Hindi Named Entity Recognizer for NER task of FIRE 2013

Rahul Sharnagat and Pushpak Bhattacharyya

IIT Bombay, Mumbai, India
{rdsharnagat@cse.iitb.ac.in,
pb@cse.iitb.ac.in}

Abstract. In this work we describe a hierarchial Conditional Random Field approach to Named Entity Recognition task. We employed the context features and suffix features for classification. Hierarchal learning for multilevel Named entity Recognition task is beneficial for getting high accuracies for secondary level tags. For a small dataset reserved for testing purpose, we achieved F1 score of 0.86 for level 1 , 0.99 for level 2 and 0.96 for level 3.

1 Introduction

Named Entity Recognition(NER) is the process of locating a word or a phrase that references a particular entity within a text. Identification of named entities is crucial for various application like information extraction, machine translation, cross lingual information access system to attain high accuracies. For Indian languages, NER is a difficult task. This can be attributed to various reasons like no concept of capitalization in Indian languages and richer morphology. Many Indian languages are agglutinative and lack of availability of resources makes the task more difficult.

Various approaches have been tried for sequence labeling like HMM[1], Maximum Entropy based models[2], Support Vector Machines and CRF[3][4] etc. In this work, we use CRF as our base model. Since the problem is not single sequence problem, we trained a hierarchical CRF model where on each level a new model is trained based on result of previous model. We observed that for the tags on higher level, only the words and label in previous layer are important for classifying the word on that level. We chose Hindi Dataset for experiments.

2 Problem Definition: Defining the task

The task is about tagging sentences for Named Entities with a set of possible tags to identify the category of the named entity. Let say we have a set of tagged sentence with POS tag and Chunking information, we need to annotate the sentence by detecting and classifying named entities in it. There are three tags for each word. Each tag differs in coarseness of semantic definition. There are 22 tags in level 1, 44 tags in level 2 and 25 tags in level 3 for Hindi language.

The task is to learn a model from the training set which can be used to tag the sentences with multilevel named entities tags.

3 Our Approach

We propose a hierarchal learning for multilevel tagging proposed in the task. Conditional Random Field (CRF) has been extremely successful in sequence tagging problems. We employed the CRF for multilevel learning.

Input
We start with the set of labeled training data D. We remove the level 2 and level 3 from training set.

Step 1: Preprocessing
As a preprocessing step, we converted all the digits in the training to uniform representation. For example, year 2007 will become year DDDD.

Step 2: CRF training for level-1
For training, we used a hierarchal training approach where we learn the model in steps. We first learn a model for fine grained tagging of base level tags. We use following set features for learning CRF model :

- Trigram, bi-gram and uni-gram of words
- Bi-gram and uni-gram of POS tags
- Bi-gram and uni-gram of Chunk tags
- Suffix of words

Suffixes are the last 4 character of the word.

Step 3: CRF training for level-2
For second level of tagging we used the results level-1 for learn second model for tagging second level tags. We removed the POS tag and chunk information from the training set. We used trigram, bi-gram of words and bi-gram of level-1 tags are used as features to get level-2 tags.

Step 4: CRF training for level-3
For level three tags, we used the results from level-1 and level-2 and words to tag the level-3. In this training set we have only words, level 1 tags and level 2 tags. Bi-gram of words, uni-gram and bi gram of level-1 and level-2 results were used for training the model.

4 Results and Conclusion

We experiment with the FIRE 2013 Named Entity Recognition dataset for Hindi language. Dataset contains 116103 words. We reserved 5000 words for testing

purpose and trained the hierarchal CRF model on rest 111103 words.For such small dataset for testing, results were not comprehensive. Many possible tags are missing from the test dataset. Our results are tabulated below. From the results, we can observe that the results of higher level are extremely good given the tag in previous level are tagged correctly. This correlates with hypothesis we stated earlier.[1]

Levels	Precision	Recall	F1 score
1	0.9738	0.7743	0.8627
2	0.9912	0.9912	0.9912
3	1.0000	0.9231	0.9600

Table 1. Results

References

1. D. M. Bikel, R. Schwartz, and R. M. Weischedel, "An algorithm that learns whats in a name," *Mach. Learn.*, vol. 34, pp. 211–231, feb 1999.
2. A. E. Borthwick, *A maximum entropy approach to named entity recognition.* PhD thesis, New York, NY, USA, 1999. AAI9945252.
3. A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, (Stroudsburg, PA, USA), pp. 188–191, Association for Computational Linguistics, 2003.
4. J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, (San Francisco, CA, USA), pp. 282–289, Morgan Kaufmann Publishers Inc., 2001.

