A rule based POS tagger for Telugu

B. Bindu Madhavi

CALTS, HCU

- **POS tagging:** POS tagging is the process of assigning a part of speech or lexical category uniquely to each word in a sentence.
- It essentially involves the task of marking each word in a sentence with its appropriate part of speech.
- It also involves the resolution of the ambiguity of POS of a given word in the text in context

- **Rule based tagging:** It is totally based on a set of rules which decide a relevant tag for a given word. Initially a POS tagger module assigns all possible tags to each word in a given sentence later a procedure involving a set of rules selectively removes all tags but leaving one i.e. the correct tag for each word.
- It involves the formulation of context based tag assignment rules. The rules are either handcrafted or mechanically generated.

- **Tag set:** A collection of a set of optimum number of POS tags for a given language is called a POS tag set. Each language may have its own POS classification schemes in terms of nouns, verbs, adjectives, adverbs, etc.
- We adopt a tag set that is developed by IL-IL MT consortium to be applicable for all Indian languages, the tag set is developed to fulfill the needs of all Indian languages. It contains 25 tags covering all the major and minor categories of the language.
- Ex: Noun – NN, Verb – VM, Adverb- RB, etc….

- **Architecture of Rule-based tagger:**

Generally a rule-based tagger consists of three components.

1. Tokenizer
2. Morphological analyzer
3. Morphological disambiguator

- **Ex:**

Noun	**morph analyses**	**POS**
ceruvu	ceVruvu{aruvu v *AjFArWa* 2_e }/	ceruvu/VM
	ceVruvu{meku n eka *0* }/	ceruvu/NN
	ceVruvu{meku n eka *obl* }/	ceruvu/NN

- **Some disambiguation rules for illustration:**

The following are some of the POS tag disambiguating rules used in the task:

W_1 :: W_2 => W_1 :: W_2

(Where W_1 and W_2 a sequence of words in that order)

NN,VM::NN => NN::NN

UT, VM:: VM => UT: : VM

UT, VM :: NN => UT :: NN

Evaluation:

Overall Result: **WORDS**

GS POS text for comparison: 50, 094

Untagged corpus for rule based tagging: 20,154

Overall Result in Rule Based POS Tagger:

1. Identity: 17,540/20,154 = (87.02) %
2. Difference: 2614/20,154 = (12.97) %

Overall Result:	**Words**
Training text (GS POS tagged):	50, 094
Testing corpus :	20, 154

Overall Result in HMM tagger:

1. Identity: 17,013/20,154 = (84.14) %

2. Difference: 3141/20,154 = (15.58) %

Observations:

1. An HMM tagger considers the frequencies of patterns and sequence of words with their tags and computes the possibilities for assigning tags to individual words.
2. While tagging it doesn’t favor less frequented words hence there is a high probability that a word is restricted to only one category when it is ambiguous.

- The Rule-Based tagger is based on the morphology and syntax of the language hence there is a chance of identifying the errors and rectifying them.

- We can easily identify the problem and find a solution using the available linguistic knowledge. There is a possibility of reaching higher levels of accuracy in case of Rule- Based tagger.

Thank you