Natural Language Processing to Improve Information Retrieval for Meeting Minutes Written in Brazilian Portuguese

Ovídio José Francisco
Professor at Facens
Sorocaba, Sao Paulo, Brazil
ovidio.francisco@facens.br

Abstract

A multi-thematic document is composed of many subjects in a continuous text. In this context, a meeting minute concentrates very important information such as guidelines and decisions. Therefore, it is often used as information source. Most information retrieval techniques do not deal well with multi-thematic documents, once it has non-structured data, lack of metadata and little subject delimitation. Furthermore, it is hard to assign them the main subject or point to a specific snippet. Here we had two main challenges. First, knowing where a subject goes to another, and second, how to identify them. To solve these needs we used text segmentation and topic extraction methods. The text segmentation technique splits a document into segments where each part contains a coherent subject, while topic extraction methods aim to group and describe them. Many researchers evaluate these methods using long texts such as concatenation of documents or transcriptions of discourse and multipart meetings written exclusively in English. In this work, we collected a *corpus* of meeting minutes written in Brazilian Portuguese which besides being the language less studied, has a more formal and succinct style. As a result, it generates a structure formed by segments represented by descriptors and grouped by topics which adds extra information about the subject that each segment deals with. Finally, we present a method to connect the text segmentation and topic extraction methods to improve the performance of information retrieval techniques as well as provide an annotated *corpus* for this domain.

Keywords

Multi thematic, Text Segmentation, Topic Extraction, Information Retrieval.

Biography

Ovídio José Francisco starts his academic career with a technician course in Informatics from ETEC Salles Gomes (1999) followed by an undergraduate degree in System Development and Analysis from FATEC in 2014. In 2018, finished the master's course in Computer Science with Machine Learning. Began as professor in 2015 at Faculdade de Tecnologia in the city of Itapetininga. In 2019 moved to UniFacens where teaching disciplines in computer science, system development and game development courses. Currently is engaged in natural language processing and correlated projects.