IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 4, Ver. IV (July – Aug. 2015), PP 10-15
www.iosrjournals.org

Name Entity Detection and Relation Extraction from Unstructured Data by N-gram Features on Hidden Markov Model and Kernel Approach

Naincy Priya[1], Amanpreet Kaur [2]
[1](Computer Science and Engineering, Chandigarh University, India)
[2](Computer Science and Engineering, Chandigarh University, India)

Abstract: *In recent years Name entity extraction and linking have received much attention. However, correct classification of entities and proper linking among these entities is a major challenge for researcher. We propose an approach for entities and their relation extraction with feature including lexicon, n-gram and parts of speech clustering and then apply hidden markov model for entity extraction and CRF with kernel approach to detect relationship among these entities. Analysis of our model is done by precision, recall and accuracy. We have used kernel approach with Conditional random field for extracting the relation between the entities and then remove the co-reference by kernel function. The accuracy of the proposed system for entity detection is 98.03, precision is 88.80and recall is 87.50 where as accuracy of relation extraction is 87.46,precision 84.46 and recall is 82.46 which is much better than the rest existing models.*

Keywords: *n-gram; lexicon; hidden markov model; Conditional random field.*

I. Introduction

The Name entity recognition is the branch of natural language processing which comes under the category of artificial intelligence. Through NER entities like person, place, organization etc are extracted and classified from any text. While recent approaches to named entity recognition (NER) have become quite efficient and effective, there are still various issues related to proper classification of entities (e.g. entities having same name). Entities that we have taken includes person, location, organization and miscellaneous.

A. Applications of Name Entity Recognition

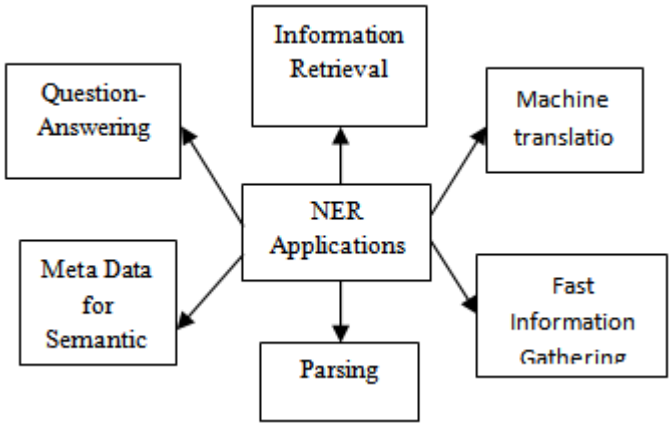

Figure 1: Applications of Name entity recognition

The way of narrating any news may vary from person to person but the main entities discussed in news remains the same. Due to which the task of finding rigid designators from news belonging to named entity type such as persons, location, and organization and miscellaneous is very important. It might be possible that two different entities have same name and thus it generates a problem of proper identification of entity. For example, name of a person say "Ram" may also be name of any organization. Similarly name of any location such as "Indira" may be the name of a person. N-gram feature is a connected sequence of n-items from a given sequence of text or speech. This feature helps us in predicting the next coming word/s of any text. N-gram with size 1 is called "unigram", with size 2 "bigram", with size 3 "trigram" and so on. When n-gram and PoS along-with lexicon feature is used with HMM model then the incorrect entity recognisation can be reduced. is very important. However the name of different entity may be same and thus it generates a problem of proper identification of entity correctly. When n-gram and PoS along-with lexicon feature is used with HMM model

then the incorrect entity recognisation can be reduced. Also when these features are used with CRF kernel approach, it helps in increasing the linking among entities.

This paper describes a system that extract n-gram, PoS and lexicon features, entities (person, location and organization) and also relationship among extracted entities (if any).

II. Related Research

There is several proposed system that extract name entities from different language. The Daljit Kaur and Ashish verma proposed a new framework using machine learning algorithm that extracts name entities on Arabic language [1]. Sudha Morwal and Nusrat Jahan applied NER on Marathi, Hindi and Urdu language [7]. Kamaldeep Kaur and Vishal Gupta used NER on Punjabi language [11]. A part from regional Hindi language, NER is applied on several International languages like English, Italian, Spanish, German, Russian, Arabic etc. NER is applied on crime reports to extract Nationality from crimes [5], extraction of crime information from police and witness reports [16].he Entity detection and relationship extraction aims at finding entities like person, place, organization etc from text or speech and finding out relationship among detected entities (if any). There are lots of works that has been done for both entity recognition, classification and tag those entities [1]. There are several online systems that help in detecting entities. These systems play an important role in gathering crime information now- a-days. It collects information like nationality of criminals, more focused questions to witnesses so that more and more information can be collected and will be useful in solving criminal cases[2][6].

NER covers variety of languages apart from English that makes its larger proportion of language independency. There have been several conferences on NER that includes languages like German, Spanish, Dutch, Chinese, and Japanese etc. NER also includes Hindi, Marathi and Urdu languages. However there are several issues with Urdu language [3] [4][7].

NER is also performed on tweets now-a-days. However tweets are small and contains several short forms, due to which proper recognisation of entities is a problem [5]. Also different type of entity may have common name which is also a problem. Thus, incorrect entity type may be recognized.

III. System Architecture

The system architecture of the proposed system is as under:

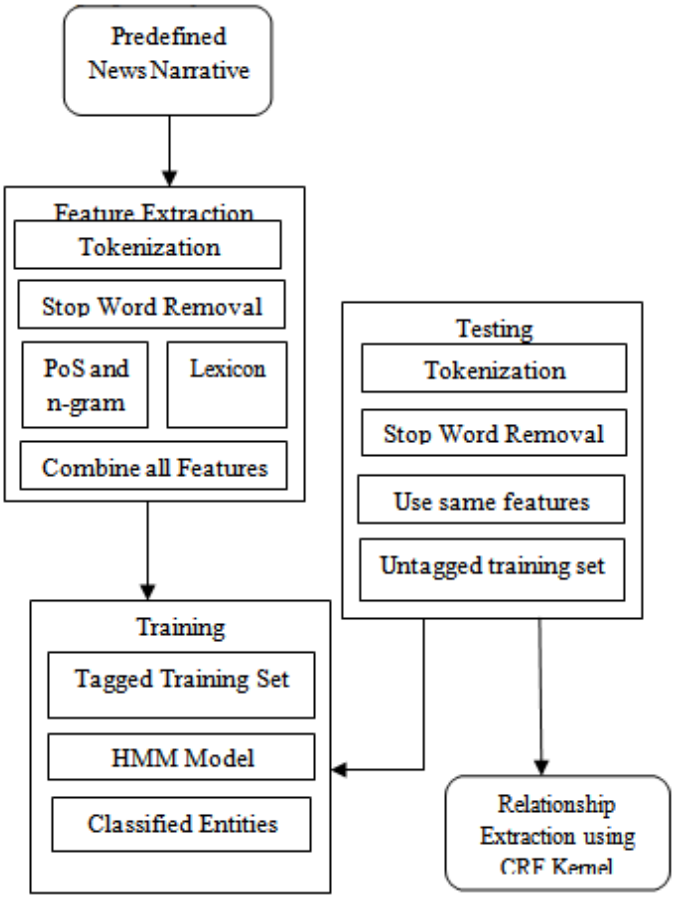

Figure 2: Diagrammatic representation of system modules

The system architecture of the proposed system is as under:

A. Our proposed system consists following four modules:
a. Feature Extraction
b. Training
c. Testing
d. Relationship Extraction

• Feature Extraction
We are extracting features like n-gram, parts of speech and lexicon. These features are extracted from news narratives after tokenization and removal of stop word. Once these features are extracted separately they are then combined.

.

• Training
The tagged training set is used for training. This set is used on HMM to train it and thus we will get classified entities.

