Tamil Shallow Parser using Machine Learning Approach

Dhanalakshmi V\textsuperscript{1}, Anand Kumar M\textsuperscript{1}, Soman K P\textsuperscript{1} and Rajendran S\textsuperscript{2}

\textsuperscript{1}Computational Engineering and Networking
Amrita Vishwa Vidyapeetham Coimbatore, India
{m_anandkumar,v_dhanalakshmi, kp_soman} @cb.amrita.edu

\textsuperscript{2}Tamil University, Thanjavur, India

Abstract

This paper presents the Shallow Parser for Tamil using machine learning approach. Tamil Shallow Parser is an important module in Machine Translation from Tamil to any other language. It is also a key component in all NLP applications. It is used to understand natural language by machine and also useful for second language learners. The Tamil Shallow Parser was developed using the new and state of the art machine learning approach. The POS Tagger, Chunker, Morphological Analyzer and Dependency Parser were built for implementing the Tamil Shallow Parser. The above modules gives an encouraging result.

Introduction

Partial or Shallow Parsing is the task of recovering a limited amount of syntactic information from a natural language sentence. A full parser often provides more information than needed and sometimes it may also give less information. For example, in Information Retrieval, it may be enough to find simple NPs (Noun Phrases) and VPs (Verb Phrases). In Information Extraction, Summary Generation, and Question Answering System, information about special syntactico-semantic relations such as subject, object, location, time, etc, are needed than elaborate configurational syntactic analyses. In full parsing, grammar and search strategies are used to assign a complete syntactic structure to sentences. The main problem here is to select the most possible syntactic analysis to be obtained from thousands of possible analyses a typical parser with a sophisticated grammar may return. This complexity of the task makes machine learning an attractive option in comparison to the handcrafted rules.

Methodology

Machine learning approach is applied here to develop the shallow parser for Tamil. Part of speech tagger for Tamil has been generated using Support Vector Machine approach [Dhanalakshmi V et.al., 2009]. A novel approach using machine learning has been built for developing morphological analyzer for Tamil [Anand kumar M et.al., 2009]. Tamil Chunker has been developed using CRF++ tool [Dhanalakshmi V et.al., 2009]. And finally, Tamil Dependency parser, which is used to find syntactico-semantic relations such as subject, object, location, time, etc, is built using MALT Parser [Dhanalakshmi V et.al., 2011].
General Framework and Modules

- The general block diagram for Tamil Shallow parser is given in Figure 1.

![Diagram](image_url)

**Figure 1. General Framework for Tamil Shallow Parser**

- **Tamil Part-of-Speech Tagger** [Dhanalakshmi V e.tal., 2009]: The Part of Speech (POS) tagging is the process of labeling a part of speech or other lexical class marker (noun, verb, adjective, etc.) to each and every word in a sentence. POS tagger was developed for Tamil language using SVMTool [Jes’us Gim’enez and Llu’ís M’arquez, 2004].

- **Tamil Morphological Analyzer** [Anand Kumar M e.tal., 2009]: Morphological Analysis is the process of breaking down morphologically complex words into their constituent morphemes. It is the primary step for word formation analysis of any language. Morphological Analyzer was developed using a novel machine learning approach and was implemented using SVMTool.

- **Tamil Chunker** [Dhanalakshmi V e.tal., 2009]: Chunks are normally taken to be non recursive correlated group of words. Chunker divides a sentence into its major-non-overlapping phrases.

176
(noun phrase, verb phrase, etc.) and attaches a label to each chunk. Chunker for Tamil language was developed using CRF++ Tool [Sha F and Pereira F, 2003].

- **Tamil Dependency Parser for Relation finding** [Dhanalakshmi V et al., 2011]: Given the POS tag, Morphological information and chunks in a sentence, this decides which relations they have with the main verb (subject, object, location, etc.). Dependency parser was developed for Tamil language using Malt Parser tool [Joakim Nivre and Johan Hall, 2005].

**Dependency Parsing using Malt Parser**

MALT Parser Tool is used for dependency parsing, which uses supervised machine learning algorithm. Using this tool dependency relations and position of the head are obtained for Tamil sentence. There are 10 tuples used in the training data that can be user define. For Tamil dependency parsing, the following features are defined and others are set as NULL and are mentioned as ‘_’ in the training data format.

- **WordID:** Position of each word in the input sentence.
- **Words:** Each word in the input sentence.
- **CPos Tag and Pos Tag:** Defines the Parts Of Speech of each word.
- **Head:** The position of the parent of each word.
- **Lemma:** The lemma of the word.
- **Morph Features:** The Morphological features of the word.
- **Chunk:** The chunk information of the word.
- **Dependency Relation:** The terminology given for each parent - child relation.

**Sample Training Data**

```
1. இவ்வோ _ <PRP> <PRP> 8 <N.SUB> _ _
2. என்றுடைவெள்ள _ <NN> <NN> 3 <D.OBJ> _ _
3. வாங்கி _ <VNAV> <VNAV> 4 <ATT> _ _
4. சைமண்டு _ <VNAV> <VNAV> 6 <VNAV.MOD> _ _
5. தழ் _ <NN> <NN> 6 <NST.MOD> _ _
6. ஜெய்டி _ <VNAV> <VNAV> 8 <V.COMP> _ _
7. என்று_ _ <PRP> <PRP> 8 <LOBJ> _ _
8. கர்வேலசண்டூர் _ <VF> <VF> 0 <ROOT> _ _
9. . <DOT> <DOT> 8 <SYM> _ _
```

For Tamil language, a corpus of three thousand sentences is annotated with dependency relations and labels using the customized tag set (Table.1). The corpus is trained using the MALT Parser tool which generates a model. Using this model the new input sentences are tested.
Table 1: Shallow Dependency Tagset

Application of Shallow Parser

Shallow parsers were used in Verbmobil project [Wahlster W, 2000], to add robustness to a large speech-to-speech translation system. Shallow parsers are also typically used to reduce the search space for full-blown, `deep' parsers [Collins, 1999]. Yet another application of shallow parsing is question-answering on the World Wide Web, where there is a need to efficiently process large quantities of ill-formed documents [Buchholz and Daelemans, 2001] and more generally, all text mining applications, e.g. in biology [Sekimizu et al., 1998].

The developed Tamil Shallow Parser can be used to develop the following systems for Tamil language.

- Information extraction and retrieval system for Tamil.
- Simple Tamil Machine Translation system.
- Tamil Grammar checker.
- Automatic Tamil Sentence Structure Analyzer.
- Language based educational exercises for Tamil language learners.

Conclusion

Shallow Parsing has proved to be a useful technology for written and spoken language domains. Full parsing is expensive, and is not very robust. Partial parsing has proved to be much faster and more robust. Dependency parser is better suited than phrase structure parser for languages with free or flexible word order like Tamil. Fully functional Shallow Parser for Tamil gives reliable results. The Shallow Parser system developed for Tamil is an important tool for Machine Translation between Tamil and other languages.

References


