Jörg Tiedemann
Prof. of Language Technology
Dep. of Digital Humanities
University of Helsinki

PhD Thesis Evaluation

Title of the Thesis: Discovering the structure of natural language sentences by semi-supervised methods
PhD Candidate: Rudolf Rosa
Faculty of Mathematics and Physics
Charles University, Prague

Rudolf Rosa's thesis presents work on cross-lingual parsing and tagging with the general aim of supporting low-resource languages that can benefit from linguistic tools and resources that are available for other languages. The focus of the work is set on annotation and model transfer with the help of parallel corpora and machine translation. The proposed methods are to a large degree language independent and have been evaluated on a substantial number of languages and data sets. The approach includes a novel measure for improved model selection and the empirical results represent state-of-the-art performance of the task.

After a careful assessment of the thesis, it is my pleasure to confirm that the work fulfils the requirements of a PhD thesis in computational linguistics including sufficient amounts of work including novel and creative solutions that are presented in a sound scientific way.

In the following, I will look into specific aspects of the thesis.

Main objectives and achievements

The overall goal of the thesis is to study the possibilities to train models that can analyze language data syntactically without the availability of annotated training data in that particular language. Modern syntactic parsers are typically trained on treebanks, i.e. data sets with verified syntactic annotation, but for most languages in the world such data sets are not available. The motivation of cross-lingual methods is to bootstrap parsers for such languages by utilizing existing tools and resources for other languages and transfer learning techniques that adapt a model to a new related task. This is in contrast to fully unsupervised approaches that try to find structure in raw data without the help of any annotation.

The thesis focuses on dependency parsing and applies a largely theory-agnostic approach that emphasizes practical techniques and methods to achieve an overall improved performance of the tools. The proposed methods are tested in well-established frameworks (datasets like HamleDT and UD with associated benchmarks and evaluation metrics) that allow proper comparisons with the state-of-the-art. The thesis includes work on delexicalized models, multi-source transfer with advanced model selection techniques and the use of machine translation for lexicalized transfer models. It also contains a study on cross-lingual part-of-speech (PoS) tagging using similar techniques as the ones proposed for cross-lingual parsing.

The main contribution of the thesis is in my opinion the introduction of the source selection method based on the novel language similarity measure that applies KL divergence over PoS n-gram distributions. This metric provides an intuitive score that allows the proper selection of the source language(s) when transferring parsers to a new language and even makes it possible (in most cases) to combine several sources in a beneficial way by applying the same score as a weight. The thesis convincingly demonstrates the robustness of the metric by extensive empirical studies.

Another contribution is the discussion of annotation styles and their influence on parsing performance. The study on adpositions provides valuable insights on the influence of annotation decisions and the performance of specific parser models.

Finally, I also value the discussion of machine translation approaches in cross-lingual transfer. The empirical results in the extensive studies presented here support the findings that simple approaches work very well or even better than more complex models that enable higher translation quality but in fact increase noise for the cross-lingual parsing/tagging tasks. The value of the PoS-based language similarity measure for multi-source setups is also an important outcome of the thesis.

Overall, the main contribution of the thesis is, thus, a detailed guideline for bootstrapping syntactic dependency parser for under-resourced languages including a novel metric for proper source selection or weighting. This “cookbook” will certainly be appreciated by researchers working on this task in the future.

Structure of the thesis

The thesis follows in general a common structure with background chapters in the beginning (describing data, tools and evaluation frameworks) and the experiments and results thereafter. The introduction is rather short and outlines the structure of the thesis rather than introducing the task, the research questions and the main hypotheses. The first background chapter summarizes the data sets that are relevant for the study and, thereafter, the main algorithms for dependency parsing (graph-based and transition-based parsing) are briefly introduced together with a section on evaluation. I would have preferred a proper and detailed introduction of the (cross-lingual) parsing task and its main approaches before presenting data sets but that is a matter of personal style.

The main contributions are presented in chapters 3-6, starting with the discussion of delexicalized parsing, the combination of multiple sources, lexicalized parsing and, finally, cross-lingual tagging. The order seems logical even though some parts require the reader to jump ahead to know, for example, how parse trees can be combined in chapter 3 and how cross-lingual tags can be obtained in chapter 5. Nevertheless, the presentation is coherent and well-structured and the conclusions bring the main findings together even though in a rather unconventional way.

Presentation and related work

The style of writing is very clear and follows scientific standards of good quality. The thesis emphasizes practical considerations and empirical studies rather than theoretic questions. The candidate is well aware of related work and presents his methods successfully in the light of the background literature and related work. Theoretical and methodological grounds on parsing with its data-driven approaches, machine translation, alignment and transfer learning are mainly left as links to background literature. A reader with less knowledge in those topics would probably appreciate more details in the exposition of the thesis. However, it is easy to relate to the relevant literature with the comprehensive references given in the thesis.

The contributions of the work are well presented and empirically supported. The KL language similarity measure is convincingly motivated also from a theoretical perspective. The candidate also manages to present the substantial experimental results in a pleasant and readable way, which is not easy considering the number of benchmarks that are necessary to demonstrate the capabilities of the algorithms.

Methodology and experimental setup

The study follows standards in empirical research. Data sets are properly described and the experiments and algorithms are presented in great detail to make it easy to replicate the results. This is also supported by data sets that are openly available, at least most of them. The candidate is also careful with the selection of development data for tuning hyperparameters and model architectures. This is very much appreciated as it is often the case that data-driven models overfit to the data used for training but also for repeated testing. In the thesis, the candidate makes sure to apply independent development data for tuning the systems whereas the final test set is only applied after that process. This way of proceeding is not always adopted by all researchers in the field, as obvious as it may seem. Well done!

As recommendations for future work I would like to add a more systematic study on the effect of predicted PoS tags in cross-lingual parsing, a comparison of cross-lingual parsing performance to learning curves of supervised parsing with increasing amounts of annotated target language data, and a pilot study on realistic settings for at least one truly under-resourced language. Those experiments would be very interesting complements to the present work and further strengthen the claims of the thesis. Nevertheless, I would like to stress that the work presented in the thesis is sufficient as it is for the requirements of a PhD in computational linguistics.

Summary

The thesis represents an important work in the field of cross-lingual parsing dedicated to the development of linguistic tools for under-resourced languages. The work presents an essential guide book for bootstrapping dependency parsers from existing resources and introduces novel methods to improve the quality of transfer models. The candidate demonstrates his abilities to work in a scientific and creative way and provides replicable approaches that are highly relevant for the research community.

Jörg Tiedemann Helsinki, May 27, 2018