

Named Entity Recognition for Telugu

P. Srikanth and Kavi Narayana Murthy

Department of Computer and Information Sciences

University of Hyderabad

email: patilsrik@yahoo.co.in, knmuh@yahoo.com

OUTLINE

- *Introduction*
 - *Example*
 - *Basic problems*
 - *Applications*
- *Approaches to NER*
- *NER for Indian Languages*
- *CORPUS*
- *Noun Identification*
- *Heuristic NER*
- *NER using Decision Trees and Naive Bayes Techniques*
- *Results*
- *Future work plan*
- *References*

Introduction

- NER involves identification of proper names in texts and classifying them into a set of predefined categories of interest such as
 - Person names
 - Organizations names (companies, political parties etc.)
 - Locations (cities, countries, towns and villages etc.)
 - Date and monetary expressions

Example

- “ **bi.je.pi** adhyakSuDu **adva:ni:** **alha:ba:d** nuMci po:Ti:
ce:stunna:Du.”
 - person : “adva:ni:”
 - location : “alha:ba:d”
 - organization : “bi.je.pi”
 - not-name : adhyakSuDu, nuMci, po:Ti, ce:stunna:Du.

Basic Problems in NER

- Variation of NEs:
 - “vai.es.ra:jaSe:karreDDi”, “vai.es.” etc.
- Ambiguity of NE types
 - “raMga:reDDi” <person vs place>
 - “Ta:Ta:” <person vs organization>
- Ambiguity with common words
 - “baMga:ru” <per-beg vs common word>

Applications

- Information Extraction
- Indexing in IR
- Automatic Summarization
- Machine Translation
- Intelligent Document Access

Approaches

- Rule Based Approaches:
 - rely on language experts
 - not portable
 - do not require training data
- Machine Learning Approaches:
 - require large amounts of training data
 - no annotated corpora for Telugu
 - re-trainable
 - ex: HMMs, CRFs, Maximum Entropy, Decision Trees etc.

Evaluation Metric

Precision = This measures the proportion of the assigned categories that were correct.

Recall = This measures the proportion of the correct categories that were assigned.

F-measure = $(2 * P * R) / (P + R)$

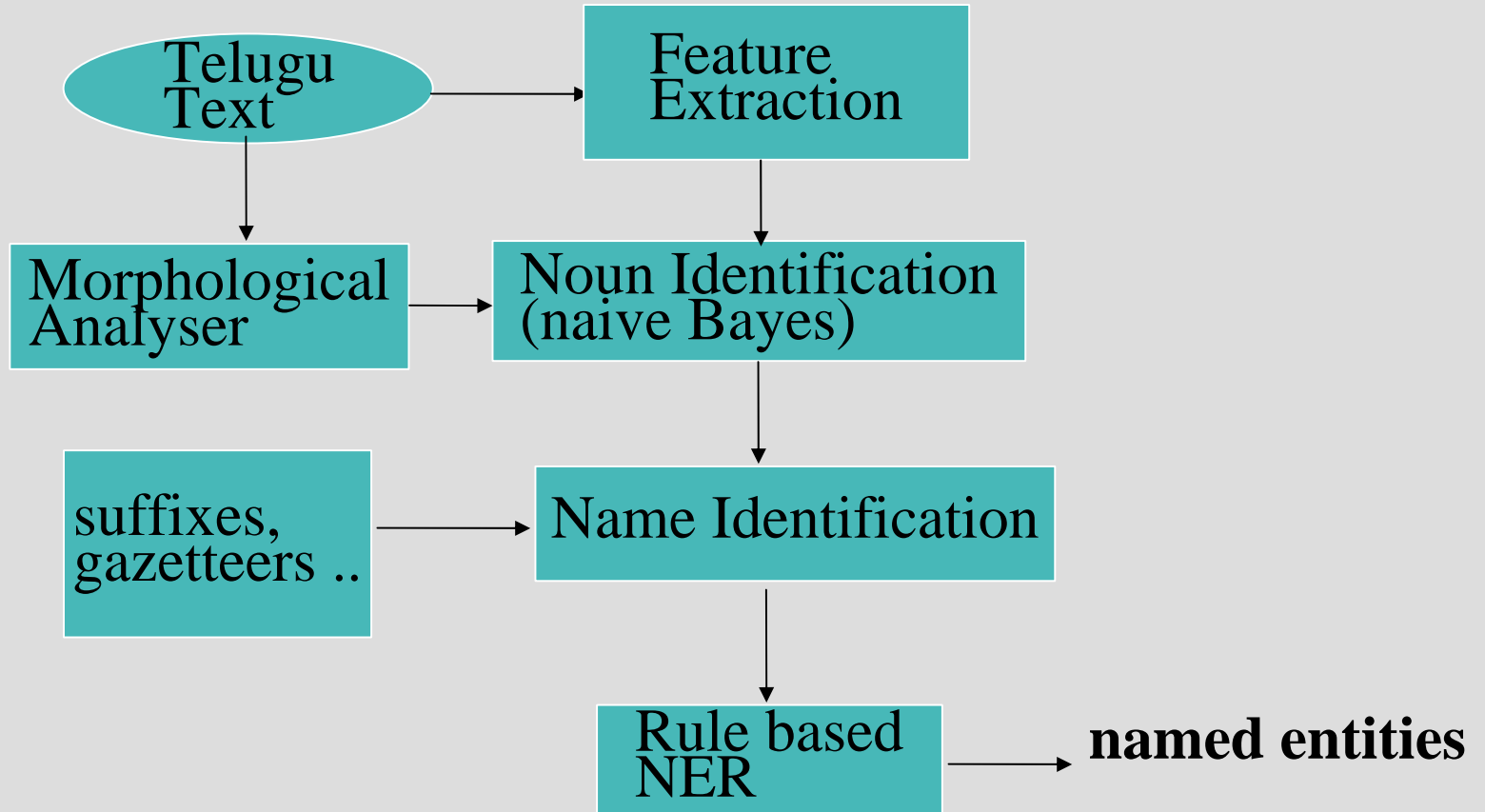
NER for Indian Languages

- Very little work done in Indian Languages
- NER for Bengali: generating the contextual patterns through bootstrapping to extract entities [1]
- NER for Telugu:
 - None exist?
 - Telugu is rich in morphology
 - NER for Telugu is a challenging task
 - Capitalisation can be exploited for English but for ILs

CORPUS

- LERC-UoH Corpus: Nearly 40 Million word Text Corpus of Telugu developed by the Department of CIS, University of Hyderabad
- One part of this Corpus - Telugu news articles from “andhraprabha” - used in this work
- 12,87,156 words, 1,95,286 unique word forms
- Training data: 26,000 words – Annotated for NEs
 - prepared semi-automatically

Structure



Noun Identification

- Named Entities are Nouns
- Morphological analyser developed here has been tested
 - Morph performance not satisfactory
- Performance on 9986 words:

<i>Actual \ Identified</i>	<i>NOUN</i>	<i>NOT-NOUN</i>
<i>NOUN</i>	3861	2111
<i>NOT-NOUN</i>	162	3852

Features for Noun Identification

- Features used:
 - suffixes of non-nouns (for elimination)
 - stop word list from a) C P Brown Telugu-English dictionary b) V Rao Vemuri Telugu-English Dictionary c) UoH Telugu-Hindi MAT dictionary d) high frequency words (> 1000 in corpus) – manually checked and filtered
 - words ending with a consonant – (Telugu words end in a vowel)
 - last word of a sentence is usually a verb – Verb Final Language
 - 97.03% sentences verb final – tested on 506 sentences
 - Grammatical category from dictionaries mentioned above
 - Small words (≤ 3 bytes) are unlikely to be Nouns
 - “idi”, “adi”, “i:” etc.
- each entity is represented as vector of feature values

Noun Identification Using naive Bayes

Accuracy of Noun Identification using the naïve Bayes classifier tested on WEKA tool kit with 10-fold cross validation:

Training set: 4000 instances

Test set-1: 1000 instances

Test set-2: 3000 instances

	<i>TEST-SET 1</i>	<i>TEST-SET 2</i>
<i>PRECISION %</i>	92.91	78.54
<i>RECALL %</i>	97.26	91.65
<i>F-MEASURE %</i>	95.03	84.59

Heuristic NER

- suffixes of named entities maintained
 - place suffixes like “ba:d” in “haidara:ba:d”
 - person suffixes like “reDDi” in “ra:ja:reDDi”
- gazetteers of locations, person surnames, person end names, organisation names maintained triggers of organisation names
 - ka:rpo:re:T, pa:rTi: etc.

Heuristic NER cont.

- Person Context List
 - for ex: “bi.je.pi adhyakSuDu adva:ni:”
- Using this Person Context List names are added to the gazetteers dynamically
- Bootstrapping
- Training data of 26,000 words developed semi-automatically

Heuristic NER: Results

➤ NER Accuracy using Heuristic based system on 5,280 words of test data:

	<i>All Names</i>	<i>PERSON</i>	<i>PLACE</i>
<i>PRECISION %</i>	82.22	80.7	84.37
<i>RECALL %</i>	70.95	66.23	81.88
<i>F-MEASURE %</i>	76.15	72.53	83.07

NER using ML

- Training data of 19,912 words used to train the system
- each entity is represented as vector of feature values and it's class
- features gathered from the training data
 - word level features
 - suffixes, prefixes,digits, special characters
 - Dictionary
 - named entity list
 - context features
 - presence of words like “maMtri” in the context
 - morph category of focused entity as well as of the preceding word and succeeding word
 - gazetteers

NER using ML

- Training data includes
 - 816 person names
 - 437 place names
 - 193 organization names
- Test data of 5,894 words includes
 - 284 person names
 - 145 place names
 - 160 organization names

Performance of naive Bayes and C4.5 classifier:

<i>Actual\Obtained</i>	<i>person</i>	<i>place</i>	<i>organization</i>	<i>not-name</i>
<i>person</i>	131	0	0	153
<i>place</i>	4	121	0	20
<i>organization</i>	1	1	78	80
<i>not-name</i>	11	37	6	5251

<i>Actual/Obtained</i>	<i>person</i>	<i>place</i>	<i>organization</i>	<i>not-name</i>
<i>person</i>	113	0	0	171
<i>place</i>	2	116	0	27
<i>organization</i>	5	3	67	85
<i>not-name</i>	11	38	3	5253

Future Work Plan

- Prepare 50,000 words of training data through bootstrapping
- More of Machine Learning

References

- 1) 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)
- 2) Baluja, S., Mittal, V.O., Sukthankar, R.: Applying machine learning for high performance Named-Entity Extraction. *Computational Intelligence* 16 (2000) 586-596
- 3) Bh. Krishna Murthy and J.P.L. Gwynn : A Grammar of Modern Telugu, Oxford University Press (1985)
- 4) A. Mikheev, M. Moens, and C. Grover. 1999. Named Entity Recognition without gazetteers. In *Proceedings. of EACL, Bergen, Norway. EACL, 1999.*
- 5) Michael Collins and Yoram Singer. Unsupervised models for Named Entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999.*

Visit

www.LanguageTechnologies.ac.in

dhanyava:damulu
(Thank You)