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	Informatika	**Studijní obor**	Matematická lingvistika

Autor posudku	RNDr. Jana Straková, Ph.D.	**Role**	oponent
Pracoviště	Institute of Formal and Applied Linguistics		

Text posudku:

The diploma thesis "Named Entity Recognition in the Biomedical Domain" is submitted for review.

The thesis is written in grammatically correct English and was clearly written independently. The typesetting is adequate.

The reviewer's objections are manyfold, as the thesis fails to deliver in all its aspects:

First, the writing itself, although grammatically correct, is imprecise, vague and betrays only superficial understanding of the topic. This is especially felt when exact equations are given. Although all the formulas appearing in the thesis are obviously copy-paste reproductions of previously published work, there are numerous errors introduced. Also, the citation norm in academic publishing is violated constantly throughout the thesis.

Second, Chapters 2 and 3 constitute solely of previous works adoptions, to a greater degree than is usual in a Related Work section. All figures and equations are copy-pasted and the text is a close reformulation of the sentences found in the respective publications. Although referenced and cannot be considered plagiarism, there is no original contribution. Moreover, due to the large amount of introduced errors and general confusion, reading this text will do more harm than good to an inexperienced reader.

Third, there is a critical flaw in the newly created data, which makes the corpus completely useless. Namely, the named entity spans are not annotated (see Table 1 for an example). As a consequence, the task is approached and evaluated as a per-token classification task, not as a sequence classification task. Needless to say, the task dealt with is NOT named entity recognition.

Foremost, the contribution of the experiments is of no value. The author created a new dataset by automatically annotating string-matched keywords and fitted a neural model to such data. The numbers achieved by training and then testing on such corpus are meaningless. The only way to assess the real contribution would be to leverage the newly created data to match or improve results on another standard benchmark. An attempt was made to test the model (trained only on the new data) on the NCBI Disease Corpus (Doğan et al., 2014) but not a single publication was given for comparison. However, a NCBI-disease corpus leaderboard[1] can be easily found. The current state of the art is 89.71% F1-score, with all listed results reaching well above 85%. The

[1] https://paperswithcode.com/sota/named-entity-recognition-ner-on-ncbi-disease

Sequence tagging (NER) BIO encoding		Per-token classification No encoding	
However	O	However	O
,	O	,	O
in	O	in	O
the	O	the	O
acute	O	acute	O
myeloblastic	B-DISEASE	myeloblastic	DISEASE
leukemia	I-DISEASE	leukemia	DISEASE
,	O	,	O
a	O	a	O
significantly	O	significantly	O
higher	O	higher	O

Tabulka 1: Named entities tagged with a BIO tagging to keep track of sequential information and without any encoding, losing named entities span information.

baseline result published with the corpus is 81.90% (Doğan and Lu, 2012). The thesis best result falls far behind with 53%.

To conclude, I have found no contribution in the thesis.

Despite the above mentioned serious objections, the thesis fulfills the formal requirements and I therefore recommend it for defence.

Questions

1. The methodology description is confusing to such a degree that it is not clear which data was used for hyperparameter tuning and which for testing. There is a substantial suspicion, though, that the hyperparameters were tuned to fit the testing portion, because the results remain unchanged from the previously withdrawn thesis version. Can you explain which parts of the data were used for hyperparameter tuning and which for testing? How is it possible that the numbers did not change between the previous and current version? Were the results measured correctly again or was only the offending information silently removed from the text?

2. Why do you think the results are so far behind the current performance, despite using a standard architecture? Is it the method, the data or the implementation? Can you design an experiment which would validate the implementation of your models?

Detailed Review

Front Cover and Formal Introduction

The supervisor's correct titles are "doc. RNDr. Pavel Pecina, Ph.D.".

Thesis Introduction

Generally, it is maybe not the lack of research, rather than the poor navigation in the researched space. Sometimes, an irrelevant or outdated publication is cited in detail when more recent works or higher impacted ones are available. Terms are used vaguely, sometimes in a wrong or confusing context.

Should an excourse into formal linguistic theory be attempted, more research and more space is needed to cover the topic. Why was one specific lingustic theory (Halliday and Matthiessen, 2013) chosen and not some other? Why is it more relevant for your work than other formal theories? And how is it relevant?

Named Entity Recognition

Overall, this chapter is better than the Thesis Introduction, because it deals with specific rather than general notions. However, rather irrelevant or outdated papers are discovered and discussed in detail where not necessary. Again, this section, although much better than the previous one, gives the impression of a compilation of sentences selected from random publications without experienced guidance as to their relevance and impact.

There is an error/typo on p. 12, the CRF formula, instead of $P(y|z)$, it should be $P(y|x)$.

Also, exp should be typesetted as exp in mathematical mode (everywhere).

The LSTM explanation on pages 13 and 14 is confusing to such a degree that it is not easy to say whether it is correct or not, but probably it is not. Constants, vectors and matrices notations are confused.

On p. 15, equations are rewritten incorrectly from (Lample et al., 2016) on p. 15 (Y substituted for y in equation (1) of (Lample et al., 2016)). Also, although the preliminary lemmas are copy-pasted, the main resulting equation is omitted.

On p. 15, Figure 2.4 does not use the IOB schema despite claiming so.

On p. 16, binary classification definition is given for precision, recall and accuracy. How is this useful for NER, a multiclass retrieval task?

Furthermore, accuracy is not used as an evaluation measure for NER, although claimed so on p. 16.

Only one related work on NER in deep learning is given (Lample et al., 2016) and none at all in biomedical domain.

Word Embeddings

There is no contribution in this chapter, as all of its content is a close reformulation of the respective publications, including copy-pasted formulas and equations, with no effort to bring any new idea or higher-level understanding of the presented material and with confusion and errors introduced.

Methodology

This section might be splitted (it usually is) to two sections: Methodology and Results.

On p. 38, reference to PubMed (Sayers et al., 2018) missing, which is awkward, given that it was the single data source for the created corpus.

On p. 39 a standard NCBI Disease Corpus is introduced as a testing benchmark. It should be stated how large it is, how many tokens it has, what ontology of diseases was used and what is the domain coverage in comparison with the newly presented data.

There are two experimental settings (Keras and Flair), each of which uses a differently sized data split. How can the experimental results be directly compared?

Also, there is a confusion as to the test/validation splits. Which was used for hyperparameter tuning and which for testing? It is not explicitly said and the two are being used rather freely. In many figures/tables, it is not said which portion data is being referred to.

Adam optimization not referenced (Kingma and Ba, 2015).

The captions of tables and figures in Section 4 are unclear. It should be always stated which experiment is measured (Keras, Flair + modifications), what ontology (disease, genes) and which data (validation, test).

Which experiment and data is displayed in Figure 4.2? And why is the line decreasing so badly? Why is it not decreasing in Figure 4.3, if I assume it is a similar setting on different ontology? Also, Figures 4.2 and 4.3 are displayed but never referenced from the text.

On p. 43, it is said that Figure 4.6 and 4.7 are similar to the first experiments. How similar? In what way? And which first experiments? Did you mean Figures 4.2 and 4.3?

On p. 47, you claim the BERT word embeddings F1 score to be in 80th percentile, while you mean to say it is above 80% (percentile vs. percentage).

Reference

Rezarta Islamaj Doğan and Zhiyong Lu. An improved corpus of disease mentions in PubMed citations. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 91–99, Montréal, Canada, June 2012. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1 – 10, 2014. ISSN 1532-0464.

M.A.K. Halliday and C.M.I.M Matthiessen. *Halliday's Introduction to Functional Grammar*. Routledge, 2013.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

Eric W Sayers, Richa Agarwala, Evan E Bolton, J Rodney Brister, Kathi Canese, Karen Clark, Ryan Connor, Nicolas Fiorini, Kathryn Funk, Timothy Hefferon, J Bradley Holmes, Sunghwan Kim, Avi Kimchi, Paul A Kitts, Stacy Lathrop, Zhiyong Lu, Thomas L Madden, Aron Marchler-Bauer, Lon Phan, Valerie A Schneider, Conrad L Schoch, Kim D Pruitt, and James Ostell. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 47(D1):D23–D28, 11 2018. ISSN 0305-1048.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 12. 8. 2021

Podpis: