Web-Based Machine Translation for Phrases from English to Tamil Languages using PoS Tagging Method

Kommaluri Vijayanand
Department of Computer Science
Pondicherry University
kvixs@yahoo.co.in

INTRODUCTION

- The process of assigning the PoS label to words in a given text is said to be PoS Tagging - An imp aspect of NLP.
- Initially it is necessary to choose various PoS tags in the process of PoS identification.
- A tag set is normally chosen based on the application used for the specified language used.
- We have chosen a tag set of 30 for Tamil, in the domain of Tourism where the tourist need for general enquiry.
- The complexity in PoS tagging task is to choose a tag for the word after resolving the ambiguity in case of a word which appear with different PoS tags in different context.
- We had applied both rule based and statistical based approaches for PoS tagging in the present work.
- Statistical language model is adopted towards assigning the PoS tags and exploited the role of morphological context in choosing PoS tags.

LITERATURE

- Taggers can be characterized as rule-based or stochastic. Rule-based taggers use hand-written rules to distinguish the tag ambiguity. Stochastic taggers are either HMM based, choosing the tag sequence which maximizes the product of word likelihood and tag sequence probability, or cue-based, using decision trees or maximum entropy models to combine probabilistic features. Abundant of work had been carried out on POS tagging for English. The initial algorithm for automatically assigning part-of-speech was Rule based. The ENGTWOL tagger (Voutilainen, 1995) is a rule based tagger which is based on two-stage architecture. There were also Transformation-Based Tagging, an instance of the Transformation-Based Learning, a machine learning approach. But all these works has been done for English and a few European languages.
- There has not been much work done in PoS tagging for Tamil. A likely reason is that Tamil is rich in morphology and most of the information for PoS tagging is available as inflections. As a result of this lot of works are being done on Tamil morpher.

PARTS OF SPEECH IN TAMIL

- Tamil is a morphologically rich language with relatively free word order characteristics and Tamil words are built on more than one morphological suffix. Often the number of suffixes is 3 and could exceed up to 13.
- The sequence of the morphological suffixes attached to a word in determining the PoS tag. We have identified had enumerated about 65 PoS tags that are commonly used in conversation with the general public.
- In Tamil, noun grammatically marks number and cases and nouns consist of eight cases.
- Morphological derivatives of Tamil noun could be Stem-Noun + [Plural Marker] + [Oblique] + [Case Marker].
- Similarly, morphological derivative of Tamil Verb is StemVerb + [Tense Marker] + [Verbal Participle Suffix] + [Auxiliary verb] + [Tense Marker] + [Person, Number, Gender].
- Moreover, adjective, adverb, pronoun, postposition could be included as stems that take various suffixes. In this work, we have used a tagged corpus of 211 words, which have been tagged manually.
- Tamil being a Morphological rich language, the Morph analyser itself can identify the part-of-speech in most of the cases.

PARTS OF SPEECH IN TAMIL

- Morph analyser is a tool that splits a given word into its constituent morphemes and identifies their corresponding grammatical categories. But it fails to resolve some of the lexical ambiguities for which we need a PoS Tagger.
- At the first level a study on the limitations on word level analysis (Morph) would be done. Second the input requirement of various NLP applications would be studied.
- By these studies we can identify the information requirement of the applications that could not be delivered by a morphological analyser.
- Then strategies would be developed to identify the methodology by which a tagger can extract / resolve those additional information.
- PoS tagger would be needed to identify the tag for the words that could not be analysed by the morphological analyser.
- If the Morph gives multiple (ambiguous) tags for a word, then the tagger could be used to resolve the ambiguity.
- The idea is to try different combination of tagging techniques to identify the best tagging scheme for inflectional and free word order languages like Tamil.
- Transformation-Based tagging method is a hybrid-tagging scheme that uses both rule-based and stochastic techniques.
- Like the rule-based taggers, Transformation based learning is based on rules that specify what tags should be assigned to what words.
- But like the stochastic taggers, TBL is a machine learning technique, in which rules are automatically induced from the data.
- This approach would be tried initially and other techniques would be explored in due course.

