Named Entity Recognition
and
Relation Extraction

Bachelor's Thesis
submitted in partial fullfilment of
the requirements for the degree of
Bachelor of Technology
in
Computer Science and Engineering

By

NARAGAM SAI KIRAN
N100638

Under the Guidance of
Mr. Ambati Udaya Kumar
Assistant Professor
Dept. of Computer Science and Engineering
RGUKT Nuzvid

Department of Computer Science and Engineering
Rajiv Gandhi University of Knowledge Technologies, Nuzvid.
Krishna District,Andhra Pradesh - 521202.

April, 2016

Rajiv Gandhi University of Knowledge Technologies
(A.P. Government Act 18 of 2008)
RGUKT NUZVID, Krishna District - 521202

CERTIFICATE OF COMPLETION

This is to certify that the work entitled, **Entity Recognition and Relation Extraction** is the bona fied work of **NARAGAM SAI KIRAN**, ID No: **N100638**, carried out under my guidance and supervision, for the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering.

Mr. Krishna Kumar Singh
(Head of the Department)
Assistant Professor,
Dept.of CSE.

Mr. Ambati Udaya Kumar
(Project Supervisor)
Assistant Professor,
Dept.of CSE.

Rajiv Gandhi University of Knowledge Technologies
(A.P. Government Act 18 of 2008)
RGUKT NUZVID, Krishna District - 521202

CERTIFICATE OF EXAMINATION

This is to certify that the work entitled, **”Named Entity Recognition and Relation Extraction”** is the bonafide work of **NARAGAM SAI KIRAN**, ID No: **N100638** and here by accord our approval of it as a study carried out and presented in a manner required for its acceptance in the partial fulfillment of the requirement for the award of the degree of Bachelor of Technology for which it has been submitted.

This approval does not necessarily endorse or accept every statement made, opinion expressed or conclusion drawn, as a recorded in this thesis. It only signifies the acceptance of this thesis for the purpose for which it has been submitted.

Mr. Amit Patel
(Project Examiner)
Lecturer,
Dept.of CSE.

Mr. Ambati Udaya Kumar
(Project Supervisor)
Assistant Professor,
Dept.of CSE.

Rajiv Gandhi University of Knowledge Technologies
(A.P. Government Act 18 of 2008)
RGUKT NUZVID, Krishna District - 521202

DECLARATION

I **NARAGAM SAI KIRAN**, with ID No:**N100638** hereby declare that the project report entitle **Named Entity Recognition and Relation Extraction** done by me under the guidance of **Mr. Ambati Udaya Kumar,M.Tech** is submitted for the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering during the academic session August 2015 April 2016 at RGUKT Nuzvid.

I also declare that this project is a result of my own effort and has not been copied or imitated from any source. Citations from any websites are mentioned in the references.

The results embodied in this project report have not been submitted to any other university or institute for the award of any degree or diploma.

N. SAI KIRAN,
N100638.

Date : ______________
Palce: ______________

Acknowledgements

I would like to thank **Mr. Ambati Udaya Kumar** for his exemplary guidance, monitoring and constant encouragement throughout the course of this thesis. I express my sincere thanks to him for making the resources available at right time and providing valuable insights leading to the successful completion of my project. I really enjoyed his `'visionay'` discussions

I would like to thank RGUKT Nuzvid Director, faculty and staff for their valuable suggestions and discussions.

I place a deep sense of gratitude to my family members and my friends who have been constant source of information during the preparation of this project work.

N. SAI KIRAN
N100638

Abstract

Information Extraction is the task of extracting structured data out of unstructured data like natural language text as the structured data can be easily processed by a computing machine. So when we have unstructured data, we extract relevant information in structured form like tables and then use some queries to get required information.

To extract information, named entities are recognized and relation between considered named entities are extracted. Here we recognized four classes of named entities (PERSON, ORGANIZATION, Global Position Entity(GPE), LOCATION) and extracted relations using hand-written rules (with regular expressions). We used BBC news dataset[1] and got good results. All the source code and documentation are hosted on github[2]

[1] Click on `>> Download raw text files` of `Dataset: BBC` of `http://mlg.ucd.ie/datasets/bbc.html` Or `http://mlg.ucd.ie/files/datasets/bbc-fulltext.zip`

[2] `https://github.com/saikiran638/MyProjects/tree/master/FinalYearProject`

Contents

Chapter 1

Introduction

When we have large amount of (previous)data we might want to extract some useful information out of it, and use it as summary or we can predict the future events by learning from the data at hand.

Most of the time data that is available for use in un-structured form like Natural Language Text rather than structured form like tables. It is easy to extract required information or answer a question if the data we are working on has structured form.But it is difficult to handle unstructured data. Because Natural Language Processing(NLP) that works on unstructured data is still developing.

The amount of natural language text that is available in electronic form is truly staggering, and is increasing every day. However, the complexity of natural language can make it very difficult to access the information in that text[1].

If we instead focus our efforts on a limited set of questions or "entity relations," such as "where are different facilities located, " or "who is employed by what company," we can make significant progress.[1]

Here we are trying to understand the given text and find the limited relevant parts of it. This is what the researchers called as **Information Extraction**. There are two subtasks in infomation extraction those are `Named Entity Recognition` and `Relation Extraction`. We work on both of them now.

1.1 Aim

Our aim is to identify named entities and working out the relationship between them using handwritten rules with regular expressions. We may use this system for question & answering. For most of the questions often the answers be named entities. The below image 1.1 will give an intuition of what we are going to do. That is the procedure followed.

For relevant, meaningful relation detection we use some regular expressions on that tuples. To host this project on `github`[1] for ease of access and modification.

To use `BBC` news data for testing purpose .

