Information Retrieval - 2020/2021 1st Semester

Open-Domain Conversational Search

Project Report

Alexandre Correia (53298) ; Henrique Raposo (57059) ; Pedro Ferreira (52370)

NOVA School of Science and Technology, 2825-149 Caparica, Portugal
https://www.fct.unl.pt/

Abstract. The goal of an open-domain conversational search is to satisfy a user's information need, which is expressed through a sequence of conversational turns. The response from the retrieval system is a ranking of short text responses suitable for voice-interface or a mobile screen.

Keywords: Information retrieval and evaluation, Conversational search, Natural Language Processing, LMD Model, BERT model, NER model, T5 model, Precision, Recall, AP, nDCG, Precision-Recall, Re-ranking, Query Rewriting.

1 Introduction

Conversational Search is a human-computer interaction concept. Its main goal is to allow users to speak a sentence into a device and that device can answer with a full sentence, creating a voice assistant like Amazon Alexa, Siri and many others.

This project focuses on understanding the information needs in a textual conversational format and finding relevant responses using contextual information. To achieve this goal the project is divided into 3 phases: first-stage passage retrieval, natural language embeddings and conversation tracking.

2 First-stage passage retrieval

In the first stage of the development, the method used to retrieve the top most relevant documents is the Language Model with Dirichlet smoothing (LMD) which helps smoothing the frequencies of the terms in a document by increasing the length of the document by μ, enabling us to add a fractional number of occurrences to each term frequency in the document according to the frequency of them occurring in the collection, like this: $\mu * Mc(ti)$.

3 Natural language embeddings

In the second stage of the development, the model used to compute the question-passage embeddings is the Bidirectional Encoder Representations from Transformers (BERT) model, which is a Transformer-based machine learning model for Natural Language Processing (NLP).

4 Conversational Re-ranking

In the last stage of the development, the two main models used are Named Entity Recognition (NER), which identifies and classifies names in text, and Text-To-Text Transfer Transformer (T5), which replaces the pronouns by their subjects in a sentence.

5 Evaluation

5.1 Resources

Relevance judgments. Is the set that contains the ground truth of the evaluated documents from the dataset used.

ElasticSearch. Is a search engine and in this project it is configured to use the retrieval model named Language Model with Dirichlet smoothing (LMD) and is indexed with data from MSMARCO and Wikipedia.

BERT. Is a Transformer-based machine learning technique for Natural Language Processing (NLP) that has achieved state-of-the-art results in a wide variety of NLP tasks. It has been used by giant companies like Google to better understand user searches by getting the contextual relations between words (or sub-words) in a text.

NER. Is a subtask of information extraction that recognizes and classifies names in text into some predefined categories.

T5. Is a Transformer based architecture that uses a text-to-text approach. Every task is cast as feeding the model text as input and training it to generate some target text. In this project is used to replace the pronouns by their subjects in a sentence.

5.2 Experimental methodology

First-stage passage retrieval. To retrieve information, the algorithm starts by reading the dialogue turns one by one and then computes the top 100/1000 candidate responses (text passages) using the ElasticSearch search engine. Finally, using the dataset ground truth, it measures the success of the retrieval method, which uses the LMD model, by plotting graphs using the following metrics: AP, nDCG and Precision-Recall.

Natural language embeddings. To improve the results that we get from the first-stage passage retrieval, the algorithm re-ranks the top 100/1000 passages using a Logistic Regression (from scikit-learn) classifier. To train the classifier, it uses the candidate responses that are classified on the ground truth, builds the question-passage pairs, gets their corresponding CLS token embeddings from BERT and finally trains the classifier.

For the final re-ranking, it gets the CLS token of each top passage, feeds the classifier with the CLS tokens and extracts the scores, using them to re-rank/re-order the passages. In order to measure the success of the methods implemented so far, some graphs are plotted using the same metrics (AP, nDCG and Precision-Recall).

In the training task of the classifier, we experimented giving different values to its parameters and, from these analysis, we consider that our training set is not large enough and the ratio between relevant and non-relevant documents is not 50-50, making the classifier not to work as expected. To fix the first issue, in the classifier training task, we forced the classifier not to overfit the data by using a lower C (regularization factor) value, which was selected by experimenting some values and by watching which one improves the most the results. To fix the second issue, we repeated each relevant document four times.

