A Step towards Context Identification for Conversational Search

Vaibhav Kumar
Carnegie Mellon University, USA
vaibhav2@cs.cmu.edu

Jamie Callan
Carnegie Mellon University, USA
callan@cs.cmu.edu

ABSTRACT

The system comprises of three different components. The first component makes a decision whether to incorporate contextual information for the current query in ongoing conversation. The decision is based on the KL-divergence between the retrieved documents for the original query and whether the query consists of pronouns. The second component identifies the contextual information (if required) for the answering the current query. This identification is performed using an SVM classifier which uses BERT attention weights along with other linguistic features. Finally, the third component utilises Indri for document retrieval.

KEYWORDS

BERT, KL-Divergence, Query Expansion, Indri

1 INTRODUCTION

The TREC 2019 Conversational Assistance Track (2019) introduces a dataset for evaluation of Conversational Information Seeking (CIS) Systems. The task, as defined by the track, is to read the dialogue context for each conversation and extract vital information which is necessary for retrieving documents that can answer the current query (dialogue) under consideration. This creates challenges for the current search systems which are highly tuned to retrieve documents for queries which are self-contained i.e no additional information is required to answer them.

More formally, for a conversation C, which contains queries $q \in (q_1, q_2..q_n)$, the system should be able to generate candidate documents for each q_i by conditioning on q_i and information extracted from all preceding queries $q_1..q_{i-1}$. The retrieved documents should be capable of answering the current query. Here, all the queries q are posed as natural language sentences.

Based on the training dataset, the queries can majorly be grouped into two categories. The first category (Cat1) contains self-contained queries where no additional information may be required for answering it. On the other hand, the second category (Cat2) contains queries where contextual information is absolutely necessary for answering it. It is possible that a particular conversation can have queries belonging to both these categories or may simply contain queries from one of the two categories. Moreover, the second category of queries can be further divided into two sub-categories, namely, Cat2-exp and Cat2-imp. Cat2-exp consists of queries which have explicit contextual markers like pronouns that needs to be resolved to form a self-contained query. Cat2-imp consists of queries which does not have explicit but rather implicit (zero pronoun) markers. Examples of all the types can be seen from Table 1.

The types of the aforementioned query classes motivates the structure of the implemented system. The entire system can be sub-divided into three components. To summarise, the pipeline of the implemented system is as follows:

- If the query belongs to Cat2-exp i.e., it consists of an explicit marker (pronoun), then the pronoun is resolved using a classifier. The resolution of pronouns is done using an SVM which takes the BERT[1] attention values and other linguistic features as input. If the subsequent queries have explicit markers, then the identified context is carried over.
- If the query does not belong to Cat2-exp, then the KL-divergence of the top retrieved documents (based on the unmodified query) is used to judge whether it belongs to Cat1 or Cat2-imp. If the current query belongs to Cat2-imp, then the context identified using the classifier is appended to it. However, if it belongs to Cat1, then no additional context is appended. It is also assumed that this marks a contextual shift and that the previously identified context cannot be carried over i.e a new context needs to be generated for subsequent queries belonging to Cat2-exp and Cat2-imp.
- Finally, retrieval is performed using Indri.

Two out of the four runs utilise the pipeline mentioned above. Out of these two, one (**coref_chisft_qe**) utilises query expansion and the other (**coref_cshift**) does not. The third run (**ensemble**) is an ensemble of four different retrieval systems. Two of the four systems comprise of coref_chisft_qe and coref_chisft. The third system uses heuristics to identify the major topic of the conversation based on its first query and appends it to all subsequent queries. The fourth system discards the assumption of contextual shift and appends all possible contexts to the query which have been identified upto that point. The fourth run (**manual _indri**)simply utilises manually rewritten queries without any preprocessing and retreives using Indri.

2 SYSTEM ARCHITECTURE

This section describes the three components of the system. The first component is responsible for context identification, the second is responsible for identifying contextual shifts, and the third is responsible for retrieval.

2.1 Context Identification

Context identification is performed using an SVM classifier with a variety of features. The dataset for training the classifier is based on the training topics provided by the track. Both implicit and explicit markers were manually resolved in order to construct the training dataset. For example: In Table 1, for Cat2-imp, "about" in 1_5 was resolved to "average starting salary" in 1_4. An $80 - 20$ split was used for training and evaluation.

