CPSC 490 Senior Project Proposal: Document Cutoff Matching

Abstract

For my senior project, I will be conducting research with Professor Radev related to the creation of a multilingual parser – a search engine that takes in queries in one language and manipulates them in order to retrieve pertinent documents in other languages. These documents and results are then either returned in the first language or second language depending on the selection of the user. The main goal of my research on this project will be to create a document cutoff system that only returns a certain number of results deemed pertinent to the query, with additional work relating to improving the overall search functionality. Implementing this requires an understanding of the strategies of information retrieval – using both heuristics and linguistic modelling to discover features of documents linked to a query. More recently, researchers have successfully used deep learning to successfully predict which documents ought to show up when performing a search. However, there has not been an investigation yet as to how the task of searching in one language and finding information in another may interact with the document retrieval and cutoff process. Furthermore, while there has been some success in creating classifiers with neural networks to determine when to cut off document retrieval, there has been relatively little research into other potential methods that might yield better results (convolutional neural networks, imposing exploration/exploitation constraints, etc). In my project, I will evaluate current models and test out different deep learning strategies to see which seems best. I have some experience with artificial intelligence, but I have not built more advanced forms of neural networks. In addition, I am unfamiliar with search engine optimization
and the application of deep learning to information retrieval, topics I hope to become more comfortable with over the course of this project. Professor Radev is well-known and experienced in the field of applying methods from AI to natural language processing tasks, and his guidance will help me learn and create a better project. The deliverables we have discussed are as follows:

**Deliverables**

Initially, the primary focus of my research will be on creating the document cutoff system. The multilingual parser is part of a project that will be evaluated on the strength of its results. False positives, or documents that are retrieved by the search engine deemed unrelated to the query, will be penalized when the parser is evaluated. As a result, a crucial step in this project is making sure that the search only returns things the user would find interesting. This task is split into two steps. First, I must apply certain methods that analyze the relatedness of the results to the query. Some of the approaches I mentioned above involve using heuristics, language modelling, and deep learning. I anticipate using a combination of these methods; that is, creating a deep learning framework that accepts features related to the query as inputs and outputs a score for each document in the database. The engine will then deem the documents with the highest score as the most valuable, and place greater importance on them. Creating this framework will involve using tools like Google’s TensorFlow to create the neural network and the nltk toolkit for modelling the data, experimenting with different types of networks for better results.

Second, once I have received the scores, I have to construct some system that determines which results should be shown to the viewer. This will largely depend upon the results from step
one. If I use a classifier like softmax, which will likely be one of the first systems I test, then I can create a ranking system that also reflects the system’s confidence that the result should be included. This could allow me to dynamically decide which results should be included based off of a certain confidence threshold. Alternatively, it may suffice to use a linear classifier and determine just include/exclude from the data. I anticipate testing several models as the linguistic translation element of the search adds a wrinkle to previously accepted results.

Once these steps are done, the feature will be incorporated into the broader project. I may work on other aspects, including the contextual translation of the search, web programming for the final presentation, and any other steps necessary to improve the functionality of the project.