Conversational Re-ranking. To further improve the results that we get from the previous phases of the project development, the algorithm uses four approaches to try to improve the user questions:

Method 1	**Method 2**
Concatenate the first user questions to each turn of the dialog.	Concatenate the entities of the first user utterance to each turn of the dialog.
Method 3	**Method 4**
Re-writing each question using the past utterances using the T5 model.	Re-writing each question using the T5 model and then concatenate the entities of the re-written question.

As a last step, it measures the success of all the methods implemented in this project by comparing them together. This is done by plotting graphs using the same metrics used in the previous phases.

5.3 Experimental results

After retrieving the top candidate responses, some graphs are plotted to measure the success of the different implemented methods in the dataset.

Average Precision (AP)

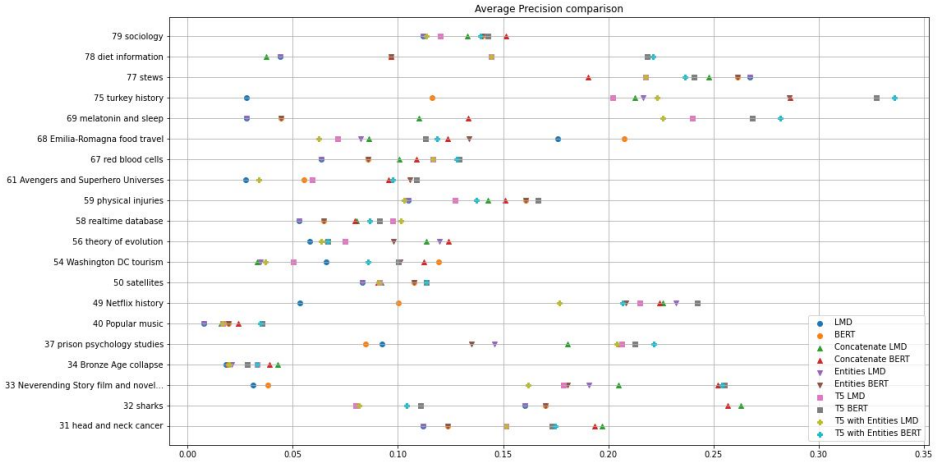

Fig. 1. AP comparison of all the conversations

From **Fig. 1.**, it's possible to see that the conversation with the highest value for the LMD model is the 77th (around 0.27), which talks about "stews", and, in contrast, the one with the lowest value is the 40th (around 0.01), which is about "Popular music". In **Fig. 1.**, it is also possible to verify that the majority of the conversations have a value closer to the 40th conversation (between 0.02 and 0.1), which means that most of the documents retrieved were not relevant.

Doing the same analysis for the BERT model, we can see that the conversations with the greatest value is the 77th conversation (around 0.26), and, by contrast, the conversations with the smallest values are the 33rd, 34th, 40th and 69th (between 0.02 and 0.04). We can also check that most of the results are between 0.06 and 0.21.

Finally, checking the results of the methods implemented in the third phase, we can see that the results improve for the majority of the conversations. This is easily verifiable in the 33rd conversation, which talks about "Neverending Story film and novel adaptation", in the 49th conversation, which is about "Netflix history", and also in the 75th conversations, which goes around "turkey history". On the other hand, a few conversations, like the 68th, which talks about "Emilia-Romagna food travel", mark the third phase methods with a lower performance than the LMD and the BERT models.

From this examination, we can conclude that BERT has a better performance than the LMD and that the third phase methods perform better than the BERT in most of the conversations. The first conclusion is due to the fact that LMD looks at how many times a word is in a document and BERT tries to understand the similarity of the questions and the passages. As we can see, both the LMD and BERT are not able to identify pronouns, and so, the metrics are lower. The reason for the second conclusion is that the third phase methods improve the user questions by concatenating the first user question

along the conversation or by replacing the pronouns on each question according to the previous ones along the conversation, helping to correctly understand the meanings behind pronouns.

