Title
Extracting Global Entities Information from News

Author
Xia, Chen

Publication Date
2019

UNIVERSITY OF CALIFORNIA

Los Angeles

Extracting Global Entities Information
from News

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Chen Xia

2019

ABSTRACT OF THE THESIS

Extracting Global Entities Information
from News

by

Chen Xia
Master of Science in Computer Science
University of California, Los Angeles, 2019
Professor Kai-Wei Chang, Chair

There is a ton of news generated every day. However, it is more than anyone can analyze. Since reading every single news is impossible, presenting key information extracted from news can highly improve the efficiency for accessing massive knowledge pieces. And people nowadays analyze news by applying various kind of natural language processing technique both of linguistic approach and machine learning approach. Different methods can interpret news in different ways. Various natural language processing approaches, such as semantic role labeling and coreference resolution, have been applied in information extraction and we applied the technique for analyzing news. We extracted global entities information from news and presented a demo to visualize different aspects of news data. To better extract key information from massive news, we focused on semantic roles in the news data, and the relationship between them in the news flow.

The thesis of Chen Xia is approved.

Guy Van den Broeck

Junghoo (John) Cho

Yizhou Sun

Kai-Wei Chang, Committee Chair

University of California, Los Angeles

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to express my gratitude to my awesome advisor Prof. Kai-Wei Chang of Henry Samueli School of Engineering at UCLA for the useful discussions, comments, and engagement through the learning process of this master thesis. He consistently allowed this thesis to be my own work, but steered me in the right direction whenever I needed it.

I would also like to acknowledge Prof. Yizhou Sun, Prof. Junghoo (John) Cho and Prof. Guy Van den Broeck of Henry Samueli School of Engineering at UCLA as the readers of this thesis, and I am gratefully in debted to them for their very valuable comments on this thesis.

I would also like to thank Taboola company for prodiving the datasets .and the support on the project.

I would also like to thank my family and friends, who have supported me throughout entire process, both by listening to my enthusiasm towards the knowledge and having my back all the time.

CHAPTER 1

Introduction

1.1 Motivation

Looking back at the history, the ability to communicate through words and language is a heritage from human ancestors. After paper was invented, human beings acquire the ability to write down information and preserve knowledge better enough to pass down generation by generation. People gradually observe this world by standing on the shoulder of their ancestors. In recent decades, the way people acquire information has greatly changed since the development of information technology, from book, newspaper to various kinds of media like electronic devices, text and website. However, one thing remains unchanged is that, most of them communicate through natural language. Nowadays countless news data describing different aspects of events are flowing through the internet and media. With this overwhelmed information, the biggest problem becomes how can people find a way to gather information efficiently from the unordered or even unstructured data?

To analyze the news better, an intuive way is to apply different kind of natural language processing technique to extact key information from news. Different types of methods can interpret different aspects of news. Current general process of natural language understanding includes lexical analysis, syntactic parsing and semantic analysis. Lexical analysis focuses on lexical meaning and part-of-speach at word level. While syntactic parsing plays an important role in telling what is an acceptable structure for the target language, semantic analysis features the meaning of the text. Among which, semantic role labeling, also know as shallow semantic parsing, labels the semantic role of words or phrases in the sentence such as the subject, verb and object.

