The Named Entity Recognizer Framework

Manika Nanda
Department of CSE/IT, Madhav Institute of Technology & Science
Gwalior (M.P.)
mani.nanda26@gmail.com

Abstract- Name Entity Recognition (NER) has been emerged as one of the Natural Language Processing (NLP) technology. This paper presents a Name Entity Recognition System for English and Hindi language. The English language mixes case text so there is presence of some clues such as initial capitalized letters clearly indicates the presence of name entities like name, place etc. but Hindi language doesn't provide such clues, so it is difficult to identify name entities in Hindi. In order to overcome this problem we store corresponding Hindi text of English words in our database. For this we build our own database which contains places, names and organization entities with their respective sub-categories as well as their Hindi transliteration. Since it's not possible to store all the numbers and dates in the database, we solve this type of problem with the help of already defined different date formats and patterns as well as matcher function.

Keywords- NER, NLP

I. INTRODUCTION

Named entity recognition is the process of extracting or identifying the named entities present in the input text. Named entity recognition system finds named entities such as person, location, organization, number, date etc. in the given text. Named entities are useful in many applications such as relation extraction, question-answering, automatic indexing of books etc. NER enables an IE system to recognize and classify information units in an unstructured text into pre-defined categories. Before we know about Named entity recognition we must know the following terminology.

A. TERMINOLOGY:

It includes entity and named entity. By using this terminology we clearly understood the Named entity recognition.

1) **Entity:** An entity is a real world object which is distinguished from other objects. Entities are something that exists by itself, although it need not be of any material existence. In different fields entity have different meaning such as
In DBMS, entity is either a thing in the modeled world or a drawing element in an Entity Relationship Diagram.
In System, entity refers to Component.
In the field of IE, entity refers to any noun such as book, location, organization names and numerical entities date, number etc.

2) Named Entity: Named entity refers to any object name in the real world. In the field of Information extraction, Named entities mainly refers to Person, Location, Organization names and Numerical entities Time, Money, Date and Number respectively. Named entity recognition (NER) is also known as entity identification and entity extraction. It is the process of identifying and recognizing the named entities such as person names, organization names, location names and numeric data entities such as date, number etc. It is a subtask of Information extraction. It is one of the important tasks in Information Extraction research area. Information Extraction (IE) is the process of extracting the relevant data from the available documents. NER is one of the important activities in natural language processing area pertaining to Information Extraction, Question Answering, Data mining, and Database Querying etc. Generally Named entity recognition tasks could find the seven types of entity names respectively. They are described as follows with their respective annotated tags.

PERSON
LOCATION
ORGANISATION
DATE
TIME
MONEY
PERCENT

Among these entity names we are finding only date number, person, location and organization names. Finding these entity names is very difficult when compared to the other entity names.

II. PROPOSED WORK:

A. TOKENIZATION

The first step of NER system consists of breaking a stream of an input English text up into meaningful elements called tokens where each token is either a word or something else like a number or a punctuation mark. Therefore, we

© 2014, IJIRAE- All Rights Reserved

need first to keep the sentence boundaries where the sentence is something that ends with a full stop '.', a question mark'?' or an exclamation mark '!', since the system is only able to tag entities on a token-by-token basis.

B. HANDLING DATE AND NUMBERS:

This module is independent from database. In this module we handle all the patterns of date with the help of already defined Date Pattern Formats like dd/MM/yyyy, dd-MMM-yyyy, dd-MM-yyyy, dd.MM.yyyy.

The numbers are handled with the help of java.util.regex.Pattern for pattern matching with regular expressions. A regular expression is a special sequence of characters that helps us match or find other strings or sets of strings, using a specialized syntax held in a pattern. They can be used to search, edit, or manipulate text and data.

C. HANDLING PHRASES

In this module we handle all the phrases of name and organization i.e. if two or more than two consecutive words are initialized with capital letter and form a phrase then we combine all these words with " – " and then consider it as a single token.

E.g : Indian-Railways.

If an input text contains a word "Dr", "Miss" , "Mrs.". , "Ms", "Prof", "Er", "Lect", "Mr" then the word follows it must be a person entity, so we combine one of the word that exists in an input text from all the words above with the next word with the help of "-"and tag it with a person entity and we also provide hindi transliteration of the hypenated word that is given by the transliterate class.

