Universität Potsdam, Karl-Liebknechtstr. 24-25, 14476 Potsdam/OT Golm

Humanwissenschaftliche Fakultät
Department Linguistik

Faculty of Mathematics and Physics
Charles University, Prague

Angewandte Computerlinguistik

Prof. Dr. Manfred Stede

Telefon: +49-331-977-2691
Telefax: +49-331-977-2087
Sekretariat: +49-331-977-2950

Datum: August 31, 2018

Review of the dissertation „Coreference from the cross-lingual perspective“ by Michal Novák

The thesis by Michal Novák targets the phenomenon of coreference in conjunction with its corresponding computational task of coreference resolution from a bilingual perspective. Specifically, it studies different ways of exploiting linguistic annotations in a parallel corpus for improving the resolution performance. The work pleasantly exemplifies the Prague viewpoint of tackling natural language processing not merely as a machine learning problem on some data set, but with an ultimate interest in paying attention to linguistic detail and thereby gaining new insights into the mechanics of human language. The foundation for this approach is the thorough annotation of corpora by means of the multi-layer tectogrammatical framework, which allows for a particularly “deep” modelling of linguistic information. Novák’s thesis exemplifies this by handling coreference not just on the level of surface words (as mainstream computational linguistics does) but on said tectogrammatical representations. Thus the objective of the work is both theoretical (how does coreference “work” in Czech and/versus English) and practical (how can automatic coreference resolution be improved in general, and specifically in a cross-lingual setting).

The thesis is clearly structured and well-written. (One mild exception at the beginning of Section 6 will be mentioned below.) The author describes his work and states the results at the right level of detail, and in a succinct manner with no extra verbosity.

Chapter 2 first provides an apt summary of the tectogrammatical representation of sentences and the handling of coreference information. This is followed by a characterisation of the linguistic properties of the relevant referring expressions in Czech and English: central pronouns, relative pronouns, zeroes, and others (demonstratives, noun groups, named entities). For the rest of the thesis, the overall coreference problem is then restricted to pronouns and zeroes, which is a sensible choice, since the resolution of nominal expressions requires rather different approaches and knowledge sources.

In the review of related work in **Chapter 3**, one notices that the examined systems for monolingual coreference resolution tend to be a bit old (references largely stop in 2015), while some more recent work (which is often NN-based) is merely mentioned in passing. For the purposes of this thesis, however, this is not so problematic, because the author does not aim at producing state-of-the-art resolution results for English (which is the language that the vast majority of research centers on).

Instead, the approach here is to use the well-established Stanford resolvers, which were developed several years ago, as a point of comparison just to make sure that the present work is not entirely off the mark. This is a viable strategy given the goals of the thesis.

In section 3.2, where the immediately-relevant work on cross-lingual approaches is being discussed, the survey appropriately includes the recent papers, and I do not know of any work that would be missing here. The author divides the approaches into delexicalization, projection, and multilingually-informed (or joined) resolution. The latter two constitute the context for the thesis project, and the inspirations include the work by Chen and Ng (2014), who gained 2-3 F-points on Chinese pronoun resolution by automatically translating to English and exploiting features on that side. A similar performance gain had been achieved by Mitkov and Barbu (2003), who added some contrastive-linguistic knowledge to a rule-based pronoun resolver.

Chapter 4 is concerned with the technical basis of the work: the corpora and analysis tools, and the decision on how to evaluate the performance of a coreference resolver. As for corpora involved, the situation is not as easy to survey as in many other theses (e.g., one corpus used throughout), but given the thesis topic and the context of the research group, it is quite clear that a variety of sub-/corpora are needed for developing and testing different aspects of this work. Importantly, corpus creation was one part of the thesis contributions, as the author participated in building the PCEDT Coref corpus and the PAWS subcorpus, which is manually annotated for the various types of information needed in the thesis.

The computational processing is based on the Treex framework, which performs the preprocessing needed for the author's new coreference module. It already includes a module for that purpose, against which the author's will be compared. Similarly, a first version of cross-lingual alignment is implemented here, which is based on straightforward surface token intersective alignment via MGIZA++; transfer to the tectogrammatical nodes; and a rule-based postprocessing for handling zeroes. On the English side, the coreference resolver for comparison is the Stanford system, as mentioned above, which implements three alternative methods for agglomerative clustering of mentions.

