

Comparative Analysis of the Performance of CRF, HMM and MaxEnt for Part-of-Speech Tagging, Chunking and Named Entity Recognition for a Morphologically rich language

Manish Agarwal*, Rahul Goutam*, Ashish Jain*, Sruthilaya Reddy Kesidi*, Prudhvi Kosaraju*, Shashikant Muktyar*, Bharat Ambati and Rajeev Sangal

Language Technologies Research Center

International Institute of Information Technology

Hyderabad, AP, India - 500032

{manish.agarwal, prudhvi.kosaraju, rahul.goutam,
shashikant.muktyar, sruthilaya.kesidi, ambati}

@research.iiit.ac.in, ashishjain@students.iiit.ac.in,
sangal@iiit.ac.in

Abstract

In this paper, we present a comparative analysis between three methods for statistical part-of-speech(POS) tagging, chunking and named entity recognition(NER) for a morphologically rich language, Hindi, using a large annotated corpus. The methods explored are Conditional Random Fields(CRF), Hidden Markov Models(HMM) and Maximum Entropy Model(MaxEnt). We further propose an iterative approach as a method to improve the results. To the best of our knowledge, there is no previous work on comparative analysis of statistical POS tagging, chunking and NER in Hindi using the three methods when a large manually annotated corpus is used. The maximum POS tagging, chunking and NER accuracies for CRF, HMM and MaxEnt achieved are (94.00%, 91.70%, 56.03%), (92.96%, 89.23%, 48.21%) and (92.88%, 85.48%, 49.09%) respectively. Our work shows that CRF performs consistently better than HMM and MaxEnt for all of the three above-mentioned tasks.

1. Key Words

POS tagging, Chunking, NER and Iteration

2. Introduction

The objective of POS tagging is to assign part of speech tags to natural language text based on both its definition and its context. Chunking is the task of identifying and segmenting the text into syntactically correlated word groups. Named Entity Recognition (NER) seeks to locate and classify entities in a text into predefined categories such as the names of persons, organizations, locations, expressions

of times, quantities, etc. All 3 tasks are important sub-components of natural language analysis and information extraction. Various approaches to all three tasks have been explored, but they can be divided into two major categories, rule based and statistical. Three of the major statistical techniques applied are Conditional Random Fields, Hidden Markov Models and Maximum Entropy Models.

While considerable work has been done involving each technique for the three tasks(see related work), there is no comparative analysis of the three techniques on a large training corpus for a morphologically rich language like Hindi. Our work presents a comparative analysis of statistical POS tagging, chunking and Named Entity Recognition(NER) for Hindi using the three techniques - Conditional Random Fields (CRF, [4]), Hidden Markov Model (HMM, [3]) and Maximum Entropy Model (MaxEnt). We further propose an iterative method which can be used to mutually improve the performance of two related tasks. In this approach, the features used in task 1 can be implicitly used in the machine learning of task 2 and vice versa.

The data sparseness problem is the major motivation behind our iterative approach where the features used in task 1 can be implicitly used in task 2 and vice versa by performing iteration between the two tasks. The data is sparse in terms of the number of instances of each class of individual tags. We have explained it further in section 3.

3. Related Work

Lafferty et al. [4] proposed a conditional random field framework for POS tagging using the PENN treebank corpus. They showed that CRF outperforms HMM and HMM outperforms MaxEnt, which was attributed as a conse-

⁰The ordering among the star marked authors doesn't mean anything and they have contributed equally to this work

