NLP Project

CSE4022

Lab Assessment – 3
Slot: - L43-44

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Team mate 1: Mihir Agarwal
Reg No: 18BCE2526
Team Mate 2: Shashank Shukla
Reg No. 18BCE2522

Sentence extraction in recognition textual entailment task

Prof. RAJESHKANNAN R

Abstract

Recognizing textual entailment (RTE) is a task that predicts whether a text fragment can be inferred from another text fragment. In this project, we tackle the RTE problem using sentence extraction to cover semantic variation and then extracting subject, predicate and object from each sentence without using external resources like Wordnet. Finally, a similarity function is used to predict entailment relation. In the sentence extraction phase, we used sentence detection, extract sentence in subordinate clause, prepositional phrase and passive sentence. Our system has potential to give accuracy which is comparable to other systems that are not using external resources.
Dataset used is the Third Pascal Recognizing Textual Entailment Challenge (RTE-3) dataset. It has approximately 800 pairs of text (T) and hypothesis (H) with labels as True or False showing whether T entails H or not.
For sentence extraction, following methodology is followed. First part is preprocessing, where the parse tree is generated using the Stanford NLP library. After this step part of speech and parse tree are used for sentence extraction. Then part of sentence is extracted i.e., subject, predicate and object using part of speech tag and syntactic parse tree. At last feature extraction is done and a classifier is used to predict entailment. In this process, TF-IDF is used for word weighing and a feature table is formed. After feature extraction, any classification algorithm can be used to classify whether the text and hypothesis are entailed or not. Naive Bayes and WEKA can also be used in the reference paper.

Introduction

Recognizing Textual Entailment (RTE) has become an important natural language processing task in recent years. The RTE goal is to detect entailment relation between two snippet text pairs <T (text), H (hypothesis)>. T entails H, if H can be inferred from T using common knowledge.

The following is an example extracted from the first RTE challenge dataset showing Text (T) entails Hypothesis (H).

> T: The body of Satomi Mitarai was found by a teacher after her attacker returned to class in bloody clothes.
> H: Mitarai's body was found by a teacher after her killer returned to their classroom covered in blood.

