Wikipedia Named Entity Linking

SURBHI GUPTA - 201505533
KISHAN VADALIYA - 201505621
VIVEK VISHNOI - 201505504

INTRODUCTION

Objective of the project is to identify named entities and link them with wikipedia data.

Named Entity-

Named entities are "atomic elements in text" belonging to "predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.". **Named entity recognition (NER)** is the task of identifying such named entities.

Categories of named entities -

Basic categories include the following:

1. Names
   - Organization
   - Person
   - Location
2. Time
   - Date
   - Time
3. Numbers
   - Money
   - Percent

However, the following may also be considered as categories/subcategories:

- Distance
- Speed
- Age
- Weight
- City
- Country
- State/Province
- River

Applications of Named Entity Recognition -

With the enormous quantity of data available via the internet and other electronic sources it is no longer feasible for human beings to process this data to identify useful information. Computers are now necessary for finding this information, and Named Entity recognition plays a significant part.

Not only are named entities useful for identifying where a piece of information may be located (i.e. a description of a person may be found in text near his or her name), but a named entity itself may be the answer to a particular question. For this reason named entity recognition is particularly useful in Question Answering. Other applications are in domains molecular biology, bioinformatics, social media.

NAMED ENTITY RECOGNITION

Steps involved in named entity recognition -

DATASET -

Conll 2003 dataset is used for training and testing of model.
http://www.cnts.ua.ac.be/conll2003/ner/

Dataset concentrate on four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups.
The CoNLL-2003 shared task data files contain four columns separated by a single space. Each word has been put on a separate line and there is an empty line after each sentence. The first item on each line is a word, the second a part-of-speech (POS) tag, the third a syntactic chunk tag and the fourth the named entity tag. The chunk tags and the named entity tags have the format I-TYPE which means that the word is inside a phrase of type TYPE. Only if two phrases of the same type immediately follow each other, the first word of the second phrase will have tag B-TYPE to show that it starts a new phrase. A word with tag O is not part of a phrase.
Here is an example:

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC

TRAINING OF MODEL-

We need to form strings of named entities from training data. For that we will use IOB tags concept used in training data.

I - inside of chunk
B - Beginning of chunk
O - Outside chunk

Form a string of those words which are at beginning and inside of chunk. Store strings and their given named entities in a dictionary along with their count of occurrences. Also store counts of chunk ids and their corresponding named entities. These counts will be used in the **Classifier model.**

For viterbi algorithm of POS tagging, store counts of tag-tag and word-tag which will help in calculating transition probabilities and emission probabilities.

TESTING OF MODEL-

Read all the words from testing data and concatenate them to form a text data. Tokenize data and apply part-of-speech tagging using Viterbi algorithm. Chunking of data is done to obtain Noun phrases. We calculate the probability of these noun phrases of being a named entity. For this, we have stored naive bayes probabilities during training of model.

P1 = Probability(I-LOC | Europe) = Probability(Europe | I-LOC) * Probability (LOC)
P2 = Probability(I-ORG | Europe) = Probability(Europe | I-ORG) * Probability (ORG)
P3 = Probability(I-PER | Europe) = Probability(Europe | I-PER) * Probability (PER)

Final named entity of “Europe” = max(P1, P2, P3)

LINKING THROUGH WIKIPEDIA-

Dbpedia data is used for linking with wikipedia and SPARQLWrapper is used for querying this data.

RESULTS - ACCURACY - 71%

Output of the code -

```
NAMED ENTITY  -
King_(chess)

WIKIPEDIA LINKING -
http://dbpedia.org/resource/King (chess)
http://dbpedia.org/resource/ملك (شطرنج)
http://dbpedia.org/resource/König (Schach)
http://dbpedia.org/resource/Rey (ajedrez)
http://dbpedia.org/resource/Roi (échecs)
http://dbpedia.org/resource/Re (scacchi)
http://dbpedia.org/resource/キング (チェス)
http://dbpedia.org/resource/Koning (schaken)
http://dbpedia.org/resource/Król (szachy)
http://dbpedia.org/resource/Rei (xadrez)
http://dbpedia.org/resource/Король (шахматы)
http://dbpedia.org/resource/王 (國際象棋)
------------------------------------------------------------------------------------------------
```

NAMED ENTITY -
West

WIKIPEDIA LINKING -
West
غرب
Westen
Oeste
Ouest
Ovest
西
West
Zachód
Oeste
Запад

NAMED ENTITY -
Port

WIKIPEDIA LINKING -
Hafen
Port
Port
港湾
ميناء
Puerto
Haven
Port wodny
Porto (transporte)
Порт
港口

NAMED ENTITY -
Christmas

WIKIPEDIA LINKING -
Christmas
عيد الميلاد
Weihnachten
Navidad
Noël
クリスマス
Natale
Kerstmis
Boże Narodzenie
Natal

Рождество Христово
圣诞节

NAMED ENTITY -
East

WIKIPEDIA LINKING -
East
شرق
Osten
Este
Est
Est
東
Oost (windstreek)
Wschód
Leste
Восток
東

NAMED ENTITY -
Squad

WIKIPEDIA LINKING -
http://dbpedia.org/resource/Squad
http://dbpedia.org/resource/حظيرة (وحدة عسكرية)
http://dbpedia.org/resource/Gruppe (Militär)
http://dbpedia.org/resource/Squadra (unità militare)
http://dbpedia.org/resource/分隊
http://dbpedia.org/resource/Groupe de combat
http://dbpedia.org/resource/Drużyna (wojsko)
http://dbpedia.org/resource/Geweergroep
http://dbpedia.org/resource/Отделение
http://dbpedia.org/resource/Grupo de Combate
http://dbpedia.org/resource/班