[1] This project with its source code and documentation available at: `https://github.com/saikiran638/MyProjects/tree/master/FinalYearProject`

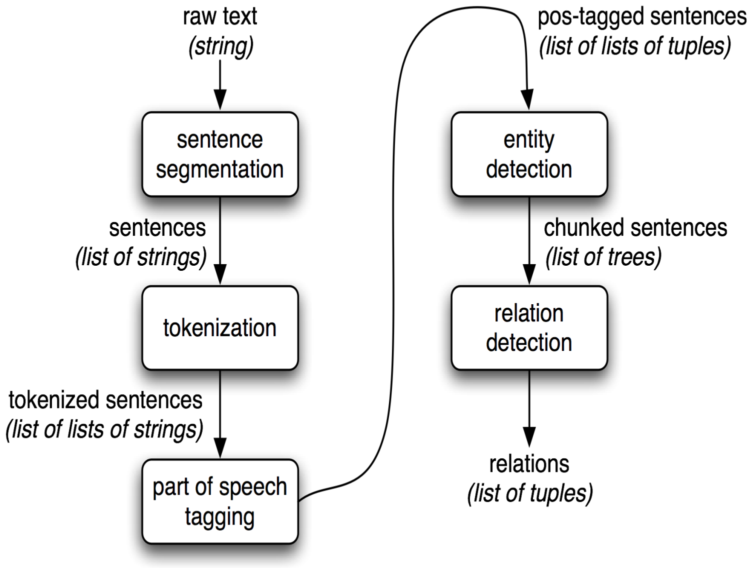


Figure 1.1: Simple Pipeline Architecture for an Information Extraction System[1]

1.2 Technologies Used

Language of choice is Python as it eases development with high-level data structures and modules built-in. Most important modules we used are :

- `nltk` (Natural Language Tool Kit module)

  We used Python's Natural Language Tool Kit for implementation. It has good documentation and tutorials[2]. It allows convenient access[3] of corpus in different languages, and has many natural language processing methods implemented for better performance. We can use them for better results.

- `re` (Regural Expression module)

  Provides convenient methods to write and test regular expressions. We used it to write rules while extracting relations.

[2] http://www.nltk.org/book

[3] http://www.nltk.org/howto

Chapter 2

Named Entity Recognition

2.1 Introduction

Named Entity Recognition is an important sub task of Information Extraction, in this we are going to find and classify (into different classes like PERSON,ORGANIZATION and LOCATION etc.) the concrete names.

We are interested in *Named* Entity Recognition. Because not all entities are attached with a name (specific). For the literature survey on named entity recognition, please refer[9].

2.2 Named Entity Recognition as Tagging

Bikel et. al[1] mapped the Named the Entity Recognition problem very directly into tagging problem. Tagging problem is to determine a tag to a particular word in the given text. Tagging problem requires a set of tags and considerable amount of tagged corpus with the same set of tags. Bikel et. al considered nearly seven name classes and a NOT-A-NAME tag as set of tags.

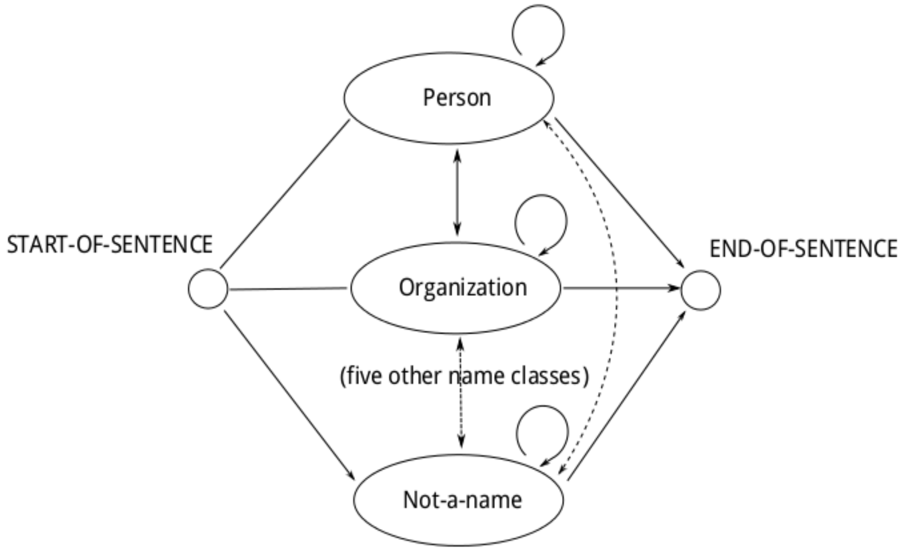
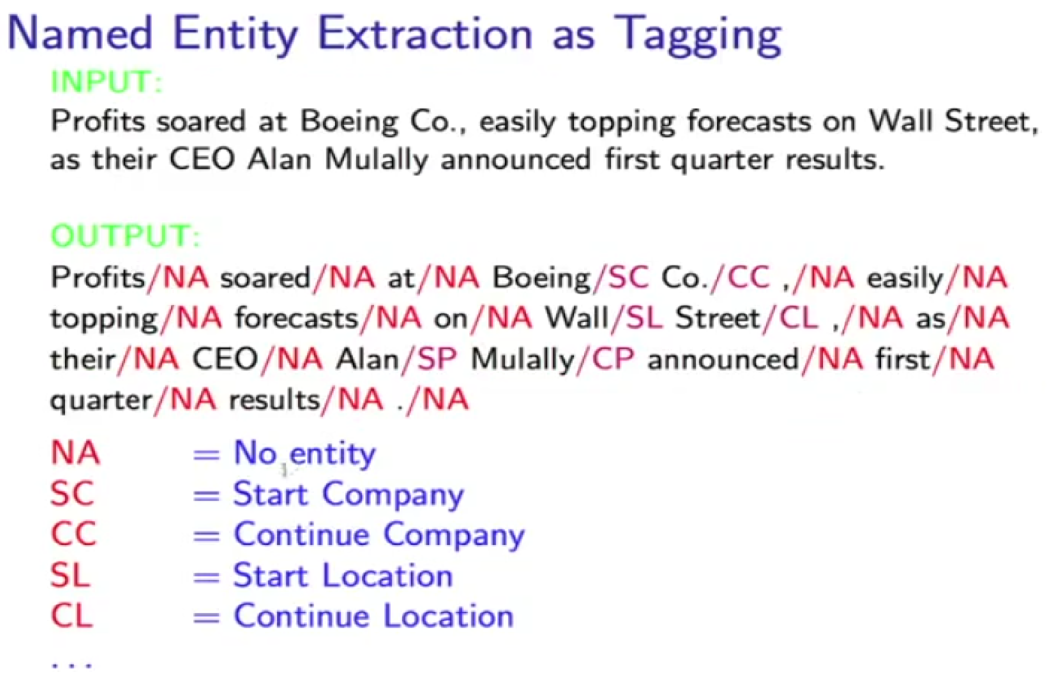


Figure 2.1: Stage Diagram of NER as Tagging by Bikel et. al

[1] http://ilk.uvt.nl/~toine/research/bikel-1999.pdf



Named Entity Extraction as Tagging

INPUT:

Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT:

Profits/NA soared/NA at/NA Boeing/SC Co./CC ,/NA easily/NA topping/NA forecasts/NA on/NA Wall/SL Street/CL ,/NA as/NA their/NA CEO/NA Alan/SP Mulally/CP announced/NA first/NA quarter/NA results/NA ./NA

NA	= No entity
SC	= Start Company
CC	= Continue Company
SL	= Start Location
CL	= Continue Location
...	

