Named Entity Recognition for Telugu News Articles using Naïve Bayes Classifier

SaiKiranmai Gorla Sriharshitha Velivelli N L Bhanu Murthy Aruna Malapati

Birla Institute of Technology and Science, Pilani, Hyderabad, India
{p2013531, f20130847, bhanu, arunam} @ hyderabad.bits-pilani.ac.in

Abstract

The Named Entity Recognition (NER) is identifying name of Person, Location, Organization etc. in a given sentence or a document. In this paper, we have attempted to classify textual content from on-line Telugu newspapers using well known generative model. We have used generic features like contextual words and their part-of-speech (POS) to build the learning model. By understanding the syntax and grammar of Telugu language, we propose morphological pre-processing of the data and this step yields us better accuracy. We propose some interesting language dependent features like post-position feature, clue word feature and gazetteer feature to improve the performance of the model. The model achieved an overall average F1-Score of 88.87% for Person, 87.32% for Location and 72.69% for Organization.

1 Introduction

News providers and publishing companies generate humongous amount of unstructured textual content on daily basis. This content is not of much use if there are no tools and techniques for searching and indexing the text. Named Entity Recognition (NER) is an important task in Natural Language Processing (NLP) to figure out the named entities in text documents. Named Entities (NEs) are usually proper nouns like name of Person, Organization, Location etc in text documents. Named Entity Recognition can naturally be applied to news articles to identify named entities in those articles. Knowing these named entities in each article help in categorizing the news articles in defined position and empower smooth information detection. NER task was first presented at MUC-6 in 1995 [GS96] and since then, the task has undergone several transitions beginning from the rule based approaches to the currently used Machine learning techniques. The performance of NER task for different languages depends on the properties of the language.

In English, capitalization feature play an important role as NEs are generally capitalized in this language. The capitalization feature is not available for Indian Languages (IL) which makes the task more challenging. In this paper, we attempt to get some insights and results of NER for Telugu language. The challenges in NER specific to Telugu language are: a) no capitalization b) two words in English can be mapped to one word in Telugu. *Example: in Delhi (English):* ఢిల్లీలో *Dhilli**lO*** where (లో) lO is an post-position marker c) absence of part-of-speech tagger d) free word ordering.

In this paper, a generative model is proposed for NER task using Naïve Bayes classifier. The following features have been considered for training the model - contextual word, part-of-speech tag, gazetteer as a binary feature, post-position feature and clue word feature. We are mainly interested in classifying a given word to one of the named entities namely Person, Location and Organization. The results obtained from the proposed approach are comparable to other competitive techniques for Telugu langauge.

The rest of the article is organized as follows. We discuss related work in Section 2 and illustrate dataset in Section 3. The methodology and evaluation metrics are presented in Section 4. In Section 5 and Section 6 we propose features to build Naïve Bayes Classifier and discuss the results. The conclusions of our study are summarized in Section 7.

In: D. Albakour, D. Corney, J. Gonzalo, M. Martinez, B. Poblete, A. Vlachos (eds.): Proceedings of the NewsIR'18 Workshop at ECIR, Grenoble, France, 26-March-2018, published at http://ceur-ws.org

2 Related Work

The NER task, can be approached in two ways: by hand-crafted rules and statistical machine learning techniques [Sar08]. A rule-based approach for NER tasks require patterns which can describe the internal structure and contextual rules which give clues for identification and classification. An example of such rules can be a street name if phrase ends with the word 'X 'proceeded by preposition word "Y ", where 'X 'can be "street "and 'Y 'could be 'in 'from sentence such as 'The Apple store in jail street in hyderabad'.

Some of the ruled-based systems include FASTUS [AHB+95] which uses regular expressions to extract Named Entities (NEs). LaSIE and LaSIE II [HGA+98] uses look up lists of NE to identify NEs. The ruled-based systems are efficient for domain specific like biological domain where certain formulation in terminology. Some biological NER task include [ARG08]. The limitation of ruled-based approach is that they require expert about the knowledge of the language and domain. These knowledge resources take time to build and not transferable to other domains. Hence, NER has been solved using machine learning approaches.

