Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, pages 66–72
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

A Deeper Study on Features for Named Entity Recognition

Malarkodi C. S. and Sobha Lalitha Devi
AU-KBC Research Centre
MIT Campus of Anna University
Chromepet, Chennai, India
sobha@au-kbc.org

Abstract

This paper deals with the various features used for the identification of named entities. The performance of the machine learning system heavily depends on the feature selection criteria. The intention to trace the essential features required for the development of named entity system across languages motivated us to conduct this study. The linguistic analysis was done to find out the part of speech patterns surrounding the context of named entities and from the observation linguistic oriented features are identified for both Indian and European languages. The Indian languages belongs to Dravidian language family such as Tamil, Telugu, Malayalam, Indo-Aryan language family such as Hindi, Punjabi, Bengali and Marathi, European languages such as English, Spanish, Dutch, German and Hungarian are used in this work. The machine learning technique CRFs was used for the system development. The experiments were conducted using the linguistic features and the results obtained for each languages are comparable with state-of-art systems.

Keywords: features, ner, CRFs, POS patterns, named entities

1. Related Work

Named Entity Recognition (NER) is defined as the process of automatic identification of proper nouns and classifies the identified entities into predefined categories such as person, location, organization, facilities, products, temporal or numeric expressions etc. Even though named entity recognition is a well-established research filed and lot of research works are available for various languages, to the best of our language no work was found on the deeper analysis of features required for named entity system across languages.

Initially the term NER was defined in Message Understanding Conference (MUC), when the structured information about company and defense related activities needed to be extracted from the unstructured text. It was noticed that the main information units to be extracted are named entities (Grishman et al. 1996). The very first research work in NER was done by Lisa F. Rau, who developed the system to recognize company names using hand-crafted rules. In MUC-7, five out of eight systems were generated using rule based method (Chinchor 1998). Nadeau et al. (2007) has reported fifteen years of research carried out in the field of entity recognition.

Gutiérrez et al. (2015) developed a Spanish NE system using CRF. The dataset was obtained from CONLL 2002 shared task. Ekbal et al. (2008) worked on a Bengali named entity recognition using CRF. Ekbal et al. (2009) contributed NER systems for Hindi & Bengali using CRF framework. Kaur et al. (2012) built an NE system for Punjabi language. Bindu & Sumam Mary (2012) used CRF based approach for identifying named entities in Malayalam text.

Khanam et.al. (2016) has worked on the Named Entity Identification for Telugu Language using hybrid approach. Sobha et al. (2007) developed a multilingual named entity system to identify the place names using Finite State Automaton (FSA). Vijayakrishna & Sobha (2008) focused on the Tamil NER for tourism domain which consists of nested tagging of named entities. Malarkodi & Sobha (2012a) built a NE system for Indian languages like Tamil, Telugu, Hindi, Marathi, Punjabi and Bengali using CRF. Malarkodi et al. (2012b) discussed the various challenges, while developing the NE system in Tamil language. Sobha et al. (2013) has participated in ICON NLP tool contest and submitted the test runs for 5 Indian languages and English.

Patil et al. (2016) reported a work on NER for Marathi using HMM. Jaspreet et al. (2015) contributed Punjabi NER using 2 machine learning approaches namely HMM and MEMM. Antony et.al. (2014) constructed the NE system for Tamil Biomedical documents using SVM classifier. Lakshmi et.al. (2016) has worked on the Malayalam NER using Fuzzy-SVM and it is based on the semantic features and linguistic grammar rules. Jiljo et.al. (2016) used TnT and Maximum Entropy Markov model for NE identification in Malayalam data. The proposed methodology yields 82.5% accuracy.

Bojórquez et al. (2015) worked on improving the Spanish NER used in the Text Dialog System (TDS) by using semi-supervised technique. Zea et.al. (2016) developed a semi-supervised NE system for Spanish language. Athavale et al. (2016) described a Neural Network model for NER based on the Bi Directional RNN-LSTM. In order to identify the mentions of medications, Adverse Drug Event (ADE) and symptoms of the diseases in the clinical notes (Florez et al. 2018) proposed the character-level word representation methods which can be used as an input feature to neural network model called LSTM.

