NLP Report

Problem Description

The project focuses on Information Extraction from a textual corpus. The relevant information is mentioned in terms of template and attributes. The corpus which is not understood by machine then goes through some processing to produce outputs in given templates format that can be extracted and understood by machine. We have chosen **News articles** as our **domain** for this task.

Proposed Solution

We are using a heuristic based approach to extract relevant details from the information. We are creating certain rules that take advantage of the general structure of a news article. Also we utilize the help of the features that derived using tokenization, lemmatization, POS tagging, Dependency Parsing and Wordnet resources to help in heuristic formation.

Full Implementation Details

TASK 1: Forming Templates

- Acquisition(Buyer,seller,Price,Industry)
- Natural Disaster(type, death, date, country)
- Awards(recipient, award, date, industry)
- Murder(perpetrator, victim, instrument, location, date)
- Epidemic(name, affected number, country,communicable)
- Sports Transfers(Player, Transfer team, Amount, time)
- Phone releases(company,model,date,location)
- Policy(name, country, type, year, policy maker)
- Sports Injuries (player, injury, sports, country/team)
- stock_information(company name, loss, stock status, period of time)

TASK 2:

- CORPUS: News articles from various sources

Programming Tools

Task 3:

1)Tokenization

Package used: NLTK package in python

- Regexp tokenizer to differentiate between articles.
- Sentence tokenizer to differentiate between sentences.
- Word tokenizer to differentiate words in sentence.

2)Lemmatization

Package used: NLTK package in python

Wordnet lemmatizer from nltk.stem

3)POS

Package used: NLTK package in python

Averaged_perceptron_tagger – downloaded

nltk.pos_tag

4)Dependency Parsing

Package used: NLTK package in python

Stanford dependency parser

5)Wordnet

Package used: NLTK package in python

Nltk corpus wordnet for synonym, antonym, hypernym, hyponym of word senses

Task 4:

Named Entity Recognition:

Technology used: spaCy

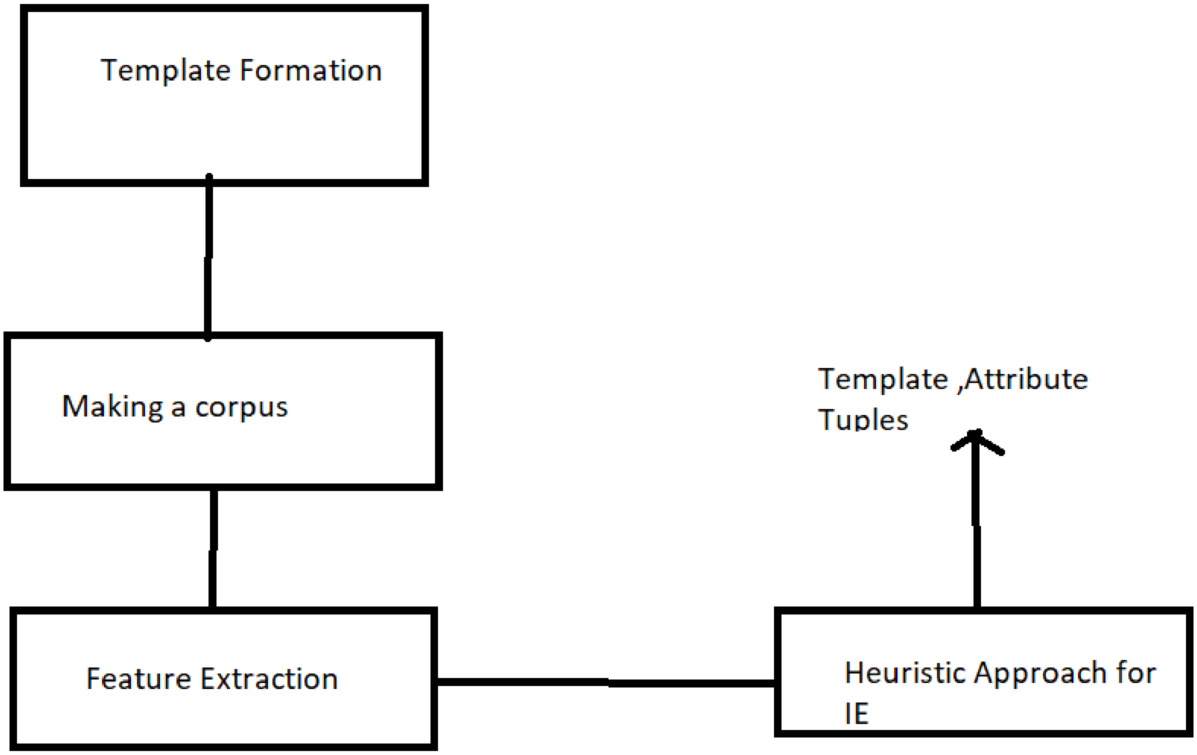
As most of our templates contains attributes such as person, organization, Location, money, time etc , it is convenient to form heuristics once these features are extorted and are mapped to attributes.

We then utilize POS tags to get root words similar or related to templates and their attributes, and form rules for processing and extracting information from the corpus.

Heuristic Based Approach:

Based on certain structural details of sentence formation in the articles of news, we devise few strategies to extract the template related to our specified templates.

