ISM@FIRE -2013 Named-Entity Recognition (NER) Indian Languages Task

Dinesh Kumar Prabhakar, Gopashree Panda, and Sukomal Pal

Department of Computer Science and Engineering,
Indian School of Mines, Dhanbad,
Jharkhand, India
{dinesh.nitr, gopashreepanda15, sukomalpal }@gmail.com

Abstract. Named-Entity Recognition (NER) is the process of identifying and tagging the proper nouns in a given text document. These proper nouns can be names of any person, place, organization, disease or artifact etc. In this approach, we have made a list of different types of named-entities and saved in different files. For a given input text file, after doing word-split we made a database match for each word. If there is any match for a specific word then it is tagged as that type of named-entity.

Keywords: NER, POS, NLP

1 Introduction

In Name-Entity Recognition, input is a set of text documents. The NER system needs to process these text files and identify the named-entities. Grammatically, a named entity is a proper noun present in a sentence. Initially the words of a sentence are divided into a number of phrases. These can be noun phrases, adjectives, prepositions, verbs, articles etc. From these, the noun-phrases are picked and for each noun-phrase the proper-noun is identified.

This proper-noun can be a name of a person, a place, an organization, a disease or a natural disaster etc. Each noun is tagged as per its category in 1st level tagging. Further these categories can be divided into smaller sub-categories, such as a place can be a country or a state or a city, an organization can be a business organization or an educational or a political organization etc. Thus a noun is tagged as per its sub-category in the 2nd level tagging.

Next we have discussed the problem description in Section 2, then our detail approach for the solution discussed in Section 3 and we have have summarized in section 4 conclusion.

2 Task Description

Corpus for this task is given in the form of text files in column format. There are six columns in each text file where in the first column word, second column parts-of-speech(POS) tag, third column chunk tag, fourth, fifth and sixth columns are for three levels of named-entity tags.

Input for this task is given in the same column format of text files where each sentence is already divided into phrases. First word of each noun-phrase is identified with B-NP (Beginning-Noun Phrase) level and those inside with I-NP (Inside-Noun Phrase) level. The task is to identify the type of each noun phrase and then tag them with 1st, 2nd and 3rd level NE (Named Entity) tag in 4th, 5th and 6th column respectively.

e.g. Agra NNP B-NP B-LOCATION B-PLACE B-CITY

2 Methodology

Our approach is mainly based on maintaining different files for different categories and sub-categories of named-entities and then making database runs to identify type of each word present in input text.

Input is given as formatted text files where each sentence is broken into noun-phrases, verbs, articles etc. and beginning of each phrase is marked with B- and I- tags.

We have followed the steps given below:

Step 1*:* Find all possible types of named-entities, make a list of words under each category and save them in different files.

[We have found 21 different types of Named-Entity in the training corpus hence we create 21 different file to store entities of each type in respective file. This step we have performed based on 4th column of training corpus documents.]

Step2: Take a *word* input from test document with POS and Chunk tags and compare tall three columns value with first three columns separated entities file entries along with *word*.

Step 3*:* If match found consider it as matched entity type and copy the NE tags of matched entity and place it to an output file.

3 Analysis

We have gone through 80 different training corpus text file and found there are 21 different types of Named-entity like name of person, organization, place etc.. Hence we create 21 different files for there entities. In these files words (entity) will be stored along with POS, Chunk and NE tags.

Our system work work properly if the word is only entity. There are some limitations of our system. As we didn't consider the sub-category of entity while separating the entities entries it may give improper tags in 2nd and 3rd levels of NE tagging. And it may not work properly for some of phrase since we didn't used window in submitted run.

4 Conclusion

This writeup describes the some basics of NER related to given task followed by Methodology used and Analysis of our system. As result has not come so we can't say anything about how worthy is this system. But we have stated limitations of our system. Once result of this task will come will look into the possibilities to improve our system output accuracy. Meanwhile we are trying to reduce the shortcomings of our system.

References

1. David Nadeau, Satoshi Sekine: A survey of named entity recognition and classification National Research Council Canada / New York University, 2007
2. Lev Ratinov and Dan Roth: Design challenges and misconceptions in named entity recognition.
3. The Stanford Natural Language Processing Group http://nlp.stanford.edu/downloads/CRF-NER.shtml