International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 2 Issue 12, December-2013, pp: (91-94), Available online at: www.erpublications.com

An english name entity recognition system using cart

Abhilash Kumar Srivastava[1], Krishnendu Ghosh[2]
[1,2]School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, India

Abstract: In the current study at this paper, the different approaches of developing one NER system is discussed. This paper discuss related work about Name Entity Recognition System in English Language. A little database collection of 200 sentences contains 3080 words. The features selection and generations are suggested to capture the Name Entity. The proposed work is expected to predict the Name Entity of the focus words in sentence with high accuracy with the help of the suitable knowledge acquisition techniques.
Index Terms :Features; Classifier; CART; Database

I. INTRODUCTION

The task of Named Entity Recognition (NER) allows identifying proper names as well as temporal and numeric expressions, in an open-domain text. NER systems proved to be very important for many tasks in Natural Language Processing (NLP) such as Information Retrieval and Question Answering tasks. Existing NER systems have been constructed using mainly knowledge based or linguistic, and machine learning approach. The Knowledge based or linguistic approach is basically a rule-based approach which uses a set of handcrafted rules are designed and defined by human experts ,especially linguists. This model considers a set of patterns consisting of grammatical, syntactic, linguistic and orthographic features in combination with dictionaries. Machine learning approaches inherently supports rule-based systems or use sequence labeling algorithms to collect knowledge from a collection of training examples.
The semi-supervised is relatively recent technique which is basically a bootstrapping approach and involves a small degree of supervision in form of a set of seeds for knowledge acquisition. Unsupervised learning technique is a clustering technique based on the similarity of context, lexical patterns and other relevant features collected from lexical resources like WordNet.
Named entity recognition is used in many applications throughout several domains and fields. Named Entity Recognition system is an essential component of complex information extraction system. A named entity tagger serves as a preprocessing step of machine translation system.
The paper is organized as follows: related literatures are discussed in section II. The database for the name entity recognition module is discussed in the section III followed by the proposed work in section IV and direction towards conclusion and future work in section V.

II. RELATED WORK

Recognizing the named entity is useful to derive the relation and effective searching of data. Existing approaches for developing NER systems are classified into two broad categories:

(i) Knowledge Engineering and Linguistic Approach and (ii) Machine Learning Based Approaches.

A. **Knowledge Engineering And Linguistic Approach**

The knowledge engineering approach uses a set of rules written manually or defined by human experts or linguists. A comparative study states that such technique creates a better result for a specific domain. However, manual creation of rules is tiresome, expensive and still almost impossible. This approach has one advantage of the ability to efficiently identifying the named entities in the text documents. For developing effective and accurate rule-based techniques, huge experience and grammatical knowledge is required for the target language or domain. This approach is purely language or domain specific.

B. **Machine Learning Based Approach**

This approach can further be classified in three approaches:

1) **Supervised Learning**: The current dominant technique for addressing the NER problem is supervised learning. SL techniques include Hidden Markov Models (HMM), Decision Trees, Maximum Entropy Models (ME), Support Vector Machines (SVM), and Conditional Random Fields (CRF). These are all variants of the SL approach, which typically feature a system that reads a large annotated corpus, memorizes lists of entities, and creates disambiguation rules based on discriminative features. A baseline SL method that is often proposed consists of tagging test corpus words when they are annotated as entities in the training corpus [6]. The performance of the baseline system depends on the vocabulary transfer, which is the proportion of words, without repetition, appearing in both training and testing corpus. They report a transfer of 21%, with the repetition of much as 42% of location names, but only 17% of organizations and 13% of person names.
Vocabulary transfer is a good indicator of the recall (number of entities identified over the total number of entities) of the baseline system, but it is also a pessimistic measure since some entities are frequently repeated in documents. Mikheev et al.precisely calculated the baseline system recall on the MUC- 7 corpus. They report a recall of 76% for locations, 49% for organizations, and 26% for persons, with precision ranging from 70% to 90%. Whitelaw and Patrick report consistent results on MUC-7 for the aggregated enamex class. For the three enamex types together, the recognition precision is 76% and the recall is 48% [7].

2) Semi Supervised Learning: The term"semi-supervised" (or "weakly 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. For example, a system aimed at "disease names" might ask the user to provide a small number of example names. Then, the system searches for sentences that contain these names and tries to identify some contextual clues common to the five examples. Then, the system tries to find other instances of disease a name appearing in similar contexts. The learning process is then reapplied to the newly found examples, so as to discover new relevant contexts. By repeating this process, a large number of disease names and a large number of contexts will eventually be gathered.. Here are some examples of SSL approaches uses lexical features implemented by regular expressions in order to generate lists of book titles paired with book authors. When a given Web site is found to contain seed examples, new pairs can often be identified using simple constraints, such as the presence of identical text before, between, or after the elements of an interesting pair.

Collins and Singer parse a complete corpus in search of NE pattern candidates. A pattern is, for example, a proper name (as identified by a part-of-speech tagger) followed by a noun phrase in apposition (e.g.,"Maury Cooper, a vice president at S&P"). Patterns are kept in pairs spelling, context where"spelling" refers to the proper name and"context" refers to the noun phrase in its context. Starting with an initial seed of spelling rules: (e.g., rule 1: if the spelling is"New York" then it is a Location; rule 2: if the spelling contains"Mr." then it is a Person; rule 3: if the spelling is all capitalized then it is an organization), the candidates are examined. Candidates that satisfy a"spelling" rule are classified accordingly, and their"contexts" are accumulated. The most frequent contexts found are turned into a set of contextual rules [7]. Cucchiarelli and Velardi use syntactic relations (e.g., subject-object) to discover more accurate contextual evidence around the entities. Again, this is a variant of Riloff and Jones mutual bootstrapping. Interestingly, instead of using human generated seeds, they rely on existing NER systems (called"early NE classifier") for initial NE examples. Pasca also use techniques inspired by mutual bootstrapping. However, they innovate by using Lin's distributional similarity to generate synonyms-or, more generally, words belonging to the same semantic classallowing pattern generalization. For instance, in the pattern "X was born in November," Lin's synonyms for "November" are March, October, April, Mar, Aug., February, Jul, Nov., etc., thus allowing the induction of new patterns such as "X was born in March." One of Pasca et al.'s contributions is to apply this technique to very large corpora (100 million Web documents) and demonstrate that starting from a seed of 10 sample facts (defined as "person" type entities paired with "year" type entities, standing for the person's year of birth), it is possible to generate one million facts with a precision of about 88%. An existing NS classifier can be improved using bootstrapping methods and selection of unlabelled data [6].

3) Unsupervised Learning: The typical approach in unsupervised learning is clustering. For example, one can try to gather NEs from clustered groups based on context similarity. There are also other unsupervised methods. Basically, the techniques rely on lexical resources (e.g., WordNet), on lexical patterns, and on statistics computed on a large unannotated corpus. Here are some examples: Alfonseca and Manandhar study the problem of labeling an input word with an appropriate NE type. NE types are taken from WordNet (e.g., location¿country, animate¿person, animate¿animal, etc.). The approach is to assign a topic signature to each WordNet Synset by merely listing words that frequently co-occur with it in a large corpus. Then, given an input word in a given document, the word context (words appearing in a fixed-size window around the input word) is compared to type signatures and classified under the most similar one. In Evans, the method for Identification of hyponyms/hypernyms described in the work of Hearst is applied to identify potential hypernyms of capitalized word sequences appearing in a document. For instance, when X is a capitalized sequence, the query "such as X" is searched on the Web and, in the retrieved documents, the noun that immediately precedes the query can be chosen as the

X hypernym. Cimiano and Volker, Hearst patterns are used, but this time, the feature consists of the number of occurrences of words like “city such as,” ”organization such as,” etc. Shinyama and Sekine observed that NEs often appear in several news articles synchronously, whereas common nouns do not. They found a strong correlation between being an NE, and appearing intermittently and simultaneously in multiple news sources. This technique allows for identifying rare NEs in an unsupervised manner, and it can be useful when combined with other NER methods. Pointwise Mutual Information and Information Retrieval is used as a feature to assess that a named entity can be classified under a given type. PMI-IR, developed by Turney, measures the dependence between two expressions using Web queries. A high PMI-IR means that expressions tend to co-occur. Etzioni et al. create features for each entity candidate (e.g., London) and a large number of automatically generated discriminator phrases, like”is a city,” ”nation of,” etc.

III. DATABASE

A corpus has been collected for developing a baseline NER system for English data. The text data have been collected from different English newspapers namely, The Time of India and The Hindu. There are 200 sentences contain 3080 words. 90% of the corpus is used as training data while, the remaining data is used for testing the performance of the baseline NER system.

IV. IMPLEMENTATION OF THE PROPOSED NER MODEL

In the current study, a preliminary attempt has been made to develop a baseline NER system. for classifying the named entities corpus is collecting from the different newspapers which will be treated as the database or manually annotated corpus. The Name Entity Recognition works in two steps (i) training phase and (ii) testing phase. In training, preprocessing of the annotated corpus will be done and then features are generated. The features, based on the classes and the learning algorithm are used to trained the classifier. In testing phase, preprocessing of annotated corpus takes place and then features are generated. Classifier takes the features of the annotated corpus and performance evolution of the system takes place by comparing with the predicted output. Hence, the proposed contains four building blocks: (i) Baseline System, (ii) Preprocessing, (iii) Features Generation and Selection, (iv) Training (v) Testing.

A. Baseline System

The baseline system is developed to get acquainted with the development and other relevant issues and thus, figure out the accuracy performance of a NER system for training and testing.