The various shared tasks conducted for Named Entity Recognition are discussed in this section. In 2002, CONLL shared task about NER was focused on Spanish and Dutch (Tjong et al. 2002). The CONLL 2003 offered dataset for English and German (Tjong et al. 2003). The NERSSEAL shared task of IJCNLP-2008 was organized for 5 Indian Languages namely Hindi, Bengali, Oriya, Telugu and Urdu (Singh, 2008). In 2013 AU-KBC has organized NER shared task as part of Forum for Information Retrieval for Evaluation (FIRE), to create a benchmark data for Indian Languages. The dataset was released for 4 Indian Languages like Bengali, Hindi, Malayalam, and Tamil and also for English. The various techniques used by the participants are CRF, rule based approach and list based search (Pattabhi & Sobha 2013). The 2nd edition of NER track for IL has organized as part of FIRE 2014 for English and 3 IL namely Hindi, Malayalam, and Tamil. The main focus of this track is nested entity identification. The

participants have used CRF and SVM for system development (Pattabhi et al. 2014).

2. Language Families Used

The Indian languages belong to different language families; most of the Indian languages come under Indo-Aryan and Dravidian language families. Indo Aryan language family is a sub-branch of Indo-Iranian family which itself is a sub-family of Indo-European language family. Mainly the Indian languages like Hindi, Bengali, Marathi, Punjabi comes under Indo-Aryan family and Indian languages like Tamil, Telugu, Malayalam, and Kannada belongs to a Dravidian language family. The European languages also have several language families. The languages like German, Dutch and English belong to Germanic families, Spanish language constitutes a Romance language family and Hungarian comes under Uralic language family.

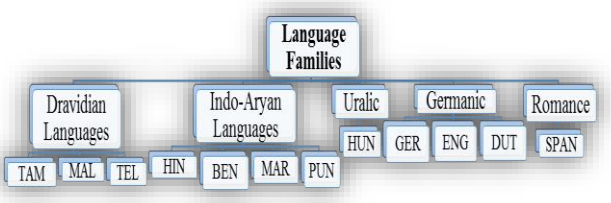

Figure 1 : Language Families Used in this work

3. Corpus Statistics

In this section, we discuss the corpus we have used for the study. The corpus for Tamil, Malayalam, and English was collected using an automated crawler program. The dataset developed as part of FIRE 2013 NER shared task and national level projects such as Cross Lingual Information Access (CLIA) are used for English, Tamil, Hindi, Malayalam, Bengali, Marathi and Punjabi. The corpus statistics of the Dravidian Languages are given in Table 1. The Tamil dataset consists of 13k sentences, 200K tokens and 27k named entities. The Malayalam corpus consists of 64,345 tokens, 5k sentences and 11k named entities. The Telugu corpus has 2K sentences, 43,062 tokens and 9,104 named entities.

Languages	Tokens	Sentences	NEs
Tamil	2,04,144	13,571	27,498
Telugu	43,062	2,150	9,104
Malayalam	64,345	5,107	11,380
Bengali	52,024	4,030	2,690
Hindi	1,90,236	14,098	21,498
Marathi	73,523	6,138	6,036

Table 1: Corpus statistics (Indian Languages)

The NE corpus used for Spanish and Dutch languages is obtained from CONLL 2003 NER shared task. The Spanish and Dutch corpus contains person, location, organization and miscellaneous NE tags. For German language, the GERMEVAL NER shared task data has been utilized. The German NE corpus has 12 NE tags and mainly has four classes. The number of tokens and named entities in the English dataset are 200K and 25K respectively. The Spanish and Dutch corpus consists of 300K and 200K tokens. The numbers of named entities in Spanish and Dutch dataset are 23,148 and 27,390. The German dataset consists of 500K tokens, 31K sentences and 33,399 NEs. The Hungarian corpus has 400K tokens and 7,068 named entities.

Languages	Tokens	Sentences	NEs
English	2,56,426	14,002	25,671
Dutch	2,47,820	15,316	27,390
Hungarian	4,44,661	27,673	7,068
German	5,91,005	31,298	33,399
Spanish	3,17,637	10,238	23,148

Table 2: Corpus statistics (European Languages)

The details of the POS tagset are explained in this section. The BIS POS tagset was used for Indian Languages. The Tamil POS tagger developed by Sobha et al. (2016) works with an accuracy of 95.16% (Sobha et al., 2016). The Brills POS tagger (Brill et al., 1992) is used for this task. The dataset used for German are preprocessed with Stanford POS tagger (Manning et al., 2014). The Spanish and Dutch dataset are obtained from the CONLL shared task are already tagged with POS information.