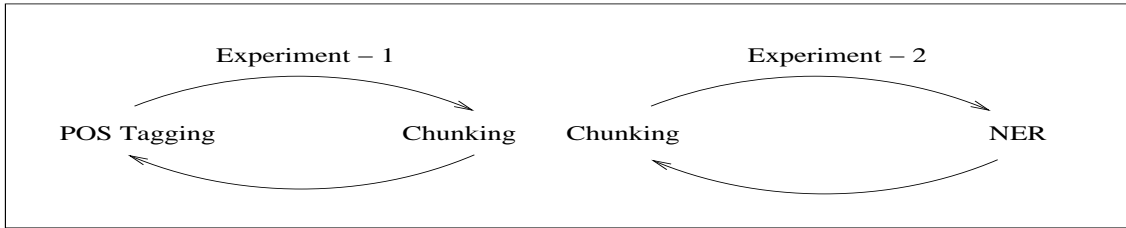


Figure 1. Iteration: Two experiments are performed, Experiment-1 for iteration between POS and Chunking, and Experiment-2 between Chunking and NER

quence of the label bias problem. Avinesh and Karthik [5] described POS tagging and chunking for Indian languages (Hindi, Telugu and Bengali) using CRF and Transformation Based Learning (TBL) techniques. They showed the importance of using morphological information during training for Hindi. Dalal et al. [1] used similar morphological and lexical features in POS tagging for Hindi using HMM. However, due to the different datasets being used by each of them, no conclusion can be drawn as to which technique performs better. Ratnaparkhi [6] presented a maximum entropy based statistical model for Part-of-Speech tagging using the PENN treebank corpus. He demonstrated the importance of using contextual information to improve the accuracy of the tagger.

Karthik et al. [2] described machine learning approaches with language specific heuristics to handle the task of Named Entity Recognition for Indian Languages. They have outlined some key linguistic issues in Indian languages like Hindi that need to be tackled in this task and are the motivation behind the use of language specific heuristics. These include the agglutinative nature of Hindi, no graphical cues like capitalization, spelling variation, etc.

4. Approach

For POS tagging and chunking, we tried out different combination of features like morphological information (root word, gender, number, person), previous word, previous POS tag, etc. These features were used and each of the three techniques was applied. The best obtained results and the corresponding feature sets are reported in Section 4.2.

4.1. NER

Generally features like capitalization, decimals, affixes, context, POS tags, etc. are used for NER. However, as described by Gali et. al, capitalization cannot be used for Indian languages like Hindi. Hence, apart from the language specific features described in [2] the previous part, certain other features were also used which are described below:

The best obtained result and the corresponding feature set is reported in Section 4.2.

Models	Features
S_no binary	If the word is beginning of the sentence and is a number
dgf_dig	If the word is a number
Dgf_4	If the word is 4 digit number

Table 1. NER Features

4.2. Iteration

Traditional cascaded paradigms treat tasks like POS tagging, chunking and NER as mutually exclusive tasks. However, this need not remain so for tasks that are closely related. An alternative way of approaching these tasks is what we term iterative design.

When we perform iteration between two tasks, the output of task1 is added to the feature set of task2 and vice versa. Thus, in the first iteration, task1 is trained independently and its output is included in the feature set for the training of task2. In the second iteration, the output of task2 from the previous iteration is included in the feature set for the training of task1. The output of task1 is then included in the feature set for training of task2. In this way, the process continues for a finite number of iterations. Using the above technique, features used in task1 are being implicitly used in task2 and vice versa. This is a method to deal with the data sparseness problem where a feature, which could not be directly used in the training of task2 due to sparseness but was used in task1, is now indirectly being used via iteration of task 1 with task2.

Iteration was performed between POS tagging-chunking and chunking-NER as shown in Figure 1. The motivation behind iteration between POS tagging and chunking is that while chunking is certainly improved by using POS tags as features (as shown in section 4.2), POS tagging can also benefit from using chunk information as it will help determine the boundaries of compound nouns which cannot span across multiple chunks. The motivation behind performing iteration between chunking and NER is that named entities should not span across multiple chunks and, thus, the chunk boundary information is a useful feature for the task of NER.