Figure 2.2: NER as tagging

They have come up with a Hidden Markov Model[6] based tagger(which is an example of generative model of learning[2]) to tag the given text with named entity tags. They used hand-tagged corpus to train their model(Hidden Markov Model) and considered some word-features to deal with low-frequency words. Figure 2.1 gives an idea of their work, it is the state diagram of the model they proposed. The model assigns a tag(state) to current word, and next word has different probabilities to be assigned with different tags. The next word will be given a tag that has maximum probability to be assigned. These probabilities will be the parameters of the model which are learned[4] from the tagged corpus .

Figure 2.2 gives a better intuition of named entity recognition as tagging. According to this example multi-word names can be easily identified and grouped. No entity(NA) tag is equivalent to NOT-A-NAME tag.

The major problem with named entity recognition as tagging is that we need huge amount of hand-labeled corpus, with named entity classes as labels. This corpus can't be used for another purpose. But the another approach to NER, which consists of part-of-speech tagging and chunking requires part-of-speech tagged corpus which can used for many other purposes like machine translation.

There are many advanced ways to solve the tagging problem some of them are Maximum Entropy Markov Models, Perceptron taggers etc.
To know about these other methods, please refer [3]

[2] For clear understanding please refer Abstract and Introduction of [10]

[3] For more details: Tagging Problems, and Hidden Markov Models of [7] and POS Tagging of [8]

2.3 Named Entity Recognition with PoS tagging & Chunking

Now we have lot of part of speech tagged corpora (especially for english) as we can use it for machine translataion and many other applications. Here we are going to use PoS tagging for NER.

After having natural language senetences with their underlying *tag sequences* we group the tags into named entities.

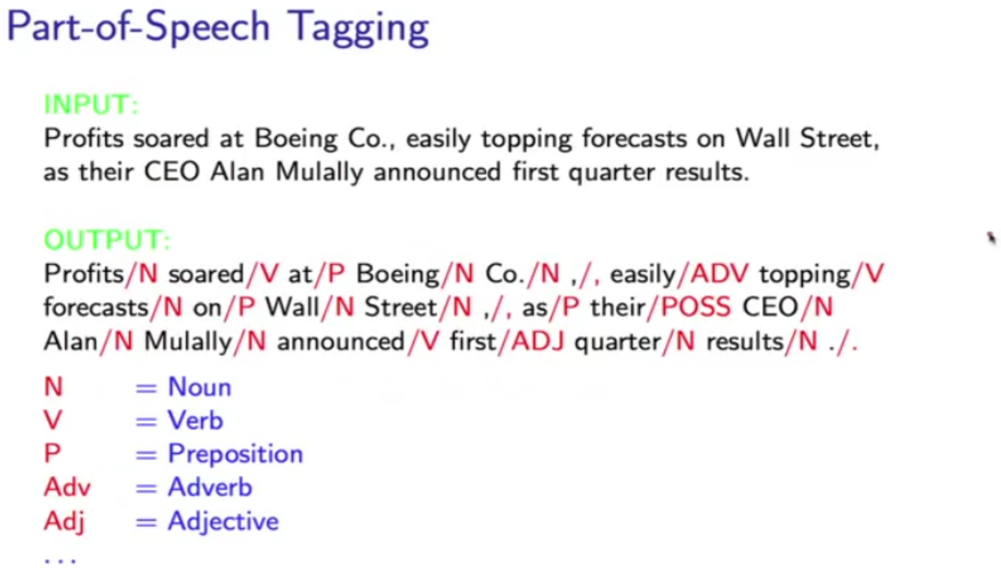
2.3.1 PoS Tagging

Part of speech tagging[4] problem is to determine the parts of speech of a particular instance of word. The intuition of PoS tagging is presented in below image 2.3[5].

Tags may vary depending on corpus we are dealing with. For example, tag set of The Brown Corpus[6] and P.O.S tag set of The Penn Treebank[7]. To check in NLTK, execute and *nltk.help.brown_tagset()*,*nltk.help.upenn_tagset()* respectively for the brown corpus tagset and Penn Treebank tagset.

If we want to write a tagger then we need *large amount of labled corpus*[8]. Which in this case are Penn Tree Bank tagged corpus or Brown corpus. For more insights on writing a parts of speech tagger in NLTK, please refer[2]

For theoritical understanding of a tagger. We learned how a Hidden Markov Model tagger[6] works. Here for this problem we used NLTK's default implementation of tagger(*nltk.tag()*) as it is recommeneded for better results.



Part-of-Speech Tagging

INPUT:
Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT:
Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ,/, as/P their/POSS CEO/N Alan/N Mulally/N announced/V first/ADJ quarter/N results/N ./.

N	= Noun
V	= Verb
P	= Preposition
Adv	= Adverb
Adj	= Adjective
. . .	

Figure 2.3: Part-of-Speech Tagging

[4] For more details: Tagging Problems, and Hidden Markov Models of [7] and POS Tagging of [8]

[5] Slide from The Tagging Problem of [7]

[6] https://www.comp.leeds.ac.uk/ccalas/tagsets/brown.html

[7] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

[8] For more insights of Machine Learning techniques I feel, [4] is a good source

2.3.2 Named Entity Chunking

After tagging comes the named entity chunking[9]. We'll group the pos tags into named entities (if possible intuit its class). Chunker is also a tagger that is trained on some corpus. One of the most useful sources of information for NP-chunking(Noun Phrase-chunking) is part-of-speech tags. This is one of the motivations for performing part-of-speech tagging in our information extraction system

Here for this problem we used NLTK's default implementation of named entity chunker (*nltk.ne_chunk()*) as it is recommended for better results. It can be used for multi- class(PERSON, LOCATION, ORGANIZATION and GPE) or binary class(NP).