[1] Results are rounded off to 4 decimal places.

Appendix: Accuracies of different levels

Tags	Precision	Recall	F1 score
B-ARTIFACT	0.9000	0.7500	0.8182
B-COUNT	1.0000	0.8095	0.8947
B-DATE	1.0000	0.5000	0.6667
B-DAY	1.0000	1.0000	1.0000
B-DISEASE	1.0000	1.0000	1.0000
B-FACILITIES	1.0000	1.0000	1.0000
B-LIVTHINGS	1.0000	0.3000	0.4615
B-LOCATION	1.0000	0.8889	0.9412
B-MATERIALS	1.0000	0.7273	0.8421
B-MONTH	1.0000	0.5000	0.6667
B-ORGANIZATION	1.0000	0.5000	0.6667
B-PERIOD	1.0000	1.0000	1.0000
B-PERSON	0.9444	0.4722	0.6296
B-QUANTITY	1.0000	0.8889	0.9412
B-TIME	1.0000	1.0000	1.0000
B-YEAR	0.9474	0.9474	0.9474
I-ARTIFACT	0.8000	0.8889	0.8421
I-COUNT	1.0000	1.0000	1.0000
I-DISEASE	1.0000	1.0000	1.0000
I-FACILITIES	1.0000	1.0000	1.0000
I-LIVTHINGS	1.0000	0.6667	0.8000
I-LOCATION	1.0000	1.0000	1.0000
I-MATERIALS	1.0000	1.0000	1.0000
I-ORGANIZATION	1.0000	0.5000	0.6667
I-PERIOD	1.0000	1.0000	1.0000
I-PERSON	0.7500	0.3333	0.4615
I-QUANTITY	1.0000	1.0000	1.0000
I-TIME	1.0000	1.0000	1.0000
Total	0.9738	0.7743	0.8627

Table 2. Accuracies for level 1 tagging

Tags	Precision	Recall	F1 score
B-CHEMICAL	1.0000	1.0000	1.0000
B-FOODMAT	1.0000	1.0000	1.0000
B-HUMAN	1.0000	1.0000	1.0000
B-INDIVIDUAL	0.9697	1.0000	0.9846
B-INSTITUTE	1.0000	1.0000	1.0000
B-MANMADE	1.0000	1.0000	1.0000
B-MEDICINAL	1.0000	1.0000	1.0000
B-NONHUMAN	1.0000	1.0000	1.0000
B-PLACE	1.0000	1.0000	1.0000
B-PVT	1.0000	1.0000	1.0000
B-SYMPTOMS	1.0000	1.0000	1.0000
B-TREATMENT	1.0000	1.0000	1.0000
I-CHEMICAL	1.0000	1.0000	1.0000
I-FOODMAT	1.0000	1.0000	1.0000
I-HUMAN	1.0000	1.0000	1.0000
I-INDIVIDUAL	1.0000	1.0000	1.0000
I-INSTITUTE	1.0000	1.0000	1.0000
I-MANMADE	1.0000	1.0000	1.0000
I-NONHUMAN	1.0000	1.0000	1.0000
I-PLACE	1.0000	1.0000	1.0000
I-PVT	1.0000	1.0000	1.0000
I-SYMPTOMS	1.0000	1.0000	1.0000
I-TREATMENT	1.0000	1.0000	1.0000
Total	0.9912	0.9912	0.9912

Table 3. Accuracies for level 2 tagging

Tags	Precision	Recall	F1 score
B-CITY	1.0000	1.0000	1.0000
B-NATION	1.0000	1.0000	1.0000
B-RELPLACE	1.0000	1.0000	1.0000
I-NATION	1.0000	1.0000	1.0000
I-RELPLACE	1.0000	1.0000	1.0000
Total	1.0000	0.9231	0.9600

Table 4. Accuracies for level 3 tagging