• Testing
In testing we perform all activities similar to feature extraction and here we will have an untagged training set. This set when used on HMM model (trained) we will get the classified entities and its type. And similarly by using this set on CRF kernel approach, linking between entities can be found.

• Relationship Extraction
The various entities extracted may be related to each other. For this relationship extraction, Conditional Random Field with kernel is used. (if they have relation).

B. Algorithms
a. Feature extraction:

Algorithm 1 (Algorithm for entity extraction, classification and training)

Input: unstructured text of news narratives.
Output: Extract the entities of person, place and organization.
1. For I= length (doc)
2. For I=0 to length (doc)
3. for doc: split in sentences
X tokenization (sentence)
Y stop word removal
Extract Part of speech(Y)
N-gram(Y)
Lexicon features(Y)
End

End
4. Combine all features (n-gram+ Lexicon+ parts of speech)
5. For I= 0 to Len (doc)
Train HMM with kernel
End.
6. Model Model of HMM with Kernel.
7. For (I= 0 to I <length (doctest . sentences))
{
Extract features according to Step 3
Input in HMM with Kernel model.
}
12. Entity extraction and classified.

b. Relationship Extraction

Algorithm 2 (Entity relationship extraction algorithm)

Input: unstructured news narrative text
Output: Extraction of relationship among entities.
1. For I= length (doc)
2. For I=0 to length (doc)
3. for doc: split in sentences
X tokenization (sentence)
Y stop word removal
Extract Part of speech(Y)
N-gram(Y)
Lexicon features(Y)
End
End
4. Combine all features (n-gram+ Lexicon+ parts of speech)
5. For I= 0 to Len (doc)
Train CRF Kernel model
6. End.
7. Model Model of CRF with Kernel.
8. For I=0 to I < length (doc. sentence)
{
Input CRF kernel model
}
9. Output relationship extraction between entities.

IV. Result

By applying features like n-gram, parts of speech, and lexicon on HMM and CRF kernel approach models, the name entity extraction and linking becomes more accurate. By using these features on above mentioned models we get and improved system. When our proposed model is compared with existing model, the precision, accuracy and recall of our system results more.

- **SVM versus Proposed model for entity extraction and classification:**

	No. of Phrases	Correctly detected	Correctly Classified	Accuracy	Precision	Recall
SVM	350	310	305	96.50	84.65	81.58
Proposed Model	350	310	310	98.03	88.80	87.50

Table1. SVM versus Proposed model for entity extraction and classification

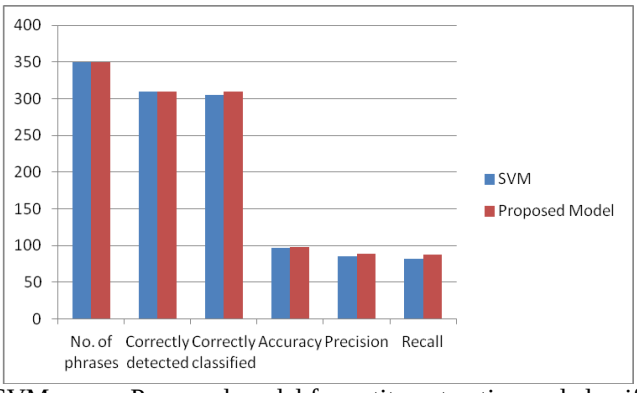

Fig. 3 SVM versus Proposed model for entity extraction and classification

Here, from above table we can say that the proposed system outperforms when compared with existing model.

