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 this project 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, Precision, Recall, nDCG, AP, Precision-Recall.

1. **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: retrieval and evaluation, natural language embeddings and conversation tracking.

1. **Retrieval and Evaluation**
	1. **Experimental methodology and Resources**

To retrieve information we started by reading the dialogue turns one by one and then computed the top 100/1000 candidate responses (text passages) using the Elasticsearch search engine. Finally, using the dataset ground truth, we measured the success of the used retrieval model by plotting graphs using the following metrics: AP, nDCG and Precision-Recall.

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

* 1. **Baseline Retrieval Method**

The retrieval method used is 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: $μ\*Mc(ti)$. This way we avoid having terms in a document with 0 of frequency.

* 1. **Experimental results**

After retrieving the top candidate responses we plotted some graphs to measure the success of the LMD model in the dataset.

 **Average Precision (AP).** The first metric we used to plot is AP which is used when there is full knowledge of the relevance judgements. It is calculated as the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved [1].

$AP=\frac{1}{\#relevant}\*\sum\_{k\in \{set of positions of the relevant docs\}}^{}p@k$[2]

| **Fig. 1.** AP comparison of all the conversations | **Fig. 2.** AP comparison on conversation 31 | **Fig. 3.** Mean AP per turn score |
| --- | --- | --- |

As we can see in **Fig. 1.**, the conversation with the highest AP is the 77th with a value of around 0.27. The conversation that has the lowest AP is the 40th with a value of around 0.01. The majority of the conversations have a value between 0.02 and 0.1.

The AP is a metric that traces the area below the precision-recall curve (analysed below) which means that most of the documents retrieved by the LMD model had a good precision on the 77th conversation, but had a low precision on the 40th conversation. At the same time, for the majority of the conversations, the precisions of the documents retrieved were close to the one measured in the 40th conversation, and so, not really good.

In general, along the turns, as we can check for the conversation 31 (**Fig. 2.**) and for all the conversations turns (**Fig. 3.**), the first turn tends to be the one with the highest AP which is normal because the first questions are usually less specific, but the following turns tend to become more particular, which means a lower AP.

 **Normalized Discounted Cumulative Gain (nDCG).** The second metric we used to plot is nDCG which is used when there is no full knowledge of the relevance judgments and there are different levels of relevance (nonbinary). It is based on the Discounted Cumulative Gain (DCG) which is obtained by summing all the divisions between the relevances and the logarithm of the relevances of the documents on the *m* top ranked search results. The nDCG is calculated by the quotient of the DCG with the bestDCG [1].

$nDCG\_{m}=\frac{DCG\_{m}}{bestDCG\_{m}}$with $DCG\_{m}=\sum\_{i=1}^{m}\frac{rel\_{i}-1}{log\_{2}(1+i)}$[2]
where *m* is the number of top search results to consider

| **Fig. 4.** nDCG comparison of all the conversations | **Fig. 5.** nDCG comparison on conversation 31 | **Fig. 6.** Mean nDCG per turn score |
| --- | --- | --- |

Similar to the AP plot for all the conversations, **Fig. 4.** presents the evaluation of the retrieval information using the nDCG. We can see that the nDCG was a little bit better evaluation metric than the AP as the percentages were generally higher, because it considers the level of relevance contrasting with the AP which only considers if the document is relevant or not.

In general, along the turns, as we can check for the conversation 31 (**Fig. 5.**) and for all the conversations turns (**Fig. 6.**), the first turn tends to be the one with the highest nDCG which is normal because the first questions are usually less specific, but the following turns tend to become more particular, which means a lower nDCG. The nDCG presents a more volatil variation of the precision and differs from AP as we can see on the turn 5 of the mean nDCG (**Fig. 6.**).

By analysing **Fig. 5.**, which shows the evolution of the nDCG along the 31st conversation, we can see that the first turn of the conversation got the highest score with a value of 1.0 which means that it got all the relevant documents for the turn.

**Precision-Recall.** The precision metric corresponds to the fraction of retrieved documents that are relevant, while the recall metric corresponds to the fraction of relevant documents that are retrieved. The combined use of the two metrics can provide us a quick overview of the balance of the system/method used [1].

For *n* search results we calculated the 10 points of the precision (*p@k*) and the recall (*r@k)*, as follows:

$k\in [0.1\*n, 0.2\*n, ..., 0.9\*n, 1.0\*n]$

And then interpolated the precision and the recall using *numpy.interp* to compute the precision values corresponding to the recall values in $[0.0, 0.1, ..., 0.9, 1.0]$.

| **Fig. 7.** Mean Precision-Recall of all the conversations | **Fig. 8.** Precision-Recall on conversation 31 |
| --- | --- |

From **Fig. 7.** it is possible to analyse that the precision is higher when the recall value is lower and vice-versa, being the highest variation between the 0.0 and 0.2 values of the recall. Analysing the plotted lines of the individual conversations (e.g.: **Fig. 8.**) we confirm the variation of these values as the precision-recall starts with a higher precision value of around 0.38, then decreases as the recall increases, and finally stabilizes around 0.19 after the recall hits the 0.2 mark.

This was the expected behaviour as there is usually a tradeoff between precision and recall values. When the precision value increases the recall value goes down and vice-versa. These graphs will allow us to quickly compare the different methods in order to choose the more desired one for the system being evaluated/developed.

* 1. **Conclusion**

### The evaluation metrics are important instruments to determine the success of the information retrieved by the search method. The AP is used when there is full knowledge of the relevance judgments and is good to determine the evaluation in a smoother way (binary relevances) while the nDCG is used when we don’t have a full knowledge of the relevance judgments and has a more volatile calculation as it considers the different document relevances (non-binary relevances). The precision-recall is the most complete way to determine the desired method for the system being evaluated as sometimes you need the most relevant results first (e.g.: on a web search) and other times you prefer all the documents related to the subject (e.g.: on an article search), even if the precision is lower.

**References**

1. C. D. Manning, P. Raghavan and H. Schütze, “Introduction to Information Retrieval”, Cambridge University Press, 2008.
2. Class Lectures of Information Retrieval, NOVA School of Science and Technology.