Machine learning approaches can be classified into three different approaches: Supervised learning, Semi-supervised learning and Unsupervised learning. In Supervised learning (SL), labeled training data with features is given as an input to the model, which can classify new data. Some of the SL algorithms are Support Vector Machine [TC02], Condidtional Random Field [ANC08], Hidden Markov Model [SZS+04], Neural Network [KT07], Decision tree [FM09], Naïve Bayes [MH05] and Maximum Entropy Model [CN02]. In semi-supervised learning the model makes use of both labeled and unlabeled data. The popular Semi-supervised learning in NER are boot-strapping [Kno11] and Co-training [CS99]. Most of the Unsupervised learning approaches in NER are clustering and distributional statistics using similarity functions.

In Indian Languages considerable amount of work has been done in Bengali, Hindi. Ekbal et.al [EB08] developed an NER system for Bengali and Hindi using SVM. These systems use different contextual information of words in predicting four NE classes, such as Person, Location, Organization and miscellaneous. The annotated corpora consists of 122,467 tokens for Bengali and 502,974 tokens for Hindi. The system has been tested with 35K and 60K tokens for Bengali and Hindi with an F1-score 84.15% and 77.17% respectively. Ekbal et.al [EB09] developed the NER system using CRF for Bengali and Hindi using contextual features with an F1-Score of 83.89% for Bengali and 80.93% for Hindi.

A very small amount of work is done in Telugu NER. Srikanth and Murthy [SM08], have used part of LERC-UoH Telugu corpus where CRF based Noun Tagger is built using 13,425 words. This has been considered as one of the feature for rule-based NER system for Telugu mainly focusing on identifying Person, Location and Organization without considering POS tag or syntactic information. This work is limited to only single word NEs. Praneeth et.al [SGPV08] build CRF based NER system with language independent and dependent features. They have conducted experiments on data released as a part of NER for South and South-East Asian Languages (NERSSEAL) [1] competition with 12 classes and obtained F1–score of 44.89%.

3 Dataset and Pre-processing

3.1 Corpus

Telugu Newspaper corpus is generated by crawling through newspaper websites[2] [3]. The corpus is annotated with three NE classes namely Person, Location, Organization and one not named entity class. The annotation was verified by Telugu linguists. The annotated data consists of 54,457 words out of which 16,829 are unique word forms. The number of named entities in the corpus are 2658 Persons, 2291 Locations and 1617 Organizations.

3.2 Morphological Pre-processing

Morphology is the study of word formation: how words are formed from smaller morphemes. A morpheme is the smallest part of a word that has grammatical information or meaning.

For Example: The word *trainings* has 3 morphemes in it: *train_ing_s*

As discussed in Section 6.1, Telugu is a highly inflectional and agglutinating language and hence it makes all sense to perform morphological pre-processing. In this work, we perform morphological pre-processing to only Nouns in the dataset because most of the NEs are Nouns.

For Example:

1. హైదరాబాదులో (haidarAbAdlo) = హైదరాబాదు_ లో
2. బిజెపికి (bijepiki) = బిజెపి_ కి
3. కవితకు (kavitaku) = కవిత_ కు

We would like to explore the significance of this morphological pre-processing step and hence put up results with and without this pre-processing step. The results unarguably signifies the importance of this step.

[1] http://ltrc.iiit.ac.in/ner-ssea-08/
[2] http://www.eenadu.net/
[3] http://www.andhrajyothy.com/

4 Methodology & Evaluation Metrics

4.1 Methodology

We have considered 11 features for every word in a sentence and classify each word to one of the three named entities namely Person, Organization, Location and one NNE class(Not a Named Entity). Thus there are $D(11)$ features for every word and each is to be classified into 4 classes say c_1, c_2, c_3 and c_4. Naïve Bayes classifier is a generative model where in posterior probability of a word belonging to a particular class, c_i where i=1 to 4, given the feature vector of the word, $(x_1, x_2, .., x_D)$, is computed by making use of Bayes theorem. Assuming the conditional independence of features given particular class, the posterior probability will be calculated as follows:

$$p(c_i|(x_1, x_2, ..., x_D)) = \frac{p((x_1, x_2, ..., x_D)|c_i)p(c_i)}{\sum_{i=1}^{4} p((x_1, x_2, ..., x_D)|c_i)}$$

$$= \frac{p(x_1|c_i)p(x_2|c_i)....p(x_D|c_i)p(c_i)}{\sum_{i=1}^{4} p(x_1|c_i)p(x_2|c_i).....p(x_D|c_i)}$$

The prior probability, $p(c_i)$, and conditional probabilities $p(x_1|c_1), p(x_2|c_2), ..., p(x_D|c_i)$ are estimated from the training data. The posterior probabilities for each of the class is computed and the word is classified into the class of maximal posterior probability.

