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

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

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.

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