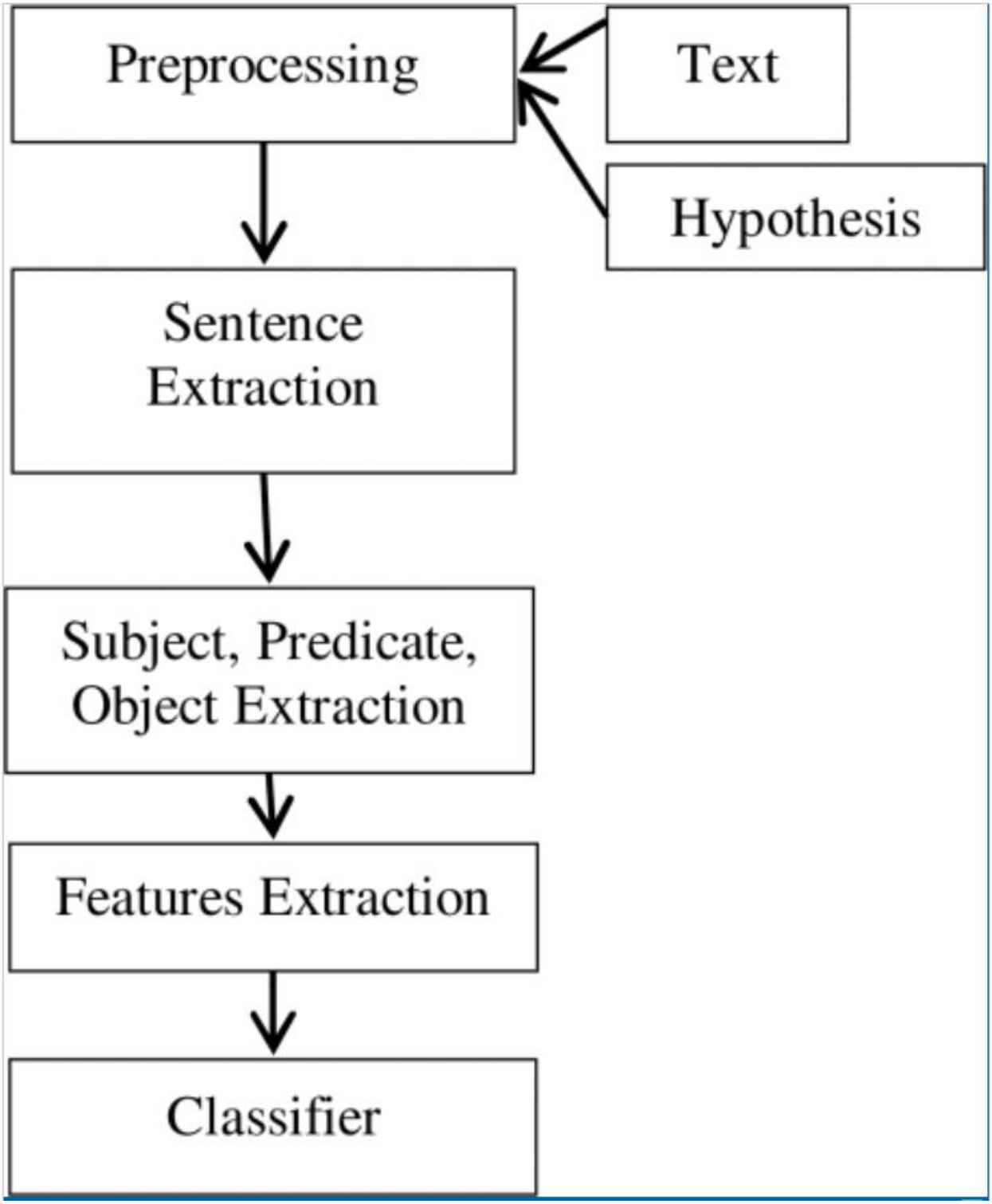
RTE has been useful in various natural language processing applications to handle variation of semantic expression, such as information extraction (IE), text summarization, question answering (QA), and machine translation (MT). In text summarization, textual entailment (TE) can be used to remove sentence redundancy. RTE also can be employed in automatic information credibility assessment tasks. Information is assumed to be more credible if there are other independent sources confirming it and the information is consistent with ground truth information. RTE then can be used to compare information from various sources to check whether it confirms with each other and is consistent with ground truth. Another important factor in information credibility is to assess the sources of information or informant credibility.

There are three groups of information in the statement map: Focus, Conflict and Evidence. Focus is a group of information that is related to the query. Conflict contains information which contradicts with Focus. Both Focus and Conflict groups have Evidence groups which contain information to support each group.