The best model obtained for each task individually were used in the iterative experiments. For POS tagging-

Models	Features	NER (F measure)
CRF	Word, S_no, dfg_dig, dfg_4, suffix(length 4) prefix(length 4), current words pos tag	56.03%
Maximum Entropy	Word, suffix(length 4), prefix(length 4), previous word, Word, suffix(length 4), prefix(length 4), category, next word, previous word, current word's pos tag	48.21%
HMM	Word, current word's pos tag	49.09%

Table 2. NER: Best Features and Accuracies

chunking, there are two ways to start the iterative method. We can start by chunking and use the output in POS tagging or vice-versa. POS tagging was chosen as task1 as POS tags are part of the feature set used to obtain the best chunking model. Similarly, for chunking-NER, chunking was chosen as task1 as the accuracy of the best NER model was too low and using NER output as feature for chunking decreased its accuracy significantly in the first iteration itself.

The results obtained for iteration between POS tagging-chunking and chunking-NER are reported in Section 4.2.

4.3. 4 - > 2 tag scheme

System uses this tagging scheme in HMM in chunking. In this tagging scheme, firstly chunking is done using 4 tag scheme (STRT, CNT, STP and STRT_STP) then system converts them into 2 tag scheme (STRT and CNT), as explained in [7]. System uses *B* for *STRT*, *I* for *CNT*, *E* for *STP* and *BE* for *STRT_STP*.

5. Experimental Setup

5.1. Data

For POS tagging, chunking, and POS tagging-chunking iteration experiments, the data consisted of 190K words and was divided as 150K for training and 40K for testing. For NER experiments and chunking-NER iteration experiments, the data consisted of 400K words and was divided into 320K for training and 80K for testing.

The tagset used for POS tagging and chunking has 26 POS tags and 11 chunk tags. The tagset used for the NER task has 12 tags, are shown in Table 3. In addition to these, in case of chunking and NER, we added the prefix *B* or *I* to indicate the beginning of boundary or inside boundary. For example, if the tag is *XX*, the first word will have tag *B-XX* and the subsequent words, including the last word, will have tag *I-XX*.

5.2. Results

The best obtained accuracies and the corresponding feature sets for POS tagging, chunking and NER are shown in Table 2, 4 and 5 respectively.

For iteration between POS tagging and chunking, we observed a slight improvement in the accuracy of POS tagging

NER tag	Description
NEP (Person)	'Orhan Pamuk' or 'Mark Twain'
NED (Designation)	'Chairman' (as in 'Chairman Mao') or just 'The Chair'
NEO (Organization)	'State Government' or 'Microsoft'
NEA (Abbreviation)	'IBM' (or I.B.M.) or 'CRF'
NEB (Brand)	'Fanta' or 'Windows'
NETP (Title-Person)	'Mr.' or 'Shri' or 'Mahatma'
NETO (Title-Object)	'The Seven Year Itch' or 'American Beauty'
NEL (Location)	'Delhi' or 'New Delhi'
NETI (Time)	'10th July', '1968', '5 pm'
NEN (Number)	'Fifty five', '3.14', 'one lakh'
NEM (Measure)	'five kilos', 'three days', 'seven years'
NETE (Terms)	'Horticulture', 'Conditional Random Fields'

Table 3. NER tagset

from 93.70% to 94%. There was no significant improvement in the accuracy of chunking.

The results for iteration between chunking and NER are shown in Table 6.

6. Error Analysis

The introduction of contextual features and morphological features (root, gender, number, person, suffixes and prefixes) helped in increasing the accuracy of POS tagging and chunking. Iteration between POS tagging and chunking further helped improve the results as expected.

A detailed error analysis and tag-wise precision/recall calculation for POS tagging showed that the maximum confusion existed between *NNP-NN* and *NN-JJ* with the F-Measure values being 85.69% and 87.02% respectively for CRF.

