CS 6320.501 Natural Language Processing

Final Project Report

Topic:

Information Extraction Application using NLP features and techniques

Team Name: Twisters

Team Members:

Parth Shirolawala (pjs170230)

Tej Patel (txp172630)

Domain: Crime Events

Task 1:

Creating a set of Information Templates for the task of filling them by extracting information from the Corpus.

The following templates were made based on the events that happen in the world every day.

Templates:

1) Murder (Suspect, Victim, Method, Location, Time)
2) Robbed (Suspect, Item, amount, location, Time)
3) Kidnap (Suspect, Hostage, Ransom, Location, Time)
4) Molest (Suspect, Victim, Relation, Location, Time)
5) Arson (Arsonist, Property, Loss, Location, Time)
6) Fraud (Suspect, conspiracy, Location, Time)
7) Manslaughter (Suspect, Victim, Method, Location, Time)
8) Abuse (Person, Drug, Location, Time)
9) Terrorism (Person, Target, Type of terrorist activity, Location, Time)
10) Cybercrime (Person, Target, Type, Location, Time)

Task 2:

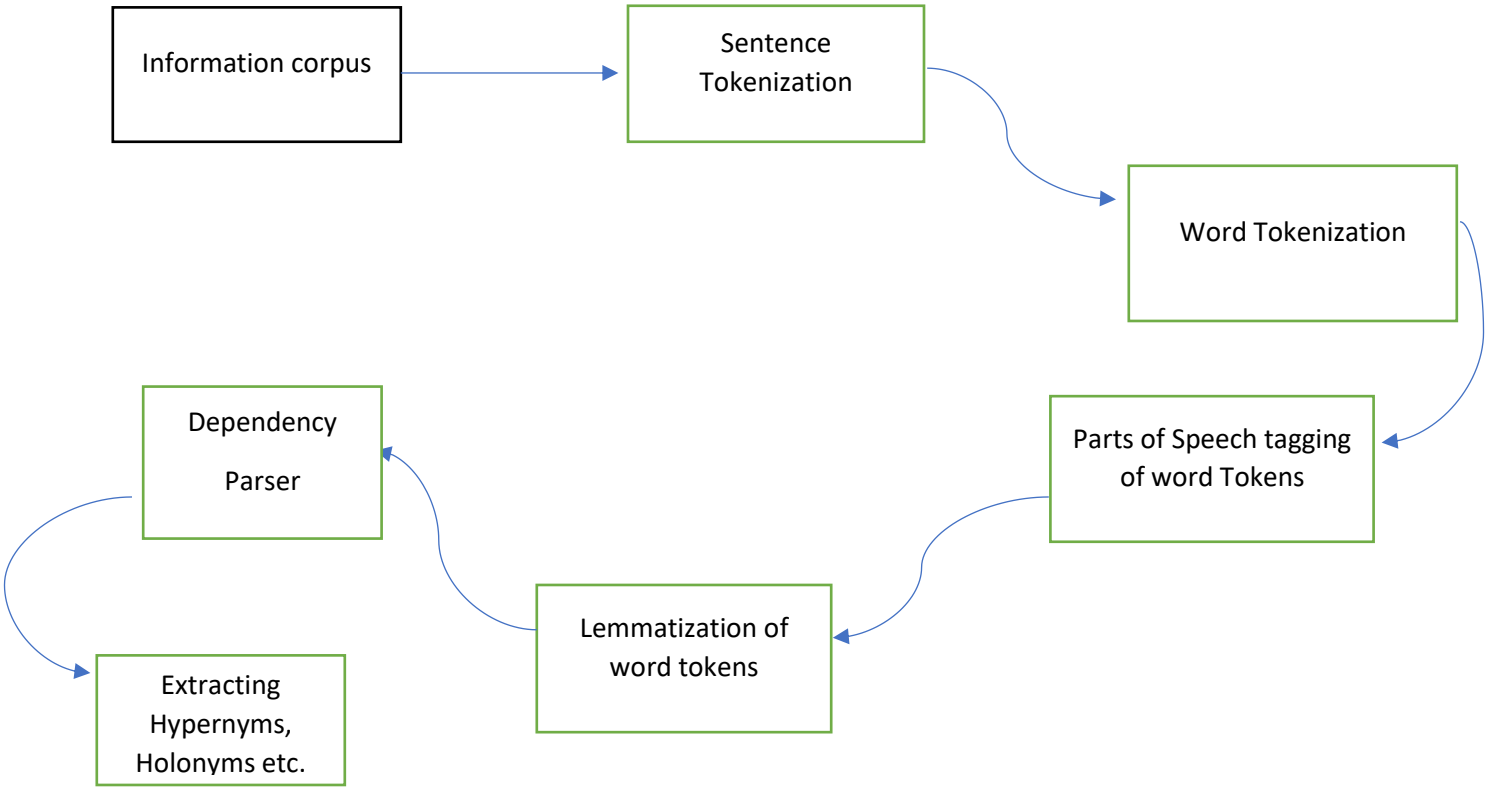
The next task was to build a corpus from which this information extraction will take place and fill each template based on the activity happened.