The algorithm is implemented in C++. It is applied 50 times on the data. In each round, 70% of the sentences are randomly chosen for training and the remaining 30% are considered for testing. The results provided in the tables in Section 5 and Section 6 are the average of 50 rounds.

4.2 Evaluation Metrics

The standard evaluation measures like Precision, Recall, F1–score are considered to find out the prediction accuracies of proposed model.

$$Precision(P) = \frac{c}{r}$$

$$Recall(R) = \frac{c}{t}$$

$$F1 - Score = \frac{2 * P * R}{P + R}$$

where r is the number of NEs predicted by the model, t is the total number of NEs present in the test set and c is the number of NEs correctly predicted by the model.

5 Contextual features and Naïve Bayes Classifier

Orthographic features (capitalization or digits), suffix, preffix, NE specific words, gazetter features, POS etc. are generally used for NER. In English, capitalization feature play an important role as NEs are generally capitalized in this language. Unfortunately this feature is not applicable for the Indian languages.

The contextual word and POS features are used to build the prediction model. For window size of 3, contextual features are the current word (w_0), previous word (w_{-1}) and the next word (w_{+1}). The corresponding POS features are part-of-speech of (w_0) and (w_{-1}) and (w_{+1}) represented by pos_0, pos_{-1} and pos_{+1} respectively. The experiments was also repeated for window size 5. The Precision, Recall and F1–score remain more or less the same.

Let us consider the following example: సీత(NNP) తన(PRP) దుస్తులు(JJ) ఇష్టపడ్డారు(VB) (Sita likes her dress). For window size of 3, the contextual word and POS tags of the current word దుస్తులు are తన (w_{-1}), ఇష్టపడ్డారు (w_{+1}), PRP (pos_{-1}), VB (pos_{+1}).

We show that the six features are conditionally independent given the class label where TAG can be Person, Location, Organization and NNE (not a named entity).

1. The features w_i and pos_i are independent.

   $$p(w_i|pos_i, TAG) = p(w_i|TAG)$$

   Most of the times a word can be tagged with only one POS tag. It cannot have different POS tags. The chances of a word being tagged under different POS tags based on context is very rare. For example, consider the word "*John*". Its POS tag is proper noun and it is the only possible POS tag for this word. So conditioning a word on its POS tag will not change its probability.

2. The features w_i and w_j $(i \neq j)$ are independent.

   $$p(w_i|w_j, TAG) = p(w_i|TAG)$$

   Telugu being a free order language, any word can occur before/after a particular word. The probability of occurrence of any word before/after a particular word is same. It can also be seen as words occur with uniform probability.

3. The features w_i and pos_j $(i \neq j)$ are independent

   $$p(w_i|pos_j, TAG) = p(w_i|TAG)$$

   Since the condition is true for the words, it will definitely be true for their POS tags.

4. The features pos_i and pos_j $(i = j)$ are independent.

$$p(pos_i|pos_j, TAG) = p(w_i|TAG)$$

The posterior probability can be represented as:
$p(w_i = Person|w_{-1}, w_0, w_{+1}, pos_{-1}, pos_0, pos_{+1}) = p(w_{-1}, w_0, w_1, pos_{-1}, pos_0, pos_1|Person)p(Person)$

Applying chain rule to the likelihood probability term:
$p(w_{-1}, w_0, w_{+1}, pos_{-1}, pos_0, pos_{+1}|Person) =$
$p(w_{-1}|w_0, w_1, pos_{-1}, pos_0, pos_1, Person) \times$
$p(w_0|w_1, pos_{-1}, pos_0, pos_1, Person) \times$
$p(w_1|pos_{-1}, pos_0, pos_1, Person) \times$
$p(pos_{-1}|pos_0, pos_1, Person) \times p(pos_0|pos_1, TAG) \times p(pos_1|Person)$

Posterior probability after applying conditional independence on the features will be:
$p(Person|w_{-1}, w_0, w_{+1}, pos_{-1}, pos_0, pos_{+1}) =$
$p(w_{-1}|Person) \times p(w_0|Person) \times p(w_1|Person) \times p(pos_{-1}|Person) \times p(pos_0|Person) \times p(pos_1|Person) \times p(Person)$

As the conditional independence holds good for all features, we train the model with 70% of the data and test on the remaining 30% of the data. The average prediction accuracies of several runs have been reported in Table 1.

