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, LMD Model, BERT model, Precision, Recall, AP, nDCG, Precision-Recall, Re-ranking.

# Introduction

Conversational Search is a human-computer interaction concept. Its main goal is to allow the user 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.

# 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: $μ\*Mc(ti)$.

# Natural language embeddings

In the second stage of the development, the model used to compute the sentence-pair embeddings is the Bidirectional Encoder Representations from Transformers (BERT) model, which is a Transformer-based machine learning model 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.

## Evaluation

## Resources

**Relevance judgments.** This is a set that contains the ground truth of the evaluated documents from the dataset used.

**ElasticSearch.** This service is configured to use the retrieval model named Language Model with Dirichlet smoothing (LMD) and is indexed with data from MSMARCO and Wikipedia.

**BERT.** This deep learning model is pre-trained using BookCorpus and Wikipedia.

## 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 used retrieval model by plotting graphs using the following metrics: AP, nDCG and Precision-Recall.

**Natural language embeddings.** To improve the results that we got 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 query-passage pairs, gets their corresponding CLS token embeddings from BERT and finally trains the classifier, with cross-validation to select the best C (regularization factor) value, using those embeddings.

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 this re-ranking, some graphs are plotted using the same metrics (AP, nDCG and Precision-Recall).

# Experimental results

After retrieving the top candidate responses, some graphs are plotted to measure the success of the LMD model compared with the Re-ranking in the dataset.

### Average Precision (AP)

 

| **Fig. 1.** AP comparison of all the conversations | **Fig. 2.** AP comparison on conversation 68 |
| --- | --- |

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 talks 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 Re-ranking, we can see that the conversations with greatest values are the 34th, 68th and 77th conversations (between 0.5 and 0.55), while the conversation with the smallest value is the 69th (around 0.15) and most of the results being between 0.3 and 0.5.

From this examination, we can conclude that the Re-ranking presents better AP values than the LMD model. This is due to the fact that re-ranking will look for the similarity of the question and the passage words/meanings and, therefore, getting better results than LMD. This is easily verifiable in the 34th conversation where the gains were about 0.5 between the two methods.

In addition, along the turns, as we can see with conversation 68 (**Fig. 2.**), the first turn, which is, in this case, “What cuisine is Emilia-Romagna famous for?”, tends to be the one with the highest value. This is because the first questions don’t usually have pronouns while in the following ones the use of those terms increases, increasing the difficulty to identify the subject, thus resulting in a lower value.

The same rationale can be applied to the re-ranking results. Although the Re-ranking performs better than the LMD model, the re-ranking is not also able to provide a bigger increase of the AP in the questions that have pronouns. This is something that needs to be improved, and will be in future developments as we can see in the table below with the 31st conversation where the improvements were not as big as in the majority of the conversations turns.

|  | **Is it treatable?**With pronoun | **What causes throat cancer?**Without Pronoun |
| --- | --- | --- |
| **LMD Model** | 0.0 | 0.12 |
| **Re-ranking** | 0.02 | 0.35 |
| **Future Developments** | > 0.02 | > 0.35 |

 **Table. 1.** Average Precision per turn on conversation 31 with and without pronouns

### Normalized Discounted Cumulative Gain (nDCG)

 

| **Fig. 3.** Mean nDCG comparison of all the conversations | **Fig. 4.** nDCG comparison on conversation 68 |
| --- | --- |

Similar to the AP plot for all the conversations, **Fig. 3.** presents the nDCG measure for the LMD model which has its highest peak at the 77th conversation, which talks about “Stews”, and its smallest peak at the 69th conversation, which talks about “Melatonin and Sleep”, while the majority of the points are between 0.05 and 0.2.

Doing a similar analysis, but for the Re-ranking, the highest value is measured in the 50th conversation, which talks about “Satellites”, in contrast, the lowest value is obtained in the 69th conversation (equal to the LMD Model). From **Fig. 3.** we can also check that the majority of them are between 0.36 and 0.58.

From this evaluation of the results, for the same reasons discussed on the examination of the AP, we can say that the Re-ranking presents better nDCG values than the LMD model. This is easily verifiable in the 50th conversation where the gains were about 0.5 between the two methods.

In addition, similarly to what happened on the AP (**Fig. 2.**), the first turn usually is the one with the greatest value and the following ones tend to have smaller values. But, as we can check by comparing the **Fig. 2.** with **Fig. 4**, the nDCG usually evaluates the success of the retrieved documents higher than the AP does. This is because the nDCG metric also considers the level of relevance of the document, while the AP only considers if the document is relevant or non-relevant.

Finally, the reason for the difference spotted between the LMD model and the Re-ranking in the nDCG measure is the same that was explained in the AP measure.

### Precision-Recall

###

| **Fig. 5.** Mean Precision-Recall of all the conversations | **Fig. 6.** 50th Conversation Precision-Recall |
| --- | --- |

In **Fig. 5.** we can see the differences between the mean precision-recall curves for all the conversations in both models. While the LMD model starts with precision values of around 0.1 of precision and abruptly decreases to around 0.03 as the recall values increase, the Re-ranking starts with values of precision around 0.15 and decreases to around 0.03 as the recall values increase in a smoother way. After analysing **Fig. 5.**, it’s possible to conclude that, in general, the precision-recall values for the Re-ranking are better than the LMD precision-recall values.

Both **Fig. 5.** and **Fig. 6.** are the practical representations 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.

# 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. Evaluation metrics allow us to verify and determine the characteristics of each model and how they perform given different examples and data.

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

**Average Precision (AP).** The results using the Re-ranking method after training the classifier are better than the results using only the LMD model.

**Normalized Discounted Cumulative Gain (nDCG).** Like the AP it presents better results in the Re-ranking model, 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 Re-ranking presents a more smooth curve that only starts decreasing when the recall values reach 0.6 or more.

**Use of pronouns in questions/queries.** The Re-ranking 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.

**Cross-Validation.** The use of cross-validation to train the classifier has a visible impact in the gains of the precisions scored by the AP and the nDCG measures.

**Natural language embeddings.** The use of BERT 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.

**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.