B. Preprocessing

Preprocessing is the module where the raw data is processed in order to get the features which can be used for training or testing data efficiently. We remove errors, like, errors due to spelling, punctuations and spacing. After manually correcting the spelling errors, automatically the other issues are also rectified in this phase. The words are thoroughly converted in a tabular form according to their classes and features.

C. Features Selection and Generation

In this phase, the features are generating according to the words and their classes. The current study, as it is in the preliminary attempt, uses a set of very common, widely popular and easily extractable features. A total of 7 features are taken to build feature. They are:

(i) Next word,
(ii) Previous word,
(iii) Number of apostrophe present in the word,
(iv) Suffix of the word,
(v) Prefix of the word,
(vi) Number of capital letter in the word,
(vii) First letter of the word capital or not.

D. Training

Classification and regression tree (CART) is used to predict classes and features for the test corpus. It is a simple, linear classification tool which approximates the best prediction for any test patterns based on the knowledge acquired in course of the training procedure. In building the classifier, classification and regression tree (CART) based algorithm is suggested for training. In this approach, using the 5 features of the focus words for training data, automatically a binary decision tree is developed. Internal nodes and leaves of the binary decision tree represent the condition and a prediction respectively. By considering the instances of the training data, CART asks binary questions on the features. Starting from the root node at

each node, CART algorithm selects the most predictive feature from the feature set and the best possible question to achieve accurate classification of the training data.

E. Testing

In the process of testing, CART offers prediction by comparing the features for the test case with the trained decision tree. While predicting the Name Entity for a test word, the corresponding features are generated and compared with the trained decision tree. Starting from the root node asking a question on individual features, the features of the test word will lead to a leaf which offers an approximated output. The features are already generated in the feature generation phase. The trained CART is feeded with features are test data and corresponding approximations are noted. The predictions are later compared with the correct named entity to perform evaluation of the current system.

V. PERFORMANCE EVALUATION

The accuracy of the proposed named entity recognition system is evaluated using an objective analysis. The analysis is carried out by determining the correct named entity percentage for the test words. Being trained on 2,772 words and tested on 308 words using 7 features, the named entity recognition system predicts the senses with 66.88% accuracy. The performance of the proposed model is discussed in Table I.

TABLE I: Objective Evaluation Of Cart Based Named Entity Recognition Model

Predicted entity	Correctly predicted entity	Accuracy
308	206	66.88%

VI. CONCLUSIONS

Most of the existing works on NER systems are developed considering some limited domains and textual genres. On the other hand, exhaustive systems have been constructed only for the highly-researched languages like English or Mandarin. The rule-based systems can effectively classify the named entities but, building an accurate rule-set for unrestricted domain is almost impossible. When supervised learning is used, a prerequisite is the availability of a large collection of annotated data. Such collections are available from the evaluation forums but remain rather rare and limited in domain and language coverage. Recent studies in the field have explored semisupervised and unsupervised learning techniques that promise fast deployment for many entity types without the prerequisite of an annotated corpus.NER systems are developed generally using machine learning based approaches. Hence, the performance of such systems is highly dependent on the size of the training data. Most of the proposed NER systems achieved poor performance in some specific domains. There are plenty of scopes for working towards developing NER systems in such domains or languages. For these, linguistic features are to be explored as well as the suitable classifier should be found out.

REFERENCES

[1]. A. Mikheev, M. Moens, C. Grover, “Named Entity Recognition without Gazetteers”, in proceedings Conference of Europian Chapter of the Association for Computational Linguistics, 1991.
[2]. Brill, “Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging”, in proceedings of the Third Workshop on Very Large Corpora, 1995.
[3]. M. S Bindu, S. M. Idicula, “Name Entity Recognizer Employing Multiclass Support Vector Machines for the Development of Question Answering System ”,International Journal of Computer Applications, July 2011, volume 25, no. 10, pp. 0975-8887.
[4]. D. Kaur, V. Gupta, “A survey of Named Entity Recognition in English and other Indian Languages”,IJCSI International Journal of Computer Science Issues, vol. 7, Issue 6, November 2010.
[5]. D. Nadeau, “Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision”, PhD Thesis, University of Ottawa, 2007.
[6]. L. Berger, S. A. Della Pietra and V. J. Della Pietra, “A maximum entropy approach to natural language processing”,Computational linguistics,22(1):39 72, March 1996.
[7]. M. Bikel, R. Schwartz, R. M. Weischedel, “An Algorithm that Learns Whats in a Name”, BBN Systems & Technologies, 70 Fawcett Street, Cambridge MA.
[8]. M. Stevenson and R. Gaizauskas, “Improving Named Entity Recognition using Annotated Corpora”, in proceedings of the LREC Workshop ”Information Extraction meets Corpus Linguistics”, Athens, Greece,2000.
[9]. M. Bikel, S. Miller, R. Schwartz, R. M. Weischedel, “Nymble: A High Performance Learning Name Finder”, BBN Corporation, 70 Fawcett Street, Cambridge MA.