4. Features used for Named Entity Recognition

The part of speech patterns frequently occurred in the context of named entities are analyzed for each language and the results are discussed in this section. We analyze the corpus to arrive at the most suitable word level features for identifying the NE which can be used for machine learning purposes. We have taken a window of three words and identified the most frequent grammatical and typographical feature that occurs. The distribution of each feature in each language is given in detail below.

4.1 Analysis of common Linguistic features

In Tamil corpus, the named entities occurred at the beginning of the sentence in 3,776 instances and in 2,056 instances named entities occurred at the end of the sentence, punctuations preceded the NE in 6,222 times and 4,596 times punctuations succeed the NE, common nouns preceded the NE in 6,274 times and succeeded the NE in 9,038 times, proper nouns occurred before NE in 2,250 instances and after NE in 2,868 instances. The postpositions occurred before NE in 999 instances, adjectives occurred before NE in 1418 instances and conjunction occurred before NE in 384 times. The verbal participle preceding the named entities in 716 instances and the relative participle verbs preceded the NE in 1,007 times. The finite verbs succeed the NE in 998 instances, postpositions, adverb, and adjectives occurred at 1131, 980 and 878 instances respectively.

The Malayalam corpus has the following distribution. The named entities occurred at the beginning of the sentence in 1,062 instances and in 72 instances named entities occurred at the end of the sentence, punctuations preceded the NE in 818 times and 751 times punctuations succeed the NE, common nouns preceded the NE in 1,281 times and succeeded the NE in 1,794 times, proper nouns occurred before NE in 774 instances and after NE in 939 instances. The postpositions occurred before NE in 209 instances, adjectives occurred before NE in 201 instances and conjunction occurred before NE in 85 times. The verbal participle preceding the named entities in 82 instances and the relative participle verbs preceded the NE in 238 times. The finite verbs succeed the NE in 628 instances, postpositions, adverb, and adjectives occurred at 192, 173 and 273 instances respectively.

In Telugu corpus, the named entities occurred at the beginning of the sentence in 776 instances and in 17 instances named entities occurred at the end of the sentence, punctuations preceded the NE in 588 times and 610 times punctuations succeed the NE, common nouns preceded the NE in 3722 times and succeeded the NE in 4641 times, proper nouns occurred before NE in 540 instances and after NE in 615 instances. The postpositions occurred before NE in 450 instances, adjectives occurred before NE in 315 instances and conjunction occurred before NE in 153 times. The verbs preceding the named entities in 1541 instances and the relative participle verbs preceded the NE in 78 times. The verbs succeed the NE in 1307 instances, postpositions, adverb and adjectives occurred at 665, 263 and 256 instances respectively.

The Hindi corpus has the following distribution. The named entities occurred at the beginning of the sentence in 5,290 instances and in 1,201 instances named entities occurred at the end of the sentence, punctuations preceded the NE in 2281 times and 2070 times punctuations succeed the NE, common nouns preceded the NE in 1862 times and succeeded the NE in 1307 times, proper nouns occurred before NE in 1055 instances and after NE in 753 instances. The postpositions occurred before NE in 3536 instances, adjectives occurred before NE in 611 instances and conjunction occurred before NE in 1844 times. The verbs preceding the named entities in 349 instances and the relative participle verbs preceded the NE in 412 times. The verbs succeed the NE in 876 instances, postpositions, adverb and adjectives occurred at 915, 536 and 436 instances respectively.

In Punjabi corpus, the named entities occurred at the beginning of the sentence in 1267 instances and in 831 instances named entities occurred at the end of the sentence, punctuations preceded the NE in 499 times and 475 times punctuations succeed the NE, common nouns preceded the NE in 1,119 times and succeeded the NE in 684 times, proper nouns occurred before NE in 304 instances and after NE in 304 instances. The postpositions occurred before NE in 1363 instances, adjectives occurred before NE in 553 instances and conjunction occurred before NE in 227 times. The verbs preceding the named entities in 99 instances and the relative participle verbs preceded the NE in 176 times. The verbs succeed the NE in 361 instances, postpositions, adverb and adjectives occurred at 3,211, 136 and 158 instances respectively.