- **SVM versus Proposed model for entity linking:**

	Number of relation	Correctly detected	Accuracy	Precision	Recall
SVM	125	107	87.46	84.46	82.46
Proposed model	125	110	85.34	82.45	79.5

Table2. SVM versus Proposed model for entity linking

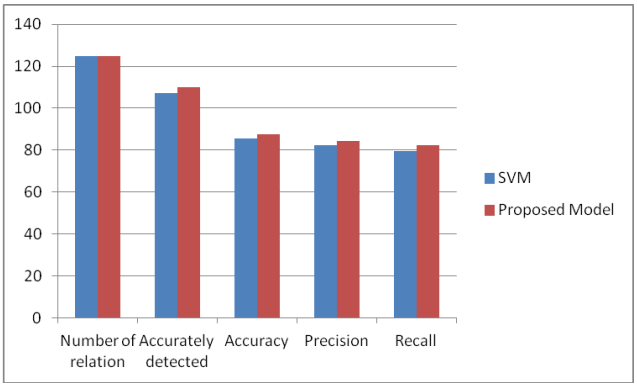

Fig. 3 SVM versus Proposed model for entity linking

V. Conclusion and Future-scope

Name entity recognition is a system of computer application that automatically recognize name entities from any text like news, reports etc. This process is mainly used in information retrieval, indexing, search engines etc. Due to similarity between person, place or organization, false classification of entities might arise which lowers the accuracy.

There are many approaches which are used for name entity recognition and linking. This paper presents name entity reorganization by using Hidden Markov model (HMM) and linking between entities by Conditional random field with kernel function, which makes impact on overlapping information of features. Classification of entities and there linking is an important task which is performed through HMM and CRF respectively, by taking the advantage of n-gram feature which used in feature set.

News narrative dataset has been taken for validation of the approach. In news narrative dataset, news data set has been used which provide training for big spectrum. Using this data set, good result has been founded. The precision value is 84.64 only when using lexicon feature set and its value increases by using features like N-gram, Part of speech (POS) i.e. 93.28. Accuracy and recall also significantly improve to conditional random field 98.99% and 96.77%. We conclude that kernel approach increases precision value with the help of n-gram features.

The Web is growing every day and new named entities appear continuously. As a result, it is not possible to keep a POS tagger up-to-date. The most frequent error made by a PoS tagger is the assignment of the tag “common noun" to a word denoting a proper noun. In future, disambiguation-based new entity filtration is recommended. The goal of new entity recommendation is to recommend all or most of the NEs that are not yet registered in knowledge base.

References

Journal Papers:

[1]. Daljit Kaur and Ashish Verma, Name Entity Recognition by New Framework using Machine learning Algorithm, IOSR Journal of Computer Engineering, Volume 16, Issue 5, Ver. IV, September-October 2014.
[2]. Navneet Kaur AulaKh and Er.Yadwinder kaur , Review Paper on Name Entity Recognition of Machine Translation, International journal of advanced Research in computer science and software engineering, Volume 4, Issue 4, April 2014.
[3]. Roman Prokofyev, Gianluca Demartini, and Phillipe Curde Mauroux Effective Named Entity Recognition for Idiosyncratic Web Collections, International World Wide Web Conference Committee, April 7–11, 2014.
[4]. Abhishek Gattani, Digvijay S. Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, AnHai Doan, Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-Based Approach, International Conference on Very Large Data Bases, Vol. 6, No. 11, 26th- 30th August, 2013.
[5]. Abdulrahman Alkaff and Masnizah Mohd, Extraction of Nationality from Crime News, Journal of Theoretical and Applied Information Technology, Vol. 54 No.2, 20th August 2013.
[6]. Wanxiang Che, Mengqiu Wang, Christopher D. Manning and Ting Liu, Named Entity Recognition with Bilingual Constraints, Association for Computational Linguistics, 9-14 June 2013.
[7]. Sudha Morwal and Nusrat Jahan, Name Entity Recognition Using Hidden Markov Model (HMM): An Experimental Result on Hindi, Urdu and Marathi Languages, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013.
[8]. Delip Rao, Paul McNamee and Mark Dredze , Entity linking: Finding extracted entities in a knowledge Base, Springer 2013.
[9]. Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, Xiangyang Zhou , Joint Inference of Named Entity Recognition and Normalization for Tweets, Association for Computational Linguistics, 8-12 July 2012.
[10]. Zaiqing Nie, Ji - Rong Wen and Wei Ying Ma, Statistical Entity Extraction from Web, Proceedings of IEEE, Vol.100, no. 9, 2012.
[11]. Kamaldeep Kaur and Vishal Gupta, Name Entity Recognition for Punjabi Language, International Journal of Computer Science and Information Technology & Security, Vol. 2, No.3, June 2012.
[12]. Sudha Morwal, Nusrat Jahan, and Deepti Chopra, Name Entity Recognition using Hidden Markov Model (HMM), International Journal on Natural Language Computing, Vol. 1, No.4, December 2012.
[13]. Ming Liu, Rafael A Calvo, Anindito Aditomo and Luiz Augusto Pizzato , Using Wikipedia and Conceptual Graph Structures to Generate Questions for Academic Writing Support, IEEE Transactions On Learning Technologies, Vol. 5, No. 3, July-September 2012.
[14]. Peter Exner and Pierre Nugues, Entity extraction: from unstructured text to DBpedia RDF Triples, The Web of Linked Entity Workshop, 2012.
[15]. Giuseppe Rizzo and Raphael Troncy, NERD: Evaluating Name Entity Recognition, 2011.
[16]. Chih Hao Ku, Alicia Iriberri and Gondy Leroy,Crime Information Extraction from Police and Witness Narrative Reports, IEEE International Conference on Technologies for Homeland Security, 12-13 May, 2008
[17]. Jie Tang, Duo Zhang and Limin Yao, Social Network Extraction of Academic Researchers, Seventh IEEE International Conference on. IEEE, 2007.
[18]. Jie Tang, Duo Zhang and Limin Yao, Social Network Extraction of Academic Researchers, Seventh IEEE International Conference on. IEEE, 2007.
[19]. Richard C. Wang and William W. Cohen, Language Independent Set Expansion of Name Entities using the Web, Seventh IEEE International Conference on. IEEE, 2007
[20]. David Nadeau and Satoshi Sekine, A survey of name entity recognition and classification, Lingvisticae Investigationes , 2007.
[21]. Natalia Ponomareva, Paolo Rosqo, Ferran Pla and Antonio Molina, Conditional random Fields vs. Hidden Markov Models in a biomedical named entity recognition task, Recent Advances in Natural Language Processing, RANLP. 2007.
[22]. Fuchun Peng and Andrew McCallum, Accurate Information Extraction from Research Papers Using Conditional Random Fields, Information processing & management42.4 (2006): 963-979.
[23]. Raymond J. Mooney and Razvan Buneseu, Mining Knowledge from Text Using Information Extraction, Volume 7, ACM SIGKDD explorations newsletter 7.1, 2005.
[24]. Phil Blunson , Hidden Markov Models, Lecture notes, August, 2004.
[25]. William W. Cohen and Sunita Sarawagi, Exploiting Dictionaries in Named Entity Extraction: Combining Semi Markov Extraction Processes and Data Integration methods, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.,ACM, 2004.
[26]. Radu Florian, Abe Ittycheriah Hongyan Jing and Tong Zhang, Name Entity Recognition through Classifier Combination , Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, Association for Computational Linguistics, 2003.
[27]. Michael Chau, Jennifer J. Xu and Hsinchun Chen, Extracting Meaningful Entities from Police Narrative Reports, Proceedings of the 2002 annual national conference on Digital government research. Digital Government Society of North America, 2002.
[28]. Nigel Collier, Chikashi Nobata and Jun-ichi Tsujii, Extracting the Names of Genes and Gene Products with a Hidden Markov Model, Proceedings of the 18th conference on Computational linguistics-Volume 1.Association for Computational Linguistics, 2000.
[29]. Michael Mateas and Phoebe Senger, Narrative Intelligence, American Association for Artificial Intelligence, 1999.
[30]. Paul Mc Kevitt, Derek Partridge and Yorick Wilks, Approaches to natural language discourse processing, Artificial Intelligence Review , Springer 1992.
[31]. Daljit Kaur and Ashish Verma , Survey on name entity recognition used machine learning algorithm, 1991.