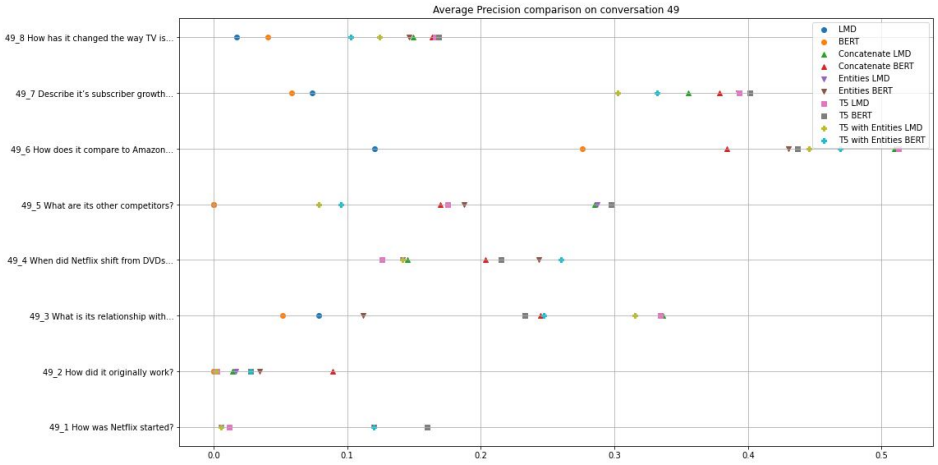

Fig. 2. AP comparison on conversation 49

To analyse in more detail the impact of the pronouns in questions, we now look at some conversation turns. The conversation 49, presented in **Fig. 2.**, is a good example of the improvement on using the third phase methods.

As we can see, the turns that have the highest improvements are the 6th ("How does it compare to Amazon Prime Video?") and the 7th ("Describe it's subscriber growth over time."). There are also some turns, like the 2nd ("How did it originally work?"), which have a really low improvement. In the majority of the turns, there usually are some improvements, as is the case of the rest of the turns in this conversation. This means that most of the questions re-writings attempts were successful, ones better than others.

LMD	**BERT**	**Concatenate BERT**	**Entities BERT**	**Entities LMD**
0.12	0.28	0.39	0.42	0.43
T5 BERT	**T5 with Entities LMD**	**T5 with Entities BERT**	**Concatenate LMD**	**T5 LMD**
0.43	0.44	0.47	0.51	0.515

Table. 1. Average Precision per turn on conversation 49th with and without pronouns

Table. 1. shows us the evolution of the AP using the different methods in a question that has a pronoun, such as the 6th turn of the 49th conversation.

If we analyse the second turn of the 49th conversation, we see that the third phase methods did not have a significant performance improvement. This behaviour has a reason: the third phase methods try to improve the identification of the subjects behind the pronouns but, in this turn, they have difficulty in identifying the subject. This is visible on the result of the T5 model - original: "How did it originally work?" ; new: "How was Netflix started?" - where it replaced the user question by its first question, not correctly replacing the pronoun, resulting in low results. An alternative used in the third phase methods for trying to identify the subject is the Concatenate BERT that concatenates the first user question to every turn of the conversation. As we can see, this is the best result in this turn, as it did not replace the current user question, but instead concatenated its first question, resulting in a better documents retrieval.

Doing a similar analysis for the 6th turn of the conversation 49, we can perceive the opposite result from the one we get in the second turn. While in the 2nd turn the T5 methods didn't express a better performance than the LMD and BERT models, the sixth turn shows a clear improvement of the performance, this is due to the fact that the T5 model - original: "How does it compare to Amazon Prime Video?" ; new: "How does Netflix compare to Amazon Prime Video?" - is now able to correctly identify the subject in the query, being in this case the best methods to use.

Normalized Discounted Cumulative Gain (nDCG)