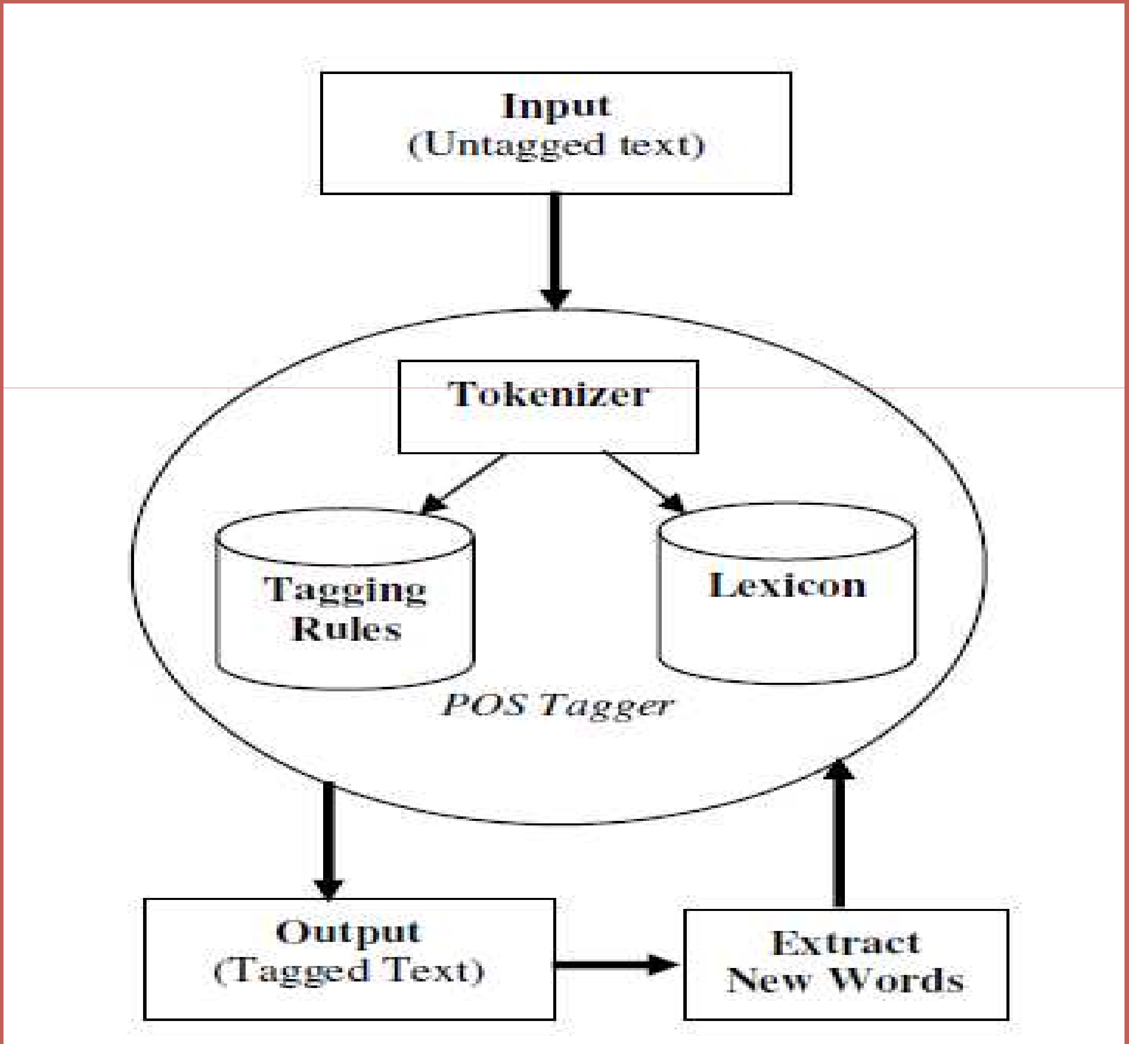
THE PoS TAGGING SYSTEM

- The present system works on the three important modules namely the tokenizer, tagging rules and a lexicon.
- The system receives the input which is the untagged text and passes into the tokenizer where it the sentence is tokenized into lexical units.
- Lexicon is used to retrive the matches for each lexical unit.
- After applying the tagging rules,Parts of Speech is identified and thus PoS tagging is done.

The algorithm

- Accept the input text from the dialogue box.
- Tokenize the input text into lexical units.
- Search for the tokens in lexicon for a match.
- If a match is not found, mark those tokens.
- Tag all tokens using the rules from the rule-base if there exist multiple tag.
- Retrieve the tagged output text.
- Extract those marked tokens from the tagged output.
- Insert those new words in lexicon.
- Add rule for that new word.
- Translation of phrases will be done based on the PoS tagged text. As new words and rules are added into the system, the system can be said to be used as the state of the art technology in learning and updating the knowledge.

PoS tagging system



CONCLUSION

- As this is an initial attempt to develop a Web based interface, we came across various problem and challenges as discussed in the paper.
- However we could find out the solutions for various problems we faced.
- We are continuously updating the lexicon and adding up the rules towards making the system more effective.

Thank You

- Queries, Suggestions, Questions, Enquiries, Doubts…….?

- WELCOME Please