Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Shadasha Williams		
Název práce	Named Entity Recognition in the Biomedical Domain		
Rok odevzdání	2021		
Studijní program	Computer Science	**Studijní obor**	Mathematical Linguistics

Autor posudku	doc. RNDr. Pavel Pecina, Ph.D.	**Role**	vedoucí
Pracoviště	Institute of Formal and Applied Linguistics		

Text posudku:

The submitted thesis of Shadasha Williams deals with the task of Named Entity Recognition (NER) applied to the texts from the biomedical domain. NER in this domain is very specific due to the variety of types of the entities (e.g., diseases, drugs, body parts). The goal of the thesis was to explore this challenging research field (inc. state-of-the-art methods, datasets, etc.) and to deliver a solution that would perform not only recognition (*aids* vs. *AIDS*) but also disambiguation of mentions that link to different entities (*LT* for *lung transplant* vs. *liver transplant*). The author focuses on two types of entities (disease/gene), designs and performs several experiments with recent NER methods and several types of word embeddings on data from PubMed, automatically tagged by a string-matching algorithm using two sets of entities (disease/gene) from existing resources.

The thesis is written in English on 55 pages (the main content) structured into four chapters. Chapter 1 is a broad introduction to the thesis, Chapter 2 reviews the task of NER, Chapter 3 describes existing approaches to word embeddings. The final Chapter 4 presents the author's main contribution (experiments). The thesis also includes a list of references, tables, and figures.

From the language point of view, the text is well written, with only occasional grammatical errors or typos (e.g. missing words, sentence fragments).

The structure is not very well balanced, though. Mainly Chapter 4 (entitled Methodology) contains description not only of the methodology but also experiments, results, and conclusions. The captions of figures and tables are often (mostly in Chapter 4) not provided with complete information. Moreover, the information is also missing in the main text (e.g. specification of training/validation/test data splits, see bellow). The references are (to some extent) not complete and/or not chosen properly (e.g. outdated). The last two pages contain an empty list of abbreviations and an empty Appendix A.1 entitled First Attachment.

Contentwise, the work is not optimal either. The introduction in Chapter 1 is too broad, contains not very relevant information (e.g. Figure 1.1 on language stratification or the discussion on Swanson's ideas about "undiscovered public knowledge" in Section 1.3.1). Chapters 2 and 3 are theoretical, containing overview of NER and word-embedding methods. The way, how this is presented is not appropriate. The level of knowledge required to understand the text is very high

(many terms are not explained) and even for a reader with such a level of knowledge, the text is not clearly written, difficult to follow and even confusing. For a non-expert, the text is probably not useful at all.

The content of Chapter 4 is also not adequate. Mainly, the description of the experiment settings and results is not sufficient. E.g the data description and how it was split for training/test/validation is not clear (does the second paragraph on Page 39 describe the training data part or the entire data set? What do the numbers in the first paragraph on Page 42 refer to? What does the last sentence mean? "I used the first 10 files as the data for this experiment." Which experiment?). The figures/tables are not correctly referred to and it is not clear which experiment they correspond to. What does "and increasing the data" in the captions of Figure 4.2 and 4.3 mean?

The main issue of the work is the methodology. Basically, the models used in the experiments were trained to "mimic" the Aho-Corasick string-matching algorithm. The author does not provide any reasoning/hypothesis why/how this could be an effective approach for NER (eventually better than Aho-Corasick). No comparison of this approach with other (state-of-the-art) methods is provided. As mentioned above, the data split is also an issue. It is not clear, how it was done and (more importantly) whether the experiments were conducted correctly.

The last issue is that one of the goals of the thesis (regarding NE disambiguation) was not met. Basically, no effort was done in this direction. The author claims that she "found that very few diseases have this type of ambiguity" but they do exist (mainly abbreviations, see e.g. *Davis N. Medical abbreviations that have contradictory or ambiguous meanings. In ISMP Medication Safety Alert! Acute care edition!, 2020*) and more importantly, the author was not limited to this type of NEs in her work and other types of NEs could have been examined too.

Questions:

1) How did you split the data for training/validation/test in all the experiments? Why the data split was not the same for all experiments? How can you compare their results?

2) Can you compare your results with the current state of the art on the NCBI dataset? What is the performance of the Aho-Corasick algorithm applied to the NCBI benchmark?

3) What is the main finding/outcome of your work. Is your approach, effective?

Despite the issues mentioned above, the thesis fulfills the formal requirements and I recommend it for defence and hope the above mentioned questions will be answered.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 28. 8. 2021 Podpis: