Named Entity Recognition: A Review

Arshdeep Singh[1], Jyoti Rani[2], Kuljot Singh[3]

[1,2] Department of Computer Science & Engg., PTUGZS Campus, Bathinda, Punjab, India

[3]Department of Computer Science & Engg., CIET, Fatehgarh Sahib, Punjab, India

[1]arshrandhawa@gmail.com, [2]csejyotigill@gmail.com, [3]kuljotsingh88@gmail.com

***Abstract*— Named entities are phrases that represent person, location, number, time, measure, organization. Named Entity Recognition is the task of identifying and classifying named entities into some predefine categories. This paper gives a brief introduction to Named Entity Recognition. It also summarizes various approaches for Named Entity Recognition like Hidden Markov Model, Maximum Entropy Markov Models, Conditional Random Field, Support Vector Machine, Decision Trees and Hybrid approaches. Named Entity Tagsets defined for MUC-6, CoNLL 2002 and 2003 and IJCNLP-2008 shared tasks are also discussed. Different NER features in context to identification and classification of named entities have also been reviewed.**

I. INTRODUCTION

Named Entities (NE) are phrases that contain person, organization, location, number, time, measure etc. Named Entity Recognition (NER) is the task of identifying and classifying the Named Entities into predefine categories such as person, organization, location, etc in the text. It aims to extract and to classify rigid designators in text. In Computational Linguistics taxonomy, NER falls within the category of Information Extraction (IE), which deals with the extraction of specific information from given document. NER is an IE task which aims to locate named entities in a particular document. It is also considered as the core of natural language processing (NLP) system. It is also known as entity identification and entity extraction. It was introduced in the sixth Message Understanding Conference (MUC-6) [1] in 1996. Later the concept of NER was also studied as shared task in CoNLL-2002 and CoNLL-2003. Proper identification and classification of named entities (NEs) are very difficult and pose a very big challenge to the NLP researchers. The level of ambiguity in NER makes it difficult to attain human performance. The process of NER involves two tasks which is firstly the identification of proper names in text, and secondly the classification of these names into a set of predefined categories of interest, such as person names, organizations (companies, government organizations, committees, etc), locations (cities, countries, rivers, etc), date and time expressions. Figure 1 shows a general data flow of NER in which features are extracted from the training file and the test file. Features are basically the attributes of the words and helps in identifying relationships between words. The feature extraction can be done by using any NER methodology like Rule-based, Machine Learning which includes Hidden Markov Model, Maximum Entropy, Conditional Random Field etc. or Hybrid approach. Then a few features are selected from the features extracted in the previous stage. These features are further used by the system for learning process.

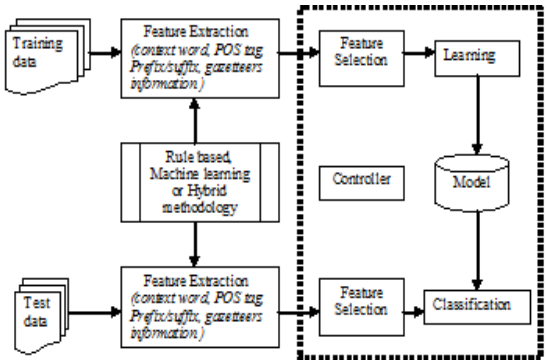

Figure 1. General Data Flow of NER

The learning process then creates a model for the system which is further used for recognizing and classifying named entities in the test data [2].

II. APPROACHES TO NER

In order to develop an NER system various approaches were proposed which includes Rule based / Handcrafted Approach, Machine Learning /Automated / Statistical approach, and Hybrid approaches.

A. The Rule based / Handcrafted Approach

List Lookup Approach: NER system uses gazetteers to classify words. We just have to create a suitable list in the gazetteer. It is simple, fast and language independent. It is also easy to retarget as we just have to create lists. The biggest disadvantage of this approach is that the gazetteers have to maintained and updated regularly. Moreover, Gazetteers cannot resolve ambiguity. Linguistic Approach: NER system uses some language based rules and other heuristic to classify words. It needs rich and expressive rules and gives good results. It requires an advanced knowledge of grammar and other language related rules. This is purely a language dependent approach which requires a thorough knowledge and understanding of the Language under consideration [15].

B. Machine Learning Based Approach / Automated Approach

Hidden Markov Models (HMM): It is a generative model. The model assigns a joint probability to paired observation and label sequence. Then the parameters are trained to maximize the joint likelihood of training sets. P(X, Y) = Π i P(Xi , Yi) P(Yi ,Yi-1) It uses forward- backward algorithm, Viterbi Algorithm and Estimation-Modification method for modeling. Its basic theory is elegant and easy to understand. Hence it is easier to implement and analyze. 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. In recent years several statistical methods based on supervised learning method were proposed [16]. Maximum Entropy (MaxEnt) 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. 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. 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 structure or we can start with fully connected model and let the training procedure decide a good structure. Borthwick et. al. investigates exploiting diverse knowledge sources via maximum entropy in named entity recognition [exploiting borthwick][3]. 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 known 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. Random field: Let G = (Y, E) be a graph where each vertex YV is a random variable. Suppose P(Yv | all other Y) = P(Yv |neighbors(Yv)), then Y is a random field. Let X = random variable over data sequences to be labeled Y = random variable over corresponding label sequence. "Definition Let G = (V,E) be a graph such that Y = (Yv) v∈V , so that Y is indexed by the vertices of G. Then (X,Y) is a conditional random field in case, when conditioned on X, the random variables Yv obey the Markov Property with respect to the graph: P(Yv|X,Yw, w ≠ v) = P(Yv |X,Yw, w ∈ v), where w ∈ v means that w and v are neighbors in G [4]. Support Vector Machine (SVM): SVM is one of the famous supervised machine learning algorithms for binary classification in all various data set and it gives the best results where the data set is a few, and with extended algorithms it can be used in multi-class problems. To solve a classification task by a supervised machine learning model like SVM, the task usually involves with training and testing data, which consists of some data instances. Each instance in the training set contains one "target value" (class labels, where class label 1 for positive and class label -1 for negative target value) and several "attributes" (features). The goal of a supervised SVM classifier method is to produce a model which predicts target value of the attributes. For each SVM, there are two data set namely, training and testing, where the SVM used the training set to make a classifier model and classify testing data set based on this model with use of their features. Takeuchi et al. presented a named entity recognition system based on support vector machines [5]. Decision Tree (DT): DT is a powerful and popular tool for classification and prediction. The attractiveness of DT is due to the fact that in contrast to neural network, it presents rules. Rules can readily be expressed so that human can understand them or even directly use them in a database access language like SQL so that records failing into a particular category may be tree. Decision Tree is a classifier in the form of a tree structure where each node is either a leaf node-indicates the value of the target attributes(class) of expressions, or a decision node that specifies some text to be carried out on a single attribute value with one branch and sub-tree for each possible outcome of the text. It is an inductive approach to acquire knowledge on classification. A tagging of unknown proper names system with Decision Tree model was proposed by Bechet et. al. [6].

C. Hybrid Model Approach

In the approach is to combine rule-based and machine learning-based methods, and make new methods using strongest points from each method. In this family of approaches Mikheev et. al. proposes a Hybrid LTG system [17], Sirihari et. al. introduce a Hybrid system by combination of HMM, MaxEnt, and handcrafted grammatical rules[16]. Although this type of approach can get better result than some other approaches, but the weakness of handcraft Rule-base NER remains the same that is when there is a need to change the domain of data.

III. NAMED ENTITY TAGSET

One of the important tasks while developing an NER system is to define a Named Entity Tagset which contains notations used for various NEs. The NE Tagset released in MUC-6 had three NE tags : ENAMEX (person, location and organization), TIMEX(date and time) and NUMEX (money and percent)[1]. The Tagset introduced in CoNLL-2003 Shared Task includes four tags : persons (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC)[21].

Tag	Name	Description
NEP	Person	James Bond, Amandeep Singh Randhawa
NED	Designation	Director, Manager
NEO	Organization	Muncipal Corporation
NEA	Abbreviation	INC, NLP
NEB	Brand	Adidas, Verka
NETP	Title-Person	Dr. , Mr.
NETO	Title-Object	Pride and Prejudice
NEL	Location	Chandigarh, California
NET	Time	22nd May, 1987
NEN	Number	500, 3.14
NEM	Measure	15 kg, Rs. 2000
NETE	Terms	Maximum Entropy

Table 1. Named entity tagset

Table 1 shows the tagset used in IJCNLP-08workshop on NER for South and South East Asian (SSEA) Languages [18]. In contrast to earlier tagsets defined in MUC-6 and CoNLL2003, this Tagset contains 12 tag notations.

IV. NER FEATURES

Feature selection plays a crucial role in identification of Named Entities. Various Language independent and Language dependent features are used to effectively identify and classify Named Entities in the text. Different possible combinations of features are also used to enhance the accuracy of NER systems. Some of the features used to identify named entities are:

A. Context word feature

Previous and next words of a particular word have been used as a feature. In our work we have experimented on word window 5and 7.

B. Word suffix

Word suffix information is helpful to identify NEs. For eg. A Word suffix, of length 1 to 4 characters, of the current and/or the surrounding word(s) was experimented as a feature.

C. Word prefix

Prefix information of a word is also helpful. Word prefix, of length 1 to 4 characters, of the current and/or the surrounding word(s) was experimented as feature.

D. Digit features

Several binary valued digit features have been defined depending upon the presence and/or the number of digits in a token (e.g., token containing digits only, token containing four digit, token containing two digits), combination of digits and punctuation symbols (e.g., token consisting of digits and comma, token consisting of digits and periods), combination of digits and symbols (token consists of digit and slash, token consisting of digits and hyphen, token consisting of digits and percent-ages). These binary valued features are helpful in recognizing miscellaneous NEs, such as time expressions, measurement expressions and numerical numbers etc.[19].

E. Parts of Speech (POS) Information

Parts of Speech (POS) of the current and/or the surrounding word(s) are also considered as features. For e.g. , we can use a highly coarse-grained tagset with 9 tags which are NN(Noun), PN(Pronoun), AJ(Adjective), AV(Adverb), Preposition(PP), Conjunction(CJ), Interjection(IJ) and PT(Postposition). Although POS tagger is very helpful in tagging the data but the success of the task is limited by the accuracy of this tagger. The wrong tags were manually corrected for NER task.

F. Named Entity Information

The NE tag of the previous word is also considered as a feature. This is the only dynamic feature in the experiment.

G. Gazetteer Lists

These lists have been used as the binary valued features . If the current token is in a particular list then the corresponding feature is set to 1 for the current and/or the surrounding word(s), otherwise, set to 0.

H. Person-Prefix

This is useful for detecting person names. This feature is set to 1 for the current and the next two words.

I. First-Name

This list contains the first name of person names. This feature is set to 1 for the current word.

J. Middle-Name

This list contains the middle name of person names. This feature is set to 1 for the current word and the previous word.

K. Last-Name

This list contains the last name of person names. This feature is set to 1 for the current word.

L. Location-Name

This list contains the location names and this feature is set to 1 for the current word.

M. Month name

This contains the name of all the different months of any calendar. The feature `Month-Name' is set to 1 for the current word.

N. Day Name

This contains the name of all the seven different days of the calendars. The feature 'Day Name' is set to 1 for the current word.

O. Common location word list

This list contains the words (e.g. road, lane etc.) that are part of the multiword location names and usually appear at their end.

P. Designation words

A list of common designation words (e.g., Neta, Mr. etc.) are usually pre-pared. This helps to identify the position of person names.

V. EXISTING NER SYSTEMS

The concept of NER was coined at MUC-6 in 1995. Following this track and to motivate the research in named entity recognition various NER systems were developed. In this section we have discussed some of the early systems.

- ABNER: Biomedical named entity recognizer [7].
- ANNIE: Information extraction package with NER capabilities [8].
- Balie: Baseline implementation of named entity recognition [9].
- ESpotter: A domain and user adaptation approach for named entity recognition on the Web [10].
- KeX: A simple Knowledge Extraction tool [11].
- Minor Third Collection of Java classes for storing text, annotating text, and learning to extract entities and categorize text [12].
- Carabao Morpho Logic Mixed dictionary-based and heuristics-based named entity recognition for single words only [13].
- OSCAR: Chemical Entity Recognizer [14].

REFERENCES

[1] R. Grishman, Beth Sundheim, "Message Understanding Conference - 6: A Brief History", In the Proceedings of the 16th International Conference on Computational Linguistics (COLING), Morgan Kaufmann publishers,

Center for Sprogteknologi, Copenhagen, Denmark , August 5-9, 1996, Pages 466 - 471.

[2] Amandeep Kaur, Gurpreet Singh Josan, Jagroop Kaur, "Named Entity Recognition in Punjabi Language – A Conditional Random Field Approach", In the Proceedings of International Conference on Natural language Processing (ICON-09), Hyderabad, India, 2009.

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

[4] K. Takeuchi and N Collier, "Use of support vector machines in extended named entity recognition", In the Proceedings of the sixth Conference on Natural Language Learning (CoNLL-2002), Association for Computational Linguistics Publishers ,Taipei, Taiwan, China, August 31 and September 1, 2002 pages 119-125.

[5] F. Bechet, A. Nasr and F. Genet, "Tagging Unknown Proper Names Using Decision Trees", In proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics Publishers ,Hong Kong, China, October 1-8, 2000, pages 77-84.

[6] D.M. Bikel, S. Miller, R. Schwartz, R, Weischedel, "NYMBLE : a High- Performance Learning Name-finder", In the proceedings of fifth conference on applied natural language processing, Morgan-Kaufmann Publishers , Washington, D.C., 31 March - 3 April, 1997, pages 194-201.

[7] R. Sirhari, C. Niu, W. Li, "A Hybrid Approach for Named Entity and Sub-Type Tagging" Proceedings of the sixth conference on Applied natural language processing ,Acm Pp. 247 - 254 , 2000.

[8] A. Mikheev, C. Grover, M. Moens, "Description OF THE LTG SYSTEM FOR MUC-7", In Proceedings of the seventh Message Understanding Conference (MUC-7), 1998.

[9] Asif Eqbal, Sivaji Banyopadhyay," Named Entity Recognition Using Appropriate Unlabeled Data, Post-processing and Voting", In the proceedings of Informatica 34,2010,pages 55-76.

[10] Erik F. Tjong Kim Sang, Fien De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition", In the proceedings of CoNLL-2003, Edmonton, Canada, May 31 and June 1, 2003.