COREFERENCE RESOLUTION

Report submitted in fulfilment of the requirements
for the Exploratory Project of

Second Year B.Tech.

by

Piyush Kumar Maurya

under the guidance of
Dr Ravindranath Chowdary C

Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, 221005, India
May 2017

Dedicated to

My parents and teachers

DECLARATION

I certify that

1. The work contained in this report is original and has been done by myself and the general supervision of my supervisor.
2. The work has not been submitted for any project.
3. Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.
4. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Place: IIT (BHU) Varanasi
Date: 03/05/2019

Piyush Kumar Maurya
B.Tech. student
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

CERTIFICATE

This is to certify that the work contained in this report entitled "Coreference Resolution" being submitted by Piyush Kumar Maurya (Roll No. 17075041), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bona fide work of our supervision.

Place: IIT (BHU) Varanasi
Date: 03/05/2019

Dr Ravindranath Chowdary C
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr Ravindranath Chowdary C for constant guidance and support by devoting his valuable time throughout the project. I express special thanks to my parents for motivativating me from time to time.

Place: IIT (BHU) Varanasi
Date: 03/05/2019 Piyush Kumar Maurya

ABSTRACT

Coreference resolution has been a Natural Language Processing task of interest and research since decades. The task consists of finding all expressions that refer to the same entity in a text. It is an important step for many higher level Natural Language Processing tasks that involve understanding of language such as document summarization, question answering, and information extraction.

Although, a lot of research has been put towards building accurate coreference resolution systems, but a very accurate system has not been build till now. It is because the task requires an understanding of a lot of world knowledge. This project is an outcome of a small attempt towards tackling these challenges.

CONTENTS

List of Figures

List of Tables

List of Symbols

1. Regular expressions in tag patterns

<tag>	A POS tag
?	tag occurs zero or one time
*	tag occurs one or more times
+	tag occurs more than once

INTRODUCTION

Overview

Coreference resolution is the task of linking all expressions that refer to the same entity in a text.[1] The importance of coreference resolution for the entity detection task, i.e. identifying all mentions of entities in text and clustering them into equivalence classes, has been well recognized in the natural language processing community. The first step in the task requires detecting *Mentions, which* can include noun phrases which may be nested within other noun phrases or they can be pronouns. To perform this detection of all such mentions is one and this process is known as mention detection. Mention detection involves passing the text through a *NLP pipeline* which performs a sequence of operations like sentence splitting, word tokenization, parts of speech tagging, named entity recognition. These operations provide necessary information which is further used for mention detection. After mention detection, these mentions are passed through multiple sieves each of which creates links between coreferreing mentions based on some rules. The order in which these sieves are applied affects the accuracy of the formed links. The output of these steps is clusters of mentions, each *cluster* referring to an *entity*.

Motivation of Research Work

Besides, many attempts the task remains a big challenge to the world. Due to its subtle requirements. Use of complex world knowledge, sentence patterns such as fallacies which are easy to understand by us humans but unimaginably difficult to be understood by a machine. An inherent interest in NLP is the motivation towards making a small contribution towards solving such challenges.

Organization of the report

This report consists of a brief introduction to the task and the solution presented in the project. The figures represent the screenshots of the working demo. All facts and figures are related to the outcome of the project.

DATASET

OntoNotes gold dataset is used throughout this project for training various models. It has been annotated for parts of speech tags, named entities, sentence parse, speaker information and coreference links. Together with this, a standard list of proper names and common nouns information is used for identifying gender and number.

CC	Coordinating conjunction	RB	Adverb
CD	Cardinal number	RBR	Adverb, comparative
DT	Determiner	RBS	Adverb, superlative
EX	Existential there	RP	Particle
FW	Foreign word	SYM	Symbol
IN	Preposition or subordinating conjunction	TO	to
JJ	Adjective	UH	Interjection
JJR	Adjective, comparative	VB	Verb, base form
JJS	Adjective, superlative	VBD	Verb, past tense
LS	List item marker	VBG	Verb, gerund or present participle
MD	Modal	VBN	Verb, past participle
NN	Noun, singular or mass	VBP	Verb, non-3rd person singular present
NNS	Noun, plural	VBZ	Verb, 3rd person singular present
NNP	Proper noun, singular	WDT	Wh-determiner
NNPS	Proper noun, plural	WP	Wh-pronoun
PDT	Predeterminer	WP$	Possessive wh-pronoun
POS	Possessive ending	WRB	Wh-adverb
PRP	Personal pronoun		
PRP$	Possessive pronoun		

T1. Table of *POS Tagset* used in the project

PRE-PROCESSING

Before performing the resolution task, useful information required for the task is extracted and saved for reference. Text is first sentence tokenized using nltk sent_tokenizer, further word tokenization is done by a word tokenizer which uses a set of rules based on regular expression, then parts of speech and named entity tags are extracted from corresponding libraries and associated with each of these words and the words are then stored as ***Token*** objects. Storing tokens as objects facilitates the further work.