For this we picked articles from New York Times and integrated into our corpus to build a huge data set. There was no structured information available for the events we are extracting from the corpus.

Total words in our Corpus: ~ 58,000

Task 3:

Task was to build deeper NLP pipeline to extract NLP based features from the natural language statements present in the corpus.



The above is a flow for our NLP pipeline. Talking more about it step wise,

- First step was to do sentence boundary detection, basically tokenizing each sentence in the available corpus.

  Tools used: NLTK sentence tokenizer after removing the Unicode characters.

  Input: "One teen is dead, and two others have been injured following a shooting at an Alabama mall prior to Black Friday shopping. Authorities say the shooting happened about 9:30 p.m. Thursday at the Riverchase Galleria in Hoover, a nearby suburb of Birmingham."

***Output*:**

"One teen is dead and two others have been injured following a shooting at an Alabama mall prior to Black Friday shopping."

"Authorities say the shooting happened about 9:30 p.m. Thursday at the Riverchase Galleria in Hoover, a nearby suburb of Birmingham."

- Secondly, each sentence is tokenized into words.

  Tool used: NLTK word tokenizer

  Input: “One teen is dead, and two others have been injured following a shooting at an Alabama mall prior to Black Friday shopping.”

  Output:
  ["One", "teen", "is", "dead", "and", "two", "others", "have", "been", "injured", "following", "a", "shooting", "at", "an", "Alabama", "mall", "prior", "to", "Black", "Friday", "shopping", "."]

- The next step is to assign parts of speech tag to all the tokens obtained in above step.

  Tool Used: NLTK pos tagger

  Input:
  ["One", "teen", "is", "dead", "and", "two", "others", "have", "been", "injured", "following", "a", "shooting", "at", "an", "Alabama", "mall", "prior", "to", "Black", "Friday", "shopping", "."]

  Output:
  [["One", "CD"], ["teen", "NN"], ["is", "VBZ"], ["dead", "JJ"], ["and", "CC"], ["two", "CD"], ["others", "NNS"], ["have", "VBP"], ["been", "VBN"], ["injured", "VBN"], ["following", "VBG"], ["a", "DT"], ["shooting", "VBG"], ["at", "IN"], ["an", "DT"], ["Alabama", "NNP"], ["mall", "NN"],

["prior", "RB"], ["to", "TO"], ["Black", "NNP"], ["Friday", "NNP"], ["shopping", "NN"], [".", "."]]

- Next step was to lemmatize the words in the sentences, we did that by passing the word and it's POS tag as input.

  Tool Used: NLTK word Lemmatizer tool.