For chunking, the maximum confusion was between *B-NP* and *I-NP*. We believe the reason behind it is that the taggers were not able to recognize the chunk boundaries. Thus, as *NP* chunk is, by far, the most frequent chunk in the test data, the boundary of *NP* chunk is not correctly identified in many cases. Output tag sequences like [*B-YY I-YY I-XX*] were also observed in many cases. This shows that

Iteration	CRF		Max-Ent		HMM	
	chunk acc	NER-F	chunk acc	NER-F	chunk acc	NER-F
Iteration-0	91.70%	55.56%	90.32%	48.21%	84.92%	44.69%
Iteration-1	91.60%	55.53%	90.28%	47.79%	85.35%	44.36%
Iteration-2	91.60%	55.53%	90.40%	46.62%	85.34%	44.32%
Iteration-3	91.60%	55.53%	90.35%	46.01%	85.34%	44.08%

Table 6. Chunk-NER Iteration: Accuracies

Models	Features	POS Accuracy
CRF	Word, suffix(length 4), prefix(length 4), root, category, gender, number, person	93.70%
Maximum Entropy	Word, suffix(length 4), prefix(length 4), next word, previous word, category, previous words pos tag	92.96%
HMM	Word, category	92.88%

Table 4. POS Tagging: Best Features and Accuracies

Models	Features	Chunking Accuracy
CRF	Word, current pos tag	91.70%
Maximum Entropy	Word, suffix(length 4), prefix(length 4), previous word, next word, current word's pos tag, previous word's pos tag	89.23%
HMM	Word, current pos tag (using 4 to 2 tagging scheme)	85.48%

Table 5. Chunking: Best Features and Accuracies

the taggers were not able to learn the rule that a new chunk has to start with a B-XX tag where XX is the chunk type.

Chunk-NER iteration did not show an improvement in the results as expected. We believe this is due the data sets being used for training the chunk model and NER model, which were different. Furthermore, as explained in section 3.2, our primary motivation behind the iterative approach was to implicitly give features to a task. We have tried to achieve this by including the output of one task in the feature set of another. However, this helps only when the output of each individual task is moderately accurate, which is not true in this case. As shown in section 4.2, for CRF, the best accuracy for chunking is 91.7% and for NER is 56%. We think NER has too low accuracy to be iterated with chunking as the errors in the output tags also get propagated to the next iteration. We believe that selective propagation of features to the next iteration by filtering out errors can produce better results.

7. Conclusion and Future work

In this paper, we have presented a set of syntactic and semantic features which together give the best results for POS tagging, chunking and NER for morphologically rich language, Hindi, using three different techniques. We have also introduced an approach of iteration between POS tagging - chunking and chunking-NER towards obtaining better results for the individual tasks. In future, we plan to further refine the approach of iteration by selective propagation of features. We also plan to conduct experiments on a single data set for chunk and NER where we hope to obtain better results.

References

- [1] A. Dalal, K. Nagaraj, U. Swant, S. Shelke, and P. Bhat-tacharyya. Building feature rich pos tagger for morphologically rich languages: Experience in hindi. In *ICON, 2007*, 2007.
- [2] K. Gali, H. Surana, A. Vaidya, P. Shishtla, and D. Sharma. Aggregating machine learning and rule based heuristics for named entity recognition. In *Workshop on NER for South and South East Asian Languages, IJCNLP 2008*, 2008.
- [3] F. Jelinek. Speech recognition by statistical methods. In *Proceedings of the IEEE, Vol. 64, 532-556, April 1976.*, 1976.
- [4] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceeding ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.
- [5] A. PVS and K. G. Part of speech tagging and chunking using conditional random fields and transformation based learning. In *Proc. of SPSAL2007, IJCAI, India, 21-24.*, 2007.
- [6] A. Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Empirical Methods in Natural Language Processing (1996)*, pp. 133-142, 1996.
- [7] A. Singh, S. M. Bendre, and R. Sangal. Hmm based chunker for hindi. In *Proc. the Second International Joint Conference on Natural Language Processing, 11-13 October, 2005, Jeju Island, Republic of Korea*, 2005.