Radboud University at TREC CAsT 2021

Hideaki Joko, Emma J. Gerritse, Faegheh Hasibi, and Arjen P. de Vries

Radboud University, Nijmegen, The Netherlands
hideaki.joko@ru.nl, f.hasibi@cs.ru.nl,
emma.gerritse@ru.nl, a.devries@cs.ru.nl

Abstract. This paper describes Radboud University's participation in the TREC Conversational Assistance Track (CAsT) 2021 for the manually rewritten utterances. We propose an entity-enriched BERT-based retrieval model, where entity information is injected into the BERT model, and compare it to the regular BERT-based retrieval model. We annotate the manually resolved user utterances with named entities using an entity linker, and inject both text and entity representations into our entity-enriched BERT-based retrieval model. We present our experimental setup, results, and analysis of helped and hurt queries when using entity information.

1 Introduction

With the growing popularity of personal assistants such as Google Assistant and Alexa, conversational systems are becoming more important. In conversational systems, entity information is known to be useful to understand user utterances and generate natural and meaningful responses [1,5,6,10]. Despite the importance of entity information in conversational systems, the research on utilizing entity information in conversational document retrieval has so far been limited.

In the paper, we report on our participation in the TREC 2021 Conversational Assistance Track (CAsT), which addresses the task of document retrieval in conversational context. Our main objective is to understand the effect of entity information on BERT-based retrieval models for this task. To this end, we utilize entity information in an entity-enhanced BERT-based retrieval model. We map entity embeddings into the same space as a BERT model and use this model to rank documents based on monoBERT re-ranking method [7]. We use an entity linker to identify named entities mentioned in the conversation utterances and inject entity embedding, along with BERT wordpiece embeddings to our BERT-based retrieval model. Our results suggest that including only named entity information in our retrieval model does not bring added value compared to a regular BERT-based retrieval model.

2 Conversational Assistance Track

CAsT is defined as a traditional document retrieval task, where queries are issued in a conversational context [2]. The specific task, referred to as conversational document retrieval task, is as follows. Given the user utterances

$$U_n = \{u_1, u_2, ..., u_i, ..., u_n\}, \tag{1}$$

the system is required to find a set of relevant documents D_i for every user utterance u_i from the document collection D. Each user utterance may contain ambiguity, have references to other utterances, or be dependent on the response of previous utterances. CAsT distinguishes two submission groups: *raw* and *manually rewritten* utterances. In the latter group, the dependency and ambiguity of user utterances are manually resolved. Each manually rewritten utterance u_i^* contains all necessary information to obtain relevant documents D_i for a turn. This, however, does not necessarily imply that conversation context is not important for retrieval anymore; even if the user utterances are self-contained, conversational context makes it easier to understand the current user utterances. In our participation, we focus on manually rewritten queries.

3 Method

Our retrieval model is an entity-enhanced BERT-based retrieval model, a hybrid between monoBERT [7] and E-BERT [9]. We utilize Wikipedia2Vec [11], which jointly embeds words $\mathbb{L}_{word}$ and entities $\mathbb{L}_{ent}$ using the lookup function $E_{Wikipedia}$ into the vector space $\mathbb{R}^{d_{Wikipedia}}$, where $d_{Wikipedia}$ are the dimensions of the Wikipedia2Vec embeddings. Following E-BERT [9], we map the entity embeddings from Wikipedia2Vec into the BERT wordpiece vector space $\mathbb{R}^{d_{BERT}}$ using a linear transformation. The linear mapping $\mathbf{W} \in \mathbb{R}^{d_{BERT} \times d_{Wikipedia}}$ is obtained by minimizing:

$$\sum_{x \in \mathbb{L}_{word} \cap \mathbb{L}_{WP}} ||\mathbf{W} \cdot E_{Wikipedia}(x) - E_{BERT}(x)||_2^2, \tag{2}$$

where E_{BERT} is the lookup function, mapping BERT wordpiece dictionary $\mathbb{L}_{wp}$ to $\mathbb{R}^{d_{BERT}}$, with d_{BERT} being the dimension of the BERT input embeddings. Using the found linear mapping $\mathbf{W}$, we obtain transformed entity embeddings using:

$$E_{en}(x) = \mathbf{W} \cdot E_{Wikipedia}(x), \tag{3}$$

where x represents an entity. The transformed entity embeddings are then injected to the BERT model.

We employ this entity-enriched BERT model in our retrieval model and perform pointwise re-ranking based on monoBERT. The input to our model consists of entity-annotated query-document pairs, with the annotations predicted by REL [4]. These queries and documents are then tokenized where all entities are embedded using Equation 3 and all other tokens using E_{BERT}.

The model is fine-tuned on the MS MARCO passage ranking collection, using 600k randomly selected query-document pairs with a batch size of 64, a learning rate of 10^{-6}, and 40k warm-up steps. For Wikipedia2Vec, we used the pre-trained embeddings from [3].