The POS and NER taggers are based on ***Hidden Markov Model***. This model is probabilistic. The probability of occurrence of words with corresponding tags is known as *emission probability*, and for each tag there will be a number of words with different emission probabilities. Further, each tag can be followed by other tags, probability of this transition in the text is known as *transition probability*. These probabilities are obtained from the dataset and stored after training for reference during tag extraction. Prediction of tags is done by the application of ***Viterbi algorithm***.

Accuracy obtained:

	POS Tagger	*NER Tagger*
Training set	*94.93%*	*97.54%*
Testing set	*92.27%*	*95.00%*

T2. Accuracy measures of the POS and NER taggers

MENTION DETECTION

Mentions include noun phrases, pronouns and named entities. This requires an extensive use of the information extracted in the previous step. ***Noun phrase extraction*** is done by a process called *chunking* in which the required tokens are selected based on their parts of speech tags tag patterns.[2] The tag pattern used in the project is:

<<PDT>?<DT>?<JJ,CD,PRP$><<NN>+<SYM,CC,POS>*<NN>*>+>+*

The noun phrases thus detected are rescanned for any named entities which went left undetected.

PERSON	People, including fictional.	LAW	Named documents made into laws.
NORP	Nationalities or religious or political groups.	LANGUAGE	Any named language.
FAC	Buildings, airports, highways, bridges, etc.	DATE	Absolute or relative dates or periods.
ORG	Companies, agencies, institutions, etc.	TIME	Times smaller than a day.
GPE	Countries, cities, states.	PERCENT	Percentage, including ”%“.
LOC	mountain ranges, water bodies etc.	MONEY	Monetary values, including unit.
PRODUCT	Objects, vehicles, foods, etc.	QUANTITY	Measurements, weight or distance.
EVENT	Named hurricanes, battles, wars, etc.	ORDINAL	“first”, “second”, etc.
WORK_OF_ART	Titles of books, songs, etc.	CARDINAL	Numerals

T3. Table of named entity tags used

As for pronouns, only personal pronouns, ie pronouns marked as <PRP> tag are used as mentions.
Each mention is passed through a gender, number and animacy identification step and this information is stored as a ***Mention*** object.
Animacy detection involves use of *POS* and *NER* tags. Mentions marked as personal pronouns or *PERSON* are marked as animate.

LINKING OF MENTIONS

Mentions which refer to the same entity are *coreferent mentions*. These mentions are linked by passing them through several sieves. Each sieve uses its own set of rules link a mention to an appropriate mention occuring before it in the text. The order in which the sieves run is crucial. The sieves in the order of their priorities are[3]:

a. **Speaker Identification**: <*I, me, my, mine*> uttered by same speaker are coreferent. While <*you, your, yours*> uttered by same speaker refers to the previous speaker.

b. **Exact string match**: It links mentions which are exact match of each other.

c. **Approximate string match**: It links mentions which match after removal of stopwords.

d. **Precise constructs**: This sieve links mentions if they follow a specific pattern such as, mentions which are appositives, role appositives, nominative, or one of them is acronym of other.

e. **Strict head match**: It links the mention to a candidate entity if all non-stop words of the mention are included in the candidate entity, all of the head word match the head words of the candidate entity and all modifiers of the mention are same as that of the candidate mention. This sieve is followed by two of its variants, one without the modifier match and another without the word inclusion property.

f. **Proper head match**: This sieve forms a link if all of the head words of the mention are included in the head words of the candidate entity and all modifiers of the mention are same as that of the candidate mention. This is done in conjunction of another rule which checks that mentions have same named entities, in case they are marked as one.

g. **Relaxed head match**: This sieve is more relaxed as compared to previous two. It forms link if any of the head words of the current mention matches that of the entity.

h. **Pronominal resolution**: Finally pronouns are linked to nominal mentions or other pronouns. This step uses gender, number, animacy and sentence distance information to link the mentions.

TESTING

Testing involves calculation of precision, recall and f1 measures. These values are defined as:

$$\text{Precision} = \frac{true\ positive}{true\ positive + false\ positive}$$

$$\text{Recal} = \frac{true\ positive}{true\ positive + false\ negatives}$$

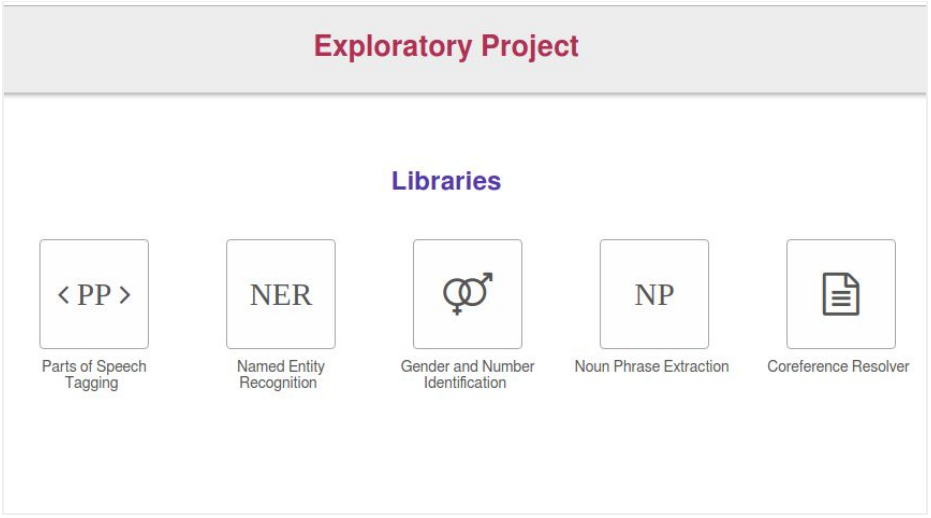
$$\text{F1} = \frac{2 * precision * recall}{precision + recall}$$