In Bengali corpus, the NE distribution is as discussed here. The named entities occurred at the beginning of the sentence in 630 instances and in 312 instances named entities occurred at the end of the sentence, punctuations preceded the NE in 288 times and 204 times punctuations succeed the NE, common nouns preceded the NE in 561 times and succeeded the NE in 908 times, proper nouns occurred before NE in 197 instances and after NE in 199 instances. The postpositions occurred before NE in 120 instances, adjectives occurred before NE in 148 instances and conjunction occurred before NE in 239 times. The verbs preceding the named entities in 208 instances and the relative participle verbs preceded the NE in 25 times. The verbs succeed the NE in 290 instances, postpositions, adverb and adjectives occurred at 159, 280 and 238 instances respectively.

The Marathi corpus has the following distribution. The named entities occurred at the beginning of the sentence is 967 instances and in 488 instances named entities occurred at the end of the sentence, punctuations preceded the NE in 609 times and 566 times punctuations succeed the NE, common nouns preceded the NE in 1956 times and succeeded the NE in 1879 times, proper nouns occurred before NE in 348 instances and after NE in 219 instances. The postpositions occurred before NE in 14 instances, adjectives occurred before NE in 114 instances and conjunction occurred before NE in 466 times. The verbs preceding the named entities in 475 instances and the relative participle verbs preceded the NE in 38 times. The verbs succeed the NE in 419 instances, postpositions, adverb and adjectives occurred at 212, 253 and 392 instances respectively.

In English corpus, the NE distribution is as discussed here. The named entities occurred at the beginning of the sentence in 1,014 instances and in 2,078 instances named entities occurred at the end of the sentence, punctuations preceded the NE in 1549 times and 2078 times punctuations succeed the NE, common nouns preceded the NE in 1239 times and succeeded the NE in 1289 times, proper nouns occurred before NE in 745 instances and after NE in 823 instances. The prepositions occurred before NE in 2794 instances. The determiners preceding the named entities in 1425 instances. The verbal participle preceding the named entities in 156 instances. The finite verbs succeed the NE in 680 instances, prepositions and conjunctions occurred at 1195 and 774 instances respectively.

The Spanish corpus has the following distribution. The named entities occurred at the beginning of the sentence in 2046 instances and in 2,131 instances named entities occurred at the end of the sentence, punctuations preceded the NE in 2404 times and 7010 times punctuations succeed the NE, nouns preceded the NE in 1123 times and succeeded the NE in 204 times. The prepositions occurred before NE in 7231 instances. The determiners preceding the named entities in 3993 instances. The verbs succeed the NE in 2060 instances, prepositions and conjunctions occurred at 2116 and 1648 instances respectively.

In Dutch corpus, the NE distribution is as discussed here. The named entities occurred at the beginning of the sentence in 5605 instances and in 2787 instances named entities occurred at the end of the sentence, punctuations

preceded the NE in 4142 times and 10321 times punctuations succeed the NE, nouns preceded the NE in 2627 times and succeeded the NE in 3174 times. The prepositions occurred before NE in 5146 instances. The determiners preceding the named entities in 4142 instances. The verbs succeed the NE in 4657 instances, prepositions and conjunctions occurred at 2062 and 1411 instances respectively.

The German corpus has the following distribution. The named entities occurred at the beginning of the sentence in 2033 instances and in 128 instances named entities occurred at the end of the sentence, punctuations preceded the NE in 966 times and 2535 times punctuations succeed the NE, common nouns preceded the NE in 5012 times and succeeded the NE in 3886 times, proper nouns occurred before NE in 321 instances and after NE in 608 instances. The prepositions occurred before NE in 5869 instances. The determiners preceding the named entities in 7166 instances. The finite verbs succeed the NE in 3140 instances, prepositions and conjunctions occurred at 3075 and 2059 instances respectively.