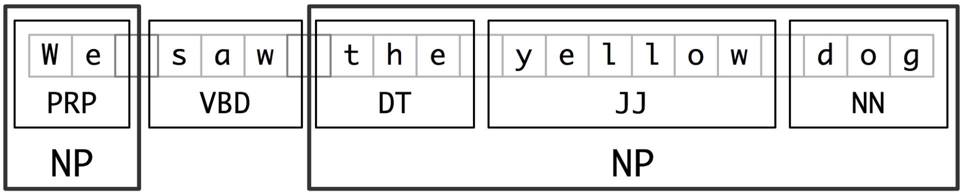


Figure 2.4: Chunking

For more insights on develping a chunker, please refer [1]. Here it described how to create a basic chunker and a chunker that can learn from data for good performance.

Morphology of words to identify Noun Phrases. And to identify the class of Noun Phrase(Named Entity) the chunker will use context of the Noun Phrase. There are many formats to represent Named Entities, those are IB(Every token is **I**n the chunk or **B**egining of the chunk), IOB(Every token is **I**n the chunk or **O**ut of the chunk or **B**egining of the Chunk),tree representation etc.

[9] For more details: refer [1]

Chapter 3

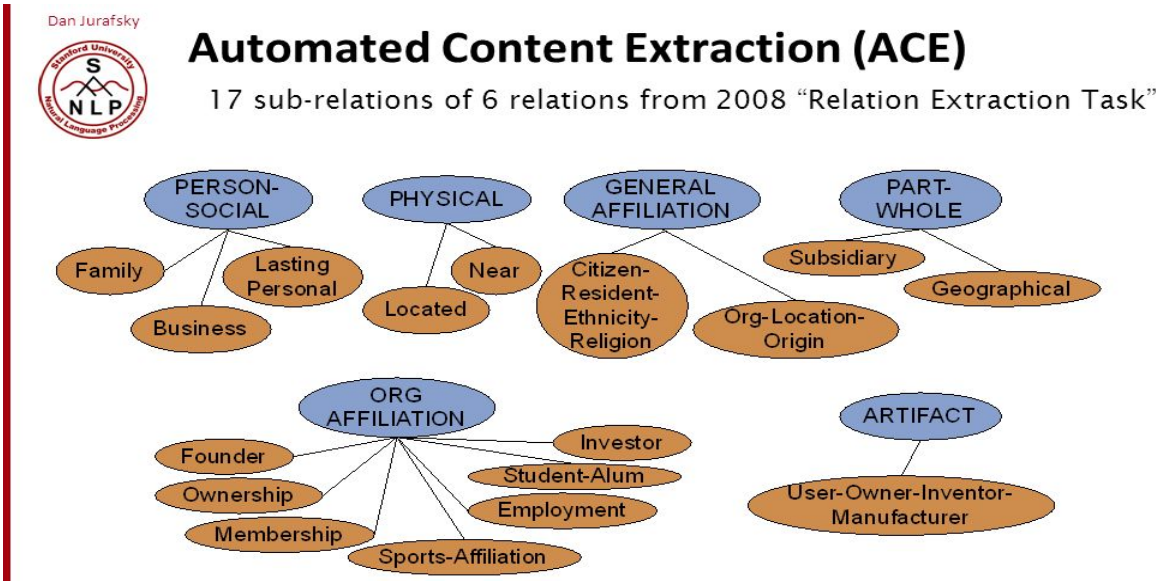
Relation Extraction

Relation Extraction is an important component of Information Extraction. Using the Named Entities and clever patterns we extract relation. These rules can get high precision as they are specific.

We will focus on the simpler task of extracting **relation triples**. Relation triples are of the form `(Named Entity,Relation,Named Entity)`. We use patterns to find whether `Relation` between those `Named Entities` is meaningful and relavent. Procedure is clearly explaned in 3.1.1

We will use `relextract` module of `Python's NLTK` for this. This is rule based relation extraction because we are using hand-written rules. We can create new structured knowledge bases by relation extraction. Questions that are asked in natural language can be converted in to a query to a structured knowledge base.

So, Here is a question for us. *Which relations should we extract ?* It depends on how many **classes of entities** we are able to extract in *Named Entity Recognition.* And here we extracted four classes of names. A set of relations comes from the *Automated Content Extraction (ACE)* task.



Dan Jurafsky

Automated Content Extraction (ACE)

17 sub-relations of 6 relations from 2008 "Relation Extraction Task"

Figure 3.1: ACE Relation set

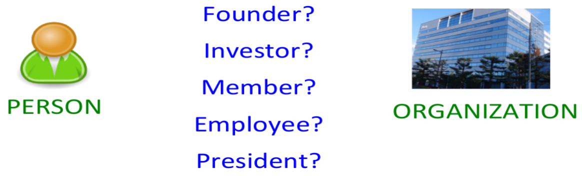


Figure 3.2: Relation between PERSON and ORGANIZATION named entity classes

Figure 3.1[1] shows ACE relations. Figure 3.2[2] gives basic intuition of relation between entities.

3.1 How to extract relations ?

We can use:

- Hand-written patterns
- Supervised,semi-supervised and unsupervised machine learning

We are using *only* Hand-written patterns to extract relations as it is simplest way. Here we are trying to `extract relations`[3] between *specific entities*.

We used **regular expressions** [5] in Python to write patterns to extract relations.

3.1.1 Procedure followed

- First identify the named entities (POS tagging then Chunking).
- Then we group a Noun phrase with its left context.
  Now we'll have document as list of tuples.
  Ex: `[(String1,Named Entity1), (String2,Named Entity2) , (String3,Named Entity3) , ... ]`
  Here `Named Entity1,Named Entity2,Named Entity3` are tree representations of Noun Phrase.
  Now we take two consecutive tuples and add them to create *semi*relation dictionaries.
  That dictionary contains **key**,**value** as described in table 3.1
- Apply hand-written rules on filler,right context and left context to extract relations

For more information on relation extraction, please refer [1] for more information. The rules we wrote for this project can be found here[4]

3.1.2 Positives of Hand-written rules

- String patterns tends to be high-precision.
- Works well for specific domains/entities.

