Group 5

Ana Carolina Baptista, ist195529
ana.p.baptista@tecnico.ulisboa.pt

António Martinho Marçal, ist195735
antonio.marcal@tecnico.ulisboa.pt

Ådne Tøftum Svendsrud, ist1108703
adne.svendsrud@tecnico.ulisboa.pt

ABSTRACT

This report delves into the development and evaluation of an Information Retrieval (IR) system designed to tackle keyword extraction and text summarization challenges. Leveraging the BBC News Summary dataset, our study embarked on a two-phased approach: initially employing unsupervised techniques for feature extraction and summarization using clustering approach, followed by an exploration of supervised methods to refine the summarization process through relevance feedback. The project showcases methodologies integrating traditional IR models to enhance the accuracy and efficiency of information retrieval. Our findings contribute to the understanding of effective strategies in IR systems, demonstrating the potential of combining unsupervised and supervised learning for improved content summarization and keyword extraction.

Keywords

Information Retrieval. Keyword Extraction. Text Summarization. Unsupervised Learning. Supervised Learning. Clustering Approach. Relevance feedback.

1. INTRODUCTION

In the first part of Project II, our goal was to investigate whether the clustering technique would be useful for the summarization and word extraction task developed in Project I.

The implemented solution is a sophisticated natural language processing (NLP) system focused on keyword extraction (based on the paper "Task 5: Single document keyphrase extraction using sentence clustering and Latent Dirichlet Allocation" by Claude Pasquier from the Institute of Developmental Biology & Cancer of University of Nice Sophia-Antipolis.) and document summarization. It uses advanced techniques including tokenization, stopword removal and punctuation to prepare data. Using the BERT model to generate embeddings (vector representations) of sentences allows the semantic context to be captured efficiently, facilitating the grouping of sentences with similar meanings.

The choice of the ideal number of clusters is made through silhouette evaluation, seeking to optimize the cohesion and separation between sentence clusters. This step is crucial to ensure that the grouping adequately reflects the different ideas or topics present in the document. After identifying the clusters, the system selects the most representative sentences from each cluster to compose the document summary, based on the Euclidean distance to the cluster's centroid, which effectively captures the essence of the original text without redundancy.

For keyword extraction, the code employs an innovative approach that also uses the BERT model to generate word embeddings. These representations are then clustered using the K-Means algorithm, with the number of clusters adjusted to maximize the silhouette score, similar to the sentence clustering process. Keywords are selected based on relevance to sentence clusters, which ensures that we capture meaningful terms that represent the document's main topics and ideas.

Additionally, the system includes an evaluation function that measures the effectiveness of summarization by comparing the generated summaries with reference summaries using metrics such as F-Measure, Precision, and AUC. This evaluation is essential to understand the system's performance and to make adjustments that can improve the quality of the abstracts and extracted keywords.

2. Clustering-guided summarization and efficacy of the IR system

Clustering significantly alters the effectiveness of the information retrieval (IR) system, especially in the activities developed in this project, such as keyword extraction and sentence summarization. It divides the document into relevant topics, grouping sentences that are similar to each other. This facilitates the understanding and comprehension of texts, covering everything that is important and minimizing redundancy, which is crucial to avoid excessive repetition of information and ensure that the document is concise.

With clusters representing distinct topics, it becomes easier to extract keywords or key phrases from each cluster to summarize the document's content. These keywords or key phrases provide an overview of the topics covered in the document, allowing for quick identification and retrieval of relevant information.

Furthermore, clustering provides an organized structure for the document, facilitating navigation and exploration of the content. Clusters can be presented as categories or sections, allowing users to navigate directly to topics of interest. This enhances the user experience and increases the effectiveness of the information retrieval system.

3. The impact of sentence representations, clustering choices and rank criteria on summarization

When we group sentences into clusters so that those that are redundant or very similar are together, we can significantly improve the quality of the summary we are trying to generate. This happens because, when grouping similar sentences, we can choose only the most representative ones from each group, avoiding unnecessary repetitions in the final summary.

For this grouping to be effective, we need a way of representing sentences that truly captures their essence and allows similar sentences to be identified as close to each other. The BERT model is known for its ability to generate rich, contextualized representations of text, which makes it a good

choice for this task. However, a version of BERT specially tuned for understanding sentences, known as Sentence-BERT, could be even more effective, as it is optimized for this specific type of comparison.

Within the same cluster, the way we choose the most representative sentence is generally based on its proximity to the cluster's centroid, calculated using Euclidean distance. However, other methods of calculating similarity, such as cosine similarity, may offer better results by focusing more on the orientation of vector representations than on their magnitude, thus capturing the similarity in a more subtle way between sentences.