Regarding evaluation, the author argues convincingly that the widely-used measures MUC, CEAF and B3 are not very meaningful for the task that is being tackled in this thesis. Here, he is largely in agreement with the thesis work of Tuggener (2017). The proposed alternative scoring method, the Prague anaphora score, closely resembles the one that was independently suggested by Tuggener. The difference to that scheme is a more tolerant correctness criterion for response links, where the Prague score accepts any mention that is in the key chain of the anaphor. While this initially seems a bit weak, it follows quite logically from the inclusion of the abundant zero referring expressions in Czech and from the decision not to handle nominal expressions. One thing that I missed in this section (though it is not a crucial point) is a comparison to other recent proposals on scoring methods, such as that of Moosavi/Strube at ACL 16.

In **Chapter 5**, the author presents an analysis of the correspondences of English and Czech referring expressions in a 1000-sentence subcorpus, where the alignments were manually annotated. The result is statistical information on what types of referring expressions map to each other in the two languages – with the slight caveat of the data resulting from a single-direction translation, as the authors points out. This study, which represents joint work with the author's colleague Anna Nedoluzhko, yields very useful insights into the contrasts between the two languages.
As a minor note: On p. 62, I don't follow how the percent figures for the personal pronouns match the frequencies in the table; for example, English personal pronouns to Czech zeroes: 135/187 = 72%, while the text states 57%. Maybe I am missing something?
(likewise for "it" on the following page where I don't see the "more than 50%")

The contrastive analysis provides the background for proposing a new method for computing the cross-lingual alignment in **Chapter 6**. I was temporarily a little confused because the PAWS subcorpus is re-introduced "from scratch" here, although it was already covered in Chapter 5. More

generally, in this part of the thesis the writing is somewhat more focusing on the genesis process of the work rather than on the final outcome, i.e., it could be a little streamlined.
The aligner is trained on these 1000 PAWS sentences, employing as features the “old” alignment from the CzEng CR pipeline, morphosyntax and dependency information, semantic roles, and the category of the referring expression. In addition, various combinations of these features are added. Here, I would have liked some retrospective feature ablation tests in order to see the relative contributions. Performance-wise the approach is however very successful, as the results are considerably better than those of the old aligner alone.

The “kernel” of the thesis is presented in **Chapter 7**, which develops a new coreference resolution system that can utilize information from an aligned parallel corpus. This work builds on the earlier implementations done in the research group, but makes significant new contributions. The underlying model is mention-ranking, hence a middle way between the classical mention-pair and entity-based models. Anaphoricity detection is done jointly with antecedent selection, instead of a separate step.

The feature set for the resolver is fairly standard: location and distance, plus some corpus/ontology knowledge: noun-verb collocations, WordNet hyperonyms, NE types. As hyperparameters, the ranker uses the size of the window for selecting candidates, the inclusion (or not) of material following the anaphor, and morphosyntactic filtering.
For Czech, the new system clearly beats the earlier one, while for English the benefit is much less pronounced on the CoNLL data. On this data, the deterministic and neural Stanford system beat the author's system by 5-6 points, which he attributes to "more advanced approaches" and a lack of optimization of the hyperparameters in his system. This seems plausible to me, because indeed the main goal of the thesis is not to achieve maximum performance on English. The author then also measures the performance differences separately for the various types of referring expressions, where it turns out that zero reconstruction proves to be a main factor for the observed improvements. He points out another potential reason for the advantages of the two Stanford resolvers on the CoNLL dataset, viz. the possibility of different annotation rules in the PCEDT and OntoNotes corpora. It would be interesting to see whether this hypothesis can be verified in the respective annotation guidelines, but the author didn’t take this step.

The core of the chapter is the experiment of adding coreference information from the "other" language of the aligned parallel corpus. This idea is based on the hypothesis that morphosyntactic features can be exploited for computing agreement with potential antecedents on the other language, and those results can be projected back to the target language. Specifically, the additional features are gained from the aligned nodes, and the author experiments with using on the one hand all the standard monolingual features as mentioned above, and on the other hand just the +/-coreferent information from annotations in the other language. These are obtained by an automatic resolver, so the setting is that of a "real-world” scenario.