[1] Slide from: https://class.coursera.org/nlp/lecture/138

[2] https://class.coursera.org/nlp/lecture/139

[3] For more information on relation extraction, please refer Week 4 - Relation Extraction of [8]

[4] https://github.com/saikiran638/MyProjects/blob/master/FinalYearProject/RelationRules.py

KEY	**VALUE**
filler(text between two Named Entities)	String2 which is leftcontext of Named Entity2
lcon(left context of the relation)	String1 which is leftcontext of Named Entity1.
objclass(Class of object of the relation)	Root of the Named Entity2 tree structure
objsym(Normalized text of object with no white space	Normalized object with underscore in palce of space.
objtext	objtext
rcon(right context of the relation)	String3, which is right context of the Named Entity 2
subjclass(Class of subject of the relation)	Root of the Named Entity1 tree structure.
subjsym(Normalized text of suject with no white space)	Normalized subject with underscore in palce of space.
subjtext	subject text
untagged_filler	filler with no POS tags

Table 3.1: Explanation of Key,Value pairs of semirelation dictionary

3.1.3 Negatives of Handiwritten rules

- It's difficult to think of all possible patterns.
- We don't want to fix the entities for relation extraction.

Chapter 4

Observations and Results

4.1 Observations

Named Entity Chunker provided in Python's NLTK (*nltk.ne_chunk*) considers morphology of words while chunking. Make sure that words are not normalized(HMMTagger requires words to be normalized) while chunking.

Trained HMMTagger(for PoS tagging) available in Python's NLTK with *treebank* tagged corpus and tried chunking but performance is not as good as NLTK's Recommended PoS tagger (*nltk.tag*). So, we used NLTK's Recommended PoS tagger for tagging sentences. It was found that NLTK's current recommended tagger is 'Averaged Perceptron Tagger'(It might change over time).

4.2 Results

This project is hosted on `github`[1] , you can access source code and documents of it.

We used `BBC's news data sets`[2]. In this dataset news is classified under business, entertainment, politics, sport and tech. Each category contains many individual files. We merged all the files into one file (We merged all individual files under category business into one file `testdatabusiness.txt` and politics into `testdatapolitics.txt`, available on github page).

The relation extraction results for BBC **politics** news(`testdatapolitics.txt`):

```
====== Relations of PERSON and ORGANIZATION ====
[PERSON: u’Carl/NNP Emmerson/NNP’] , from the [ORGANIZATION: u’Institute/NNP’]
[PERSON: u’David/NNP Redvers/NNP’] , 34 , from [ORGANIZATION: u’Hartpury/NNP’]
[PERSON: u’John/NNP Bourn/NNP’] , head of the [ORGANIZATION: u’NAO/NNP’]
[PERSON: u’Andrew/NNP Hogg/NNP’] , spokesman for the [ORGANIZATION: u’Medical/NNP Foundation/NNP’]
[PERSON: u’Veritas/NNP’] ’ deputy leader . [ORGANIZATION: u’UKIP/NNP’]
[PERSON: u’Tony/NNP Beddow/NNP’] , from the [ORGANIZATION: u’Welsh/NNP Institute/NNP’]
[PERSON: u’Ieuan/NNP Wyn/NNP Jones/NNP’] , leader of the [ORGANIZATION: u’Plaid/NNP Cymru/NNP’]
[PERSON: u’Simon/NNP Sweetman/NNP’] , from the [ORGANIZATION: u’Federation/NNP’]
[PERSON: u’Hutu/NNP’] leader . The five-year [ORGANIZATION: u’Department/NNP’]
[PERSON: u’Kayitesi/NNP Blewitt/NNP’] , founder of the [ORGANIZATION: u’Survivors/NNPS Fund/NNP’]
[PERSON: u’Galloway/NNP’] was expelled from the [ORGANIZATION: u’Labour/NNP’]
[PERSON: u’Massoud/NNP Shadjareh/NNP’] , from the [ORGANIZATION: u’Muslim/NNP Safety/NNP Forum/NNP’]
[PERSON: u’Mike/NNP’] Hobday , from the [ORGANIZATION: u’League/NNP Against/NNP Cruel/NNP Sports/NNP’]
[PERSON: u’Neill/NNP’] , editor of union-backed [ORGANIZATION: u’Hazards/NNP’]
[PERSON: u’David/NNP Rose/NNP’] , Chief Executive of [ORGANIZATION: u’Hereford/NNP Hospitals/NNP’]
[PERSON: u’Bob/NNP Neill/NNP’] , leader of the [ORGANIZATION: u’London/NNP Assembly/NNP Conservatives/NNPS’]
[PERSON: u’Winston/NNP Churchill/NNP’] told us - from the [ORGANIZATION: u’Baltic/NNP’]
```

[1] https://github.com/saikiran638/MyProjects/tree/master/FinalYearProject

[2] Click on `>> Download raw text files` of `Dataset: BBC` of http://mlg.ucd.ie/datasets/bbc.html Or http://mlg.ucd.ie/files/datasets/bbc-fulltext.zip

```
[PERSON: u’Veritas/NNP’] ’ deputy leader . [ORGANIZATION: u’UKIP/NNP’]
[PERSON: u’Veritas/NNP’] ’ deputy leader . [ORGANIZATION: u’UKIP/NNP’]
[PERSON: u’Graham/NNP Lane/NNP’] , leader of the [ORGANIZATION: u’Labour/NNP’]
[PERSON: u’Carl/NNP Emmerson/NNP’] , from the [ORGANIZATION: u’Institute/NNP’]
[PERSON: u’Maeve/NNP Sherlock/NNP’] , chief executive of the [ORGANIZATION: u’Refugee/NNP Council/NNP’]
[PERSON: u’Adams/NNP’] , from the [ORGANIZATION: u’UK/NNP’]
[PERSON: u’Anne/NNP Weyman/NNP’] , chief executive of the [ORGANIZATION: u’Family/NNP’]
====== Relations of PERSON and PERSON =======
[PERSON: u’Blunkett/NNP’] ’s ex-lover ’s nanny . [PERSON: u’Sir/NNP’]
[PERSON: u’Pound/NNP’] said his wife [PERSON: u’Maggie/NNP’]
[PERSON: u’Sandra/NNP’] , daughter [PERSON: u’Larissa/NNP’]
====== Relations of PERSON and LOCATION =======
[PERSON: u’Neil/NNP Coppendale/NNP’] , from [LOCATION: u’West/NNP Sussex/NNP’]
[PERSON: u’Welsh/NNP’] , was born in [GPE: u’Melbourne/NNP’]
[PERSON: u’Andrew/NNP Elliot/NNP’] , 42 , from [GPE: u’Bromesberrow/NNP’]
[PERSON: u’Richard/NNP Wakeham/NNP’] , 34 , from [GPE: u’York/NNP’]
[PERSON: u’Budget/NNP Chancellor/NNP Gordon/NNP Brown/NNP’] will deliver his [GPE: u’Budget/NNP’]
[PERSON: u’Michael/NNP Ferguson/NNP’] to be released unescorted from [GPE: u’Carstairs/NNP’]
[PERSON: u’Nick/NNP Griffin/NNP’] - who lives near [GPE: u’Welshpool/NNP’]
[PERSON: u’Terry/NNP Griffiths/NNP’] , like Mr Howard from [GPE: u’Llanelli/NNP’]
[PERSON: u’Feroz/NNP Abbasi/NNP’] , from [GPE: u’London/NNP’]
[PERSON: u’Feroz/NNP Abbasi/NNP’] , from [GPE: u’London/NNP’]
[PERSON: u’Tony/NNP Blair/NNP’] seems to have disappeared from [GPE: u’Labour/NNP’]
[PERSON: u’Labour/NNP’] on issues from [GPE: u’Iraq/NNP’]
[PERSON: u’Budget/NNP Chancellor/NNP Gordon/NNP Brown/NNP’] will deliver his [GPE: u’Budget/NNP’]
[PERSON: u’Brown/NNP’] was born in [GPE: u’Glasgow/NNP’]
====== Relations related to DISTANCE =======
```

