LINBER: A RETRIEVAL-BASED CONVERSATIONAL ASSISTANT USING ENTITY LINKING AND BERT

YUCHENG LI, YUXIANG SUN, JUN ZHANG, PEI HUO, YAN YANG, LIANG HE

Department of Computer Science, East China Normal University, 3663 N. Zhongshan Rd., Shanghai 200062 China

1. INTRODUCTION

We developed a retrieval-based conversational assistant named LINBER. LINBER was featured by entity linking module and Bert re-ranking process. LINBER perform Conversational Assistance Track (CAsT) by the fellowing five modules:

- (1) Coreference resolution module
- (2) Keywords extraction module
- (3) Entity linking module
- (4) Retrieval module
- (5) BERT re-ranking module

1.1. **Coreference Resolution.** Our coreference resolution processure is done by AllenNLP¹ coreference resolution model. This model is a end-to-end neural model which consists of a word embedding layer, a bi-lstm layer and a feed-forward layer. Experiments shows that this model access the state-of-the-art performance on many benchmark.

1.2. **Keywords.** Finding keywords of utterances is a crucial part in our retrieval. We produce retrieval query by combining keywords extracted in the whole dialogue history. Rather than use all words in dialogue, use keywords may involve less noise in our query. We realize keywords extraction module by three keywords algorithms: Tf-idf, TextRank and RAKE.

Tf-idf calculate term frequency and inverse document frequency which is used to obtain the importance of words. Term Frequency refers to the number of times a word appears in the document, and inverse document frequency means that the fewer documents the keyword W is contained, the better categorization ability the keyword W has. Tf-idf can be obtained by multiplying B and C. Term Frequency tf and inverse document frequency idf is given by:

(1)
$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

(2)
$$\operatorname{idf}_{i} = \log \frac{|D|}{|\{j : t_{i} \in d_{j}\}}$$

¹https://allennlp.org/

LINBER: A RETRIEVAL-BASED CONVERSATIONAL ASSISTANT USING ENTITY LINKING AND BERT

$$(3) tf - idf = tf * idj$$

TextRank is a graph-based sorting algorithm for text. It uses co-occurrence information between words within a document to extract keywords. It can extract the key words and phrases of a given text, and obtain the key sentences of the text with the automatic abstracting method.

RAKE is a very efficient algorithm. The keywords extracted by RAKE are not single words, but also possible phrases or professional terms. Therefore, RAKE can extract more relevant result which is useful for searching answers.

1.3. Entities Linking. It is important to link mentions to entities in order to find precise entity names in dialogue. Due to the dataset we used is TREC CAR which is a copy of English wikipedia, link mentions in dialogue to wikipedia entities is a good choices. We chose TAGME² to achieve this goal. The fast speed, on-the-fly searching and the good performance are the three main advantages of TAGME.

1.4. **Retrieval.** We use elasticsearch³ as our base retrieval tools. Each retrieval query is based on keywords and linked entities in the former dialogue history. What's more, each retrieval step will produce a tf-idf based similarity score which indicate how similar the query and each response are.

1.5. **BERT Re-ranking.** After getting the candidate responses, we use a BERT model to compute the relevance score between the question and response. BERT has shown its powerful performance in NLP tasks. We use a pre-trained BERT base model⁴, and fine-tuned it on MSRP dataset⁵. The final re-ranking is based on the mix of tf-idf based similarity score (see section 1.4) and BERT relevance score.

2. Result

We inspect the result of our approach mainly on four metrics: NDCG@5, NDCG@1000, MAP@5 and MAP@1000.

Statistics are shown in the fellowing table:

| | NDCG@5 | NDCG@1000 | MAP@5 | MAP@100 |
|-------------------------------|--------|-----------|--------|---------|
| LINBER | 0.2447 | 0.4116 | 0.0397 | 0.1907 |
| TABLE 1. The result of LINBER | | | | |

- ⁴https://github.com/google-research/bert
- $^{5} https://www.microsoft.com/en-us/download/details.aspx?id{=}52398$

²https://tagme.d4science.org/tagme/

³https://www.elastic.co/