In Hungarian corpus, the NE distribution is as discussed here. The named entities occurred at the beginning of the sentence is 2175 instances and in 26 instances named entities occurred at the end of the sentence, punctuations preceded the NE in 878 times and 1861 times punctuations succeed the NE, nouns preceded the NE in 845 times and succeeded the NE in 2053 times. The postpositions occurred before NE in 148 instances. The determiners preceding the named entities in 1788 instances. The finite verbs succeed the NE in 751 instances, prepositions and conjunctions occurred at 220 and 439 instances respectively.

We have analysed the corpus for the various part of speech which is associated with the named entities. In the window of three, the following are the grammatical features that occurred. Also, the Typographical features also arrive through the analysis. From the above analysis, we arrived at the following points

- In Dravidian languages Tamil, Telugu and Malayalam the most commonly occurring pattern for NE are
- 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

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

- In Indo Aryan Languages Hindi, Bengali, Marathi, and Punjabi the most commonly occurring pattern for NE are
    - 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.
- Typological patterns are
    - NEs at the beginning of the sentence
    - NEs at the end of the sentence
    - Punctuations followed NEs
    - NEs Occurring after punctuations
- In European Languages English, Hungarian, Spanish, Dutch, and German the most commonly occurring pattern for NE are
    - 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
- Typological patterns are
    - NEs at the beginning of the sentence
    - NEs at the end of the sentence
    - Punctuations followed NEs
    - NEs Occurring after punctuations

5. Experiments & Results

In this section, the results obtained by each feature combinations are discussed in detail. The experiments are conducted for each language is given in the table. The machine learning technique CRFs was used for the system development.

Languages	PRE	REC	F-M
Tamil	80.12	83.1	81.58
Malayalam	70.63	74.82	72.66
Telugu	69.4	57.25	62.74

Table 3: Results for Dravidian Languages

Languages	PRE	REC	F-M
Hindi	81.05	83.13	82.07
Bengali	82.78	89.31	85.92
Punjabi	80.54	83.45	81.96
Marathi	78.32	87.32	82.57

Table 4: Results for Indo-Aryan Languages

Languages	PRE	REC	F-M
English	84.32	80.35	82.28
Spanish	86.13	84.37	85.24
Dutch	90.3	92.23	91.25
Hungarian	83.84	85.21	84.51
German	81.41	72.99	76.97

Table 5: Results for European Languages

The linguistic feature yielded the precision and recall of 80.12% and 83.1% for Tamil, 70.63% precision and 74.82% recall for Malayalam and 69.40% precision score and 57.25% recall value for Telugu. The f-score obtained by Dravidian languages are 81% for Tamil, 72% for Malayalam and 62.74% in Telugu.

The results obtained Indo-Aryan languages using linguistic feature are discussed in this section. The precision and recall achieved for Hindi is 80.12% and 83.1% respectively. Bengali has obtained the f-score of 85%. Punjabi scored the precision of 80.54% and recall of 83.45%. Marathi has achieved the f- measure of 82.57%.

The results obtained European languages using linguistic feature are discussed in this section. The precision and recall achieved for English is 84% and 80% respectively. Spanish has obtained the f-score of 85%. Dutch scored the precision of 90.3% and recall of 92.23%. Hungarian has achieved the f- measure of 84.57%. German has obtained the precision of 81%, recall of 72% and f-measure of 76% respectively.

The different feature combinations shown in Table 3-5 clearly show that all the linguistic features used in the present system have the capability to improve the system's performance. The results show that the feature combinations presented in this work yields reasonable results not only for Indian Languages but also for European languages. By using linguistic features alone, we have achieved reasonable scores for languages belong to different language families.