The relation extraction results for BBC **business** news (`testdatabusiness.txt`):

```
====== Relations of PERSON and ORGANIZATION ====
[PERSON: u’Yukos/NNP’] ’ owner [ORGANIZATION: u’Menatep/NNP Group/NNP’]
[PERSON: u’Paul/NNP Sheard/NNP’] , economist at [ORGANIZATION: u’Lehman/NNP Brothers/NNPS’]
[PERSON: u’Rick/NNP Egelton/NNP’] , deputy chief economist at [ORGANIZATION: u’BMO/NNP’]
[PERSON: u’Sri/NNP Mulyani/NNP Indrawati/NNP’] , State Minister for [ORGANIZATION: u’National/NNP Development/NNP’]
[PERSON: u’David/NNP Naude/NNP’] , economist at [ORGANIZATION: u’Deutsche/NNP Bank/NNP’]
[PERSON: u’Hannes/NNP Wittig/NNP’] , telecoms analyst at [ORGANIZATION: u’Dresdner/NNP Kleinwort/NNP Wasserstein/NNP’]
[PERSON: u’Ed/NNP Silliere/NNP’] , analyst at [ORGANIZATION: u’Energy/NNP Merchant/NNP’]
[PERSON: u’Takashi/NNP Yamanaka/NNP’] , an economist with [ORGANIZATION: u’UFJ/NNP Bank/NNP’]
[PERSON: u’Norbert/NNP Reithofer/NNP’] , a member of the [ORGANIZATION: u’BMW/NNP’]
[PERSON: u’Brad/NNP Wernle/NNP’] , from [ORGANIZATION: u’Automotive/JJ News/NNP Europe/NNP’]
[PERSON: u’Simon/NNP Wheatley/NNP’] , from [ORGANIZATION: u’Goldman/NNP Sachs/NNP’]
[PERSON: u’Bill/NNP Armstrong/NNP’] , a retail analyst at [ORGANIZATION: u’CL/NNP’]
[PERSON: u’Patrick/NNP Juchemich/NNP’] , auto analyst at [ORGANIZATION: u’Sal/NNP Oppenheim/NNP Bank/NNP’]
[PERSON: u’Stuart/NNP Quint/NNP’] , an analyst at [ORGANIZATION: u’Gartmore/NNP’]
[PERSON: u’James/NNP Carrick/NNP’] , an economist with [ORGANIZATION: u’ABN/NNP Amro/NNP’]
[PERSON: u’Michael/NNP Blythe/NNP’] , chief economist at the [ORGANIZATION: u’Commonwealth/NNP Bank/NNP’]
[PERSON: u’Hiromichi/NNP Shirakawa/NNP’] , chief economist at [ORGANIZATION: u’UBS/NNP Securities/NNPS’]
[PERSON: u’Arjan/NNP Sweere/NNP’] , an analyst at [ORGANIZATION: u’Petercam/NNP’]
[PERSON: u’Heronry/NNP Nozaki/NNP’] , an analyst at [ORGANIZATION: u’NikkoCitigroup/NNP’]
[PERSON: u’Michael/NNP Rabb/NNP’] , an analyst with [ORGANIZATION: u’Bank/NNP Sal/NNP Oppenheim/NNP’]
[PERSON: u’Avery/NNP Shenfeld/NNP’] , senior economist at [ORGANIZATION: u’CIBC/NNP World/NNP Markets/NNPS’]
[PERSON: u’James/NNP Tambone/NNP’] , who it says headed [ORGANIZATION: u’CFD/NNP’]
[PERSON: u’Arne/NNP Kristiansen/NNP’] , a spokesman for the [ORGANIZATION: u’Danish/NNP Dairy/NNP Board/NNP’]
[PERSON: u’Richard/NNP Jeffrey/NNP’] , chief economist at [ORGANIZATION: u’Bridgewell/NNP Securities/NNPS’]
[PERSON: u’Libya/NNP’] ’s oil minister , told [ORGANIZATION: u’Reuters/NNP’]
[PERSON: u’John/NNP Palmer/NNP’] , political director at the [ORGANIZATION: u’European/JJ Policy/NNP Centre/NNP’]
[PERSON: u’Libya/NNP’] ’s oil minister , told [ORGANIZATION: u’Reuters/NNP’]
[PERSON: u’Jonathan/NNP Loynes/NNP’] , chief UK economist at [ORGANIZATION: u’Capital/NNP Economics/NNP’]
[PERSON: u’Lebedev/NNP’] headed the [ORGANIZATION: u’Menatep/NNP’]
[PERSON: u’Paul/NNP Cherney/NNP’] , chief market analyst at [ORGANIZATION: u’Standard/NNP’]
[PERSON: u’Gary/NNP Thayer/NNP’] , an economist at [ORGANIZATION: u’AG/NNP Edwards/NNP’]
[PERSON: u’Robert/NNP Brusca/NNP’] , chief economist at [ORGANIZATION: u’Fact/NNP’]
[PERSON: u’Anais/NNP Faraj/NNP’] , an analyst at [ORGANIZATION: u’Nomura/NNP’]
[PERSON: u’Marc/NNP Touati/NNP’] , an economist at [ORGANIZATION: u’Natexis/NNP Banques/NNP Populaires/NNP’]
[PERSON: u’Helmut/NNP Schneider/NNP’] , director of the [ORGANIZATION: u’Institute/NNP’]
[PERSON: u’Marc/NNP Toutai/NNP’] , an economist at [ORGANIZATION: u’Natexis/NNP Banques/NNP Populaires/NNP’]
[PERSON: u’Nicolas/NNP Claquin/NNP’] , an analyst at [ORGANIZATION: u’CCF/NNP’]
[PERSON: u’David/NNP Graham/NNP’] from the [ORGANIZATION: u’FDA/NNP’]
[PERSON: u’Deutsche/NNP Boerse/NNP’] investors unhappy with its [ORGANIZATION: u’London/NNP Stock/NNP Exchange/NNP’]
[PERSON: u’Paul/NNP Richards/NNP’] , an analyst at [ORGANIZATION: u’Numis/NNP Securities/NNPS’]
```

```
[PERSON: u’Wallace/NNP Cheung/NNP’] , an analyst at [ORGANIZATION: u’DBS/NNP Vickers/NNP’]
[PERSON: u’David/NNP Cummings/NNP’] , head of [ORGANIZATION: u’UK/NNP’]
[PERSON: u’John/NNP Reade/NNP’] , an analyst at [ORGANIZATION: u’UBS/NNP’]
[PERSON: u’Digby/NNP Jones/NNP’] , director general of the [ORGANIZATION: u’UK/NNP’]
[PERSON: u’Richard/NNP Moffat/NNP’] , investment director of [ORGANIZATION: u’UK/NNP’]
[PERSON: u’Al/NNP Breach/NNP’] , an economist at [ORGANIZATION: u’UBS/NNP Brunswick/NNP’]
[PERSON: u’David/NNP Cummings/NNP’] , head of [ORGANIZATION: u’UK/NNP’]
[PERSON: u’Brunswick/NNP’] withdrawing from the [ORGANIZATION: u’Glazer/NNP’]
[PERSON: u’Chris/NNP Panayis/NNP’] , managing director of [ORGANIZATION: u’ISP/NNP’]
[PERSON: u’Rick/NNP Egelton/NNP’] , deputy chief economist at [ORGANIZATION: u’BMO/NNP’]
[PERSON: u’Lian/NNP Chia/NNP Liang/NNP’] , economist at [ORGANIZATION: u’JP/NNP Morgan/NNP’]
[PERSON: u’Sureyya/NNP Serdengecti/NNP’] , head of the [ORGANIZATION: u’Turkish/JJ Central/NNP Bank/NNP’]