The values obtained for the datasets are:

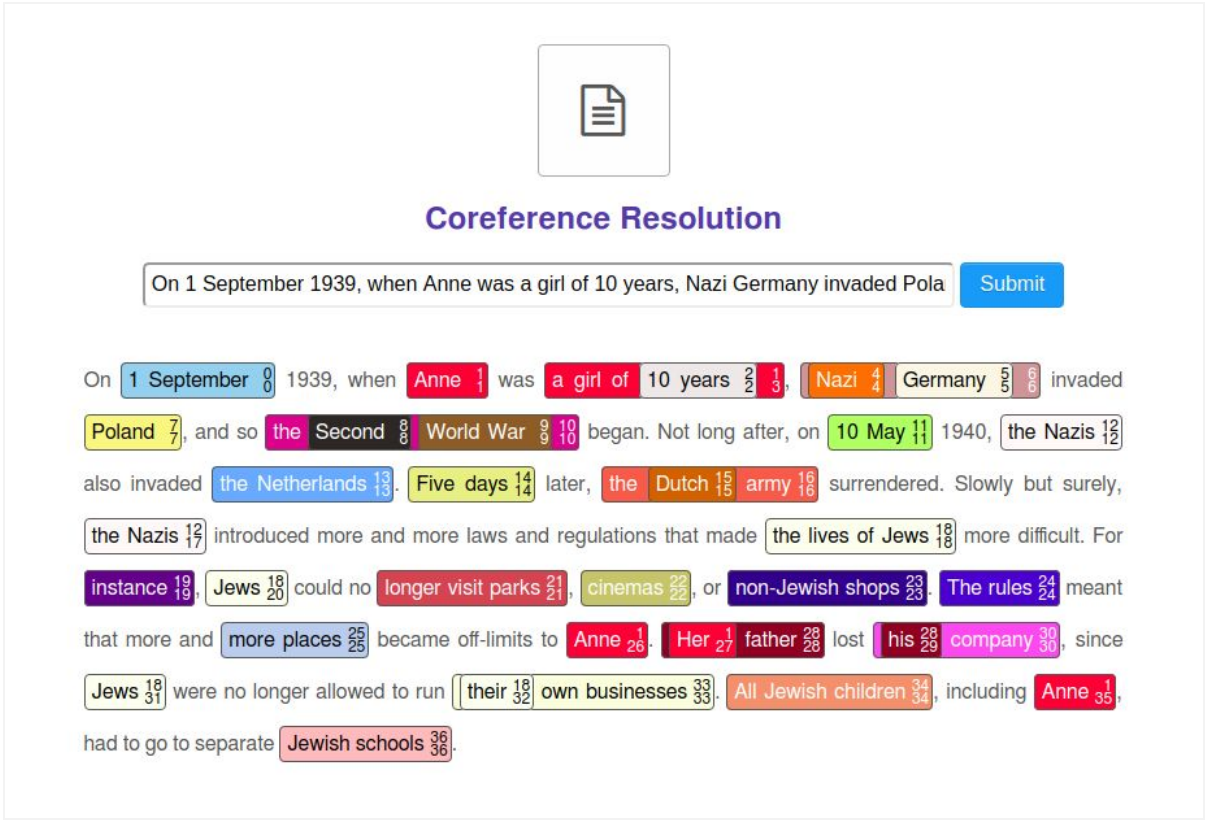
	Precision	*Recall*	*F1*
OntoNotes gold training set	*60.61*	*56.61*	*58.54*
OntoNotes gold test set	*57.41*	*53.18*	*55.22*
OntoNotes gold dev	*54.92*	*52.74*	*53.80*

T4. Precision, Recall and F1 measures of the Coreference Resolution System

INTERFACE



F1. Overview of the Interface



F2. Working demo of the Coreference Resolution System

CONCLUSIONS AND DISCUSSION

After completing the project, it can be concluded that coreference resolution is a difficult task to be performed by a machine. The task involves lots of *world knowledge* to accurately predict the correct link. The accuracy of result can be greatly affected by the preprocessing libraries used. It can be said that a highly accurate coreference resolver will be a complete solution to all *natural language processing* tasks. So there should be constant endeavours in this field to make more and more accurate systems.

BIBLIOGRAPHY

1. Coreference Resolution, The Stanford NLP Group
   https://nlp.stanford.edu/projects/coref.shtml

2. Extracting Information from Text, NLTK
   https://www.nltk.org/book/ch07.html

3. Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules
   https://web.stanford.edu/~jurafsky/pubs/coli_a_00152.pdf