Existing Systems	**Methods**	**Languages used**	**F-M**
Gayen et al. (2014)	HMM	Bengali English Hindi Marathi Punjabi Tamil Telugu	85.99 77.04 75.20 42.89 54.55 44.00 40.03
Abinaya et al. (2014)	CRF for English SVM for other languages	English Hindi Tamil Malayalam	57.81 25.53 30.75 24.91
Ekbal et al. (2009)	CRF	Bengali (LI) Hindi (LI)	77.74 77.08
Florian et al. (2003)	Stacking based approach	Spanish Dutch	79.05 74.99
Our system	CRFs	Bengali Hindi Marathi Punjabi Tamil Telugu Malayalam English Spanish Dutch German Hungarian	85.92 82.07 82.57 81.96 81.58 62.74 72.66 82.28 85.24 91.25 76.97 84.51

Table 6: Comparison with existing works

Though the present work is about multilingual named entities, we have compared our work with the existing multilingual NER works. Gayen et al. (2014) has participated in ICON NER shared task and built a named entity system for English and 6 Indian languages using HMM. In comparison with the performance reported by Gayen et al. (2014), except Bengali we have achieved the highest f-score for all the Indian languages. Abinaya et al. (2014) has participated in FIRE 2014 shared task and developed a NE system for English and 3 Indian Languages. As reported in FIRE 2014 NER task overview paper (Pattabhi et al., 2014), the results given in table 6 are obtained by Abinaya et al. for maximal entities. They have implemented CRFs for English and SVM for other languages. The present system achieved the better scores than Abinaya et al. The language Independent (LI) NE system has developed for Hindi and Bengali using CRFs by Ekbal et al. (2009). The results attained by the present work in Bengali and Hindi languages are higher than Ekbal et al. (2009). But the NE system developed using language specific features by Ekbal et al. (2009) are performing better than the present system. Florian et al. (2003) participated in CONLL 2002 NER shared task and obtained 79.05% for Spanish and 74.99% for Dutch. The present system obtained 85% and 91% f-measure for Spanish and Dutch respectively.

6. Conclusion

The different kinds of features used for the named entity recognition are discussed in this work. The linguistic analysis of POS patterns precedes and following the named entities are analyzed for each language and from the observation linguistic features for the POS patterns are identified in the proximity of NE. This helps the system to learn the structure of named entities by providing the linguistic information. The experiments are conducted for both Indian and European languages. The results shown that the linguistic features obtained state-of-art results for both Indian and European languages.

7. Bibliographical References

Abinaya, N. Neethu, J. Barathi, H.B.G. Anand, M.K. & Soman, K.P. (2014). AMRITA CEN@ FIRE-2014: Named Entity Recognition for Indian Languages using Rich Features. In Proceedings of the Forum for Information Retrieval Evaluation, pp. 103-111

Antony, J.B. & Mahalakshmi, G.S. (2014). Named entity recognition for Tamil biomedical documents. In Proceedings of the 2014 International Conference on Circuits, Power and Computing Technologies, ICCPCT-2014, pp. 1571-1577.

Athavale, V. Bharadwaj, S., Pamecha, M., Prabhu, A. & Shrivastava, M. (2016). Towards deep learning in hindi ner: An approach to tackle the labelled data scarcity. arXiv preprint arXiv:1610.09756.

Brill, Eric. (1992). A simple rule-based part of speech tagger. In Proceedings of the third conference on Applied natural language processing. Association for Computational Linguistics.

Bojórquez, Salvador, S. & Vıctor, M.G. (2015). Semi-Supervised Approach to Named Entity Recognition in Spanish Applied to a Real-World Conversational System. Pattern Recognition: 7th Mexican Conference, MCPR 2015, Mexico City, Mexico, vol. 9116, pp. 24-27.

Chinchor, N. (1998). Overview of MUC 7. In Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, Virginia.

Ekbal, A. & Bandyopadhyay, S. (2008). 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.

Ekbal, A. & Bandyopadhyay, S. (2009). A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi. Linguistic Issues in Language Technology, Vol. 2, no. 1, 1-44.

Florez, E., Precioso, F., Riveill, M. & Pighetti, R. (2018). Named entity recognition using neural networks for clinical notes. International Workshop on Medication and Adverse Drug Event Detection, pp. 7-15.

Florian, R. Ittycheriah, A. Jing, H. & Zhang, T. (2003). Named entity recognition through classifier combination. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. Association for Computational Linguistics, vol. 4, pp. 168-171

Gayen, V. & Sarkar, K. (2014). An HMM based named entity recognition system for Indian languages: the JU system at ICON 2013. ICON NLP Tool Contest arXiv preprint arXiv:1405.7397

Grishman, R & Beth, S. (1996). Message understanding conference-6: A brief history. In Proceedings of the16th International Conference on Computational Linguistics, Vol. 1

Gutiérrez, R., Castillo, A., Bucheli, V., & Solarte, O. (2015). Named Entity Recognition for Spanish language and applications in technology forecasting Reconocimiento de entidades nombradas para el idioma Español y su aplicación en la vigilancia tecnológica. Rev. Antioqueña las Ciencias Comput. y la Ing Softw, 5, 43-47.