4 Runs and Results

In this section, we describe our runs and present the results on the TREC CAsT 2020 and 2021 datasets, followed by analysis of helped and hurt queries.

4.1 Runs

We submitted three official runs:

- **RU-turn**: Uses the manually resolved form of the current turn u_i^* as input for our method; see Section 3.
- **RU-turn-finetuning**: Uses the current turn u_i^* as input, and the model is fine-tuned on CAsT 2020 dataset. We split the dataset into train, validation, and test, containing 15, 5, and 5 conversations, respectively. The settings for fine-tuning are the same as Section 3, except that batch size 8 is used.
- **RU-history**: For each turn u_i^*, uses current and previous turns $\{u*_1, ..., u*_i\}$ as input to model, where turns are concatenated with space.

In addition to our official runs, we report on the following two runs:

- **monoBERT-turn**: The model described in [7], and current run as input.
- **baseline**: The official baseline model[1], which is a T5-based retrieval model using BM25 as first stage retriever.

We use the baseline method as first stage retriever for all five runs.

4.2 Results on CAsT 2020 dataset

Similar to this year's CAsT, the 2020 dataset provides manually rewritten queries for conversations. Table 1 shows the results of our runs for CAsT 2020 dataset. The results indicate that RUIR-turn outperforms the other methods: baseline and monoBERT-turn, suggesting that information from named entities might be beneficial for conversational document retrieval.

[1] `https://github.com/daltonj/treccastweb/tree/master/2021/baselines`, accessed Aug. 2021

Table 1. Evaluation results for CAsT 2020 manually written dataset.

	MAP	MRR	Recall	NDCG@3	NDCG@5	NDCG@500
baseline	0.272	0.772	**0.508**	0.479	0.461	0.451
monoBERT-turn	0.272	0.781	**0.508**	0.496	0.481	0.451
RUIR-turn	**0.273**	**0.791**	**0.508**	**0.498**	**0.483**	**0.453**

4.3 Results on CAsT 2021 dataset

Table 2 presents the results on the CAsT 2021 dataset. Median represents the mean of officially provided median values of all submitted runs for manually rewritten submission group. We observe that monoBERT outperforms our entity-enriched models, suggesting that the injected information from named entities does not lead to improvements. Additionally, the RU-turn run outperforms the RU-turn-fine-tuning and RU-hist runs, indicating that using history as input does not improve retrieval performance. Another observation is that fine-tuning on CAsT 2020 hurts the performance. We attribute this to known BERT instabilities in few-sample fine-tuning [12].

Table 2. Evaluation results for CAST 2021 manually written dataset.

	MAP	MRR	Recall	NDCG@3	NDCG@5	NDCG@500
baseline	0.417	0.868	**0.746**	0.595	0.588	0.649
median	0.371	NaN	NaN	0.555	0.550	0.612
monoBERT-turn	**0.418**	**0.872**	**0.746**	**0.597**	**0.589**	**0.650**
RU-turn	0.390	0.829	**0.746**	0.554	0.560	0.626
RU-turn-finetuning	0.378	0.818	**0.746**	0.554	0.548	0.618
RU-hist	0.361	0.811	**0.746**	0.493	0.492	0.593

Table 3 presents the performance results on the queries from Y3 where entity annotations are available. Out of 158 queries, 48 contain at least one entity annotation (30.3%). We observe that even for the queries where entity annotations are available, monoBERT outperforms our entity-enriched BERT models.

Table 3. Results for queries with at least one annotated entities.

	MAP	MRR	Recall	NDCG@3	NDCG@5	NDCG
baseline	0.386	0.818	**0.759**	0.525	0.530	0.620
median	0.344	NaN	NaN	0.488	0.490	0.579
monoBERT-turn	**0.387**	**0.832**	**0.759**	**0.531**	**0.533**	**0.621**
RU-turn	0.364	0.801	**0.759**	0.491	0.503	0.590
RU-turn-finetuning	0.331	0.770	**0.759**	0.501	0.478	0.572
RU-hist	0.360	0.771	**0.759**	0.425	0.431	0.561

Table 4. Examples from CAsT 2021 queries where entity annotations are available. RU-turn and monoBERT columns represent NDCG@3 scores for RU-turn and monoBERT-turn. Utterance column shows manually rewritten utterances. Source and #relevant documents column shows in which collections the documents judged as relevant are found and the number of those documents. The top (bottom) five rows are the queries where RU-turn performs better (worse) than monoBERT with the largest NDCG@3 differences.