  Input:

  [["One", "CD"], ["teen", "NN"], ["is", "VBZ"], ["dead", "JJ"], ["and", "CC"], ["two", "CD"], ["others", "NNS"], ["have", "VBP"], ["been", "VBN"], ["injured", "VBN"], ["following", "VBG"], ["a", "DT"], ["shooting", "VBG"], ["at", "IN"], ["an", "DT"], ["Alabama", "NNP"], ["mall", "NN"], ["prior", "RB"], ["to", "TO"], ["Black", "NNP"], ["Friday", "NNP"], ["shopping", "NN"], [".", "."]]

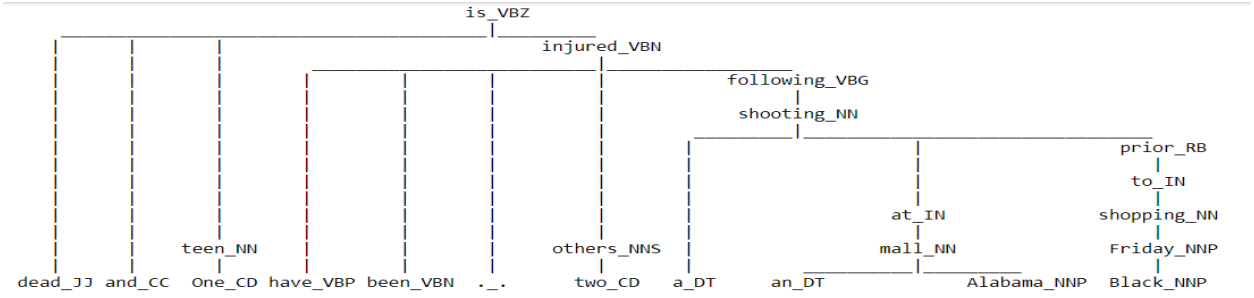
  Output:
  ["One", "teen", "be", "dead", "and", "two", "others", "have", "be", "injure", "follow", "a", "shoot", "at", "an", "Alabama", "mall", "prior", "to", "Black", "Friday", "shopping", "."]

- Next is to build dependency parser for each sentence in Corpus.

  Tool Used: Spacy English model and NLTK tree

  Input: "One teen is dead, and two others have been injured following a shooting at an Alabama mall prior to Black Friday shopping."

  Output:



- The last step is to find hypernyms, hyponyms, meronyms and holonyms of each token.

  Tool Used: PractNLPTool for word annotation and fetching verbs, from the wordnet synset extracting all the Hypernyms, hyponyms, Meronyms and Holonyms available.

  Input: 'died'

  Output:

  ```
  die.v.01
  [Synset('abort.v.02'), Synset('buy_it.v.01'), Synset('drown.v.03'), Synset('fa
  ll.v.07'), Synset('predecease.v.01'), Synset('starve.v.02'), Synset('succumb.
  v.02'), Synset('suffocate.v.05')]
  [Synset('change_state.v.01')]
  []
  []
  ```