Table 1: Naïve Bayes Classifier with contextual word and its POS features

NE classes	Precision	Recall	F1-Score
Person	82.78	89.50	86.01
Location	78.15	87.01	82.35
Organization	39.18	47.86	43.09

As discussed in Section 3.2, morphological pre-processing of each word in the dataset is considered and the results are presented in Table 2. It is interesting to observe that there is improvement after the morphological pre-processing.

Table 2: After morphological pre-processing

NE classes	Precision	Recall	F1-Score
Person	85.16	90.34	87.67
Location	80.09	91.96	85.62
Organization	42.30	52.63	46.90

Though the overall results put up decent performance, but F1-score of Organizations is not impressive. We will introduce language dependent features to improvise the overall performance and prediction accuracies of organization.

6 Language dependent features and building comprehensive Naïve Bayes Classifier

Language dependent features are used to enhance the performance of the classifier. We propose couple of langauge dependent features and they are illustrated in the below sub-sections.

6.1 Post-position (PSP) feature

Telugu is highly inflectional and agglutinating language. The way lexical forms get generated in Telugu are different. Words are formed by productive derivation and inflectional suffixes to roots or stems as explained in Section 3.2. Some of the PSP markers in Telugu are లో (lO), కు (ku), కి (ki) etc. We propose a boolean feature whose value is 1 if a Proper noun (NNP) is followed by a postpostion otherwise 0. The statistics of PSP following the NEs are shown in Table 3. We build a Naïve Bayes Classifier with contextual word and POS features along with PSP feature and average accuracies of several runs are shown in Table 4.

Table 3: Statistics on Postposition followed after a proper noun

Named Entity	No. of times NNP followed by PSP
Person	205
Location	523
Organization	32
Not a NE	15

Table 4: Naïve Bayes Classifier with contextual word and its POS features, PSP feature

NE classes	Precision	Recall	F1-Score
Person	84.42	89.85	87.04
Location	80.91	90.53	85.45
Organization	40.73	52.56	45.90

6.2 Clue words for Organization

Clue words plays an important role for identifying NEs. In this work we considered clue words for recognizing organization. Since organizations is a multi-word and they tend to end with few suffixes like మండలి (Council), సంఘం(Company), సంఘం (Community), సమఖ్యా (Federation), క్లబ్ (Club) etc. We build a Naïve Bayes Classifier with contextual word and POS features, PSP feature along with Clue word feature and average accuracies of several runs are shown in Table 5.

Since the list of suffixes are as exhaustive as possible for Telugu names, we would expect predominant increase in accuracy for organization. But that is not

Table 5: Naïve Bayes Classifier with contextual word and its POS features, PSP feature, Clue word feature

NE classes	Precision	Recall	F1-Score
Person	84.54	90.11	87.24
Location	79.60	91.84	85.28
Organization	45.72	59.79	51.81

the case here because the words like 'సంఘం (Community) 'are tagged in the corpus as organization and not a named entity equal number of times. Hence, there is not much of improvement in accuracies.

6.3 Constructing Gazetteer from Wikipedia

In this section, we explain the process of building Gazetteer for NEs from Wikipedia. Wikipedia keeps up the list of categories for each of its title. For example, the Wikipedia categories are 'Educational institutions established in 1926', Companies listed on the 'Bombay Stock Exchange' refer to the names of Organization whereas 'Living people', 'Player' refer to Person whereas 'States and territories', 'City-states' refer to Location.

The following are steps for constructing gazetteers:

- Initially we manually constructed a list of seeds for Person, Location and Organization. We then search each seed in Wikipedia and extract the categories in order to construct the list of categories (*category_ list*) for each NE class.
- In order to resolve ambiguity we remove the categories that are present in more than one NE class in the category list and call it as *Unique_ category_ lists*. For example the category list may contain 'actor', 'engineer'and 'famous' for NE Person and 'city', 'street' and 'famous' for NE Location. The category label 'famous' is removed because it is present in both NE Person and Location
- We extract list of Wikipedia titles using Telugu Wikipedia dump [4].
- Then we start searching the category labels in *Unique_ category_ lists* of each NE class in Wikipedia dump. The *Unique_ category_ lists* having maximum matches is assigned as NE class for that NE.