Furthermore, we could further improve our summary by considering the possibility of selecting more than one sentence from particularly relevant clusters, or even completely discarding clusters that are considered irrelevant for the purpose of the summary. These strategies have not been explored, but are certainly worth investigating and discussing, perhaps in a future report or analysis, as they can offer valuable insights into how to improve the quality of text summarization.

4. Anchor sentences and outlier sentences

Anchor sentences are those that capture the most information about the document, meaning that, in an ideal ranking of relevance, these would yield the highest score. This fact makes them prime candidates for the final summary and should, in theory, be selected every time during the summarization task. In particular, in the clustering scenario, anchor sentences would be great choices as centroids for the clusters and should necessarily be included in them.

However, upon inspecting our results, we came to the conclusion that our IR model is only partially successful at including anchor sentences. This can be inferred from the underwhelming results obtained in terms of precision and F-Measure, which were within a 2% difference from the baseline summarization model from Project 1.

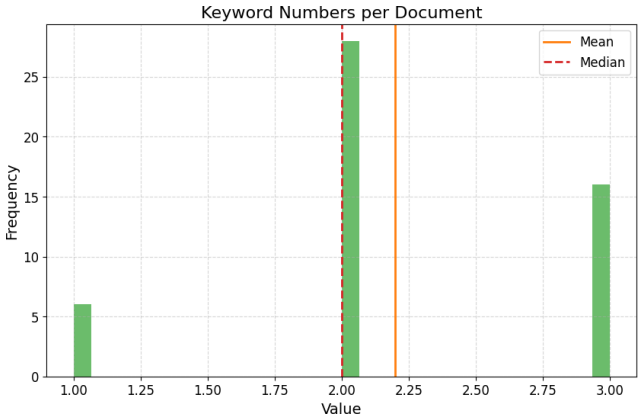
Manual inspection of some specific examples also demonstrated that sentences with the highest tf-idf scores were only included roughly 70% of the times. Although, this must not be taken as an absolute judgement of the inclusion of anchor sentences, since it is a very complex task to determine what those sentences are or even if they exist at all and tf-idf scores may not be the best way to measure a sentence's status as an anchor sentence.

As for outlier sentences, these are those that offer little in common with the other sentences, therefore complicating the clustering task. A well-implemented clustering algorithm should always identify these outliers and exclude them from the clusters, specially since the main objective of a summarization task is to find sentences that are closely related to the overall document. Being highly distinct (i.e. an outlier) is the opposite of this definition.

As mentioned before, the precision and F-measure scores were not very promising, which suggests that our IR model might not be excluding these sentences as much as it should. However, it is important to note that this iteration of the model using clustering produced significantly shorter summaries, even though we did not limit it to a fixed number of characters or sentences. This suggests that there is some possibility our model might be successfully identifying less relevant outliers and not including them in the final summary. This possibility is strengthened when we take into account that the overall performance scores were similar to before, despite the shorter summaries.

5. Keywords across documents

In order to further our understanding of these clustering solution for the keyword extraction task, we decided to compare how many keywords were being identified in each document and if that number had significant variations across multiple categories and individual texts. For that purpose, we developed to plot below:



Due to computational limitations, we were only able to extract keywords for 50 documents, distributed evenly across all genres, within the provided deadline. Despite this setback, we believe the results to be somewhat representative of the performance of the algorithm.

Analyzing the plot, we can see that at least for the dataset used, our algorithm always extracted between 1 and 3 words, with the average being around 2.2, since for the clear majority of documents, 2 keywords were identified. This numbers are directly tied to the number of clusters selected, which suggests that the algorithm is very unlikely to identify more than 3 clusters. This is due to the short length of the news articles, which include a small number of sentences only. Therefore, the number of keywords is appropriate for the task at hand.

Manually inspecting some of the extracted words, we decided to compare them, putting special focus on their similarity. The list below demonstrates the outputs for some randomly selected documents:

- ['taster', 'hold', 'work'],
- ['confirmed', 'spring'],
- ['current','bbc'],
- ['coalition', 'dries', 'live'],
- ['telco', 'outside-in', 'centres']

As can be easily observed, the keywords extracted are all quite different from each other, showing an extremely low level of similarity with other key words in the same document and even across documents. This low redundancy may be due in large part to the extensive pre-processing applied to the documents, which makes sure to normalize words so that similar terms are interpreted as the same, and consequently included only once in the results for the keyword extraction.