Jaspreet, S. & Gurpreet, S.L. (2015). Named entity recognition for Punjabi language using Hmm and Memm. In Proceedings of the IRF International Conference, Pune, India, pp. 4-7.

Jiljo, Pranav, P.V. (2016). A study on named entity recognition for malayalam language using tnt tagger & maximum entropy markov model. International Journal of Applied Engineering Research, 11(8), pp. 5425-5429.

Kaur, A. & Josan, G.S. (2015). Evaluation of Named Entity Features for Punjabi Language. Procedia Computer Science 46, 159-166.

Bindu, M.S. & Idicula, S.M. (2011). Named entity identifier for malayalam using linguistic principles employing statistical methods. International Journal of Computer Science Issues (IJCSI), 8(5), 185.

Khanam, M.H. Khudhus, M.A. & Babu, M.P. (2016). Named entity recognition using machine learning techniques for Telugu language. In Proceedings of the 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 940-944.

Lakshmi, G. Janu, R.P. & Meera, M. (2016). Named entity recognition in Malayalam using fuzzy support vector machine. In Proceedings of the 2016 International Conference on Information Science (ICIS), IEEE, pp. 201-206.

Malarkodi, C.S., Pattabhi, R.K. & Sobha L. (2012). Tamil NER–Coping with Real Time Challenges. In Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages(MTPIL-2012), COLING, pp. 23-38.

Malarkodi, C.S & Sobha L (2012). A Deeper Look into Features for NE Resolution in Indian Languages. In Proceedings of the Workshop on Indian Language Data: Resources and Evaluation, LREC, Istanbul, pp. 36-41.

Manning, C, Surdeanu, M, Bauer, J, Finkel, J, Bethard, S & McClosky, D 2014, 'The Stanford CoreNLP natural language processing toolkit', Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60

Nadeau, D. & Sekine, S. (2007). A survey of named entity recognition and classification. Linguisticae Investigationes, Vol. 30, no. 1, pp. 3–26.

Patil, N., Patil, A.S. & Pawar, B.V. (2016). Issues and Challenges in Marathi Named Entity Recognition. International Journal on Natural Language Computing (IJNLC), Vol. 5, pp. 15-30.

Pattabhi, R.K. & Sobha, L. (2013). NERIL: Named Entity Recognition for Indian Languages @ FIRE 2013–An Overview. FIRE-2013.

Pattabhi, R.K., Malarkodi C.S., Ram V.S. & Sobha L. (2014). NERIL: Named Entity Recognition for Indian Languages @ FIRE 2014–An Overview. FIRE-2014.

Singh, A.K. (2008). Named Entity Recognition for South and South East Asian Languages: Taking Stock. In Proceedings of the IJCNLP-08, pp. 5-16.

Sobha L, Malarkodi, C.S, & Marimuthu, K. (2013). Named Entity Recognizer for Indian Languages. ICON NLP Tool Contest.

Sobha, L. Pattabhi RK Rao, & Vijay Sundar Ram, R (2016). AUKBC Tamil Part-of-Speech Tagger (AUKBC-TamilPoSTagger 2016v1). Web Download. Computational Linguistics Research Group, AU-KBC Research Centre, Chennai, India.

Tjong Kim Sang, E.F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In Proceedings of the CONLL-2002, Taipei, Taiwan.

Tjong Kim Sang, E.F & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Vol. 4, pp. 142-147.

Vijayakrishna, R & Sobha, L. (2008). Domain focused Named Entity for Tamil using Conditional Random Fields. In Proceedings of the workshop on NER for South and South East Asian Languages, Hyderabad, India, pp. 59-66.

Zea, J.L.C., Luna, J.E.O., Thorne, C. & Glavaš, G. (2016). Spanish ner with word representations and conditional random fields'. In Proceedings of the sixth named entity workshop, pp. 34-40.