E.g : Dr-Manika

If two more consecutive words are initialized with capital letter and form a phrase then we combine all these words with "-" and consider it as a single token and then we check if it ends with

Ltd","Limited","Corporation","Corp","University".If yes,we tag it with an organization entity and we also provide hindi transliteration of the hypenated word that is given by the transliterate class . E.g : Abmtc-Limited

D. POS TAGGING

Parts of speech tagger is used for identifying and recognizing the entity names. It determines each word in the document as noun, proper noun, verb etc. according to the grammar rules.

E. SPLITTING ON THE BASIS OF "/"

After getting the tagged output we split it on the basis of "/" with the help of already defined Split function in order to separate words from their respective tags so as to get each token separately.After getting each token separately we fetch the previous words of the tags only if its tag is NN or CD or JJ so as to work further on fetched words.

English NER:

In this module we implement NER for English Language i.e. Identifies named entities like name, organization, location etc. After tagging and splitting we separate the nouns and with the help of database we compare each noun of an input string with the entity names present in the database and then finally fetched all the noun entities with their appropriate sub-entity i.e. name, place or organization. Here we created our own database. Database consists of collection of large number of entities of names, places, organization with columns words, entity (sub category) and its Hindi transliteration.

Steps:-

1. First of all we check if word contains tag "NN".
2. If yes then we fetch its previous word and stored it into a database table "filled".
3. Compare the entity stored in table filled with entities stored in another table "ambiguious" or "used".
4. After that we check if corresponding to that word the previous word is "in or from or to" then check the recent word in "ambiguious" table whether it is place or not.If yes then fetch it and if not check it in another table "used" to fetch its respective sub entity.
5. If the previous word is not "in or from or to" then check the recent word in "ambiguious" table whether it is name or not. If yes then fetch it as name entity else go to another table "used" to check whether it is name or place or organization.
6. If the tag of the word is "CD" its a number and hence we compare this word with the pre-defined pattern of the numbers .

F. BILINGUAL DATABASE

Here bilingual means we maintain our database in both languages i.e. For Hindi and English. In our database there are mainly three columns words, their Hindi transliteration and their entities i.e. name, places or organization.Our database contains mainly three tables filled, used, and ambiguous. Ambiguous table consists of those words that act as both name and place entities.Filled table consists of those words that are nouns present in our given input text .Used table consists of almost all entities that are names, place , organization and their respective Hindi transliteration .

© 2014, IJIRAE- All Rights Reserved

G. HINDI NER

In this we directly fetch Hindi named entity corresponding to that of English from the database.

III. EVALUATION

In our project we evaluate our tool on the basis of Precision, recall and F-measure. In this way we come to know how well our NER tool is doing in extracting named entities.

A. PRECISION

Precision is the ratio of the number of items of a certain named entity type correctly identified to all items that were assigned that particular type by the system.

B. RECALL

Recall measures the number of items of a certain named entity type correctly identified, divided by the total number of items of this type. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision.

C. F-MEASURE

The F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision P and the recall R of the test to compute the score. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. It combines Recall (R) and Precision (P) using the formula .The traditional F-measure or balanced F-score (F1 score) is the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

In our project we test our system on 100 sentences and according to that the result is as follows:

D. PIE-CHARTS:

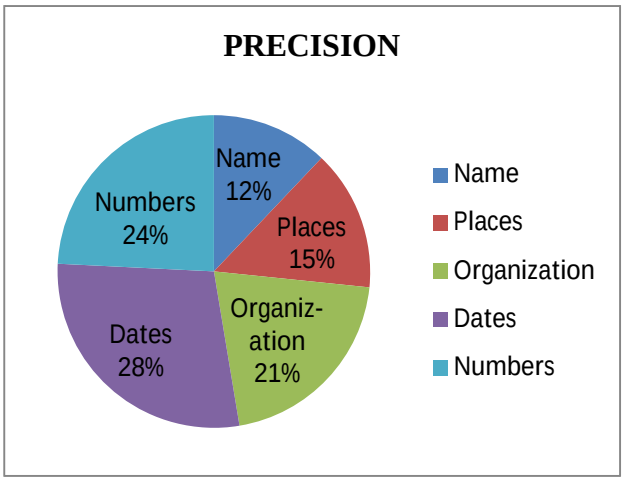

Fig 1

CATEGORIES	PRECISION	RECALL	F-MEASURE
NAME	0.35	0.5	0.41
PLACE	042	0.6	0.49
ORGANIZATION	0.60	0.75	0.66
DATE	0.82	0.86	0.83
NUMBERS	0.70	0.77	0.73