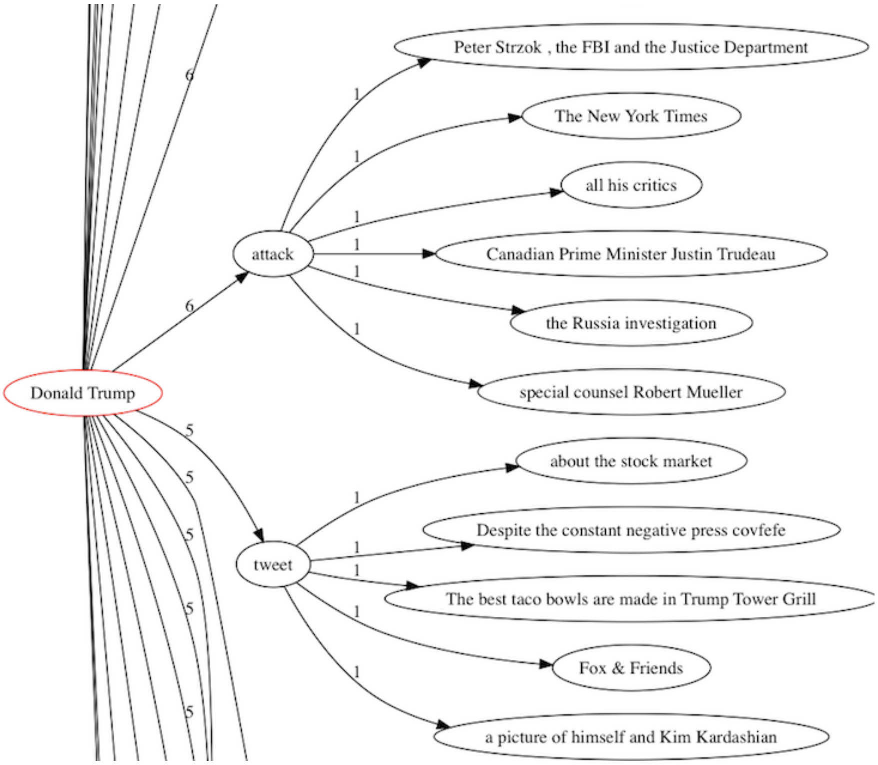

Donald Trump
attack
tweet
6
6
5
5
5
5
5
1
Peter Strzok , the FBI and the Justice Department
The New York Times
all his critics
Canadian Prime Minister Justin Trudeau
the Russia investigation
special counsel Robert Mueller
about the stock market
Despite the constant negative press covfefe
The best taco bowls are made in Trump Tower Grill
Fox & Friends
a picture of himself and Kim Kardashian

Figure 1.1: Semantic Role Labeling Tree for *Donald Trump*

It is not easy to read news title descriptions one by one to understand what a certain subject is doing. For example, *Donald Trump* has been involved in many events and he likes to tweet. Looking through all the Trump news to summarize what he is doing requires a lot of human work. However, by applying semantic role labeling to the news data involving *Donald Trump*, we can easily extract the core information such as what he is doing or he did what to whom. Looking at the semantic role labeling tree we extracted from the news in Figure 1.1, it is easy to summarize the news around him as he attacked *FBI*, *Russia investigation*, *the New York Times* and fought back towards *all his critics*. He also tweeted on *stock market*, *taco bowls in Trump Tower Grill*, *Fox & Friends*, and *a picture of Kim Kardashian and himself*. The label on edges represents how many times this kind of news appears in the database, for example, *Donald Trump* has attacked 6 times towards different people and tweeted 5 times on different topics. This provides an perspective to intuitively see what kind of verbs are related to a certain subject and also how certain subject and object get along through the news flow. Today, in the world full of changes, one thing remains unchanged is that nothing is unchangeable. By analyzing news from different time slots, we can also observe the trend of different semantic roles changing through time.

1.2 Thesis Outline

The thesis is organized as follows. Necessary background for information extraction, including semantic role labeling and coreference resolution are introduced in Chapter 2. The dataset we use as well as how we apply semantic role labeling to the news and address challenges are described in Chapter 3. In Chapter 4, the result of experiment is analyzed. In the end, our conclusions and future directions are discussed in Chapter 5.

CHAPTER 2

Background in Information Extraction

In this Chapter, we present the necessary background for the core technique we apply on news dataset, semantic role labeling and coreference resolution.