[PERSON: u’Rick/NNP Mueller/NNP’] , an analyst at [ORGANIZATION: u’Energy/NNP’]
[PERSON: u’Christian/NNP Jasperneite/NNP’] , an economist with [ORGANIZATION: u’MM/NNP Warburg/NNP’]
[PERSON: u’Suhas/NNP Naik/NNP’] , an investment analyst from [ORGANIZATION: u’ING/NNP Mutual/NNP Fund/NNP’]
[PERSON: u’Reza/NNP Moghadam/NNP’] , assistant director of the [ORGANIZATION: u’IMF/NNP’]
[PERSON: u’Michael/NNP Deppler/NNP’] , director of the [ORGANIZATION: u’IMF/NNP’]
[PERSON: u’Axa/NNP’] spokesman , told [ORGANIZATION: u’BBC/NNP News/NNP’]
[PERSON: u’Tim/NNP Congdon/NNP’] , economist at [ORGANIZATION: u’ING/NNP Barings/NNP’]
[PERSON: u’Michael/NNP Moran/NNP’] , analyst at [ORGANIZATION: u’Daiwa/NNP Securities/NNPS’]
[PERSON: u’Kurt/NNP Karl/NNP’] , economist at [ORGANIZATION: u’Swiss/NNP Re/NNP’]
[PERSON: u’Kerry/NNP’] to release supplies from the [ORGANIZATION: u’US/NNP’]
[PERSON: u’Ivo/NNP Geijsen/NNP’] , an analyst with [ORGANIZATION: u’Bank/NNP Oyens/NNP’]
[PERSON: u’Paul/NNP Collison/NNP’] , chief analyst at [ORGANIZATION: u’Brunswick/NNP’]
[PERSON: u’Ronald/NNP Smith/NNP’] , an analyst at [ORGANIZATION: u’Renaissance/NNP Capital/NNP’]
[PERSON: u’Oleg/NNP Maximov/NNP’] , an analyst at [ORGANIZATION: u’Troika/NNP Dialog/NNP’]
[PERSON: u’Miles/NNP Shipside/NNP’] , commercial director at [ORGANIZATION: u’Rightmove/NNP’]
[PERSON: u’Chen/NNP Huiqin/NNP’] , an analyst at [ORGANIZATION: u’Huatai/NNP Securities/NNPS’]
[PERSON: u’Paul/NNP Newsome/NNP’] , an insurance analyst at [ORGANIZATION: u’AG/NNP Edwards/NNP’]
[PERSON: u’Jan/NNP Egeland/NNP’] , head of the [ORGANIZATION: u’UN/NNP’]
[PERSON: u’Card/NNP’] ’s creditors have given [ORGANIZATION: u’LG/NNP’]
[PERSON: u’Lynn/NNP Franco/NNP’] , director of the [ORGANIZATION: u’Conference/NNP Board/NNP’]
[PERSON: u’Marc/NNP Gonsalves/NNP’] , an executive at [ORGANIZATION: u’Xstrata/NNP’]
[PERSON: u’Lian/NNP Chia/NNP Liang/NNP’] , economist at [ORGANIZATION: u’JP/NNP Morgan/NNP’]
[PERSON: u’David/NNP Kim/NNP’] , an analyst at [ORGANIZATION: u’Sejong/NNP Securities/NNPS’]
[PERSON: u’Gordon/NNP Lishman/NNP’] , director general of [ORGANIZATION: u’Age/NNP Concern/NNP England/NNP’]
[PERSON: u’Nick/NNP Bubb/NNP’] , an analyst at [ORGANIZATION: u’Evolution/NNP Securities/NNPS’]
[PERSON: u’Blake/NNP’] Lee-Harwood , campaigns director at [ORGANIZATION: u’Greenpeace/NNP’]
[PERSON: u’Ken/NNP Kim/NNP’] , an analyst at [ORGANIZATION: u’Stone/NNP’]
[PERSON: u’David/NNP Berson/NNP’] , chief economist at [ORGANIZATION: u’Fannie/NNP Mae/NNP’]
[PERSON: u’Michael/NNP Moran/NNP’] , analyst at [ORGANIZATION: u’Daiwa/NNP Securities/NNPS’]
[PERSON: u’Kurt/NNP Karl/NNP’] , economist at [ORGANIZATION: u’Swiss/NNP Re/NNP’]
[PERSON: u’Urban/NNP Decay/NNP’] , from [ORGANIZATION: u’LVMH/NNP’]
[PERSON: u’Anthony/NNP Pratt/NNP’] from [ORGANIZATION: u’JD/NNP Power/NNP’]
[PERSON: u’Wangli/NNP’] , a spokesman for the [ORGANIZATION: u’State/NNP Tobacco/NNP Administration/NNP Monopoly/NNP’]
[PERSON: u’Stefan/NNP Schilbe/NNP’] , analyst at [ORGANIZATION: u’HSBC/NNP Trinkaus/NNP’]
[PERSON: u’John/NNP Nettle/NNP’] , a former employee of [ORGANIZATION: u’General/NNP Mills/NNP’]
[PERSON: u’Rolf/NNP Dress/NNP’] , a spokesman for [ORGANIZATION: u’Union/NNP Investment/NNP’]
[PERSON: u’Wang/NNP Yan/NNP’] , an official from the [ORGANIZATION: u’Beijing/NNP Municipal/NNP Commission/NNP’]
[PERSON: u’Ray/NNP Neidl/NNP’] , an analyst at [ORGANIZATION: u’Calyon/NNP Securities/NNPS’]
[PERSON: u’Tim/NNP Congdon/NNP’] , economist at [ORGANIZATION: u’ING/NNP Barings/NNP’]
[PERSON: u’Digby/NNP Jones/NNP’] , director general of the [ORGANIZATION: u’UK/NNP’]
[PERSON: u’Simon/NNP Rubinsohn/NNP’] , chief economist at [ORGANIZATION: u’Gerrard/NNP’]
[PERSON: u’Frank/NNP Brown/NNP’] , global advisory leader at [ORGANIZATION: u’PwC/NNP’]
====== Relations of PERSON and PERSON =======
[PERSON: u’Viktor/NNP Pinchuk/NNP’] , son-in-law of former-President [PERSON: u’Leonid/NNP Kuchma/NNP’]
[PERSON: u’Glazer/NNP’] ’s two sons , [PERSON: u’Avi/NNP’]
[PERSON: u’Viktor/NNP Pinchuk/NNP’] , son-in-law of former-President [PERSON: u’Kuchma/NNP’]
[PERSON: u’Glazer/NNP’] ’s two sons , [PERSON: u’Avi/NNP’]
====== Relations of PERSON and LOCATION =======
[PERSON: u’Money/NN’] has moved out from [GPE: u’India/NNP’]
[PERSON: u’Bruce/NNP Misamore/NNP’] lives in [GPE: u’Houston/NNP’]
[PERSON: u’Joshua/NNP Osagie/NNP’] , a cocoa farmer from [GPE: u’Edo/NNP’]
[PERSON: u’Sergei/NNP Bogdanchikov/NNP’] . According to reports from [GPE: u’Russian/JJ’]
[PERSON: u’Alvarez/NNP’] added . Companies from the [GPE: u’United/NNP States/NNPS’]
[PERSON: u’Nanik/NNP Rupani/NNP’] , president of the [GPE: u’Indian/JJ’]
[PERSON: u’Helen/NNP Carroll/NNP’] , from [GPE: u’Portsmouth/NNP’]
[PERSON: u’Sandy/NNP Oatley/NNP’] have both resigned from [GPE: u’Southcorp/NNP’]
[PERSON: u’Mauritius/NNP’] and one from [GPE: u’Malaysia/NNP’]
[PERSON: u’Siena/NNP’] , both from [GPE: u’Italy/NNP’]
====== Relations related to DISTANCE =======
```