© 2014, IJIRAE- All Rights Reserved

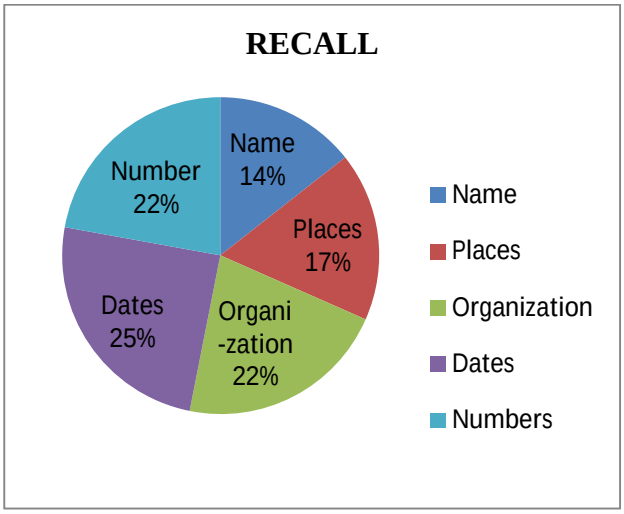

Fig 2

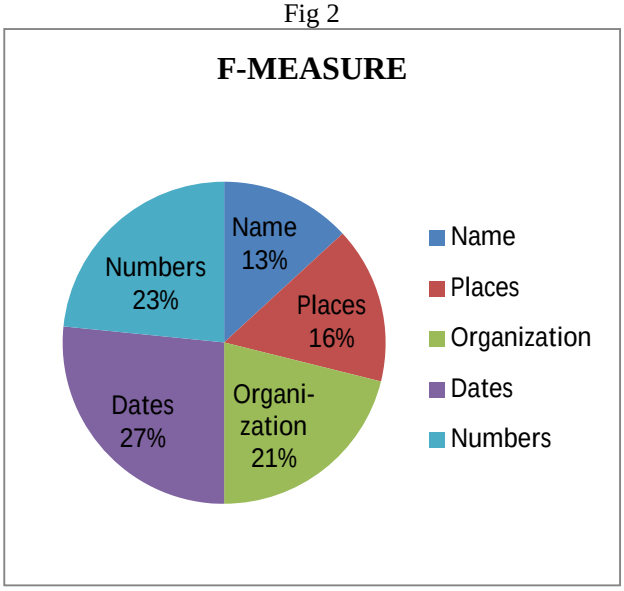

Fig 3

IV. CONCLUSION:

In this paper the bilingual NER has been developed. The NER requires a database and POS tagger as its important elements to generate correct output. In this we worked on different and almost all date patterns as a result of which our system generates correct output for date entities and hence provides us high value of F -measure.In addition to date patterns we also worked on Number entites with the help of some regular expressions and matches method as a result of which our system correctly recognizes number entities present in the input text.The recogniziton of name, place and organization entities is dependent on database , so we conclude here that our system gives us correct result in case of above three entites only if all the names , places and organizations are available in database . This means that the above three entities are 90% dependent on database . In it we also worked on ambiguous words like the words having same name for person and place so that the efficiency of the system will be increased. Our proposed NER system will be able to find efficiently all the named entities in different domains depending on enitites present in database used. Hence, we conclude that our proposed NER system have good performance in finding named entities and achieving high F-measure score with limited size corporate. The quality of NER system depends upon size of database and quality of database.

V. FUTURE WORK

The work can be extended to solve the ambiguity problem which can be solved by training the system with a large training corpus of various kinds of news, so that it contains a variety of combinations of names. The work can also be extended to solve unknown words problem which can be solved by using some lists that contain names, especially foreign names.The work can also be extended to make NER more general The work can also be extended in making more efficient transliteration rules. Since the Hindi POS tagger is not currently available to us ,we can't do so much work in Hindi NER . The work can also be further extended in case of Hindi NER.