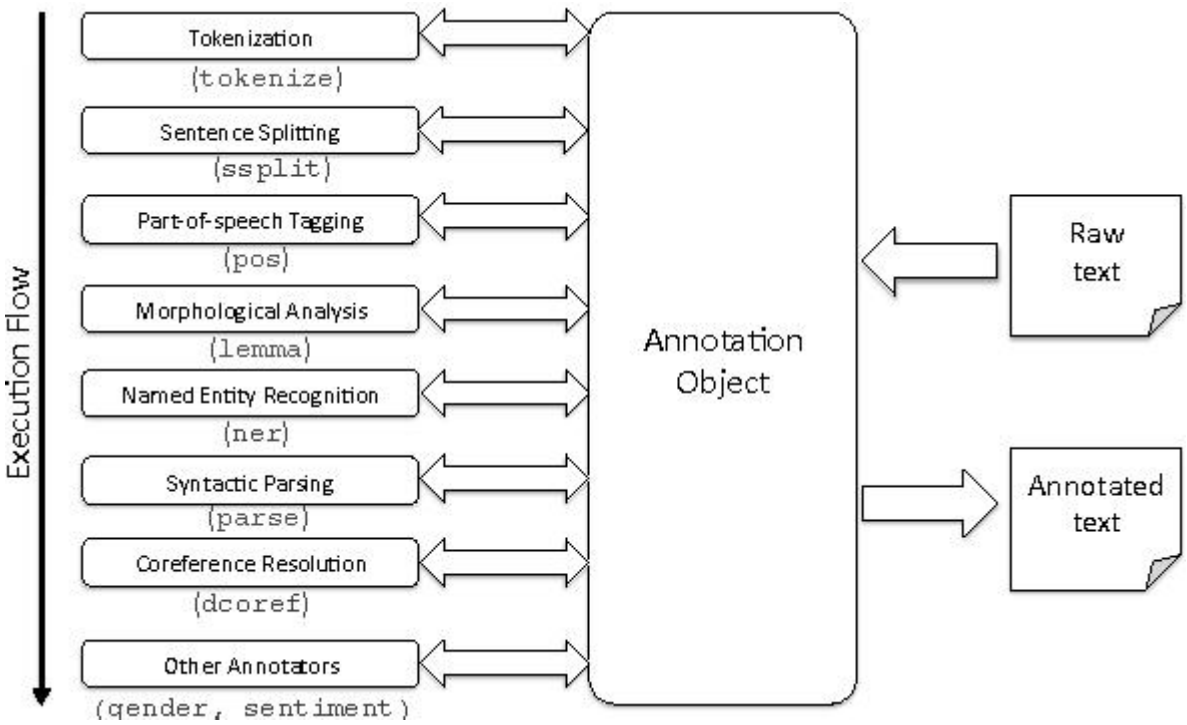
Some approaches have been employed to recognize textual entailment automatically, such as: 1) lexical similarity and syntactic alignment; 2) logic-based and 3) combination techniques. Some systems use external lexical databases such as Wordnet, DIRT, Wikipedia and verb oriented external databases like VerbNet, Framenet. Although external resources could increase accuracy, it needs more processing power and time.

Problem Statement

Architecture Diagram



Flow Diagram



The Stanford CoreNLP Natural Language Processing Toolkit execution flow

Approach and Pseudocode

The textual entailment recognizer has five modules, as shown in the above architecture diagram. First, the preprocessing module applies Stanford Parser and transforms T and H into syntactical structure in the form of a parse tree. Second, sentence extraction applied to H and T that may have more than one sentence or clause. Third, subject, predicate and object extracted from sentence. Fourth, features are extracted from sentence, subject, predicate and object and then classifier used to determine whether H entails T.

In this section we will discuss each module in more detail.

A. Preprocessing

In the preprocessing step, 3 things are done-

1. Removing symbols like “(“,”)”. ‘-”
    a) List of h and t are traversed using for loop
    b) .replace function of python is used to replace characters like ‘,’, ‘-’ with whitespace
2. Generation of syntactic parse tree
    a) Syntactic parse tree is generated using stanford parser
    b) Parser is loaded in gui version of it and english is chosen as language
    c) Text file having required text is loaded and a tree is generated for each text.
3. Generation of part of speech
    a) POS(part of speech) tag is generated using stanza of stanford NLP library.
    b) It is also extracted from a parse tree generated using the above step.
    c) In the parse tree, each word has its own tag.

B. Sentence Extraction

Sentence extraction is needed because hypothesis (H) can entail with only some part of the text (T), for example:

> T: “At the same time the Italian digital rights group, Electronic Frontiers Italy, has asked the **nation's government to investigate Sony** over its use of anti-piracy software.”
> H: “Italy's government investigates Sony.”

Part-of-speech tag and parse tree are used in the sentence extraction module. Steps in the sentence extraction for T and H are:

1. Sentence detection, split regular sentence which is separated by a period.

2. Extract subordinate clauses. “SBAR” part of speech tag used to extract subordinate clauses.

3. Extract sentences in prepositional phrases. Detected using “S” or “VP” tag which is exist in “PP”

After sentence extraction steps, the number of sentences in T or H will be increased.

C. Part of Sentence Extraction: Subject, Predicate, and Object (Future Work)

We extracted subject, predicate and object for every sentence generated in the sentence extraction module. To extract those parts of sentence, we used part of speech tag and syntactic parse tree information and applied heuristic rules.

