COREFERENCE RESOLUTION

Exploratory Project Report

(January 2019 - April 2019)

completed by

Piyush Kumar Maurya
Roll Number: 17075041
Computer Science and Engineering
Indian Institute of Technology (BHU), Varanasi

under the guidance of

Dr Ravindranath Chowdary C
Assistant Professor
Computer Science and Engineering
Indian Institute of Technology (BHU), Varanasi

ABSTRACT

Coreference resolution is the task of linking all expressions that refer to the same entity in a text. These expressions, also known as *mentions* 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*.

CONTENTS

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.

POS Tagset used in the project (Penn Treebank POS tag list):

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

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:

Training set: 94.93%

Testing set: 92.27%

NER Tagger:

Training set: 97.54%

Testing set: 95.00%

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. 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. The named entities used are:

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

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:

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:

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

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

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

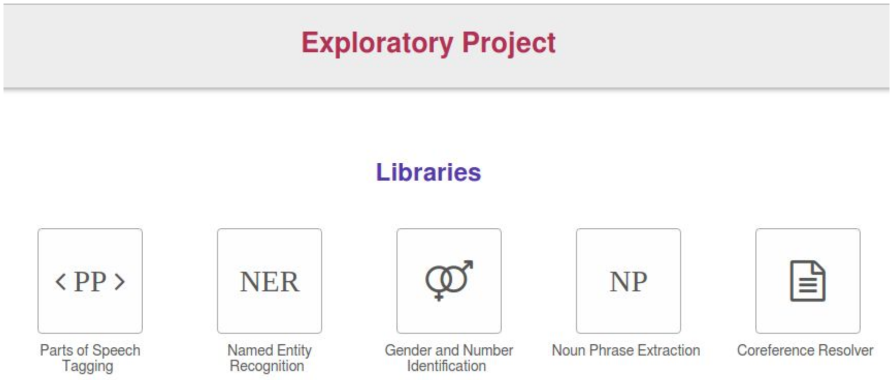
The values obtained for the datasets are:

	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*

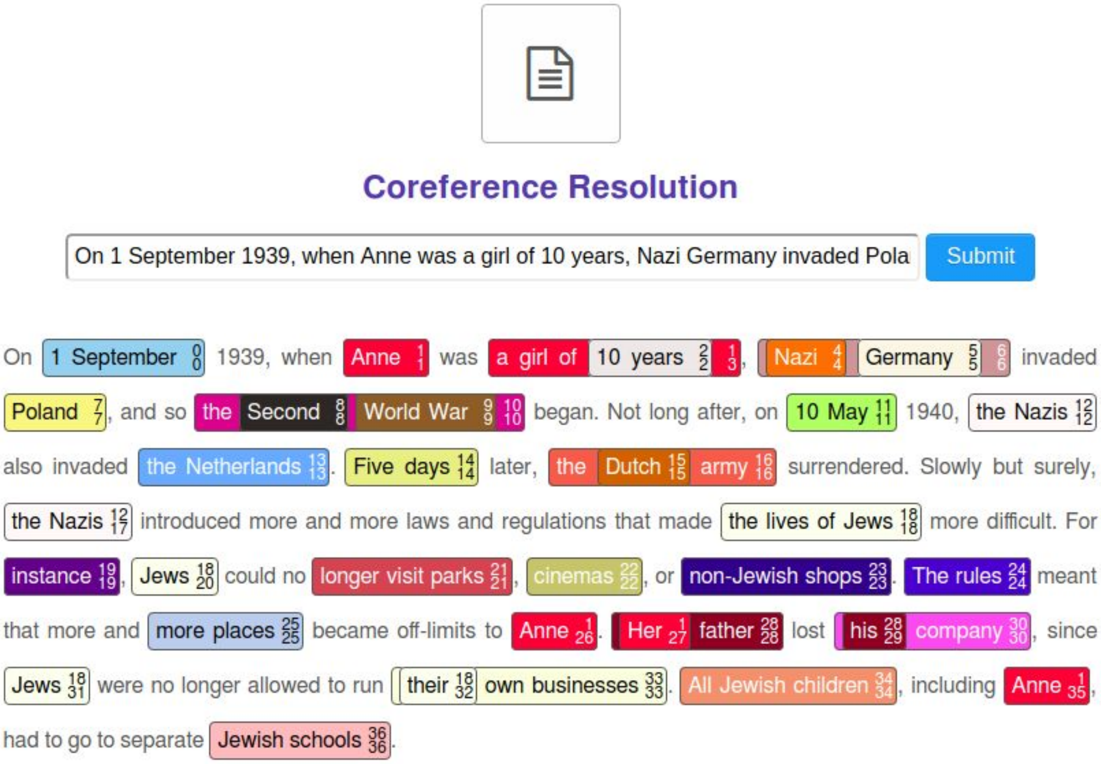
INTERFACE

Some images of the interface are shown here:

Homepage:



Coreference Resolution with coreferent entities of the text marked in same color:



On 1 September 1939, when Anne was a girl of 10 years, Nazi Germany invaded Pola

On 1 September 1939, when Anne was a girl of 10 years, Nazi Germany invaded Poland, and so the Second World War began. Not long after, on 10 May 1940, the Nazis also invaded the Netherlands. Five days later, the Dutch army surrendered. Slowly but surely, the Nazis introduced more and more laws and regulations that made the lives of Jews more difficult. For instance, Jews could no longer visit parks, cinemas, or non-Jewish shops. The rules meant that more and more places became off-limits to Anne. Her father lost his company, since Jews were no longer allowed to run their own businesses. All Jewish children, including Anne, had to go to separate Jewish schools.

CONCLUSION

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.