A survey of Named Entity Recognition in English and other Indian Languages

Darvinder kaur[1], Vishal Gupta[2]

[1] Department of Computer Science & Engineering, Panjab university Chandigarh-160014, India

[2] Department of Computer Science & Engineering, Panjab university Chandigarh-160014, India

Abstract

In this paper, a survey is done on various approaches used to recognize name entity in various Indian languages. Firstly, the introduction is given about the work done in the NER task. Then a survey is given about the work done in recognition of name entities in English and other foreign languages like Spanish, Chinese etc. In English language, lots of work has been done in this field, where capitalization is a major clue for making rules. Secondly, a survey is given regarding the work done in Indian Languages. As Punjabi is one of the Indian languages and also the official language of Punjab. In next part, survey is given on Punjabi Language regarding what work is done and what work is going on in this field.

Keywords: *Named Entity, Named Entity Recognition, Tag set.*

1. Introduction

The term "Named Entity", the word Named restricts the task to those entities for which one or many rigid designators stands as referent[22]. It is widely used in Natural Language Processing (NLP). It is the subtask of Information Extraction (IE) where structured text is extracted from unstructured text, such as newspaper articles. The task of Named Entity Recognition is to categorize all proper nouns in a document into predefined classes like person, organization, location, etc. NER has many applications in NLP like machine translation, question-answering systems, indexing for information retrieval, data classification and automatic summarization. It is two step process i.e. the identification of proper nouns and its classification. Identification is concerned with marking the presence of a word/phrase as NE in the given sentences and classification is for denoting role of the identified NE. The NER task was added in Message Understanding Conference (MUC) held in November, 1995 at Los Altos [5][18]. The various approaches of NER are namely- Rule Based, Machine Learning based which includes HMM, Maximum Entropy, Decision Tree, Support Vector Machines and Conditional Random Fields and Hybrid Approach. Although a lot of work has been done in English and other foreign languages like Spanish, Chinese etc with high accuracy but regarding research in Indian languages is at initial stage only. Here a survey of research done till now in English and other foreign and Indian languages are presented. Early systems are making use of handcrafted rule-based algorithms. While modern systems most often use machine learning techniques. Handcrafted rule-based systems usually give good results, however they need months of development by experienced linguists. Whereas machine learning techniques uses a collection of annotated documents to train classifier for the given set of NE classes. According to the specification defined by MUC, the NER tasks generally work on seven types of named entities as listed below:

- Person Name
- Location Name
- Organization Name
- Abbreviation
- Time
- Term Name
- Measure

2. Previous Work

There are several classification methods which are successful to be applied on NER task. Till now, the

research aiming at automatically identifying named entities in texts forms a vast and heterogeneous pool of strategies, methods and representations. The main approaches to NER are Linguistics approaches and Machine Learning approaches. The Linguistics approach uses rule-based models manually written by linguists. ML based techniques make use of a large amount of annotated training data to acquire high-level language knowledge. Various ML techniques which are used for the NER task are Hidden Markov Model (HMM) [7], Maximum Entropy Model (MaxEnt) [1], Decision Tree [9], Support Vector Machines [20] and Conditional Random Fields (CRFs) [11]. Both the approaches may make use of gazetteer information to build system because it improves the accuracy.

Ralph Grishman in 1995 developed a rule-based NER systems which uses some specialized name dictionaries including names of all countries, names of major cities, names of companies, common first names etc[19]. In rule-based approaches, a set of rules or patterns is defined to identify the named entities in a text. Another rule-based NER system is developed in 1996 which make use of several gazetteers like person name, organization name, location names, person names, human titles etc[21]. But the main disadvantages of these rule-based techniques are that these require huge experience and grammatical knowledge of particular languages or domains and these systems are not transferable to other languages.

Borthwick in 1999 developed a ML based system i.e. MaxEnt based system[1]. This system used 8 dictionaries. ML based techniques for NER make use of a large amount of NE annotated training data to acquire high level language techniques uses gazetteer lists. A lot of work has been done on NER for English employing the machine learning techniques, using both supervised learning and unsupervised learning. In English language, it is easier to identify NE because of the capitalization of names.