© 2014, IJIRAE- All Rights Reserved

REFERENCES

[1] Animesh Nayan, B. Ravi Kiran Rao, Pawandeep Singh, Sudip Sanyal and Ratna Sanyal-IIIT ALLAHABAD-NAME ENTITY RECOGNIZER FOR INDIAN LANGUAGES.
[2] IJCSI International Journal of Computer Science Issues, Vol. 7, November 2010-A survey of Named Entity Recognition in English and other Indian Languages
[3] http://nlp.stanford.edu/software/CRF-NER.shtml
[4] Darvinder kaur, Vishal Gupta, November 2010, A survey of Named Entity Recognition in English and other Indian Languages, Department of Computer Science & Engineering, Panjab university ,Chandigarh.
[5] Mitchell P. Marcus, Mary Ann Marcinkiewicz, Building a Large Annotated Corpus of English: The Penn Treebank, University of Pennsylvania
[6] Andrei Mikheev, Marc Moens and Claire Grover, Named Entity Recognition,Database , HCRC Language Technology Group,University of Edinburgh
[7] http://opennlp.apache.org
[8] Xiaoyi Ma, Toward a Name Entity Aligned Bilingual Corpus, Linguistic Data Consortium, 3600 Market St. Suite 810, Philadelphia
[9] Awaghad Ashish Krishnarao, Himanshu Gahlot, Amit Srinet, D. S. Kushwaha, A Comparative Study of Named Entity Recognition for Hindi Using Sequential Learning Algorithms, Motilal Nehru National Institute of Technology, Allahabad.
[10] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Irina Temnikova, Anna Widiger, Wajdi Zaghouani, Jan Žižka, Multilingual person name recognition and transliteration**,** European Commission Joint Research Centre, Bulgarian Academy of Sciences, University of Brno.
[11] Thierry Poibeau and the INaLCO Named Entity Group**,** The Multilingual Named Entity Recognition Framework, INaLCO/CRIIVI 2 rue de Lille 75007 Paris.
[12] Penny Treebank Tagset is available http://www.ldc.upenn.edu/catalog/docs/LDC95T7/c193.html.
[13] Hai Leong Chieu and Hwee Tou Ng, "Named Entity Recognition: A Maximum Entropy Approach Using Global Information," in 19th international conference on Computational Linguistics (COLING 2002), Taipei, Taiwan, 2002.
[14] Deepti Chopra, Nusrat Jahan, Sudha Morwal , hindi named entity recognition by aggregating rule based heuristics and hidden markov model in International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.6, November 2012
[15] Nusrat Jahan , Sudha Morwal and Deepti Chopra, Named Entity Recognition in Indian Languages Using Gazetteer Method and Hidden Markov Model: A Hybrid Approach , Nusrat Jahan et al./ International Journal of Computer Science & Engineering Technology (IJCSET)
[16] Dimitra Farmakiotou, Vangelis Karkaletsis, John Koutsias, George Sigletos Constantine D. Spyropoulos and Panagiotis Stamatopoulos, "RULE-BASED NAMED ENTITY RECOGNITION FOR GREEKFINANCIAL TEXTS" Institute of Informatics and Telecommunications, N.C.S.R "Demokritos" 15310 Aghia Paraskevi, Athens, Greece Department of Informatics, University of Athens TYPA Buildings, Panepistimioupolis, 157 71 Athens, Greece.
[17] David Nadeau, Satoshi Sekine , A survey of named entity recognition and classification , National Research Council Canada / New York University .
[18] Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat , Named Entity Recognition Approaches , Faculty of Computer Science & Information Technology, University Putra Malaysia, 43400 Serdong, Malaysia.
[19] Diego Moll´a and Menno van Zaanen and Daniel Smith , Named Entity Recognition for Question Answering , Centre for Language Technology Macquarie University Sydney Australia.
[20] GuoDong Zhou Jian Su Named Entity Recognition using an HMM-based Chunk Tagger , Laboratories for Information Technology 21 Heng Mui Keng Terrace Singapore 119613.
[21] Kashif Riaz , Rule-based Named Entity Recognition in Urdu , University of Minnesota Department of Computer Science Minneapolis, MN, USA.
[22] Sungchul Kim, Kristina Toutanova , Hwanjo Yu , Multilingual Named Entity Recognition using Parallel Data and Metadata From Wikipedia , POSTECH Pohang, South Korea Kristina Toutanova Microsoft Research Redmond, WA 98502 , Hwanjo Yu POSTECH Pohang, South Korea.
[23] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Irina Temnikova, Anna Widiger, Wajdi Zaghouani, Jan Žižka , Multilingual person name recognition and transliteration , European Comission Joint Research Centre , Bulgarian Academy of Sciences , Unniversity of Brno.
[24] Yufeng Chen, Chengqing Zong , Keh-Yih Su , On Jointly Recognizing and Aligning Bilingual Named Entities , Institute of Automation, Chinese Academy of Sciences Beijing, China Behavior Design Corporation Hsinchu, Taiwan, R.O.C.
[25] David Burkett Slav Petrov John Blitzer Dan Klein Learning Better Monolingual Models with Unannotated Bilingual Text , University of California, Berkeley Google Research.

© 2014, IJIRAE- All Rights Reserved