Lab 4: Information Extraction
Reflection Report

TDDE16 – Text Mining

Anonymous ID
11895

IDA
Linköping University

November 27, 2023

RQ 4.1: *In Problem 3, you did an error analysis on the task of recognizing text spans mentioning named entities. Pick one type of error that you observed. How could you improve the model's performance on this type of error? What resources (such as domain knowledge, data, compute, ...) would you need to implement this improvement?*

Looking at the error report, we quickly found that the model had a hard time to distinguish between different types of numbers. For example, sometimes numbers refers to years, and in other contexts, they refer to how much something costs (money) or percentages. Numbers can also be written with commas or punctuation, which seems to confuse the model further. The problem here is however that spaCy picks up a lot of labels that the gold data has not used. So, in the lab, we chose to exclude the label 'CARDINAL', which refers to numbers that do not fall under any other type, to improve the model.

However, we also noticed other problems that were instead linked to the way the model predicted the spans. For example, when the model predicts 'The European Commission' while the gold data span was 'European Commission'. For this, more data could be needed in order for the model to improve and understand how to predict the spans correctly.

RQ 4.2: *What does the word "context" refer to in the context of Problem 6? How does this help to disambiguate between different entities? Suggest another type of context that you could use for disambiguation.*

The context refers to the words around the mention. This provides information about the surroundings of a word and helps to distinguish different words that might be ambiguous. As exemplified in the lab, the mention 'Lincoln' can refer to the city Lincoln in Nebraska, or the person Abraham Lincoln, depending on the context.

In the lab, (in the ambiguous case) the sentence the specified mention is in is used as context to better distinguish what label should be predicted. However, other types of context could be used as well. For example, one could use n-grams, sentiment analysis or noun-phrase extraction for the same purpose. Noun-phrase extraction use part-of-speech to identify the important words and their relation to each other. Nouns are usually subjects or objects, while verbs and adjectives can explain their relations. This can then be used as context data to help disambiguate between different entities.

RQ 4.3: *One type of entity mentions that we did not cover explicitly in this lab are pronouns. As an example, consider the following sentence pair:*

> *"Ruth Bader Ginsburg was an American jurist. She served as an associate justice of the Supreme Court from 1993 until her death in 2020."*

What facts would you want to extract from this sentence pair? How do pronouns make fact extraction hard?

From this sentence pair, you would want to extract that the woman Ruth Bader Ginsburg was an *American Jurist*, and served as *associate justice of the Supreme Court* from *1993 until 2020*. What makes this difficult is the use of "She" and "her" which can create ambiguity, especially in sentences or sentence pairs which includes multiple people. Moreover, the model has to understand that "she" and "her" links back to Ruth Bader Ginsburg, to actually understand that it is Ruth Bader Ginsburg that was an associate justice of the Supreme Court from 1993 until 2020.