Unsupervised learning approaches do not require labeled training data i.e. training requires few seed lists and large unannotated corpora. In unsupervised learning, the goal is to build representations from data. These representations are then be used for data compression, classifying, decision making and other purposes. Unsupervised learning is not a very popular approach for NER and the systems that do use unsupervised learning are usually not completely unsupervised. Collins et. al[6]. discusses an unsupervised model for named entity classification by the use of unlabelled examples of data. Secondly, Koim et. al[10]. Proposes an unsupervised named entity classification models and their ensembles that uses a small-scale named entity dictionary and an unlabeled corpus for classifying named entities.

Supervised learning involves using a program that can learn to classify a given set of labeled examples that are made up of the same number of features. The Supervised learning approach requires preparing labeled training data to construct a statistical model. But supervised approaches can achieve good performance only when large amount of high quality training data is available. Supervised approaches are more expensive than unsupervised one, in terms of the time spend to pre-process the training data. Statistical methods such as HMM, Decision Tree Model and Conditional Random Fields have been used.

Hidden Markov Model is a generative model. The model assigns the joint probability to paired observation and label sequence. Then the parameters are trained to maximize the joint likelihood of training sets. It is advantageous as its basic theory is elegant and easy to understand. Hence it is easier to implement and analyze. It uses only positive data, so they can be easily scaled.

Disadvantage - In order to define joint probability over observation and label sequence HMM needs to enumerate all possible observation sequence. Hence it makes various assumptions about data like Markovian assumption i.e. current label depends only on the previous label. Also it is not practical to represent multiple overlapping features and long term dependencies. Number of parameter to be evaluated is huge. So it needs a large data set for training.

Maximum Entropy Markov Models (MEMMs): It is a conditional probabilistic sequence model. It can represent multiple features of a word and can also handle long term dependency. It is based on the
principle of maximum entropy which states that the least biased model which considers all know facts is the one which maximizes entropy. Each source state has a exponential model that takes the observation feature as input and output a distribution over possible next state. Output labels are associated with states.

Advantages: It solves the problem of multiple feature representation and long term dependency issue faced by HMM. It has generally increased recall and greater precision than HMM.

Disadvantages: It has Label Bias Problem. The probability transition leaving any given state must sum to one. So it is biased towards states with lower outgoing transitions. The state with single outgoing state transition will ignore all observations. To handle Label Bias Problem we can change the state-transition

Conditional Random Field (CRF): It is a type of discriminative probabilistic model. It has all the advantage of MEMMs without the label bias problem. CRFs are undirected graphical models (also know as random field)

which is used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes.

In Hybrid NER system, approach uses the combination of both rule-based and ML technique and makes new methods using strongest points from each method. It is making use of essential feature from ML approach and uses the rules to make it more efficient. Sirihari et. al. introduce a hybrid system by combination of HMM, MaxEnt, and handcrafted grammatical rules[24]. In the field of NER for English and other European Languages, lots of work has already been done. This is possible because of the main feature in English i.e. the capitalization of names in the text. That is why NER task is achieved with high accuracy.

Hybrid Approach: Here several Machine Learning and Rule based systems are combined to improve the accuracy of classifier. Some examples of Hybrid systems are

- MaxEnt + Rule **:** Borthwick(1999) – 92% f-measure
- MaxEnt + Rule**:** Edinburgh Univ.–93.39% f-measure
- MaxEnt +HMM + Rule: Srihari et al. (2000) – 93.5% f-measure

In this field, recent researches are focused on multimedia indexing, unsupervised learning, complex linguistics phenomena and machine translation. Lots of efforts are taken toward semi-supervised and unsupervised approaches to NER motivated by the use of very large collection of texts [8] and the possibility of handling multiple NE types [15]. Complex linguistic phenomena that are common short-coming of current systems are under investigation [17].

The term semi-supervised is relatively recent. The main technique for SSL is called bootstrapping and involves a small degree of supervision, such as a set of seeds, for starting the learning process. Recent experiments in semi-supervised NERC [15] report performance that rival baseline supervised approaches.

Features are characteristic attributes of words designed for algorithmic purpose. Following features are most often used for the recognition and classification of named entities. These are defined into three categories i.e.

- Word-level features
- List lookup features
- Document and corpus features

Word-level features describe the character makeup of words i.e. the word case, punctuation, numerical value, part-of-speech (POS) and special characters

List lookup features can be called also as the term “gazetteer”, “lexicon” and “dictionary”. It include the general list, list of entities such as organization name, first name etc. and the looking into predefined list.

Document and corpus features are defined as collection of document content and document structure. Large collection of document (corpora) are also excellent sources of features. These all features together or in different combination helps in generating effective and efficient NER system for different domains or languages.