The experiments on the PCEDT data show that cross-lingual information adds 1.9 F points on Czech and 1.5 on English, which is mostly due to improvements for personal and possessive pronouns. While this is not a dramatic improvement, it generally suggests that the method can be useful, and invites future work of comparing the approach with other language pairs.

When replacing the alignment and coreference information on the “other” language by manual annotations (from the PAWS subcorpus), not suprisingly both manual layers lead to increased performance and provide useful indication on how much can be gained by future work on trying to optimize either component.

Finally, the author contrasts in detail which coreference assignment decisions profit (or not) from having information on the aligned language. Amongst other things, this analysis reveals that for English reflexive pronouns (and only for these) the cross-lingual information in fact leads to lower results. In terms of further inguistically-oriented qualitative analysis (based on random samples), one of the insights is that the resolution of Czech personal and possessive pronouns profits most from the availability of the English side and specifically the gender information. The author attributes this to the fact that in English, gender essentially encodes animacy, which is absent in the Czech masc/fem

distinction, and it also benefits from a simpler syntactic environment in English. This and a range of similar results (or more precisely, hypotheses) distinguishes this work from much of the mainstream computational linguistics work: Beyond mere counting, the author looks at the different sets of in/correctly processed data points, tries to identify patterns, and carefully relates these to properties of the languages involved. In this sense, section 7.4.2 is in fact one of the most interesting parts of the thesis.

In **Chapter 8**, a smaller set of experiments is devoted to the alternative approach to exploiting a parallel corpus, viz. the projection method. These are run on gold trees and projections of gold coreference links. Training a corefence resolver on them thus constitutes an upper bound experiment, which is in line with the thesis' emphasis on gaining insights about the languages, rather than (more standardly) using projection to create resources in a new language. Besides, as the author notes, this upper bound experiment gives an indication of the necessary post-editing effort for a language pair.

In contrast to earlier work on projection, the author works on tectogrammatical trees, which implies that zeroes can be accounted for quite comfortably. Another consequence is that only heads are being projected, thereby sidestepping the problem of mention boundary detection. This clearly amounts to a simplification for many practical purposes, but again seem warranted by the main goals as stated above. (Still, it would be interesting to get an estimate on how difficult it is to extract mentions spans (whose boundaries accord with a specific annotation guideline) from the tectogrammatical trees.)

With F-scores around 0.5, the results are not so encouraging; the only promising figure is a relatively high precision for Cz->Eng personal, possessive and reflexive pronouns. In a qualitative error analysis, the author looks at reasons for non-aligned mentions (largely: their absence in one of the languages), mention matching, and antecedent selection. The influence of error types is quantitatively measured by calculating scores only on the correctly aligned/matched mentions, and some examples are discussed. The role of alignment quality is then measured by comparing performance on gold versus system alignments on the PAWS corpus. Gold alignment leads to slight improvements, but still, the projection method does not appear to be a promising avenue for this language pair and in the setting implemented here. In the final experiment, the Treex coreference resolver is trained on the projected gold coreference links. Here, he performance loss for the English coreference task is almost twice as big as for Czech (where it is already substantial), so that both the cross-lingual informed CR and the projection approach lead to the conclusion: With respect to coreference, English is more informative for Czech than vice versa. This "converging evidence" finding is very interesting and, again, invites comparison with other language pairs.

In summary, Michal Novák presents a very thorough study of the cross-lingual coreference problem with many experiments that not just directly concern the central goals of the project, but also various interesting sidelines. In several methodological respects, the work goes beyond what is "standard" practice in the field: Measurements are taken not just across the board of referring expressions but individually for their different types; upper bound calculations in various places are very informative for deciding what components would be worth trying to optimize; qualitative error analysis generally accompanies the quantitative experiments. Much of the added-value results from the desire to not just implement an automatic system that works well, but to also learn about the languages involved. The work clearly constitutes a highly significant and original contribution to linguistic and computational-linguistic research on coreference, whose methodological aspects are also relevant for cross-lingual research in general, and by extension to more applied research on multilingual tasks, especially machine translation. I am more than happy to recommend the work to progress to the oral defense.

Prof. Dr. Manfred Stede