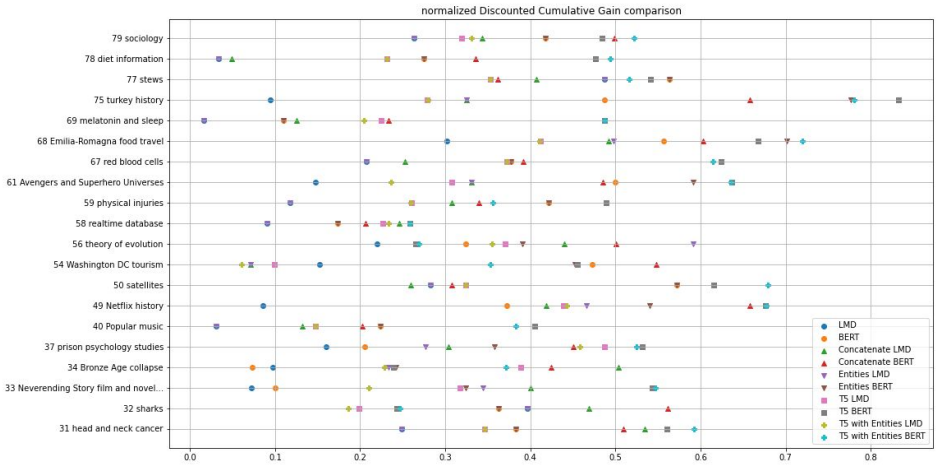

Fig. 3. Mean nDCG comparison of all the conversations

Similar to the AP plot for all the conversations, **Fig. 3.** presents the nDCG measure for all the conversations. The LMD model has its highest peak at the 77th conversation, which talks about "Stews", and its smallest peaks at the 69th, 40th and 78th conversations. The majority of the conversations are between 0.08 and 0.3.

Doing the same analysis for the BERT, the highest values are measured in the 68th, 50th and 77th conversations. In contrast, the lowest value is obtained in the 34th, which is about “Bronze Age collapse”. From **Fig. 3.** we can also check that the most of them are between 0.1 and 0.5.

To finish this first discussion of the results about the nDCG we compare the improvements obtained from the third phase methods. From **Fig. 3.**, we can say that the third phase methods have the biggest improvements in the 68th and 75th conversations. It’s also visible that the worst improvement is in the 32nd conversation, which talks about “sharks”, and that most of the conversations tend to have little improvement.

From this examination, as we did for the AP metric, we can conclude that BERT has a better performance than the LMD. We can also conclude that the third phase methods perform better than their ground methods (LMD or BERT) in most of the conversations.

By comparing **Fig. 1.** and **Fig. 3.** there is something that we notice right away: the nDCG of the different methods, for most of the conversations, has a higher value than the AP. The reason supporting this statement is that the nDCG metric considers the level of relevance of the document while the AP only observes if the document is relevant or non-relevant.

Making a deeper analysis on how the nDCG performs along the turns (**Fig. 3.**), similarly to what happened on the AP (**Fig. 2.**), for the LMD and BERT methods, the first turn usually has one of the greatest values and the following ones tend to get smaller values. The reason for this to happen is that usually the first question a user asks does not contain any pronouns, and so, the base methods, which do not use any re-writing mechanisms, can understand well what the user is looking for.

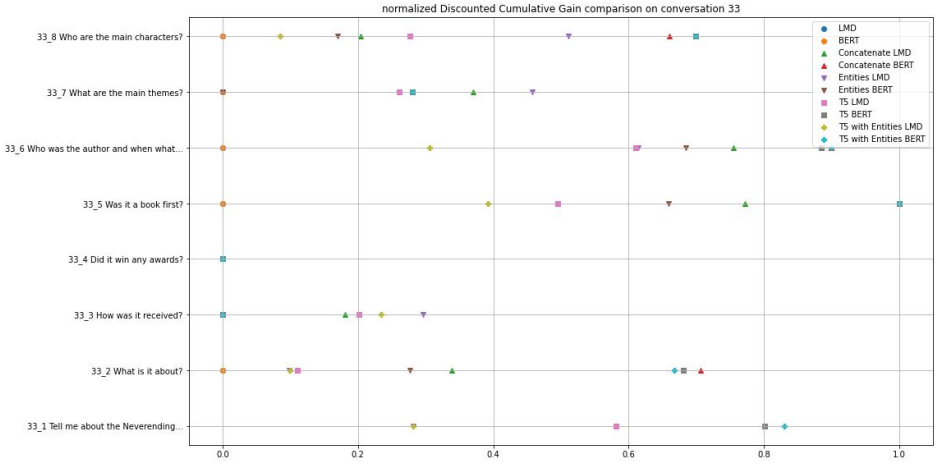

Fig. 4. nDCG on conversation 33

Analysing the differences between the third phase methods (**Fig. 4.**), we can see that the largest improvements belong to the T5 BERT method and T5 with Entities BERT method. These methods are close to each other and have the higher values in most of the conversations, but T5 BERT is generally better.