3. NER for Indian languages

NLP research around the world has taken major turn in the last decade with the advent of effective machine learning algorithms and the creation of large annotated corpora for various languages. But not much work has been done in NER for Indian languages because annotated corpora and other lexical resources have started appearing very recently in India. As common feature function like capitalization are not available in Indian languages and due to lack of large labeled dataset and lack standardization and spelling variation, so English NER cannot be directly used for Indian languages. So there arises the need to develop novel and accurate NER system for different Indian languages.

3.1 Characteristic and some problems faced by Hindi and other Indian languages

- No capitalization
- Brahmi script- It has high phonetic characteristic which could be utilized by NER system.
- Non-availability of large gazetteer
- Lack of standardization and spelling
- Number of frequently used words (common nouns) which can also be used as names are very large. “Also the frequency with which they can be used as common noun as against person name is more or less unpredictable.”
- Lack of labeled data
- Scarcity of resources and tools
- Free word order language

3.2 Some points to consider while building NER System

- Ease to change
- Portability (domains and language)
- Scalability
- Language Resources
- Cost-effective

3.3 Performance Evaluation Metrics are:

- Precision (P): Precision is the fraction of the documents retrieved that are relevant to the user's information need.
  Precision (P) = correct answers/answers produced
- Recall (R): Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.
  Recall (R) = correct answers/total possible correct answers
- F-Measure: The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is
  F-Measure = $(\beta^2+1)PR/(\beta^2R+P)$
  β **is** the weighting between precision and recall typically $\beta=1$.

When recall and precision are evenly weighted i.e. $\beta=1$, F-measure is called F_1 measure.

F_1-Measure = $2PR/(P+R)$

There is a tradeoff between precision and recall in the performance metric.

In IJCNLP-08 workshop on NER for South and South East Asian languages, held in 2008 at IIT Hyderabad, was a major attempt in introducing NER for Indian languages that concentrated on five Indian languages- Hindi, Bengali, Oriya, Telugu and Urdu.

The work regarding Telugu language is mentioned in [16]. The evaluation has reported F-Score of 44.91%. The development of a NER system for Bengali language is reported in 2008[2]. Its F-Score is 91.8%.

The work of Gali et al, in 2008 reports lexical F-Score of 40.63%, 50.06%, 39.04%, 40.94%, and 43.46% for Bengali, Hindi, Oriya, Telugu, and Urdu respectively [12]. In 2007 discussed the comparative study of Conditional Random Field and Support Vector Machines for recognizing named entities in Hindi language [4]. Indian languages are resource poor languages because of the non-availability of the annotated corpora, name dictionaries, good morphological analyzers etc. That is why high accuracy is not achievable yet.

The maximum accuracy for NER in Hindi is reported by Kumar and Bhattacharyya in 2006. They achieved an F measure of 79.7% using a Maximum Entropy Markov Model [13]. Among other Indian languages, Punjabi language still lacks behind in this field. A research work is concentrated on NER for Punjabi language.

Punjabi is the official language of the Indian state of Punjab. It is also official language of Delhi and ranked 20^{th} among the language spoken in the world [23]. Among the Indian languages, Punjabi is the one in which the lots of research is going on in this field. Due to the non-availability of annotated corpora, name dictionaries, good morphological analyzer etc. up to the required measure, Punjabi is the resource poor language like other Indian languages.

A recent research on NER for Punjabi language is done using Conditional Random Field (CRF) Approach [25]. It was aimed to develop a standalone system based on CRF approach which can be used with other NLP applications like Machine Translation, Information Retrieval etc. In this paper, 12 named entities are mentioned as in table 1.

Table 1: Named Entity Tagset

NE Tag	***Definition***
NEP(Person)	Name of a person
NEL(Location)	Name of a place, location
NEO(Organization)	Name of a political organization
NED(Designation)	Name of any designation
NETE(Term)	Name of diseases
NETP(Title-Person)	Name of title coming before the name of person
NETO(Title-Object)	Name of Object
NEB(Brand)	Brands Name
NEM(Measure)	Any measure
NEN(Number)	Numeric value
NETI(Time)	It include date, month, year etc
NEA(Abbreviation)	Name in short form

These tagset are used to tag each word in the sentence. Firstly, to find the useful features for NER task and secondly, to find the optimum feature set for the task. The various features which are applied to the NER tasks in this experiment are as follows:

Context word feature : Previous and next words of a particular word have been used as a feature. Generally, word window of size 5 or 7 is used.