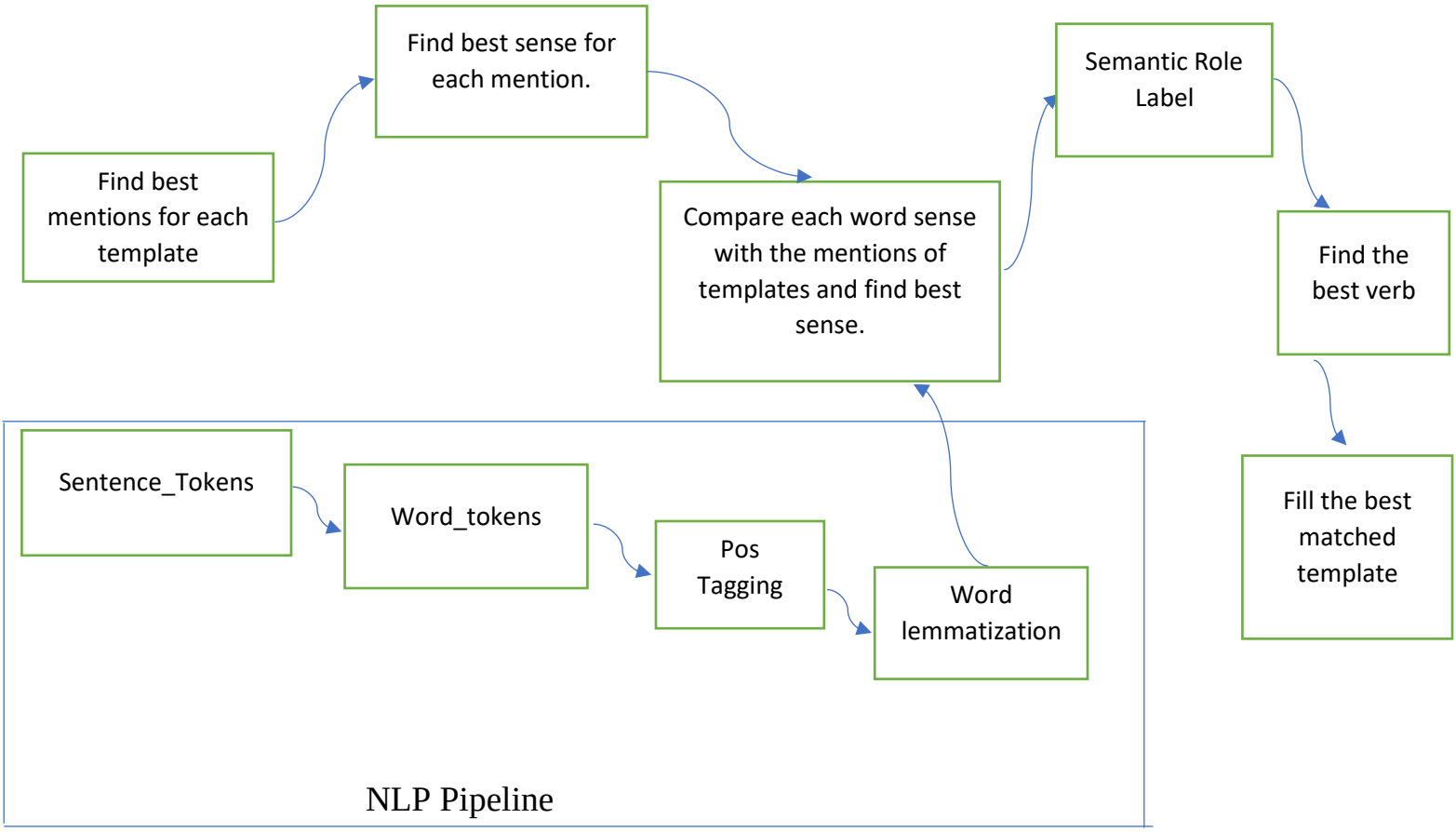
  It had a long list for all senses.

Task 4: Finally, the task was to integrate whole system and implement the program to extract information from the corpus and fill the templates by running the NLP pipeline.

We took machine learning models to do semantic role labelling in our system. After that we used heuristic approach to find the best matching verb with the pre-determined templates.

The semantic role labels were used to fill the constituents of the best matched templates.

The whole architecture for our system can be understood from the below given diagram.



- Firstly, the NLP pipeline was executed as discussed above to obtain the lemmatized words.

- Secondly, the task is to find the best mentions for each information template. In this based on our corpus we decide which mention will fit the best.

  For example: Mentions for our ‘Murder’ template will be:
  ‘murder’, ‘stab’, ‘homicide’, ‘shot’, ‘attack’, ‘kill’

- Thirdly, we found best match for each mention, we did it by getting the result using wordnet synset and then hand picking synset mention for the above mentions.

- Fourth, we found the best sense for each word.

Tool used: wup similarity tool.
With that we found for each sense of word, the best matching sense and keep track of the mention having best similarity.