We have generated a list of 7,593 Person names, 4,791 Location names, and 254 Organizations after the following the above mentioned procedure.

Example for Person name 'Mahendra Singh Dhoni (మహేంద్రసింగ్ ధోని)' belongs to categories (వర్గాలు in Telugu) as shown in Figure 1 [5].

[4]https://dumps.wikimedia.org/tewiki/

[5]https://te.wikipedia.org/wiki/మహేంద్రసింగ్ _ధోని

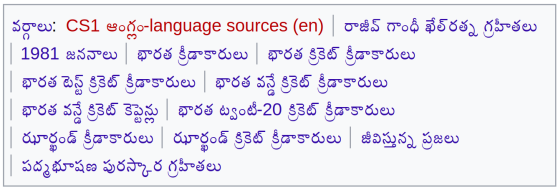

Figure 1: Example of wikipedia category in Telugu

The gazetteer feature enhanced the performance accuracies and the results are shown in Table 6. The

Table 6: Naïve Bayes Classifier with contextual word and its POS features, PSP feature, Clue word feature, gazetter feature

NE classes	Precision	Recall	F1-Score
Person	86.48	91.41	88.88
Location	84.38	90.48	87.32
Organization	63.40	85.16	72.70

F1-Score of Organization has been increased by 19% and there are impressive improvements for other NEs as well.

7 Conclusion

In this paper, we have attempted to classify Named Entities in Telugu News articles using Naïve Bayes classifier. The prediction accuracies of learning models have been enhanced significantly after the data being morphologically preprocessed as proposed in this work. The language dependent features, proposed in this paper, improve prediction accuracies where in a notable increase of 26% in the F1-score of Organization is observed. The comprehensive learning model built with contextual words and their parts of speech along with proposed language dependent features achieved an overall average F1-Score of 88.87% for Person, 87.32% for Location and 72.69% for Organization.

References

[AHB+95] Douglas E. Appelt, Jerry R. Hobbs, John Bear, David Israel, Megumi Kameyama, Andy Kehler, David Martin, Karen Myers, and Mabry Tyson. Sri international fastus systemmuc-6 test results and analysis. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995.

[ANC08] Andrew Arnold, Ramesh Nallapati, and William W. Cohen. Exploiting feature hierarchy for transfer learning in named entity recognition. In *Proceedings of ACL-08:*

HLT, pages 245–253. Association for Computational Linguistics, 2008.

[ARG08] Mark Hepple Angus Roberts, Robert Gaizasukas and Yikun Guo. Combining terminology resources and statistical methods for entity recognition: an evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA).

[CN02] Hai Leong Chieu and Hwee Tou Ng. Named entity recognition: A maximum entropy approach using global information. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[CS99] Michael Collins and Yoram Singer. Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[EB08] Asif Ekbal and Sivaji Bandyopadhyay. Bengali named entity recognition using support vector machine. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008.

[EB09] Asif Ekbal and Sivaji Bandyopadhyay. A conditional random field approach for named entity recognition in bengali and hindi. *Linguistic Issues in Language Technology*, 2(1):1–44, 2009.

[FM09] Jenny Rose Finkel and Christopher D. Manning. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150. Association for Computational Linguistics, 2009.

[GS96] Ralph Grishman and Beth Sundheim. Message understanding conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

[HGA+98] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. University of sheffield: Description of the lasie-ii system as used for muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*, 1998.

[Kno11] Johannes Knopp. Extending a multilingual lexical resource by bootstrapping named entity classification using wikipedia's category system. In *Proceedings of the Fifth International Workshop On Cross Lingual Information Access*, pages 35–43. Asian Federation of Natural Language Processing, 2011.

[KT07] Jun'ichi Kazama and Kentaro Torisawa. A new perceptron algorithm for sequence labeling with non-local features. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[MH05] Behrang Mohit and Rebecca Hwa. Syntax-based semi-supervised named entity tagging. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 57–60. Association for Computational Linguistics, 2005.

[Sar08] Sunita Sarawagi. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377, 2008.

[SGPV08] Praneeth M Shishtla, Karthik Gali, Prasad Pingali, and Vasudeva Varma. Experiments in telugu ner: A conditional random field approach. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008.

[SM08] P. Srikanth and Kavi Narayana Murthy. Named entity recognition for telugu. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008.

[SZS+04] Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 2004.

[TC02] Koichi Takeuchi and Nigel Collier. Use of support vector machines in extended named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.