In addition and as expected, the T5 methods with LMD (T5 LMD and T5 with Entities LMD) generally perform worse than T5 methods with BERT (T5 BERT and T5 with Entities BERT). The first two methods usually present lower results than the ones given by the other methods that use LMD (Concatenate LMD and Entities LMD), as we can see in the conversation 56 (**Fig. 3.**), where the Entities LMD method has the highest mean nDCG value of the conversation (around 0.59) and the Concatenate LMD (about 0.44) the third highest. This is an example of a situation where the T5 model seems to misbehave.

Looking more closely at the different plots of each conversation, we get an insight to these values. Using the **Fig. 4.** as a reference, we can see a clear improvement of the results, where the third phase methods achieved better results than the previous phases. In this plot we can spot an exception to these results: the 4th turn question - "Did it win any awards?" - where none of the methods were able to get a good result. To further confirm the previous conclusion that the T5 BERT and T5 with Entities BERT methods tend to get the best results, with close values between the two methods and even reaching nDCG value of 1 in the 5th turn question - "Was it a book first" - with a considerable gap to the other methods.

Precision-Recall

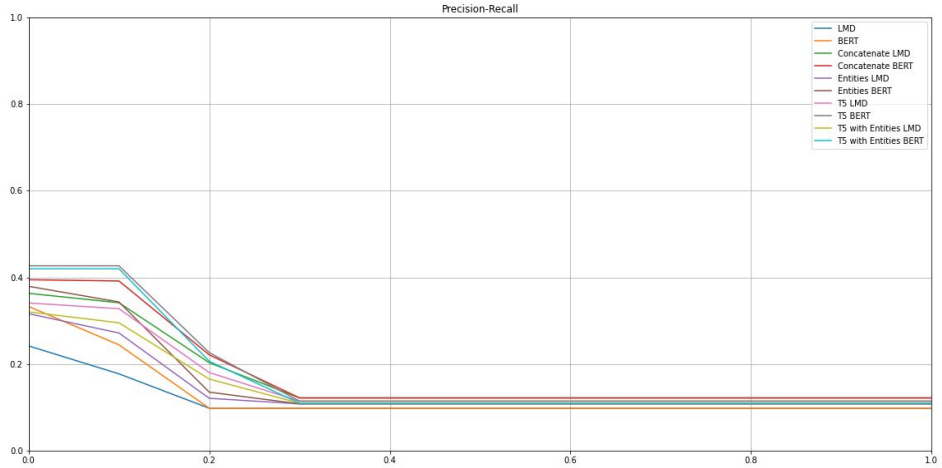

Fig. 5. Mean Precision-Recall of all the conversations

In **Fig. 5.** we can see the differences between the mean precision-recall curves for all the conversations in all the methods. While the LMD model starts with precision values of around 0.22 of precision and abruptly decreases to around 0.10 as the recall values increase, the BERT model starts with values of precision around 0.33 and decreases to around 0.10 as the recall values increase. Looking at the third phase methods, the curves with the best performance are the T5 BERT for values of recall between 0 and 0.2, and the Concatenate BERT for recall values between 0.2 and 1.

After analysing **Fig. 5.**, it's possible to conclude that, in general, the values for the BERT model are better than the LMD values and that all the curves relative to the third phase are better than the BERT model for near all the recall values, being the exception the Entities LMD and the T5 with Entities LMD for recall values between 0 and 0.02.

The T5 BERT and the T5 with Entities BERT being the best methods for recall values between 0 and 0.2 is something expected because they improve the results of the BERT model by re-writing the questions/queries with the correct subject instead of the pronouns. For recall values bigger than 0.2 it was also expected that the Concatenate BERT to behave better because with low precision the rewriting queries methods will not be able to correctly write the subject instead of the pronoun causing the T5 BERT not to be the best method, but the concatenation instead.