- Fifth, the task was to do semantic role labelling for each sentence.
  Tool used: Allen NLP
  Allen NLP uses deep lstm neural network to predict the semantic roles in a given sentence.
  For example:
  Input:
  Tom Cunnings killed Mary Jones using Ak-47 in Texas on 6Th August, 2011.

  Output:
  ARG0: Tom Cunnings
  ARG1: Mary Jones
  ARG-MNR: Using an AK-47
  ARGM-LOC: in Texas
  ARGM-TMP: on 6th August, 2011

- Lastly, after getting the SRL tags we filled the appropriate template by assigning the semantic role with appropriate template arguments.

Evaluation of Results obtained from our system:

1) Template – Murder(Suspect, Victim, Method, Location, Time)

   Input:
   "25 years old Tom Cunnings killed 15 years old Mary Jane using an ak-47 in Texas on 6th August, 2011."

   Output:
   Murder (Tom Cunnings, 15 years old Mary Jane, using a ak-47, in Texas, on 6th August, 2011)

2) Template – Robbed (Suspect, Item, Amount, Location, Time)

Input:
"A rare Dodge Challenger, valued at $320,000, has been stolen from a dealership in Melbourne's south east."

Output:

Robbed (, A rare Dodge Challenger, valued at $ 320,000, ,from a dealership in Melbourne 's south east, ,)

3) Template – Kidnap(Suspect, Victim, relation, Location, Time)

Input:
"Sam Sterling kidnapped 3 girls in London setting price of 100k dollars ransom"

Output:
Kidnap(Sam Sterling, 3 girls, setting price of 100k dollars ransom, in London,)

4) Template – Molest (Suspect, Victim, relation, location, time)

Input:
"A six-year-old mentally challenged child was allegedly molested by a cab driver in southwest Delhi on August 23, with police saying that her parents approached school administration about the incident but no action was taken."

Output:
Molest (by a cab driver, A six - year - old mentally challenged child, , in southwest Delhi, on August 23)

5) Template – Arson(Arsonist, Property, Loss, Location, Time)

Input:
"Patels burned several busses in Ahmedabad resulting in total loss of 5 crores to Gujarat Government."

Output:

Arson(Patels, several busses, resulting in total loss of 5 crores to Gujarat Government, in Ahmedabad,)

6) Template – Manslaughter(Suspect, Victim, Method, Location, Time)

Input:
"A woman will face court today with manslaughter charges for running her partner over on their driveway at a property in rural Queensland. "

Output:
Manslaughter (A woman, her partner, on their driveway, at a property in rural Queensland,)

7) Template – Abuse (Person, Drug, Location, Time)

Input:
"24 year-old drug addict overdosed on morphine in New York died at hospital during treatment."

Output:
Abuse (24 year - old drug addict, on morphine, in New York,)

8) Template – Terrorism(Person, Target, Type of terrorist activity, Location, Time)

Input:
"The ISIS terrorized United Stated by craching an airplane into world trade center on 9th September, 2011"

Output:
Terrorism (The ISIS, United Stated, by craching an airplane

9) Template – Cybercrime (Person, Target, type, location, time)

Input:
"Hackers from China attacked the Marriott International Inc in Paris on December 5, were affiliated with Chinese government intelligence gathering operation, according to sources familiar with the matter."

Output:Cybercrime (Hackers from China,the Marriott International Inc, , in Paris, on December 5)

Difficulties Faced:

Semantic role labelling was the hardest part in our system as while passing a bit complex lingual to practnlpTool gave us weird results. Usage of AllenNLP made our task a little bit easier.

In the sentences where there are numerous events happening at the same time, as in has lot of verbs, at that time it becomes tough to get the meaningful output. But handling the best sense in our system solved this problem a bit.

Now our system is eligible to handle simple, a bit complex and little more complex sentences. It can be seen from the above evaluation of results.

We have evaluated more sentences also for that we will provide the different evaluation file.