11-712: NLP Lab Report

Pradeep Prabakar Ravindran

April 26, 2013 (Final Due date)

Abstract

First draft of a report documenting my experience developing a dependency parser for Tamil.

TamilDep is a tool that performs dependency parsing of the Tamil Language. This tool would soon be open-sourced and become publicly available at https://github.com/pradeep-gnr/TamilDep. Along with the code for a dependency parser, a corpora of tamil sentences annotated with their dependency parse trees is also expected to be released. Since there are very few resources publicly available for Tamil, I hope that this software would prove useful to researchers working on Tamil natural language processing.

1 Basic Information about Tamil

Tamil is a Dravidian language that is widely spoken in the Indian state of Tamil Nadu and the north eastern part of Sri Lanka. Reports estimate that there are about 70 million native Tamil Speakers all over the world (Wikipedia (2013)). In terms of history, Tamil is also one of the oldest languages in the world with works of Tamil literature dating back to almost 2000 years ago (Wikipedia (2013)).

Grammatically, Tamil has a lot of characteristics that make it pretty different from the Germanic languages. For example, Tamil almost always follows a subject-object-verb oriented form (Ramasamy and Žabokrtskỳ (2011)). For instance, consider the following sentence.

```
Avan      angu       selgiran
(He)     (there)     (going)
```

The sentence approximately translates to 'He is going there'. Sentences where inflected verb forms occurring at the end of the sentence are relatively common in Tamil. Also, similar to languages like Czech, Tamil is relatively a free word order language. For example, the sentence that was discussed above could be expressed in multiple forms such as

```
 angu      avan     selgiran
(There)    (he)      (going)

 selgiran  avan      angu
 (going)   (he)    (there)
```

All of these sentences are perfectly valid in terms of grammatical structure. Another key characteristic of Tamil sentences is that words could be agglutinative and sometimes it is very difficult to even define what a word is. There are a number of ways in which suffixes could be added to the noun and verb stems to produce highly inflected forms.

For instance, 'has gotten over' which is a phrase in English could be expressed as a single word in Tamil. Also, there could be multiple such words that would approximately translate to the sample phrase. Some such examples are discussed below.

```
1) Nadanthathu
2) Nadanthathuvittathu
3) Nadanthayittru
```

All these words express the phrase 'has gotten over' and they are all highly inflected forms of the root word 'Nada' which roughly translates to 'happening' or 'proceeding'. Since these type of verb phrases are relatively common in Tamil, identifying 'head' words of such phrases would be a challenge from a dependency parsing perspective.

2 Past Work on the Syntax and NLP tools for Tamil

One of the most infleuntial books on Tamil grammar written in English was by Thomas Lehmann : A grammar of modern Tamil (Lehmann et al. (1989)). The book gives a very detailed analysis of several of the morphological and syntactic phenomenon that characterize written Tamil. While languages like English do not differ significantly from their spoken and written form, there are significant differences among the spoken and written forms in Tamil. Tamil has the property that sentences could be semantically valid even when certain words are omitted and spoken Tamil usually consists of a lot of missing words. Written Tamil however is more morphologically richer than its spoken version. Lehmann also notes that Tamil is relatively a head final language and that the phrasal head appears in several inflected forms at the end of the phrase (Lehmann et al. (1989)). Also, another key characteristic of Tamil is that the verbs agree with gender, tense etc. One of the key differences in tamil syntax when compared to other languages is that the language exhibits a restricted free word order. Hence there are a multitude of semantically and syntactically valid sentences that could be formed using the same set of words. From a parsing perspective - this property poses a lot of challenges because unlike English, writing context free grammar rules to characterize Tamil Syntax is a non-trivial problem. (Srinivasan CJ (2007)) discuss the challenges they face because of this issue while building a spoken dialog system for Tamil and turn to dependency parsing to aid semantic annotation of sentences.

Unlike other Indian languages like Hindi, Telugu and Bengali - Tamil does not have a lot of annotated corpora for NLP tasks. The Tamil dependency treebank ((Ramasamy and Žabokrtskỳ (2011))) which was released recently is the only publicly available dependency corpora that I could find online. The authors have also released an annotation manual that describes the rules that were used in labeling dependency edges among words. The entire corpus was annotated in the Prague dependency treebank (PDT) (Böhmová et al. (2003)) format and a detailed description of this corpora and the annotation methodology that was used will be discussed in part 3. Although this corpus is quite small (about 3000 words), the annotations are quite diverse and cover various types of sentences in Tamil. The same group also used this corpus to build a rule based and corpus based dependency parser and report a 75 % accuracy when sentences are labeled with POS tags.

There was also an earlier paper (Dhanalakshmi and Rajendran (2010)) that used morphological analysis along with various custom heuristics for syntactic parsing of Tamil to identify phrasal consituents. Besides the work done by Ramaswamy et al (Ramasamy and Žabokrtskỳ (2011)), I could find only one related paper that describes a set of Natural language tools that were created

specifically for Tamil. They have also released an open source Morphological analyzer for Tamil but unfortunately there was no dependency tree annotated corpora or software that was available for download. To the best of my knowledge, TamilDep would be the first open source dependency parser written specifically for Tamil.

3 Available Resources

[include discussion of your corpora –NAS]

4 Survey of Phenomena in [Your Language –NAS]

5 Initial Design

6 System Analysis on Corpus A

7 Lessons Learned and Revised Design

8 System Analysis on Corpus B

9 Final Revisions

10 Future Work

References

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. The prague dependency treebank. In *Treebanks*, pages 103–127. Springer, 2003.

V Dhanalakshmi and S Rajendran. Natural language processing tools for tamil grammar learning and teaching. *International journal of Computer Applications (0975-8887)*, 8(14), 2010.

T. Lehmann, Pondicherry Institute of Linguistics, and Culture. *A Grammar of Modern Tamil*. PILC publication. Pondicherry Institute of Linguistics and Culture, 1989. URL `http://books.google.com/books?id=THlkAAAAMAAJ`.

Loganathan Ramasamy and Zdeněk Žabokrtskỳ. Tamil dependency parsing: results using rule based and corpus based approaches. In *Computational Linguistics and Intelligent Text Processing*, pages 82–95. Springer, 2011.

Loganathan R Santhosh Kumar C Srinivasan CJ, Udhaykumar N. Robust dependency parser for natural language dialog systems in tamil. *5th Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 1–6, 2007.

Wikipedia. Tamil language. 2013. URL `http://en.wikipedia.org/wiki/Tamil_language`.