Qid	RU-turn	monoBERT	Utterance	Source and #relevant documents
116_5	0.521	0.000	Okay. How does the use of rhyme in Biblical poetry compare to the use of rhyme in Islamic poetry?	KILT (1)
115_2	0.965	0.665	What are the causes of mental health stigma in Africa?	KILT (2)
121_2	0.844	0.548	What caused the drought in Brazil's coffee belt region?	KILT (5), WAPO (1), MARCO (2)
118_5	0.296	0.000	What are the common entry requirements for a Bachelor of Fine Arts program?	KILT (2), MARCO (4)
115_10	0.281	0.000	I found the Indian student's approach to treating depression through social support and ...	KILT (1), WAPO (1), MARCO (2)
115_6	0.265	0.735	Is there a relationship between Asian cultural values and mental illness?	KILT (3)
117_6	0.000	0.532	Did international support from the US help with the EndSars movement?	WAPO (2)
121_3	0.339	0.893	I also heard that other Latin American countries had coffee production issues. Was the disruption due to the drought widespread?	KILT (7), WAPO (3), MARCO (6)
117_4	0.000	0.871	How did the international community respond to the EndSars protests in Nigeria?	WAPO (4)
117_1	0.000	1.000	A friend said there were massive protests in Nigeria towards the end of 2020. What happened?	WAPO (9)

4.4 Analysis

To better understand the effect of named entities on our retrieval model, we analyze queries that are most helped and hurt by our model. We calculate the differences of NDCG@3 between RU-turn, which is the best performing of our methods, and monoBERT-turn. The top-5 helped and hurt queries are shown in Table 4. RU-turn and monoBERT columns represent NDCG@3 scores for RU-turn and monoBERT-turn. The top five rows are the queries where RU-turn performs better than monoBERT. We observe that KILT [8], which is based on Wikipedia, has relevant documents for all of the top five queries, while only two KILT documents are observed in the five bottom queries. Additionally, in the queries where RU-turn's NDCG@3 scores are 0 (i.e., qid 117_6, 117_4, 117_1), only WAPO has the relevant documents. Knowing that KILT is based on Wikipedia (containing more entity related documents), and WAPO contains news articles, we observe that our entity-enriched model performs better on entity-related documents than general news articles. This finding can also explain why our model performs better at CAsT 2020 dataset, as the collections used in CAsT 2020 do not contain WAPO but only MS-MARCO and Wikipedia.

We cannot yet explain why news articles behave so differently in our entity-focused models, as it is generally believed that entities are predominant in news. Our best guess it that the finetuning setup is not well suited to adapt to the language use in the newspaper articles, and more beneficial to the style of writing on Wikipedia and the web.

5 Conclusion

We proposed an entity-enriched BERT-based retrieval model, where named entity information is injected into a BERT-based model. Our model was employed on TREC CAsT 2020 and 2021 manually rewritten utterances. We compared our model with the regular BERT-based retrieval model and analyzed the helped and hurt queries when named entity information is used. We found that injecting only named entity information is not sufficient to improve performance.

References

1. Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J.: Wizard of wikipedia: Knowledge-powered conversational agents. In: International Conference on Learning Representations (2019)
2. Gemmell, C., Dalton, J.: Glasgow representation and information learning lab (GRILL) at the conversational assistance track 2020 p. 5 (2020)
3. Gerritse, E., Hasibi, F., De Vries, A.: Graph-embedding empowered entity retrieval. In: Proceedings of the 42nd European Conference on Information Retrieval (ECIR). pp. 97–110 (2020)
4. van Hulst, J.M., Hasibi, F., Dercksen, K., Balog, K., de Vries, A.P.: Rel: An entity linker standing on the shoulders of giants. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2197–2200. SIGIR '20 (2020)
5. Joko, H., Hasibi, F., Balog, K., de Vries, A.P.: Conversational entity linking: Problem definition and datasets. In: Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21, ACM (2021)
6. Lertvittayakumjorn, P., Bonadiman, D., Mansour, S.: Knowledge-driven slot constraints for goal-oriented dialogue systems. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3407–3419 (2021)
7. Nogueira, R., Yang, W., Cho, K., Lin, J.: Multi-stage document ranking with BERT (2019)
8. Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., Riedel, S.: KILT: a benchmark for knowledge intensive language tasks. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (2021)
9. Poerner, N., Waltinger, U., Schütze, H.: E-BERT: Efficient-yet-effective entity embeddings for BERT. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 803–818 (2020)

10. Shang, M., Wang, T., Eric, M., Chen, J., Wang, J., Welch, M., Deng, T., Grewal, A., Wang, H., Liu, Y., Liu, Y., Hakkani-Tur, D.: Entity resolution in open-domain conversations. In: Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (2021)
11. Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., Matsumoto, Y.: Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In: Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 23–30 (2020)
12. Zhang, T., Wu, F., Katiyar, A., Weinberger, K.Q., Artzi, Y.: Revisiting few-sample bert fine-tuning. arXiv preprint arXiv:2006.05987 (2020)