[9] 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