2.1 Information Extraction

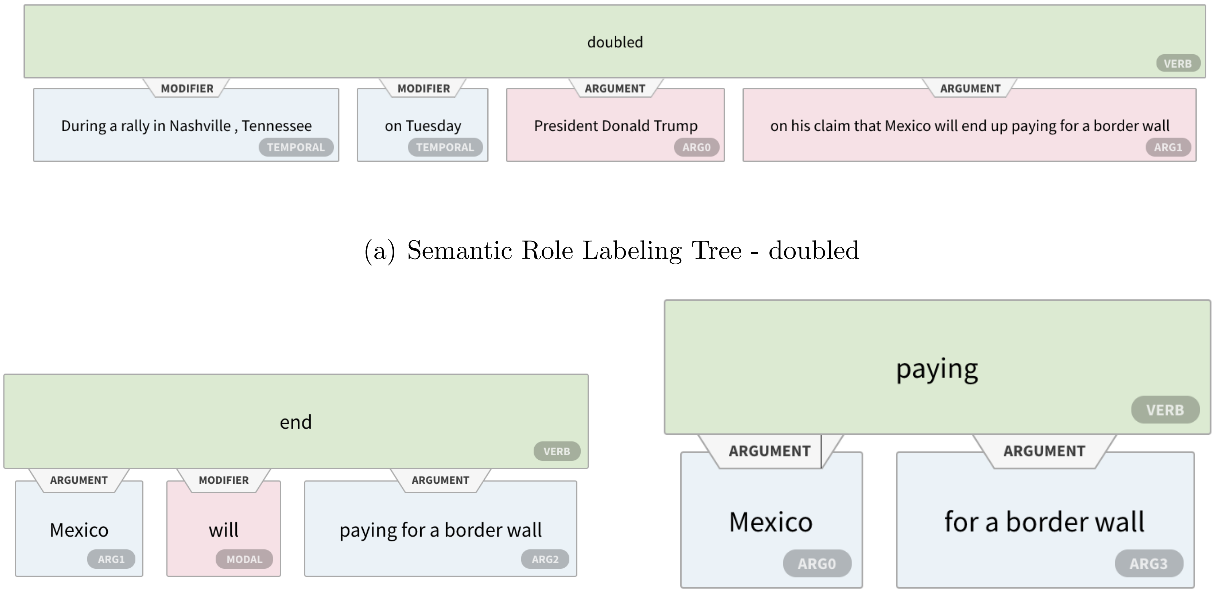
Information Extraction is a widely applied technology in natural language processing. It takes raw natural language text as input and produces structured information with constraints by certain applications. Various sub-tasks of information extraction such as named entity recognition, coreference resolution, named entity linking, relation extraction forms the building blocks of various high end Natural Language Processing tasks such as machine translation and question answering system [Sin18].

One of the approaches in information extraction is the pattern matching using regular expression. However, this approach highly relies on the rules and it better suits domain specific problems. Nowadays, another popular approach is machine learning, including decision trees, naive Bayes classifier, support vector machine, conditional random fields. They are typically more time consuming and requires labeled training dataset [Sin18].

2.2 Semantic Role Labeling

Semantic role labeling is also known as shallow semantic parsing. It can label the semantic role for words and phrases in sentences and plays an important role in applications like question answering, machine translation and information extraction. Since semantic role labeling focuses on predicates and the relation between other roles and the predicate, it explains the meaning of shallow. Shallow semantic parsing instead of deep ones has better accuracy because its unique character reduces error.

There are multiple ways to do semantic role labeling in a shallow way. Trditional semantic role labeling consists of parsing sentences, identification on arguments and the rest, classification on which arguments and disambiguation. Nowadays, as neural network become more and more prevalent, there is also research on semantic role labeling using deep learning methods. The state-of-art work, He adopted a 8-layer deep BiLSTM model to deal with the problem [HLL17]. However, they didn't apply them to multiple news data. The semantic role labeling method we applied in this thesis is from a natural language processing library Allennlp [GGN18], which reimplemented He's work.



(a) Semantic Role Labeling Tree - doubled

(b) Semantic Role Labeling Tree - end

(c) Semantic Role Labeling Tree - paying

Figure 2.1: Semantic Role Labeling Example of Allennlp Library

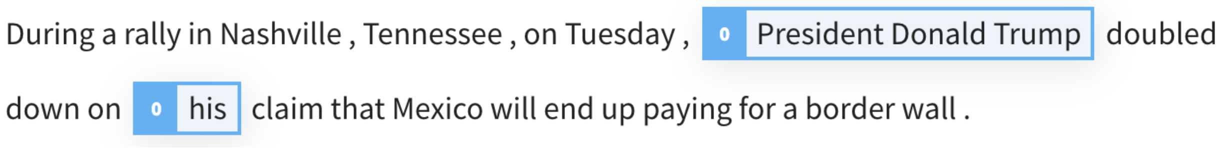
To make it clearer, take the following sentence as an example. *During a rally in Nashville, Tennessee, on Tuesday, President Donald Trump doubled down on his claim that Mexico will end up paying for a border wall.* In Figure 2.1, the semantic role labeling from Allennlp library provides three semantic role labeling trees with verbs *doubled, end, paying* as root. Each of the tree has several leaf nodes such as ARG0 (subject), ARG1 (object), TEM-

PORAL(temporal modifier), MODAL (modal auxiliary modifier). In this thesis, we mainly focus on subject, verb and object. Some modal auxiliary modifiers are also discussed shortly.