Heuristic rules-

1. To extract the subject, we choose the highest level noun phrase in the parse tree.
2. Predicate is extracted from the first verb phrase after the subject.
3. Finally, object is extracted from the highest level noun phrase after the predicate.

D. Feature Extraction and Classifier (Future Work)

We will use Term Frequency-Inverse Document Frequency (TF-IDF) for words weighting. Words with high TF-IDF have a strong relationship with the sentences they appear in.

After sentence extraction, each T and H could be transformed into several sentences. Cosine similarity is used for calculating distance between parts of a sentence (subject, predicate, object) from all H's sentences to all T's sentences.

After all features extracted from the dataset, we use a machine learning classifier like WEKA and Naïve Bayes to automatically classify whether T entails H.

Experiment and Results

A. Dataset (sample with explanation)

We used a dataset from Third Pascal Recognizing Textual Entailment Challenge (RTE-3). RTE-3 has two datasets: development set and test set. Each consists of 800 pairs text (T) and hypothesis (H), all manually annotated. RTE-3 dataset is the last RTE challenge dataset available for direct download without needing special permission.

RTE-3 pairs were taken from various sources from the web and were reviewed by three human judges. Average agreement between judges is 87.8% with Kappa level 0.75.

Given below is a screenshot of one group in the dataset which is in XML format. It has pair id, value as True or False which shows that hypothesis is entailed from text or not and task as QA which stands for Question Answering. It has two tags as t and h. T is for Text and H is for Hypothesis.

```
<pair id="625" value="FALSE" task="QA">

   <t>This year, however, the contest has taken on a new urgency as the Clinton Administration, moving to block the Pyongyang government &amp;apos;s bid to build a nuclear arsenal, has rekindled some of the passion in North Korea&amp;apos;s defiance of the West.</t>

   <h>Pyongyang is the capital of North Korea.</h>

</pair>
```

B. Partial Output

In this report, part A, B, and C of methodology are implemented.

Output of Part A -

1- Removing symbols like "(","")". '-"-

Without PreProcessing

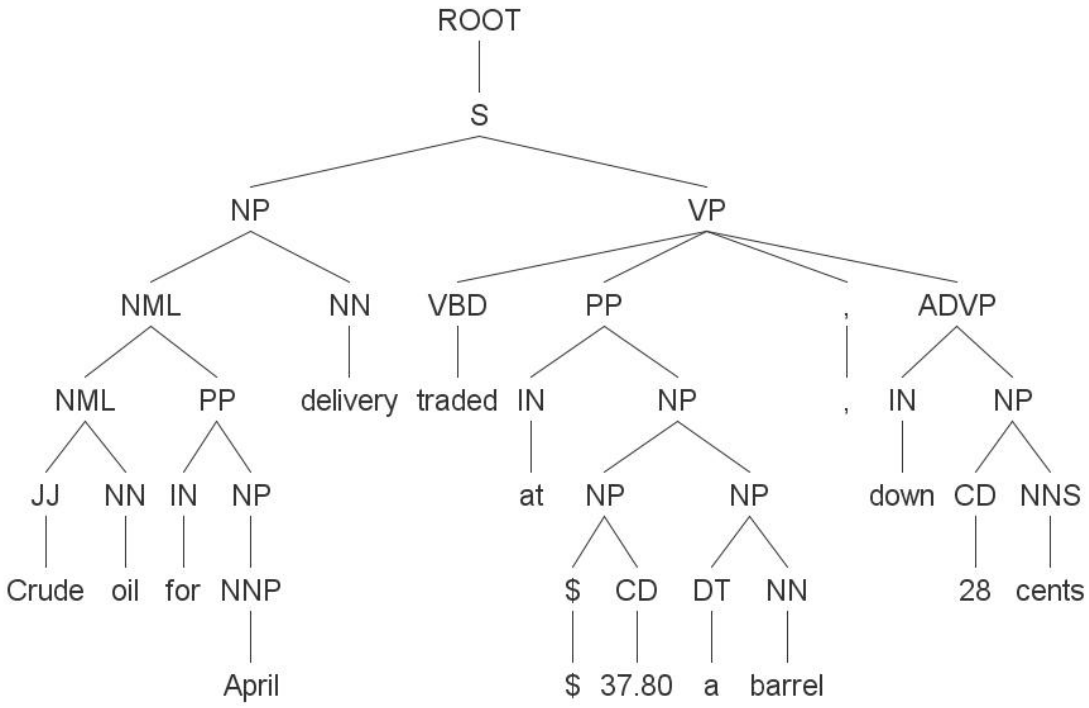
```
In [252]: t_array_all[10]
Out[252]: <t>Iraqi militants said Sunday they would behead Kim Sun-Il, a 33-year-old translator, within 24 hours unless plans to dispatch
          thousands of South Korean troops to Iraq were abandoned.</t>
```

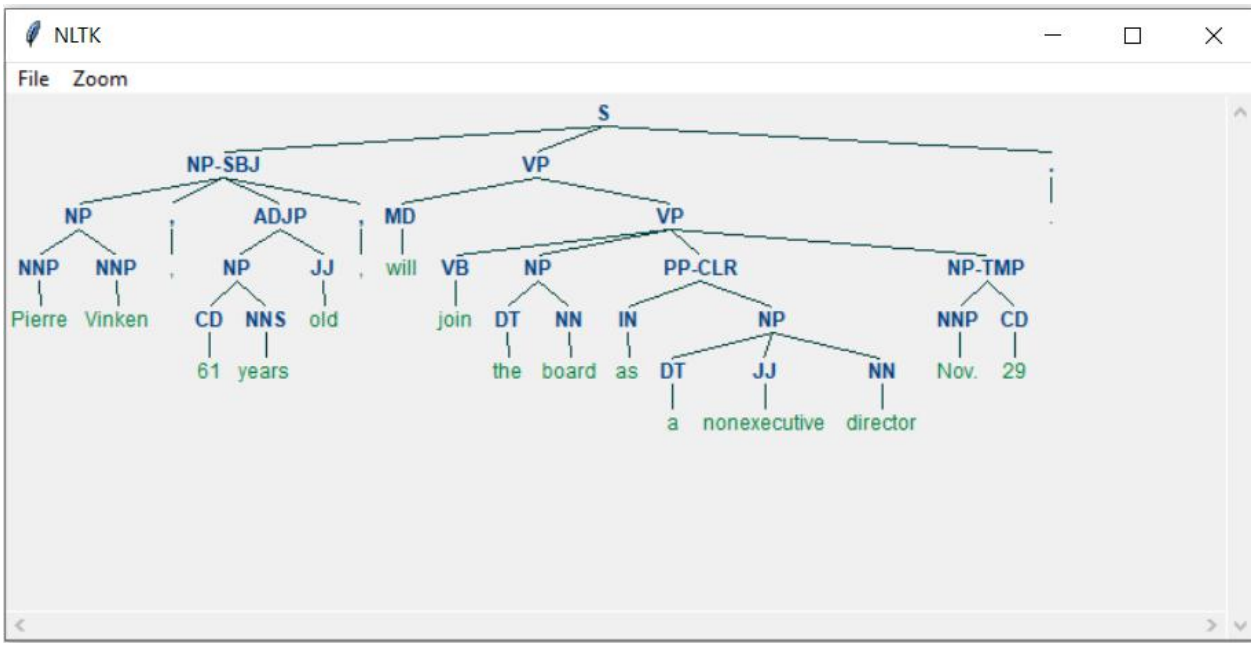
With PreProcessing

```
In [251]: t_array[10]
Out[251]: 'Iraqi militants said Sunday they would behead Kim SunIl a 33yearold translator within 24 hours unless plans to dispatch thousa
          nds of South Korean troops to Iraq were abandoned. '
```

2- Generation of syntactic parse tree

Sentance - Crude oil for April delivery traded at $37.80 a barrel, down 28 cents





(ROOT (S (NP (NML (NML (JJ Crude) (NN oil)) (PP (IN for) (NP (NNP April)))) (NN delivery)) (VP (VBD traded) (PP (IN at) (NP ($ $) (CD 37.80))) (NP (NP (DT a) (NN barrel)) (ADVP (IN down) (NP (CD 28) (NNS cents))))) (. .)))

3- Generation of part of speech -

```
word: Ebola     upos: PROPN     xpos: NNP
word: hemorrhagic       upos: ADJ       xpos: JJ
word: fever     upos: NOUN      xpos: NN
word: is        upos: AUX       xpos: VBZ
word: a         upos: DET       xpos: DT
word: fatal     upos: ADJ       xpos: JJ
word: disease   upos: NOUN      xpos: NN
word: caused    upos: VERB      xpos: VBN
word: by        upos: ADP       xpos: IN
word: a         upos: DET       xpos: DT
word: new       upos: ADJ       xpos: JJ
word: virus     upos: NOUN      xpos: NN
word: which     upos: PRON      xpos: WDT
word: has       upos: VERB      xpos: VBZ
word: no        upos: DET       xpos: DT
word: known     upos: VERB      xpos: VBN
word: cure      upos: NOUN      xpos: NN
word: .         upos: PUNCT     xpos: .
```

Output of Part B -

Using steps in part B, we extracted the following text into six sentences

“Ebola hemorrhagic fever is a fatal disease caused by a new virus which has no known cure. When a new epidemic was detected in Zaire in the spring of 1995, it was widely perceived as a threat to the West. Public attention was intense.”

Sentence	Step
Ebola hemorrhagic fever is a fatal disease caused by a new virus which has no known cure	Sentence detection
When a new epidemic was detected in Zaire in the spring of 1995, it was widely perceived as a threat to the West	Sentence detection

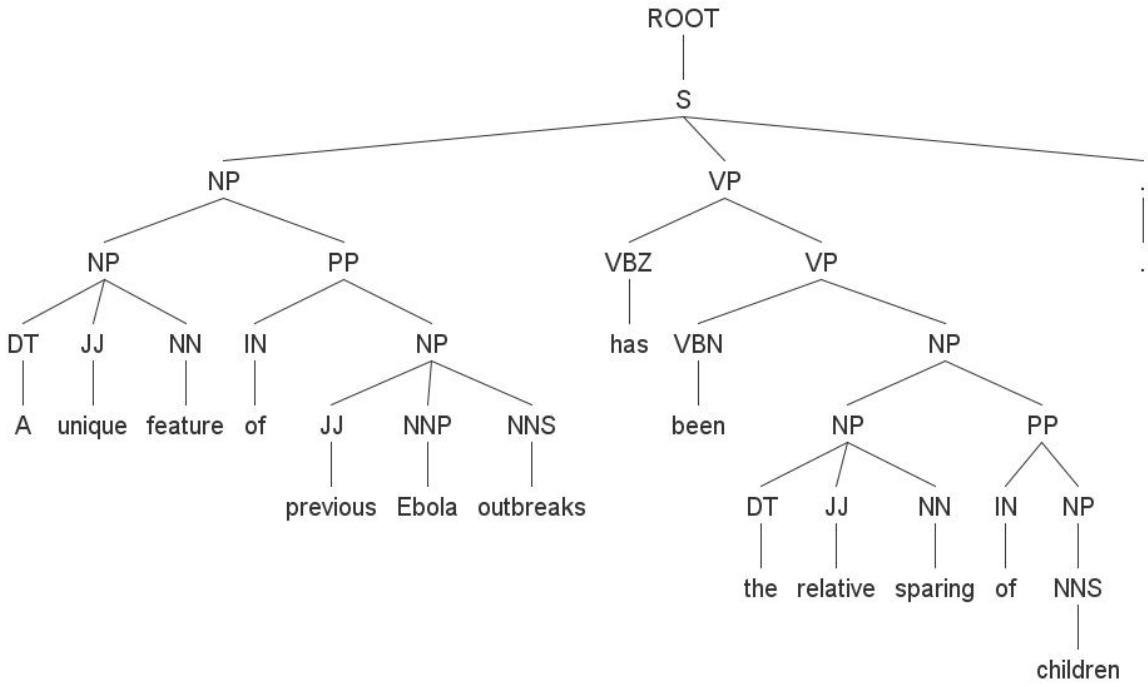
Public attention was intense	Sentence detection
a new epidemic was detected in Zaire in the spring of 1995	Subordinate clause

Output of Part C -

Example-

Sentence	Subject	Object	Predicate
A unique feature of previous Ebola outbreaks has been the relative sparing of children.	A unique feature of previous Ebola outbreaks	has been	Sparing of children
The number of the confirmed Ebola cases has risen slightly to 26 in Gabon and to 16 in Congo Brazzaville	The number of the confirmed Ebola cases	has risen	26 in Gabon and to 16 in Congo Brazzaville

It is calculated by applying given heuristic rules to the parse tree which is output of stanford parser.


ROOT
S
NP
VP
.
NP
PP
VBZ
VP
DT
JJ
NN
IN
NP
has
VBN
NP
A
unique
feature
of
JJ
NNP
NNS
been
NP
PP
previous
Ebola
outbreaks
DT
JJ
NN
IN
NP
the
relative
sparing
of
NNS
children

Conclusion

From the above implementation it can be concluded that, to analyse entailment relation between sentences, preprocessing and feature extraction are the main steps. In the presented methodology, syntactic parse tree and part of speech of tags are extracted for the given dataset.

In this method, we present our method that employs sentence extraction and part of speech extraction (subject, predicate, object) in recognition textual entailment task. For sentence extraction, we use sentence detection to extract subordinate clause, sentences in prepositional phrases.

References

1.Bar-Haim, Roy et al., "Semantic inference at the lexical-syntactic level", PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, vol. 22, no. 1, 2007.

2.Bos, Johan and Katja Markert, "Combining shallow and deep NLP methods for recognizing textual entailment", Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, 2005.

3.Dan Klein and Christopher D. Manning, "Accurate Unlexicalized Parsing", Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430, 2003.

4.Giampiccolo, Danilo et al., "The third pascal recognizing textual entailment challenge", Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing, 2007.

5.Harabagiu, Sanda and Hickl Andrew, "Methods for using textual entailment in open-domain question answering", Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006.

6.Iftene, Adrian and Balahur-Dobrescu Alexandra, "Hypothesis transformation and semantic variability rules used in recognizing textual entailment", Proceedings of the ACL-PASCAL

Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

7.Lloret, Elena et al., "A Text Summarization Approach under the Influence of Textual Entailment", NLPCS, 2008.

8.Magnini, Bernardo et al., "Entailment Graphs for Text Analytics in the Excitement Project", Text Speech and Dialogue, 2014.

9."Ministry of Defence", Joint Doctrine Publication 2–00 Understanding and Intelligence Support to Join Operations, 2011.

10.Snow, Rion, Lucy Vanderwende and Arul Menezes, "Effectively using syntax for recognizing false entailment", Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006.

11.Tatu, Marta and Dan Moldovan, "Cogex at RTE3", Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

12.Murakami, Koji et al., "Statement map: assisting information crediblity analysis by visualizing arguments", Proceedings of the 3rd workshop on Information credibility on the web, 2009.

13.Wang, Rui and Gunter Neumann, "Recognizing textual entailment using sentence similarity based on dependency tree skeletons", Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

14.I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2000.

15.Li, Baoli et al., "Machine learning based semantic inference: Experiments and Observations at RTE-3", Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

16.Malakasiotis, Prodromos and Ion Androutsopoulos, "Learning textual entailment using SVMs and string similarity measures", Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

17.Marsi, Erwin, Krahmer Emiel and Bosma Wauter, "Dependency-based paraphrasing for recognizing textual entailment", Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

Uses-
1-Plagiarism
2-Reviews of amazon which are positive(like sentiment analysis)
3-Answer checking- Answer given entails with key answer
4-Useful in interviews for checking small answers given by candidates from a large set of text.