A Hybrid Approach for POS Tagging for Relatively Free Word Order Languages

Sobha Lalitha Devi,

T.Pattabhi Rama Krishna Rao

http://nlp.au-kbc.org/

Presentation Outline

- □ Introduction
- □ Approaches to POS
- □ Current Approach for POS - Hybrid method
- □ Results
- □ Conclusion

Part-of-Speech (POS) Tagging

- Task of labeling/assigning each word in a sentence with its appropriate syntactic category
- Symbols, punctuation markers etc. are also assigned specific tag
- Here it is modeled as sequence labeling task, each word in the sequence is labeled with its POS tag

POS Tagging … Contd

- This is one of the most basic preprocessing tasks and Important for all natural language processing(NLP) tasks
- Useful in Information extraction (IE), Information retrieval (IR)
- Helps in word sense disambiguation

Tagset

- Designing of Tagset is very important
- Tagset should consider all morphosyntactic categories of the language
- Tamil is an agglutinative and morphologically rich language
- Tagset consists of 17 basic Tags and 31 sub tags

Approaches to POS

- Different Approaches
  - Rule-based approach
  - Statistical & Machine learning techniques
  - Hybrid Approach

Rule-based Approach

- Rules are hand-crafted by language experts,requires exhaustive rules to be built
- Systems built using this approach have given good results
- One of the popular rule-based systems reported accuracy of 97.5 % (Brill, 1994)
- In Indian languages, POS tagger for Tamil (Arulmozhi et al., 2004) reports precision of 92 %

Statistical & Machine Learning Techniques

- □ Statistical methods are based on probability measures
- □ Probability measures include unigram, bigram, trigram and n-grams (Charniak, 1993)
- □ Availability of large annotated corpora has given rise to use of machine learning techniques
- □ In the recent past machine learning techniques such as HMM, MEMM, CRF ’s etc have been successfully used for this task

Statistical & Machine Learning Techniques - HMM

- Hidden Markov Model (HMM) is one of the popular technique used
- Here a few assumptions are made
  - Probability of item (word) in sequence depends on its immediate predecessor (word)
  - both the observed events and hidden events must be in a sequence

Statistical & Machine Learning Techniques - HMM

- A Hidden Markov Model (HMM) is a five-tuple $H = (\Sigma , Q, q0, A, B)$ where:
  - Σ is a finite observation alphabet;
  - Q is a finite set of states;
  - $q0 \in Q$ is the distinguished initial state;
  - $A : Q \times Q$ -> $[0..1]$ is a probability distribution on state transitions
  - $A (q1, q2)$ is the probability of a transition to state q2 from state q1
  - $B : Q \times \Sigma$ $[0..1]$ is a probability distribution on state symbol emissions
  - $B(q, a)$ is the probability of observing the symbol a when in state q

HMM … Contd

- probability distributions A and B are estimated from the tagged training corpus
- Its observed that using this technique has a drawback of data sparseness.
- smoothing algorithms such as Good-Turing algorithm, TNT etc are used to overcome data sparseness

Other Machine Learning techniques - MEMM

- Maximum Entropy Model (MEMM) is a technique which is used to overcome the problem of data sparseness

- Use of MEMM has a drawback of label bias, wherein certain nodes are skipped by the system due to very low score of probability measure

Drawbacks of Statistical and Machine Learning Techniques

- □ Even though the modern techniques have reported very encouraging results for Precision but their recall scores are not good enough

- □ Machine learning algorithms fail to model a natural language fully due to inherent linguistic complexities present in a language.

Drawbacks of Statistical and Machine Learning Techniques

- □ Use of machine learning techniques require annotated corpora of large sizes
- □ In Indian languages availability of very large sizes of annotated corpus is not easy
- □ Requires smoothing algorithms, but its found that even use of smoothing algorithms don't yield significant improvement of results

Rule-based systems

- Rule-based systems perform with good accuracy
- Doesn’t require large sized annotated corpus
- Hand-crafting rules exhaustively is very difficult and time consuming

Our Approach -Hybrid method

- Here we have used hybrid approach
- Tags of the words are taken into consideration and the script encoding is not considered
- Here we first use HMM technique to tag and after that rule-based algorithm is used

Hybrid Approach … Contd

- In the HMM part of the system, no statistical method of smoothing is used, instead rule-based system is called in for smoothing
- In literature, it is shown that use of linguistic smoothing increased the results significantly (Arulmozhi et al, 2006)
- Rule based algorithm has 7 context sensitive rules & 90 lexical rules
- The context sensitive rules can be applied across all Indian languages

Hybrid Approach

- An important question arises here **how would HMM system and rule-based system be merged**?
- In the HMM system we have found mainly three problems
  - Lexical sparseness and
  - structural sparseness
  - Words are tagged wrong

Lexical Sparseness

- Lexical Sparseness
  - In lexical sparseness, the system encounters new words which are not there in the training corpus.
  - This can be overcome by using the knowledge of sentence structure, using the transition probability distribution alone and neglecting the emission probability while calculating the score

Structural Sparseness

- Structural Sparseness
  - this type of sparseness is due to all possible sentence structures not being defined in the training corpus
  - When all possible sentence structures are not available in the training corpus, then in those instances the scores in the transition probability distribution becomes zero

… Contd

- When both transition probability and emission probability both become zero in such instances the whole sentence remains untagged by the HMM system.
- In such a case where a whole sentence is not tagged, rule-based system is called by our hybrid system and the sentence is tagged

Wrongly tagged Words

- How to identify words that are tagged wrong by the system, without human intervention?
  - It has been observed that certain tags which occur in low frequency in the training corpus, within the same class such as, class of Noun which have tags NNP, NN, PRP etc are often tagged wrong

… Contd

- Hence such tags are identified by calculating the frequencies of each tag in the training corpus.
- For such low frequency tagged words the rule-based algorithm is called by the hybrid system
- This reduces most of the errors in the system

Results

Results of Rule based system

Number of Tested Words	Totally Tagged words	Correctly Tagged Words	Precision
1000	1000	920	92%
1000	1000	918	91.8%

Results

Results of HMM System

Title	No. Of Words	Words tagged	Correctly Tagged	Precision	Recall
Set 1	1565	1216	996	81.9%	63.6%
Set 2	1810	1411	1161	82.28%	64.1%
Set 3	1616	1277	1063	83.2%	65.7%

Results

Results of Hybrid system

Output of the Hybrid Tagger			
Title	**Set 1**	**Set 2**	**Set 3**
No. Of Words	1565	1810	1616
Words tagged	1565	1810	1616
Words Untagged	0	0	0
Correctly Tagged	1520	1757	1572
Wrongly Tagged	45	53	44
Precision	97.12%	97.07%	97.27%
Recall	97.12%	97.07%	97.27%

Conclusion

- A hybrid approach for POS tagging for Indian Languages is presented in this paper
- Used Linguistic smoothing for HMM instead of traditional statistical smoothing methods.
- This system can be used for any Indian language.

Thank you