2.3 Coreference Resolution

Coreference resolution is the method that can cluster mentions in the document that refer to the same entity. It has a wide application in every task involving natural language. Some of the tasks include determining who the pronoun refer to in machine translation and differentiating who exactly is the sentence talking about in question answering. It also has a wide use in information extraction. The accuracy of coreference resolution is highly important since it determines the meaning of a sentence.

Recent research on coreference resolution mainly focuses on neural coreference model as well as mention detection and mention clustering [ZSY18]. The state-of-art work done by Lee [LHL17], which provides an end-to-end method that consider all spans as mention and learn distributions over possible antecedents, achieving average 68.8 F1 score. In this thesis, the coreference resolution by library Allennlp [GGN18] is the same as Lee's. An example of the coreference resolution of this library is that, it can cluster mentions *President Donald Trump* and *his* in the sentence below because *President Donald Trump* doubled his own claim.



During a rally in Nashville , Tennessee , on Tuesday , President Donald Trump doubled down on his claim that Mexico will end up paying for a border wall .

Figure 2.2: Coreference Resolution Example of Allennlp Library

2.4 Chapter Summary

In this chapter, we mainly discussed about the necessary background of information extraction, including what semantic role labeling is and what coreference resolution is. Semantic role labeling points out the semantic role for words and phrases. It presents the key structure of sentences like who did something to whom. Coreference resolution, however, clusters mentions that refer to the same entity together. It benefits the accuracy of understanding and reduces ambiguity. The two methods help us to understand the sentence better.

CHAPTER 3

Methodology

In this chapter, we describe the dataset we use and methodology to analyze the data. There are challenges of duplicate or similar meanings in verbs and objects, affecting the methodology result. We also present solutions to those challenges.

3.1 Dataset

The news dataset we are using is from Taboola company. We have applied our technique to two major datasets. One is a topic specific Trump dataset containing more than two months of news from late April to early July 2018. The news in this dataset share the same character, which is they all involve President Trump. The other dataset is general 6 months news data from November 2018 to April 2019. The topic of 6 months dataset ranges from sports to politics, food recipes to crime. Detailed format of the two datasets is discussed as follows.

3.1.1 Trump Dataset

The Trump dataset has overall 20,833 news entries. Each news entry contains title description with punctuations, a unique article id, and probability on ten topic clusters clustered by latent dirichlet allocation, which is a three-level hierarchical Bayesian model where each item of a collection is modeled as a finite mixture over an underlying set of topics [BNJ03].

attribute	description
title description	title description for each news
article ids	unique article ids for each news
probability	probability on topic clusters

Table 3.1: Data Format of Trump Dataset

3.1.2 6 Month Dataset

The 6 month dataset contains over half a million news title descriptions from the Taboola company. The 6 month dataset has two different formats, StepContent and StepIndexingData. They are both in json format but differ in attributes.

3.1.2.1 StepContent Format

The StepContent format for the 6 months dataset includes taxonomy, first level taxonomy, unique article ids, title description, and so on. StepCotent uniquely contains title descriptions for later natural language processing. This special attribute is colored in red in the table below.

attribute	description
articleIds	unique article ids for each news
firstLevelTaxonomy	first level category
taxonomy	multiple detailed categories
title	news title
title description	title description with punctuations
...	...

Table 3.2: StepContent Format of 6 Months Dataset

3.1.2.2 StepIndexingData Format

StepIndexingData also shares attributes that the StepContent format has, like taxonomy, unique article id, title and so on. This format has uniquely hierarchical topic cluster information including cluster id, label, labelScore, taxonomy as well as topTerms for level 0 to level 4 where level 0 indicates the most general topic clusters and level 4 indicates the most specific level. This information is labeled in blue in the following table.

attribute	**description**
articleIds	unique article ids for each news
firstLevelTaxonomy	first level category
taxonomy	multiple detailed categories
title	news title
traffic	traffic for this news article
topic cluster	contains id, label, labelScore, taxonomy, topTerms, topTermsScore
...	...