Word suffix and prefix: In this feature, a length of 1 to 4 characters of the current and/or the surrounding words is taken.

Parts of Speech (POS) Information: A rule-based POS tagger developed at Punjabi University by Gill and Lehal in 2007 is used [14]. It is helpful in tagging the data but with limited accuracy. Some wrong tags are manually corrected for NER task.

Named Entity Information: It is the feature in which the NE tag of the previous word is considered. It is the dynamic feature.

Gazetteer Lists: Due to the scarcity of resources in electronic format for Punjabi language, so the gazetteer lists are prepared manually from websites and newspaper available online. Seven different lists are prepared such as:

- Person-Prefix
- First-Name
- Middle-Name
- Last-Name
- Location-Name
- Month Name
- Day Name

The F-score is calculated for the different use of features to obtain the optimal feature set. An overall F-score of 80.92% achieved for the Punjabi NER. The F-score has different value for the different NE tags. This means NER systems can be changed according to the type of NE tags required. The performance can be improved by improving gazetteer lists.

4. Conclusions

The Named Entity Recognition field has been thriving for more than fifteen years. It aims at extracting and classifying mentions of rigid designators, from text, such as proper names and temporal expressions. In this survey, we have shown the previous work done in English and other European languages. A survey is given on the work done in Indian Languages i.e. Telugu, Hindi, Bengali, Oriya and Urdu. An overview of the techniques employed to develop NER systems, documenting the recent trend away from hand-crafted rules towards machine learning approaches. Handcrafted systems provide good performance at a relatively high system engineering cost. When supervised learning is used, a prerequisite is the availability of a large collection of annotated data. Such collection are available from the evaluation forums but remain rather rare and limited in domain and language coverage. Recent studies in the field have explored semi-supervised and unsupervised learning techniques that promise fast deployment for many entities types without the prerequisite of an annotated corpus. Here also provided an overview of the evaluation methods that are in the use of NER accuracy. We have listed and categorized the features that are used in recognition of NE. The use of an expressive and varied set of features turns out to be just as important as the choice of machine learning algorithms. And finally the survey on the NER for Punjabi language is given. In it the working of an approach is explained.

5. Future work

- The performance can further be improved by improving gazetteer lists.
- Analyzing the performance using other methods like Maximum Entropy and Support Vector Machines
- Comparing the results obtained by using different approaches and calculating the most accurate approach for it.
- Improve the performance of each NE tag to make it overall more accurate.

References

[1] Andrew Borthwick. 1999. “Maximum Entropy Approach to Named Entity Recognition” Ph.D. thesis, New York University.

[2] Asif Ekbal, Sivaji Bandyopadhyay. “Bengali Named Entity Recognition using Support Vector Machine” in the proceedings of the IJCNLP-08 workshop on NER for South and South East Asian Languages, pages 51-58, Hyderabad, India.

[3] Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateshwar Poka and Sivaji Bandyopadhyay. 2008., “Language Independent Named Entity Recognition in Indian Languages” in the proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 33-40, Hyderabad, India.

[4] Awaghad Ashish Krishnarao, Himanshu Gahlot, Amit Srinet, D.S.Kushwaha, 2009. “A Comparison of Performance of Sequential Learning Algorithms on the task of Named Entity Recognition for Indian Languages” in the proceedings of 9th International Conference on computer Science. Pages 123-132. Baton Rouge, LA, USA.

[5] Charles L. Wayne. 1991., “A snapshot of two DARPA speech and Natural Language Programs” in the proceedings of workshop on Speech and Natural Languages, pages 103-404, Pacific Grove, California. Association for Computational Linguistics.

[6] Collins, Michael and Y. Singer. 1999. “Unsupervised models for Named Entity Classification”, in the proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

[7] Daniel M. Bikel, Scott Miller, Richard Schwartz and Ralph Weischedel. 1997 "Nymble: a highperformance learning name-finder" in the proceedings of the fifth conference on Applied natural language processing, pages 194-201, San Francisco, CA, USA Morgan Kaufmann Publishers Inc.

[8] Etzioni, Oren; Cafarella, M; Downey, D.; Popescu, A.-M.; Shaked, T.; Soderland, S.; Weld, D.S.; Yates, A. 2005. " Unsupervised Named-Entity Extraction from the Web: An Experimental Study" in Artificial Intelligence 165. Pages 91-134, Essex: Elsevier Science Publishers.