Chapter 5

Conclusion and Future Work

5.1 Conclusion

We extracted named entities and relations associated with them in BBC news data sets.

The data we are going to use should be a formal writing. For informal writings it won't work well as every steps assumes the formal nature of the text. More importantly the total performance this system directly depends on the performance `taggers` we are using for parts of speech tagging and `chunkers` noun phrase chunking.It is better to take taggers those performs well.

5.2 Future Work

Indian languages have very less *tagged* corpus[1] compared to English. We require large amount of *tagged corpus* to train taggers.

- To recognize named entities in Indian languages
- To apply Information Extraction for Indian languages.

[1] 2.2 Reading Tagged Corpora of [2] and http://www.nltk.org/howto/corpus.html

Bibliography

[1] **Extracting Information from Text**
http://www.nltk.org/book/ch07.html

[2] **Categorizing and Tagging Words**
http://www.nltk.org/book/ch05.html

[3] **Information Extraction in NLTK**
http://www.nltk.org/howto/relextract.html

[4] Prof.Yaser Abu-Mostafa,California Institute of Technology
Learning from Data(Online Course)
https://www.youtube.com/watch?v=mbyG85GZ0PI&list=PLD63A284B7615313A

[5] **Regular Expression in Python**
https://docs.python.org/2/howto/regex.html#regex-howto

[6] Prof. Michael Collins,Columbia University
Tagging Problems, and Hidden Markov Models
http://www.cs.columbia.edu/~mcollins/hmms-spring2013.pdf

[7] Prof. Michael Collins, Columbia University
Natural Language Processing(coursera)
https://class.coursera.org/nlangp-001/lecture

[8] **Dan Jurafsky,Christopher Manning**
Natural Language Processing(coursera lectures),
https://class.coursera.org/nlp/lecture

[9] Rahul Sharnagat
Named Entity Recognition: A Literature Survery
http://www.cfilt.iitb.ac.in/resources/surveys/rahul-ner-survey.pdf

[10] Andrew Y. Ng, Micheal I. Jordan
On Discriminative vs. Generative classifiers: A comparison of logistic regression and navie Bayes,
http://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf