Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Vinit Ravishankar
Název práce	Parsing of Texts with Code-Switching
Rok odevzdání	2018
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	David Mareček **Role** oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

The goal of this thesis is to apply a dependency parser on code-switched texts when only monolingual training treebanks are available. The second goal is to predict the points in a running text where the code is going to be switched.

The thesis is divided into 4 chapters. The first one describes prior works: Universal Dependencies project, differences between transition- and graph-based parsing, statistical and neural methods, parsing evaluation, and existing annotated corpora and methods for texts with code-switching. The second chapter describes the neural dependency parser by Dozat and Manning (2016), including the background theory of neural networks (neurons, backpropagation, gradient descent, word embeddings, recurrent networks, LSTM). The author shows the architecture of their parser and the way he reimplemented it in Py-Torch.

The third chapter is the main one describing the parsing experiments on the code-switched texts. First it shows statistics of the testing treebanks Hindi/English and Komi/Russian and sets the naive baseline using concatenated monolingual training treebanks. Then it continues by showing many methods how to improve the results: a) mapping of the word embeddings of the two languages into one space, b) introducing artificial code-switching in the monolingual treebank, c) adding the language ID to each word-embedding, d) multitask learning with additional prediction of the language on individual words, e) multitask learning with additional prediction of binary label monolingual/multilingual for each sentence, f) development weight learning.

The fourth chapter describes automatic prediction of code-switch points. It shows four different network architectures with unidirectional/bidirectional LSTM, with using or not using language IDs of words. Evaluation is done on Hindi/English and German/Turkish data.

In the conclusion, all the results are shown in one table and discussed.

The thesis is 75 pages long including 9 pages of references. It is written in English, however, it is often very hard for reading. The sentences are sometimes very long and complicated for the reader to understand them for the first time. The theoretical background is very well described in detail, however, in the experimental part (chapters 3 and 4), there are some weaknesses, for example:

- There is no examples of code-switched data, nor any examples of well or badly analyzed sentences.
- There are many tables with results during the text, however without any comparison to the previous results or baselines. The overall Table 4.6 in Conclusions have often very short labels, so the reader does not know what experiment it is.
- I was not able to understand Section 3.4.4 - Development weight learning.
- It is not described how scrambling of language IDs work.

Overall, the student proved that he can design, implement, and evaluate various neural-network architectures. He proposed a lot of improvements to the baseline methods and showed that even without training on code-switched treebank he can achieve comparable results at least according to unlabeled attachment score. Compared to the baselines, the best results were about 2% better both for UAS and LAS.

Questions and comments:

1. In code-switch points prediction, you use bidirectional LSTM applied on word embeddings. That means that the network have access to the following word and therefore code-switch prediction can be simply learned by language identification of the following word. And I assume that is why the results on unidirectional LSTM (the only one which predicts the code-switch without knowing the future) are much worse. Could you comment on it?

2. Why is the size of language ID embeddings in Section 3.4 so high? (100). I thought that there are only two possible tags for each dataset.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 4. 9. 2018

Podpis: