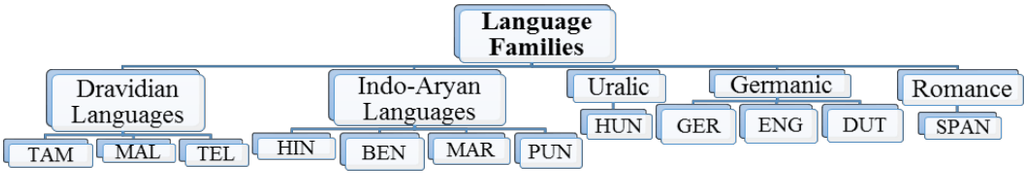
A Deeper Study on Features for Named Entity Recognition

(Malarkodi C.S., Sobha Lalitha Devi)
AU-KBC Research Centre

Objectives:

- To develop the language & domain independent Named Entity Recognition (NER) system which can identify named entities from any given dataset irrespective of the language and domain
- To analyze the various linguistic patterns surrounding the named entities across the languages which belongs to different language families

Fig. 1. Languages Used in this work



Linguistic Features Used in this work

Dravidian languages
Grammatical patterns

- RP verbs precede and follow
- Common noun precedes or follows
- Occurring after the verb
- Postpositions precede the NEs
- Verbs succeed the NEs
- Postpositions, adjectives or adverbs follow the NEs

Indo Aryan Languages
Grammatical patterns

- The common nouns, pronouns or conjunctions precedes the NEs
- Verbs precede in Bengali and Marathi
- The postpositions precede the NEs in Hindi and Punjabi
- NEs following by postposition, verbs, conjunctions or adjectives Occurring at the beginning of the sentence.

European Languages
Grammatical patterns

- Follows by verbs, common nouns or punctuations
- Prepositions, determiners or punctuations precedes
- Verbs or adjectives precede the NEs in Hungarian, Dutch and German.
- Occurring at the beginning of the sentence

Dravidian, Indo-Aryan and European Languages
Typological Patterns

- NEs at the beginning of the sentence
- NEs at the end of the sentence
- Punctuations followed NEs
- NEs Occurring after punctuations

- Machine Learning Technique Used – CRFs

Corpus Details:

- FIRE corpus – Indian Languages and English
- CONLL corpus – Dutch and Spanish

Results

Tamil, Malayalam and Telugu

- 81.58% , 72.66%, 62.74% F-M respectively

Hindi, Bengali, Punjabi, Marathi

- 82.07%, 85.92%, 81.96%, 82.57% F-M respectively

English, Spanish, Dutch, Hungarian, German

- 82.28%, 85.24%, 91.25%, 84.51% and 76.97% respectively

Conclusion

The linguistic features used in this work helps the system to learn the structure of named entities
The results shown that the linguistic features obtained state-of-art results for both Indian and European languages.