[9] Hideki Isozaki. 2001. "Japanese named entity recognition based on a simple rule generator and decision tree learning" in the proceedings of the Association for Computational Linguistics, pages 306-313. India.

[10] J. Kim, I. Kang, K. Choi, "Unsupervised Named Entity Classification Models and their Ensembles", in the proceedings of the 19th International Conference on Computational Linguistics, 2002.

[11] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data" in the proceedings of International Conference on Machine Learning, pages 282-289, Williams College, Williamstown, MA, USA.

[12] Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla and Dipti Misra Sharma. 2008.,"Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition" in the proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 25-32, Hyderabad, India.

[13] Kumar N. and Bhattacharyya Pushpak. 2006. "Named Entity Recognition in Hindi using MEMM" in the proceedings of Technical Report, IIT Bombay, India.

[14] Mandeep Singh Gill, Gurpreet Singh Lehal and Shiv Sharma Joshi, 2009. "Parts-of-Speech Tagging for Grammar Checking of Punjabi" in the Linguistics Journal Volume 4 Issue 1, pages 6-22.

[15] Nadeau, David; Turney, P.; Matwin, S. 2006. "Unsupervised Named Entity Recognition; Generating Gazetteers and Resolving Ambiguity" in the proceedings of Canadian Conference on Artificial Intelligence.

[16] Praneeth M Shishtla, Karthik Gali, Prasad Pingali and Vasudeva Varma. 2008. "Experiments in Telugu NER: A conditional Random Field Approach" in the proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 105-110, Hyderabad, India.

[17] Poibeau, Thierry. 2006. "Dealing with Metonymic Readings of Named Entities " in the proceedings of Annual Conference of the Cognitive science Society.

[18] R. Grishman, Beth Sundheim. 1996. "Message Understanding Conference-6: A Brief History" in the proceedings of the 16th International Conference on Computational Linguistics (COLING), pages 466-471, Center for Sprogteknologi, Copenhagen, Denmark.

[19] R. Grishman. 1995. "The NYU system for MUC-6 or Where's the Syntax" in the proceedings of Sixth Message Understanding Conference (MUC-6) , pages 167-195, Fairfax, Virginia.

[20] Takeuchi K. and Collier N. 2002. "Use of Support Vector Machines in extended named entity recognition" in the proceedings of the sixth Conference on Natural Language Learning (CoNLL-2002), Taipei, Taiwan, China.

[21] Wakao T., Gaizauskas R. and Wilks Y. 1996. "Evaluation of an algorithm for the Recognition and Classification of Proper Names", in the proceedings of COLING-96.

[22] http://en.wikipedia.org/wiki/Named_entity.

[23] http://en.wikipedia.org/wiki/Punjabi_language.

[24] Srihari R., Niu C. and Li W. 2000. "A Hybrid Approach for Named Entity and Sub-Type Tagging" in the proceedings of the sixth Conference on Applied Natural Language Processing.

[25] Amandeep Kaur, Gurpreet S. Josan and Jagroop Kaur."Named Entity Recognition for Punjabi: A Conditional Random Field Approach" in the proceedings of 7th International Conference on Natural Language Processing, Macmillan Publishers, India.

First Author

Darvinder Kaur is Assistant Professor in Computer Science and Engineering Department at Lovely Professional University, Phagwara. She has done M.E. in Computer Science and Engineering from University Institute of Engineering and Technology, Panjab University, Chandigarh in 2010. She has done B.Tech in Computer Science and Engineering from Guru Nanak Dev Engineering College, Ludhiana in 2008.

Vishal Gupta is Assistant Professor in Computer Science & Engineering
Department at University Institute of Engineering & Technology, Panjab university Chandigarh. He has done MTech. In computer science & engineering from Punjabi University Patiala in 2005. He was among university toppers. He secured 82% Marks in MTech. Vishal did his BTech. in CSE from Govt. Engineering College Ferozepur in 2003. He is also pursuing his PhD in Computer Sc & Engg. Vishal is devoting his research work in field of Natural Language processing. He has developed a number of research projects in field of NLP including synonyms detection, automatic question answering and text summarization etc. One of his research paper on Punjabi language text processing was awarded as best research paper by Dr. V. Raja Raman at an International Conference at Panipat. He is also a merit holder in 10_{th} and 12_{th} classes of Punjab School education board. in professional societies. The photograph is placed at the top left of the biography.