8.10 Named Entity Recognition for Telugu

Srikanth. P and Kavi Narayana Murthy, University of Hyderabad, Hyderabad

Abstract—This paper describes our on-going work on Named Entity Recognition (NER) for Telugu. NER involves the identification of named entities such as person names, location names and names of organizations. NER for Indian Languages is a challenging task. There is not much work done in NER for Indian Languages in general and Telugu in particular. Telugu, a language of the Dravidian family, is spoken mainly in southern part of India and ranks second among Indian languages in terms of number of speakers. Telugu is a highly inflectional and agglutinating language providing one of the richest and challenging set of linguistic and statistical features. There are few languages in the world that match Telugu in this regard. NLP Applications such as IR and IE are often hard because of, among other things, the large number of morphological variants for a given root. High performance morphological analyzers have eluded researchers for a long time. In this work we have used news articles in Andhraprabha, a popular Telugu daily. Named entities are generally nouns and it is therefore useful to build a noun identifier. The performance of the Telugu morphological analyzer developed here over the last many years has been evaluated from this angle. Using this in conjunction with a variety of other features we have then built a binary classifier for noun identification using naive Bayes approach. Results are included here. Seed lists of personal suffixes, location suffixes, person name contexts, personal surnames, organization names (such as names of political parties) and place a gazetteer have been developed. A heuristic algorithm for NER has been developed and tested. A training data set of 20,000 words has been developed using the NER system and manual checking. The seed lists have been refined iteratively.

I. INTRODUCTION

NAMED Entity Recognition is a Computational linguistic task in which we seek to classify every word in the corpus as falling into one of four categories: Person, Location, Organization and Not-a-name. In the taxonomy of Computational Linguistics, NER falls under the category of Information Extraction. NER emerged as one of the subtasks of the DARPA-sponsored Message Understanding Conference (MUCs). The task has important significance in the Internet search engines and is an important task in many of the Language Engineering applications such as Machine Translation, Question-Answering systems, Indexing for Information Retrieval and Automatic Summarization.

Srikanth. P and Kavi Narayana Murthy are with the Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, Andhra Pradesh, 500046 (e-mail: patilsrik@yahoo.co.in, knmuh@yahoo.com).

II. NER IN INDIAN LANGUAGES

There has been a considerable amount of work done for NER in English [1, 2] but these ideas cannot be borrowed straight away for NER in Indian languages. For example, the concept of capitalization does not work for Telugu, whereas it is very useful for English. AU-KBC Research center, Chennai [3] has designed a Biological Name Detection system for English which is based on manually developed rules that rely upon lexical information, linguistic constraints of English and the contextual information. There has not been much work done in NER in Indian Languages. [4] describes a pattern based shallow parsing system of Named Entity Recognition for Bengali. Potential contextual patterns are obtained by taking two contextual words after and before the entity that is already tagged and the patterns that cross a threshold are chosen for further tagging through bootstrapping. There does not seem to be any work on NER in Telugu. Telugu is a free word order Language. Each word in Telugu is inflected for hundreds of word forms. Telugu is primarily a suffixing Language - an inflected word starts with a root and may have several suffixes added to the right. Suffixation is not a simple concatenation and morphology of the language is very complex. According to one study (Personal Communication, Uma Maheswar Rao, CALTS, University of Hyderabad) there can be up to 10 lakh forms for a given Telugu verb root. Nouns are also highly inflected. External saMdhi adds to the complexity. This paper is about NER for Telugu.

III. APPROACHES TO NER

Approaches to NER include rule based approaches and Machine Learning techniques. Rule based techniques require substantial linguistic expertise whereas Machine Learning approach requires large amounts of training data. In this work the emphasis is on machine learning techniques.

IV. SENTENCE BOUNDARY DETECTION AND VERB IDENTIFICATION:

Named entities are all generally nouns and recognizing nouns is therefore a useful strategy. We could work by elimination. Telugu is a verb final language and segmenting paragraphs into sentences is therefore useful for recognizing verbs and eliminating them. Full stops and exclamation marks are taken as potential sentence boundary markers and segmentation into sentences effected if the last word has

4 or more characters. On sample evaluation, 506 out of the 520 sentences detected out of 506 actual sentences given were correct, giving us a precision of 97.3%. Of the 506 sentences, 12 sentences do not end with a verb and we conclude that 97.3% of the sentences end with a verb.

V. DEVELOPMENT OF TAGGED DATA

News articles generally start with a headline and the body starts with location name, month and date. A seed list of location names is extracted from this. Seed lists of personal surnames, locational names and organization names have also been developed. Lists of person suffixes such as "reDDi", "na:yuDu" etc, locational suffixes such as "ba:d", "pe:Ta" etc, and name context lists are maintained for tagging the corpus. It has been observed that whenever a context word (such as "maMtri") appears, then in many cases the following two words (consisting of a surname and person name) indicate a person name. This way we build a list of person names. After extensive experimentation over many iterations, a training data set of 20,000 words has been developed.

VI. NOUN IDENTIFICATION

It is useful to recognize nouns and eliminate non-nouns. The Telugu morphological analyzer developed here has been used to obtain the categories. A stop word list including function words has been collected from existing dictionaries and stop words are removed. Words with less than three characters are unlikely to be nouns and so eliminated. Last word of a sentence is usually a verb and is also eliminated. Digits are eliminated. Verbs are recognized based on a list of verb suffixes and eliminated. Telugu words normally end with a vowel and consonant ending words (laMDan, sTe:San, etc.) are usually nouns. Existing dictionaries are also checked for the category. Using these features, a naive Bayes classifier is built using the available tool WEKA. Results are given in the tables 1 and 2.

Table 1: Noun Identification using Morphological Analyzer

	NOUN	NOT-NOUN
Precision (%)	95.97	64.59
Recall (%)	64.65	95.96
F-measure (%)	77.25	76.2

VII. Table 2: Noun Identification using a naive Bayes Classifier

	Test-set 1	Test-set 2
Precision (%)	92.91	78.54
Recall (%)	97.26	91.65
F-measure (%)	95.03	84.59

VII. NAMED ENTITY RECOGNITION AND CLASSIFICATION

Nouns detected above are checked against the heuristics described earlier for detection and classification into various types. The system has been tested on a data set of 5280 words. Results are given table 3.

Table 3: Name identification using heuristics

	NAMES	PERSON	PLACES
PRECISION(%)	82.22	80.7	84.37
RECALL (%)	70.95	66.2	81.88
FMEASURE(%)	76.15	72.73	83.07

REFERENCES

1. Baluja, S., Mittal, V.O., Sukthankar, R.: Applying machine learning for high performance named-entity extraction. Computational Intelligence 16 (2000) 586-596
2. Collins, M., Singer, Y.: Unsupervised models for named entity classication. Michael Collins and Yoram Singer. Unsupervised models for named entity classication. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999. (1999)
3. Meenakshi Narayanaswamy, K. E. Ravikumar, V.K.S.: A biological named entity recognizer. Pacic Symposium on Biocomputing, Hawaii (Jan 2003)
4. Eqbal, A.: Named entity recognition for bengali. Satellite Workshop on Language, Artificial Intelligence and Computer Science for Natural Language Applications (LAICS-NLP), Department of Computer Engineering Faculty of Engineering Kasetsart University, Bangkok, Thailand (2006)

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.