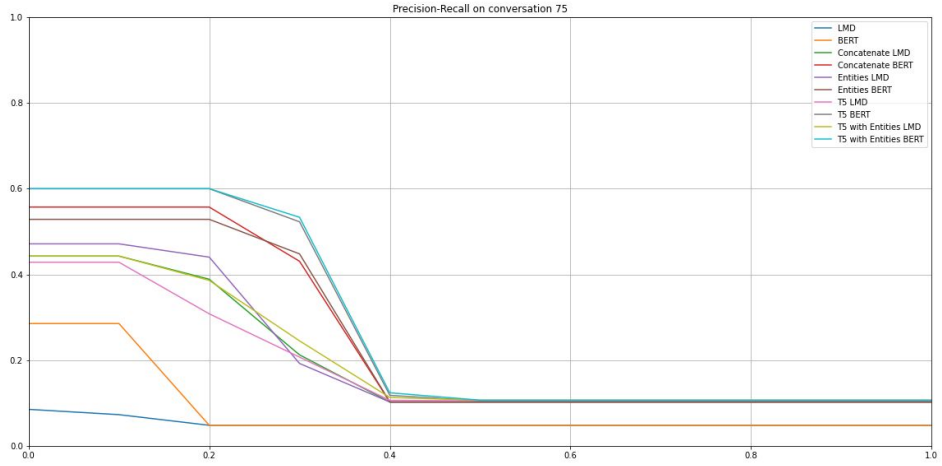

Fig. 6. Precision-Recall on conversation 75

Fig. 6. is the practical representation of the theory behind precision-recall curves which states: as the recall increases the precision continues the same or goes lower, resulting in a trade-off between the precision and the recall metrics. For instance, as expected, the T5 with Entities BERT, is the curve with better performance starting with a precision of 0.6 and decreasing for values between 0.2 and 0.5 of recall, stabilizing at the precision value at 0.12.

Similarly to **Fig. 5.**, in **Fig. 6.** the LMD is the model with the worst performance curve with a precision almost constant for all recall values. Compared to LMD, the BERT model significantly improves the results for recall values between 0 and 0.2. The best methods on the 75th conversation are, as happens in the mean precision-recall and for the same reasons, the T5 with Entities BERT and the T5 BERT.

6 Conclusions

After analysing all the results, it is possible to conclude that evaluation metrics are important instruments to determine the success of the information retrieved by the search methods and the different attempts to improve the results. Evaluation metrics allow us to verify and determine the characteristics of each model and method, and how they perform given different examples and data.

In short, there are a few conclusions that we can draw:

Average Precision (AP). In general, the results using the BERT method after training the classifier are better than the results using only the LMD model. The third phase methods are even better than the base methods (LMD and BERT models) in most of the cases.

Normalized Discounted Cumulative Gain (nDCG). Like the AP it presents better results in the third phase methods, but its results still differ from the AP due to the fact that they take in account the relevance judgments classified between 0 and 4 and not only the binary classification like the AP does to measure.

Precision-Recall. The plot of the mean of all the conversations using BERT and the third phase methods presents a higher initial value and goes down a bit smoother than the LMD.

Use of pronouns in questions/queries. The BERT model still suffers from the same problem as the LMD model, which is the difficulty in identifying the subject because of the use of pronouns in the questions/queries. The third phase methods improved the results even when the questions have pronouns. These approaches (third phase) generally perform better than LMD and BERT.

Regularization factor. In this dataset, the use of a lower regularization factor helped the classifier to better classify the documents into relevant or non-relevant.

Document ratio. From some experiments with this dataset we consider that the ratio between the relevant and non-relevant documents is not 50-50, which did not help the classifier and lead us to repeat the relevant documents four times to mitigate this ratio effect.

Natural language embeddings. The use of the BERT model to get the CLS tokens, which represents the similarity between the question and the passage, has improved the retrieval of the most important documents by helping re-ordering them.

Named Entities Recognition (NER). The use of NER to identify the entities from the questions helped to get the most important documents from ElasticSearch, as it helps boosting the results by indicating what entities are referred in the question.

Text-To-Text Transfer Transformer (T5). The use of T5 improves queries with pronouns by replacing them with their corresponding subjects, which are referred in the previous queries.

Phase 3 Methods. The implementation of some methods, with four different mechanisms, that try to better understand what the user wants, by understanding the meanings of the pronouns, has generally helped the retrieval method to get the most relevant documents.

References

1. C. D. Manning, P. Raghavan and H. Schütze, “Introduction to Information Retrieval”, Cambridge University Press, 2008.
2. R. Ferreira, M. Leite, D. Semedo, J. Magalhães, “Open-Domain Conversational Search Assistant with Transformers”, NOVA School of Science and Technology.
3. Class Lectures of Information Retrieval, NOVA School of Science and Technology.