Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Adedayo Oluokun
Název práce	Creation of a Dependency Treebank for Yoruba using Parallel Data
Rok odevzdání	2018
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	Rudolf Rosa **Role** oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

Introduction

In her master thesis, Adedayo Oluokun tries to solve a real and very current problem of devising a dependency treebank for Yoruba, a heavily under-resourced language from the Niger-Congo language family with 30 million speakers.

Positives

The author takes a very sensible approach to solving the problem, based on well established methods for cross-lingual induction of part of speech tags and dependency trees, namely the projection methods of Yarowsky et al. (2001) and of Hwa et al. (2005). She also cleverly makes use of the available datasets for Yoruba and for other languages, such as the multi-parallel Bible corpus, the manually annotated Bible texts in the Universal Dependencies treebank collection, and other UD treebanks.

The author does not devise any new cross-lingual projection methods, but she seems to have reimplemented the established methods correctly, using up-to-date tools and datasets. She also carried out a range of experiments to explore the effect of varying certain parameters of the setup, such as the choice of the source languages to use in the projection, or some of the hyperparameters of the tools.

The evaluation of the approach is reasonably thorough, including both automatically computed scores, as well as insights coming from manual inspection of the outputs and explaining some of the observed results. The results are not great in absolute terms, but are very respectable given the setting of processing a resource-poor language very distant from any of the available resource-rich languages. In fact, especially the tagging accuracies are considerably higher than what I would have expected, given my experience in the field.

Moreover, the author does not stop here, but also takes the next logical step of manually annotating a small Yoruba treebank, using it both for evaluating the cross-lingual approaches, as well as for training supervised tools and evaluating them via cross-validation. (As it often happens, the standard supervised tools trained even on this very small treebank outperform the complex cross-lingual approaches.) The treebank itself is a very valuable "by-product" of the thesis, and has already been accepted and published by the Universal Dependencies project.

The author thus covered the problem of obtaining a treebank for Yoruba quite comprehensively, taking *both* of the two established approaches, i.e. using cross-lingual methods, and using manual annotation. Moreover, she achieved all of this basically starting from scratch, with no tools and no annotated data for Yoruba.

Negatives

While the work the author carried out for her thesis is great, sadly the text of the thesis is not.

The general structure of the thesis is sensible, but the organization of the text within this structure is sometimes chaotic. There is a clear attempt to separate texts about tagging and

about parsing, but this separation is incomplete, often skipping from tagging to parsing and back again without a clear sense, e.g. in the Related work chapter. This is also true for evaluation, as there is a separate evaluation chapter, which starts off saying it will contain tagging evaluation, but then switches to talking about parsing only – tagging was already evaluated in the tagging chapter – but then suddenly does include some tagging results as well; which, moreover, are in disagreement with the tagging results reported earlier in the thesis, and it is totally unclear how these scattered evaluations relate to each other.

There is a large amount of examples, which I like, but it is not always clear how these relate to the texts, what they are illustrating, and some are really hard to understand for a non-speaker of Yoruba due to insufficiently detailed English glosses and accompanying texts (many examples are merely listed in the text without any explanation, and a few examples are not even referred from the text at all). For examples of sentence alignment and projected annotations, it is practically never explained to what extent these are correct.

This is also true for the first chapter, which introduces the Yoruba language to the reader. While I enjoyed reading this chapter very much, I did not understand many of the points – e.g. serial verbs, where it is not explained what the verbs mean if they stand on their own, if anything, or repetitive verbs, where it is unclear if they mean anything when not repeated, or "noun qualifiers", which are described as being similar to adjectives but categorized as adverbs without any explanation, or compounding with something marked "NEG" which looks like there is a negation marker somewhere but this is not explained in the text or labelled in the examples... Moreover, while the chapter does give the reader a general idea about the language and lightly motivates the choice of other source languages for projection which are similar to Yoruba in some aspects, I do not see any other clear connection between the information in this chapter and the rest of the thesis. On the other hand, some Yoruba phenomena are mentioned later in the text, especially function words which are hard to assign a part-of-speech label, but such examples are not present in this chapter.

The Related work chapter, although chaotically organized, has good content, reviewing many relevant works. However, it is hard to see a clear connection with the rest of the thesis, as the approaches that the author eventually uses are among the "early approaches"; the author also reviews many newer methods, clearly describing some of them as well suited for the setting of this thesis and outperforming the classical methods that the author eventually uses. It is thus not clear why the author did not use these newer methods, or, at least, to what extent the author believes in the potential of applying some of those methods to Yoruba in future work.

On a lower level, the text is not of a good quality either, containing many broken sentences with missing words, extra words, and presumably different than intended words, sentences which look like a start of one sentence connected to the end of a different sentence, and generally a lot of sentences with unclear meaning. Some paragraphs are somewhat confusing, jumping from one topic to another without a clear structure and message. Due to this, I had trouble understanding some parts of the thesis, including some aspects of the work that the author did.

On an even lower level, there is a number of errors in English (spelling, word order, syntactic structure) and other typos. The typographical quality of the thesis is also quite low, with a disturbing way of typesetting literature references, inconsistent and sometimes broken typesetting of tables, many tables and figures floating away from the text which refers to them, inconsistent indentation of paragraphs, bad number formatting, missing/extra spaces, visible line overflow markers, etc.

In general, the writing skills of the author are rather poor. Moreover, it seems that the thesis did not really go through any proof-reading; some parts look like they have not been proof-read even by the author.

I still believe that the work behind the thesis is great, and this is what should matter most in my opinion, but I am quite disappointed by the text of the thesis, which occasionally makes it hard or even impossible to fully understand the work done. I thus do recommend the thesis for defense, but admittedly with some reservation.

Questions for defence

I would like the author to explain the following points:

1. Some of the reported results are unclear, especially for cross-lingual POS tagging, for which you report 87.86% on page 35, and then 76.71% on page 53. Supervised POS tagging is also unclear, as on page 53, you report two numbers above 80%, but then another result of 72.76% labelled as "tested on UDPipe test data" which I have no idea what means. Could you clearly compare and contrast all the tagging accuracies together with an explanation of how the experiment settings differ for them?

2. Why do you project the POS on type level and not token level? You admit that this is the cause for mistagging of words which can belong to multiple POS labels, and this seems to be a major problem of your approach. So why stick to it? And why did you choose it in the first place?

3. On page 36, you mention that some words can have multiple POS tags – this is no surprise, many languages have words that belong to one POS in one sentence and to a different POS in another sentence. However, you go on to imply that for some words, it is often hard to decide to which POS they belong even in a fixed sentence ("the behaviour of these function words are not clear in a lot of instances"). Could you provide a few examples of sentences with words that are hard to label with POS?

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Berouně dne 4. 9. 2018

Podpis: