NLP TOOLS DEVELOPMENT FOR TAMIL LANGUAGE

Loganathan, UFAL
Overview

- Introduction
- My work involving Tamil NLP
  - English – Tamil MT
  - Tamil Morphological Analyzer
Introduction – Indian Languages

- **23** official languages
- **29** languages have more than **1 million** native speakers in India.

**Tamil**

Approx **67 million** speakers in India
Introduction – Resources for Tamil

- Tamil Editing/Unicode Support - **Available**
- Dictionary – **Available**
  - Tamil lexicon, Winslow, Fabricius, McAlpin, Kathirvelu pillai
    - Published online by Univ. Of Chicago
- Morphological Analyzer/Tagger – **Partially Available**
- Phrase Structure/Dependency Parser – **NO**
- Parallel Corpora – **NO (publicly, readily)**
  - Active Bilingual websites: www.wsws.org, www.cinesouth.com
  - Tamilnadu government schoolbooks (with English translations)
My Work involving Tamil NLP

- English – Tamil Translation System (Master’s Thesis)
- Morphological Analyzer
English – Tamil Translation System
## General Differences

### Morphological

<table>
<thead>
<tr>
<th>Noun Cases</th>
<th>Tamil suffixes</th>
<th>English words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominative</td>
<td>Ø</td>
<td>Ø</td>
</tr>
<tr>
<td>Accusative</td>
<td>ai</td>
<td>Ø</td>
</tr>
<tr>
<td>Dative</td>
<td>kku, ukku, ku</td>
<td>Ø, to, for, at</td>
</tr>
<tr>
<td>Benefactive</td>
<td>(u)kkuAka</td>
<td>for</td>
</tr>
<tr>
<td>Instrumental</td>
<td>Al</td>
<td>with, of, by</td>
</tr>
<tr>
<td>Sociative</td>
<td>Otu, utan</td>
<td>with</td>
</tr>
<tr>
<td>Locative</td>
<td>il, itam</td>
<td>in, on, among, to,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with, from</td>
</tr>
<tr>
<td>Ablative</td>
<td>iliruwTu, itamiruwTu</td>
<td>from</td>
</tr>
<tr>
<td>Genitive</td>
<td>in, utaiya, aTu</td>
<td>‘s, of</td>
</tr>
</tbody>
</table>
General Differences

- Syntactical difference

Kumar *talked* about linguistics

cumAr *moziyiyalaip paRRip pEcinAn*

தமிழ் வியியல் பற்றிய பேசின்
General Differences

- **Syntactical (complex sentences)**

  Pollution Control Authority’s regional officer said that his department is not agreeing with the central minister’s opinion that Pollution Control Authority is not functioning.

  \[\text{MC} \text{ SC} \text{ RC} -> \text{ RC} \text{ SC} \text{ MC}\]

  mAčuk kattuppattu vAriyam ceyalpaTavillai enRa
  maTTiya amaiccarin karuTTil TangkaLaTu TuRaikkku utanpAtu illai ena
  mAčuk kattuppattu vAriya aTikAri TeriviTTAr
How hard is English-Tamil MT

- The previous examples illustrate
  - Tamil -> SOV, English -> SVO
  - Tamil is a restricted free word order language
  - Tamil is agglutinative

- Difference occurs in
  - Syntactical level i.e. word ordering
  - Morphological level

- We need
  - An efficient syntax reordering module
  - Morphological generator
Approaches

- Syntax Transfer Based MT
- Statistical Machine Translation (SMT)
Syntax Transfer MT

English Text

Morph Analyzer

PoS Tagging

Source Grammar

Source Lexicon

Parser

Bilingual dictionary

Target generation

Transfer rules

Morph generator

Target Grammar

Tamil text

Syntax Transfer MT: Architecture
MT using Synchronous CFG
Source Grammar – Tree Adjoining Grammar

- Tree Generating System introduced by Aravind Joshi
- TAG – Multilevel tree rewriting system
- Basic units (Elementary trees)
  - Initial trees (Basic structures)
  - Auxiliary trees (Recursive structures)

![Diagrams of Initial and Auxiliary trees](attachment:image.png)
Definition: (Tree Adjoining Grammar). TAG consists of 5-tuples \((\Sigma, NT, I, A, S)\), where

1. \(\Sigma\) is a finite set of terminal symbols;
2. \(NT\) is a finite set of non-terminal symbols: \(\Sigma \cap NT = \emptyset\);
3. \(S\) is a distinguished non-terminal symbol: \(S \in NT\);
4. \(I\) is a finite set of trees, called initial trees, characterized as follows,
   - interior nodes are labeled by non-terminal symbols;
   - the nodes on the frontier of initial trees are labeled by terminal or non-terminals; non-terminal symbols on the frontier of trees in \(I\) are marked for substitution; usually marked as (\(\downarrow\))
5. \(A\) is a finite set of trees, called auxiliary trees, characterized as follows,
   - interior nodes are marked by non-terminal symbols;
   - the nodes on the frontier of auxiliary trees are labeled by terminal symbols or non-terminal symbols. Non-terminal symbol on the frontier of the trees in \(A\) are marked for substitution except for one node, called the foot node. The foot is node is marked with (\(*\)); the label of the foot node must be identical to the label of the root node.
TAG Operations

Substitution and Adjunction/ Ex: from XTAG Manual
TAG Derivation Structure

Many parsing algorithms were suggested, including CYK parser for TAG, Head-Corner parsing algorithm, Bidirectional parsing algorithm and more recent work on Statistical LTAG parsing.

For parsing source side, Yves Schabes algorithm was implemented in Java.
Transfer Grammar
Experiments and Results

- The entire translation system is written in Java.
- Implemented modules include LTAG parser for English, STAG system for syntax reordering of English into Tamil.
- Our system uses the same language resources developed for XTAG system for parsing the source side sentence. All XTAG related databases have been converted into Mysql format.
LTAG Tree Editor for Visualization

Collaborative effort between Amrita and CDAC
Experiments and Results

English: I met John
Tamil: இன்றைய காலம் அண்டிய ஐன்

English: He met John yesterday
Tamil: முன்னாள் காலத்தில் இன்றைய காலம் அண்டிய ஐன்

English: John is a good boy
Tamil: ஐன் ஒரு வள்ளியான இளம்

English: John is not a good boy
Tamil: ஐன் ஒரு சிறந்த இளம்

English: Mary said that John said that Ram came yesterday
Tamil: மாரி பொதுவாகவும் ஐன் பொதுவாகவும் ராம் வந்தார் இன்றைய காலத்தில் Mary பொதுவாக
Statistical Machine Translation

- EILMT English-Tamil Parallel Corpus

<table>
<thead>
<tr>
<th>Health</th>
<th>Tourism</th>
</tr>
</thead>
<tbody>
<tr>
<td>6000</td>
<td>15000</td>
</tr>
</tbody>
</table>

- Monolingual Tamil Data

<table>
<thead>
<tr>
<th></th>
<th>#Sentences</th>
<th>#Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training data</td>
<td>95464</td>
<td>&gt;1.2 million</td>
</tr>
<tr>
<td>Test data</td>
<td>1000</td>
<td>12K</td>
</tr>
</tbody>
</table>
Statistical Machine Translation

<table>
<thead>
<tr>
<th>POS</th>
<th>Suffix list</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>ai, (u)kku(aaka), aal, ooṭu, (il</td>
</tr>
<tr>
<td>Adj+Adv</td>
<td>aana, aaka</td>
</tr>
<tr>
<td>Verb</td>
<td>(kkir</td>
</tr>
<tr>
<td></td>
<td>aal</td>
</tr>
</tbody>
</table>

Monolingual data: Sensitivity to Morphology
SMT – Results for Health Corpus

a) Sensitivity to sentence length & Morphology

b) Sensitivity to training corpus size & Morphology
SMT – Results for Tourism Corpus

a) Sensitivity to sentence length & Morphology

b) Sensitivity to training corpus size & Morphology
in karnataka there are the chamundi hills.
mumbai is a cluster of seven islands.
the building is illuminated on sunday evenings.
the 12th century temple is dedicated to lord shiva.
Tamil Morphological Analyzer
NLP @ Amrita – Morphological Analyzer for Tamil

- Tamil is **agglutinative**
- The major inflectional categories in Tamil are nouns and verbs.
- Noun morphology of Tamil is simple compared to verb morphology.
- Extremely simple paradigms were used to categorize the root words.
- The lexicon includes **50000 nouns** and few hundred verbs.
- FSTs were used to build Morphological Analyzer/generator

*Figure:* Morph Generator FST
NLP @ Amrita – Morphological Analyzer for Tamil

**Morph Generator**

- Lexicon FST
- Rule FST
- $\text{TEnl}+\text{N}+\text{PL}$

**Morph Analyzer**

- Rule Invert FST
- Lexicon inv FST
- $\text{TEnlkkaL}$

**Figure:** Finite State Transducer for Morphological Processing
NLP @ Amrita – Morphological Analyzer for Tamil

Figure: Morphological Analyzer screenshot
Thank you