At the time of testing, each pronoun token in the current query is matched against all the tokens in the previous query. The classifier merely states whether to include the previous token (1) or not (0). Thus, the problem of selecting context is posed as a binary classification problem. However, a slight modification is made to

Category	Query Progression	Analysis
Cat1	15_4:What kind of problems can I expect? 15_5:Tell me about the history of linguistics as a field.	15_5 does not require any additional information
Cat2-exp	25_1:Tell me about Orca whales. 25_2:Are they really whales?	"They" in 25_2 is a pronoun (explicit marker)which needs to be resolved to Orca.
Cat2-imp	1_4:What's the average starting salary in the UK? 1_5:What about in the US?	In 15_2, "about" implicitly refers to "average starting salary". Here "about" is not a pronoun in a conventional sense.

Table 1: Example of various query categories.

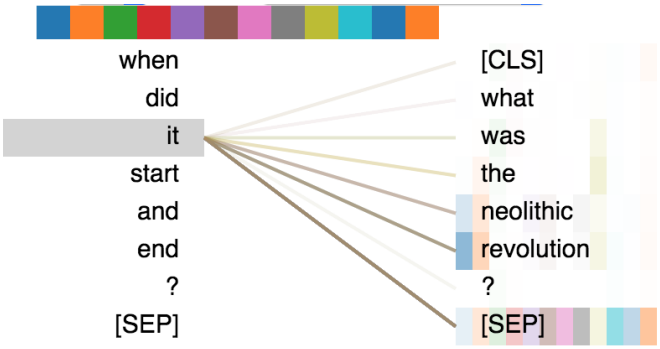

Figure 1: Attention weights of the 9th Layer of BERT. The different color boxes represent the various attention heads. On the left side query 4_2 is presented and on the right side query 4_1 is presented. It can be clearly seen that the pronoun "it" maps to "neolithic revolution".

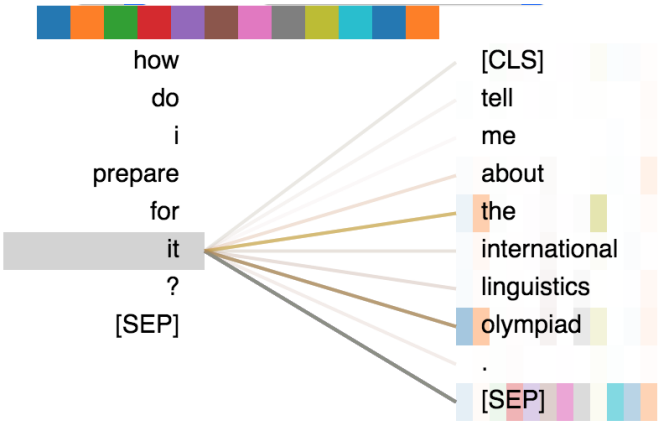

Figure 2: Attention weights of the 9th Layer of BERT. The different color boxes represent the various attention heads On the left side query 15_2 is presented and on the right side query 15_1 is presented. It can be clearly seen that the pronoun "it" maps to "olympiad".

the final result. It might be possible that the classifier selects a discontinuous set of tokens from the previous query. For example: In Table 1, for Cat2-imp, only "average" and "salary" might be selected by the classifier, while leaving the term "starting" out. In order to overcome such a phenomena, the tokens left out in between the selected tokens are added as well.

As mentioned earlier a variety of features are used by the classifier. An investigation of the BERT attention weights [1] revealed that it might be helpful in identifying the important context (resolving pronouns in particular). For example: In Figure 1, it can be observed that "it" has highest attention weights over "the neolithic revolution" and in Figure 2, "it" has highest attention values over "olympiad". This suggests that utilising the BERT attention values might help in better classification. All the experiments utilise the BERT-base-uncased models and the corresponding attention values of the 9th and the 11th Layer.

Apart from the BERT attention values, various linguistic features are used as well. These features are: POS tags of the current and the previous query tokens, binary indicator for plurality of the current and the previous query tokens, binary indicator for stopwords, binary indicator suggesting whether the two tokens appeared together in some other previous query, rank of the attention values, number of token matches between the two queries, POS tags of the neighbouring tokens.

An SVM utilsing the above mentioned features achieves a training accuracy of 89.97% and a testing accuracy of 91.10%.

2.2 Identification of Contextual Shift

If a query consists of a pronoun, then it is resolved using the method described in Section 2.1. However, if this is not the case, then there are two possibilities: either the query is of type Cat1 or of type Cat2-imp. If the query belongs to Cat1, then it is a self-contained query and demarcates a contextual shift. If the query belongs to Cat2-imp, then the action to be performed is to carry over the previous context (if available) or extract a new context. The new context is extracted based on a similar procedure as mentioned earlier with a slight modification. Since this current query has no pronouns, all its tokens are matched against the tokens of the previous query.

[1] https://github.com/jessevig/bertviz

Overall, the entire problem of identification of contextual shift can be posed as a binary classification problem once again.

