Doctoral Thesis Review

Loganathan Ramasamy: Parsing under-resourced languages: Cross-lingual transfer strategies for Indian languages

Otakar Smrž, Ph.D.

Overview

In his doctoral thesis, Loganathan Ramasamy presents his research work that explores the possibilities and limits of the automatic discovery of syntactic structure for languages without extensive amounts of linguistically annotated data and other computational resources. While keeping his approach general and abstract, he applies it to five Indian languages, namely Bengali, Hindi, Tamil, Telugu, and Urdu, as the cases of particular interest.

The author develops his research along the following three lines differing in the degree of human supervision involved in the proposed methods:

1. He designs and builds a manually annotated treebank of Tamil and compiles an English-Tamil parallel corpus.
2. He proposes and executes a computational scenario that builds syntactic parsers for the above-mentioned Indian languages based on parallel texts and the transfer of syntactic dependency relations through projection and delexicalized parsing.
3. He proposes and executes a computational scenario that uses machine translated texts for the same purpose.

In the Introduction, the author outlines the motivation for his research and compares the availability of treebanks with the distribution of languages in the world's population. In Chapter 2, he gives a thorough report on the related work, neatly summarizing the relevant state-of-the-art results in the field and interpreting them in the context of the thesis.

In Chapter 3, the author deals with the essential question about the point of the work being done: are Indian languages under-resourced? Except for Hindi, they are under-resourced. He surveys the available computational resources for ten Indian languages altogether. I appreciate this chapter the more as justifying the direction of one's research is not as often seen as it should.

Chapter 4 presents TamilTB, the author's own dependency treebank. The chapter includes a very detailed and elaborate account on morphological tags and esp. subpos, with examples and

transliteration, but not always with translation. It then provides syntactic annotation guidelines exemplifying the node attachments and labels with dependency trees taken from the treebank.

Chapter 5 gives more details about the available treebank data and the unlabeled attachment score. Chapter 6 reports on the elaborate methods and experiments constituting the parsing strategies for under-resourced languages.

Section 6.1 on dependency transfer using bitext projection culminates with the finding that parsers based on this method do not work better than supervised parsers trained on just about 10 manually annotated sentences. This result is, however, quite important. It shows the limits of the intuitive approach as well as the tremendous sensitivity of the problem to annotation conventions (most notable with the `hi`, `ta` and `te` data). Section 6.2 thus addresses the problem of different annotation styles and aims at applying the delexicalized parser approach to the selected five Indian languages. This involves training syntactic parsers on 30 treebanks and evaluating them carefully depending on fine-grained distinctions, such as the presence of syntactic harmonization or the variant of the morphological tags used in the process. Section 6.3 then reiterates both transfer approaches on the machine-translated treebank texts, for the 5 Indian languages as well as 13-18 other languages for which this approach is novel, too.

Chapter 7 examines projection and alignment errors esp. in Tamil. Chapter 8 generalizes the lessons from the different dependency transfer scenarios and provides a recipe for obtaining a syntactic parser for any natural language. The thesis concludes with a list of contributions made.

Judging on the references made throughout the thesis, I have been pleased to see that the author as well as his colleagues in the department work on the edge of the global language technology research and collaborate intensely, critically and yet heartily.

Comments

1. Your English-Tamil parallel corpus, though having a portion of news (section 3.4.2), is not used as the source of texts for the TamilTB (section 4.3). Why did you not decide to annotate the Tamil data for which there is a parallel English translation available?
2. Does the annotation of morphology and syntax in the PDT style fit the Tamil language and the common formulation of its grammar, such as that taught at schools? How does the TamilTB annotation relate to the Paninian grammar or perhaps other traditional approaches? Would it be possibly more useful to design annotation guidelines that propose specific dependency labels and attachment conventions, in order to simplify it for machines (cf. parsing accuracy of Tamil vs. other ILs) and make it more natural for those users of the language who are interested in the problem, but do not know PDT? My experience with the Prague Arabic Dependency Treebank (PADT) and the Columbia Arabic Treebank (CATiB) speaks for the annotation style of the latter, which is practically as informative but much more natural and consistent. It has just 8 (6) syntactic labels, no

auxiliaries nor special treatment of compound expressions, and annotates coordination with a more convenient pattern. http://www.elda.org/medar-conference/pdf/25.pdf

3. It seems to me that the distinction between AAdjn (adverbial adjunct), AComp (adverbial complement), and Obj (object) is subtle and conditioned by verbal valency. Could this distinction be possibly reduced in favor of more coarse-grained and consistent annotation? Similarly for Atr and AdjAtr, and the series of Aux[ACPGKVXZ] labels. The kind of attributes or auxiliaries might well be determined by the morphology and lexical identity of the nodes in question.
4. Reading that TamilTB 1.0 has been available since November 2013, I would prefer the documentation given in the thesis be updated from that of version 0.1. The claim "most of the annotation descriptions explained here are valid for the latest version, too" (page 29) does not help.
5. Although Appendix B does mention the online repository where the TamilTB treebank is located, the CD attached to the thesis only contains two data files in the treex.gz format. This makes it hardly possible for the interested reader to inspect the data. I do happen to know how to handle these files, but I did not manage to make the necessary easytreex extension run in order for TrEd to open these files for me.
   I therefore explored the online repository. While version 0.1 of the treebank is provided in several common formats and has rich documentation, version 1.0 is there only in the mst format and its web page is quite brief. This is a little confusing and user-unfriendly, again. Providing the most up-to-date data in the most common formats would be great.
6. Online, you provide a mapping for romanizing the syllable-based Tamil script. Have you considered turning it into a reusable library or script? Such a tool would be appreciated in general, I guess, at least by everyone working with TamilTB or other Tamil data.
7. On page 44, "Position 8 - Negation" might rather be "Position 9 - Mood" or "Position 9 - Polarity", and in Table 4.18, the tag for "cannot" should probably read "VR-T3SN-N".
8. Table 5.1 swaps the columns for target number of tokens and source average sentence length. The "corpus size" column should rather be titled "# sentences", as in Table 5.2.
9. I enjoyed reading about the system of pronouns in Tamil and thought about analogies in Czech word formation, e.g. "něco", "cosi", "cokoli", "lecco", "copak", "co", "to", "toto" ...
10. As to the projection algorithm on page 86, you claim that "not processing the source nodes in order may likely result in different projected structures". I do not see why or how often this should be the case, when the algorithm is based on unchanging node alignments and the Direct Correspondence Assumption. Yet if your claim holds, why do you not consider improving your algorithm to find the optimal projected structures only?
11. I have struggled with discrepancies in tagset sizes in Tables 5.2 and 6.9, esp. for Tamil.
12. Tables 6.12 shows that harmonization of treebanks may hurt the accuracy of parsers, cf. both `hi` and `ta`. Does this say anything about the fitness of the harmonized annotation style? Would other harmonization choices do better?
13. The delexicalized `ta` parser does better parsing `te` and `hi` than parsing `ta` itself. Why?
14. The thesis is written in English and is well structured and clear. However, there are quite a few places where perhaps the editing of the text resulted in errors or disfluencies, such as "with the increasing of richness in the annotation" (page 1), "choosing the source data

that are as closer to target language" (page 13), "India is one of the linguistically diverse country in the world" (page 21), "is a parallel corpora" (page 24), or "the errors could also have been resulted" (page 132).

Conclusion

The extent and quality of work presented in the thesis certainly qualify its author for the doctoral degree in the field of natural language processing.

September 8, 2014
Prague

Otakar Smrž, Ph.D.