Diploma Thesis Review

Thesis title: Universal POS Tagger

Thesis author: Duong Thanh Long

Opponent: Zdeněk Žabokrtský

Thesis description

The main aim of the thesis is to develop a tagger that is capable of applying POS information available for a resource-rich language onto a text in a resource-poor language.

The thesis is structured as follows. The introductory chapter briefly describes the task of part-of-speech tagging and the motivation for other than traditional supervised approaches. The second chapter sketches basic tagging techniques and shows examples of approaches for multilingual tagging that can be found in the literature. The third chapter is the core of the author's own contribution: it describes the newly developed tagger (called universal tagger), which is based on two steps: first, projecting POS information from the source language to the target language via GIZA++ alignment (or at least the pieces of information which can be projected relatively reliably), and second, growing the POS assignment on the target side using self training. The third chapter also presents experiment evaluation, which reaches the state of the art. The fourth chapter elaborates how the choice of the source language influences the target language tagging performance, how this choice can be optimized, and even how multiple source languages can be combined. The fifth chapter concludes. The thesis consists of 67 pages.

Comments

The thesis is well structured and is written in very clear English, with only a few language errors (usually in subject-predicate agreement such as in "where the adjective are missing...", "Keeping these alignment would..."). Formally, the quality of the text is very high too (one can find some line overflows, and several bibliographic entries would deserve more harmonizing).

As for the content of the thesis, it's a surprisingly mature research work, with high number of well designed experiments and with really excellent experimental results. The author gained inspiration from his detailed insights into related literature and builds the "research story" in a very logical way.

A minor factual correction: the author says that "We do not have any official tagset for many widely spoken languages as Telugu", but there's the ICON 2010 dataset containing a Telugu treebank, which includes POS labeling too.

In fact, my only non-trivial concern relates weak or missing error analysis. The author usually looks at the achieved results only through the lenses of one-dimensional global performance, without trying to identify main types and possible sources of tag

misclassifications. Of course, once we state the optimization criterion, we should try to mechanize the optimization process fully. But it's a shared wisdom in machine learning that visualizing the dominating error classes sometimes brings useful stimuli for designing new features. I would like to ask the author about it during the defense: when you try to analyze the tagging errors, can you see some repeated patterns and do you think such observations can be exploited somehow?

Conclusion

Duong Thanh Long has shown that he is able to analyze and implement a very complex research task. I recommend to accept the thesis for the defense.

In Prague, 29th July 2013

doc. Ing. Zdeněk Žabokrtský, Ph.D.
Institute of Formal and Applied Linguistics
Charles University in Prague