ARCHITECTURE:



Our architecture is as shown above:

1)Form Templates

10 templates with minimum 4 attributes each.

2)Collection of Corpus

Various newspaper articles were collected from websites.

3)Feature Extraction

We have performed this in the Task 3 of our implementation to help process the text so that it is machine understandable and to extract features to help with the next steps of entity detection.

This task involves Tokenization, Lemmatization , POS tagging, Dependency Parsing

We identify the type of templates that our article could pertain to and then run the Named Entity Recognition(NER) over the sentences , to extract the information like organization, Location, Person etc and then construct rules to identify the accurate attribute that these features will pertain to.

With the help of these rules and features we extract information in form of tuples.

ERROR Analysis

Template 1:

Acquisition(Buyer,seller,Price,Industry)

Example: Adobe has acquired the software company Marketo for $4.75 billion last Tuesday.

Expected Output: Acquistion(Adobe, Marketo, $4.75 billion, software company)

Actual Output : Acquistion(Adobe, Marketo, $4.75 billion, software company)

Accuracy of approximately 65%.

Template 2:

Natural Disaster(type, death, date, country)

Example: Flash floods and landslides killed at least 12 people in central Vietnam, officials said Sunday, as hundreds of troops were dispatched to clean up destroyed villages and washed out roads.

Expected Output: Natural Disaster(Floods and Landslides,12 people, Sunday, central Vietnam)

Actual Output: Natural Disaster(Floods,12,Sunday,Central Vietnam)

Accuracy of approximately 60%.

Template 3:

Awards(recipient, award, date, industry)

Example: A 33-year-old female Lisa who runs her own mobile coffee and health food business was named the overall winner of the Small Business Academy Award (SBA) awards ceremony hosted by the University of Stellenbosch Business School (USB) on Tuesday evening.

Expected Output: Awards(Lisa, Small Business Academy , Tuesday, mobile and health food buisness)

Actual output: Awards(Lisa, Business Academy Award, Tuesday, Buisness)

Accuracy of approximately 53%.

Template 4:

Murder(perpetrator, victim, instrument, location, date)

Example: Joseph, a 25-year-old was stabbed by Robert using a knife, in a scuffle in California.

Expected Output: Murder(Robert, Joseph, knife, California, N/A)

Actual Output: Murder(Joseph, Joseph, knife, California, N/A)

Accuracy of approximately 58%.

Template 5:

Epidemic(name, affected number, country,communicable)

Example: The death toll from an Ebola outbreak in eastern Congo climbed to 125 as the number of new infections accelerates, the Health Ministry said. Ebola is a " contagious disease" that can cause death.

Expected Output: Epidemic(Ebola,125,eastern Congo, contagious)

Actual Output: Epidemic(Ebola,125,Congo,contagious)

Accuracy of approximately 44%.

Template 6:

Sports Transfers(Player, Transfer team, Amount, time)

Example: LeBron James agrees to four-year, $154-million contract with Los Angeles Lakers

Expected Output: Sports Transfer(LeBron James, Los Angeles Lakers,$154-million,four-year)

Actual output: Sports Transfer(LeBron James, Los Angeles Lakers,$154-million,four-year)

Accuracy of approximately 51%.

Template 7:

Phone releases(company,model,date,location)

Example: Samsung is widely expected to unveil the Galaxy S10 at next year's Mobile World Congress at Korea, which takes place between 25 and 28 February.

Expected Output: Phone releases(Samsung, Galaxy S10, 25 and 28 Februray, Korea)

Actual Output: Phone releases(Samsung, Galaxy, 25 and 28 February, Korea)

Accuracy of approximately 44%.

Template 8:

Policy(name, country, type, year, policy maker)

Example: National Health Policy 2017 of India policy proposes free drugs, free diagnostics and free emergency and essential health care services in all public hospitals in a bid to provide access and financial protection.

Expected Output: Policy (National Health Policy, India, health care, 2017, N/A)

Actual Output: Policy (National Health Policy, India, N/A, 2017, N/A)

Accuracy of approximately 40%.

Template 9:

Sports Injuries (player, injury, sports, country/team)

Example: Atletico Madrid will be without the services of defender Filipe Luis for up to two weeks with a calf injury.

Expected Output: Sports Injuries(Filipe Luis, calf, football, Atletico Madrid)

Actual Output: Sports Injuries(Filipe Luis, calf, N/A, Atletico Madrid)

Accuracy of approximately 56%.

Template 10:

stock_information(company name, loss, stock status, period of time)

Example: Netflix stock also declined 5.2%, losing $6.5 billion in stock value, after the New York Times reported the video streaming company will pay WarnerMedia about $100 million for the right to stream Friends in 2019, citing people with direct knowledge of the matter.

Expected Output: stock_information(Netflix,$6.5 billion ,decline, N/A)

Actual Output:stock_information(Netflix,$6.5 billion,decline,2019)

Accuracy of approximately 46%.

Challenges Faced:

For a complicated sentence with lot of features in one sentence doesn’t give high accurate results.

The derivation of relationship between attributes and mapping of attributes correctly.

If the matter at hand or template is expanded over multiple sentences, it is harder to compute features and fill in the templates.

Improvements:

We could devise strategies and use machine learning and statistical algorithms in combination to obtain better and more accurate results. We could improve upon the model for better and accurate extraction.