It can be well observed that Cat1 queries contain important tokens, which if used as is can retrieve a good set of documents capable of providing an answer to it. For example: A query of category Cat1 15_5: "Tell me about the linguistics as a field", if used as is for retrieval should retrieve a consistent set of documents which contains terms related to the particular query. One can expect that the top documents, must be somehow related to the topic "history" and "linguistics" even though such documents might not provide a relevant answer. On the contrary, one can expect that a query of type Cat2-imp, will not be able to retrieve documents which are consistent in the terms of the topics they contain. For example: 1_5: "What about in the US?", the retrieved documents for this query might contain the term "US", but all the documents might not talk about one particular aspect related to "US". Some documents might talk about "politics in US", some would mention "tourism in US". Thus, there is a variety of unique terms that the top retrieved documents for a such query would contain signifying that this particular query needs some salient piece of information for better on-topic retrieval.

Based on the argument presented above, contextual shift can be identified based on the KL-divergence of the top retrieved documents. A low KL-divergence would signify that the retrieved documents correspond to similar topics and thus would denote a contextual shift. On the other hand, a higher KL-divergence would mean that the retrieved document pool consists of a variety of topics. This would denote that there is an implied context that needs to be resolved.

The average of the KL-divergence between pairs of the top 20 documents is computed. A threshold is set below which a contextual shift is identified, thus requiring no additional terms. It is to note that the KL-divergence is not symmetric and hence pairs are formed in a rank-wise sorted order i,e (d_i, d_{i+1}) is a valid pair whereas d_{i+1}, d_i is not. Here d_i represents the document at rank i.

2.3 Retrieval Using Indri

Finally, retrieval is performed using Indri in a completely unsupervised fashion. The stop words from the original query are not removed. However, if a context has been appended to the query, then stopwords from the appended context is removed before retrieval.

Run1 (coref_chisft_qe) of the system comprises of all the components as mentioned above. Adding to it, it also uses query expansion. Run2 (coref_chisft_) of the system is the same as Run1 except for the query expansion part.

Run3 (ensemble) is an ensemble of four different systems. The first two systems are the results of Run1 and Run2. The third system utilises a simple heuristic to identify the main topic of the conversation and then appends it to the subsequent queries. The heuristic is to select the longest noun phrase and adjacent adjectives to this phrase. This serves as the context for the entire conversation. The fourth system discards the component which identifies contextual shift and instead appends all possible contexts that has been identified from the beginning. The results of the four systems are then merged. The scores of the individual systems are normalized before merging.

Run4 (manual_indri) simply uses the manually rewritten queries and performs retreival using Indri.

Rund Id	P@5	P@10	P@20	P@100
coref_cshift	0.3977	0.3890	0.3460	0.1746
coref_cshift_qe	0.4289	0.4133	0.3595	0.1783
ensemble	0.4532	0.4634	0.3974	0.1994
manual_indri	0.5376	0.5231	0.4512	0.2169

Table 2: Comparison of Precision at different levels for the various runs.

Rund Id	R@5	R@10	R@20	R@100
coref_cshift	0.0424	0.0821	0.1446	0.3514
coref_cshift_qe	0.0443	0.0835	0.1485	0.3517
ensemble	0.0567	0.1007	0.1755	0.4179
manual_indri	0.0719	0.1353	0.2299	0.4975

Table 3: Comparison of Recall at different levels for the various runs.

Rund Id	ND@5	ND@10	ND@20	ND@100
coref_cshift	0.2618	0.2695	0.2684	0.3034
coref_cshift_qe	0.2816	0.2862	0.2841	0.3092
ensemble	0.3010	0.3079	0.3162	0.3583
manual_indri	0.3691	0.3784	0.3749	0.4237

Table 4: Comparison of NDCG at different cut points for the various runs.

3 RESULTS AND DISCUSSION

The results can be seen in Tables 2, 3 and 4. Table 2 presents the Precision scores at different levels for the various runs. Table 3 and 4 presents the Recall and NDCG values for the different runs.

It is quite evident that manual_indri outperforms the other systems by a significant margin. This points towards the low efficacy of the classifier responsible for selecting contextual terms. Perhaps, a good classifier is essential for a better performance on the task. Apart from this, query expansion seems to have helped. As it can be seen, the results of coref_cshift_qe is better than that of coref_cshift.

Amongst the non-manual runs, ensemble seems to have performed best. This gain could be attributed to the heuristic which might have selected contextual terms if the classifier missed out any.

In future, we would like to explore more sophisticated methods for context identification which are end-to-end trainable and work well even in dearth of training data.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).