Table 3.3: StepIndexingData Format of 6 Months Dataset

3.2 Apply Semantic Role Labeling

For each dataset, we firstly apply semantic role labeling from a natural language processing library Allennlp [GGN18] . The semantic role labeling from Allennlp provides us with labels for certain words and phrases in the sentence. The result can be visualized in a tree structure with the subject as the root, verbs as the first layer and objects as the leaf nodes.

However, semantic role labeling results can only provide an exact match for subject of interest. By searching *Trump*, other names to refer *Trump* such as *Donald Trump*, *President Trump* are also related. As a consequence, secondly, coreference resolution is applied to solve this problem and then the single tree graph transforms into a forest graph.

Besides semantic role labeling on roles and coreference resolution on subject, there are other merging challenges on object-verb edges and verb-subject edges as well. Some tricks involving modifier, negative and lemmatizing verbs are also discussed in the following sections.

3.2.1 Tree Graph for Semantic Role Visualization

We provide users with a search bar to explore role of interest. For example, if they search for *Trump*, we provide tree graph, with subject *Trump* as the root. The second layer of the tree is all of the verbs labeled together with subject *Trump*, see *blamed* and *liked* in Figure 3.1. The label on the edge represents how many times the subject *Trump* and Verb, for example *blamed*, are indicated together by labeling (in a sentence) in the whole dataset. The leaf node is the object related to the predicate.

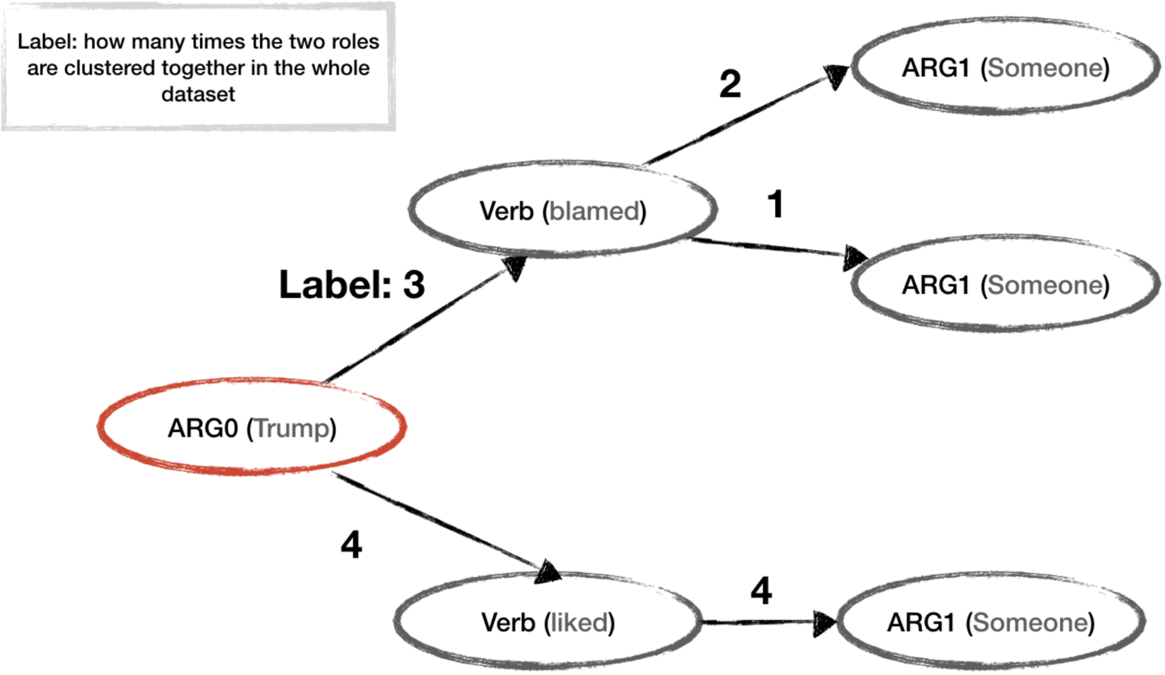


Figure 3.1: A Design of Tree Graph for Semantic Role Visualization.

Here we can indicate from Figure 3.1, Trump has performed the action three times in the Trump dataset blaming someone twice and another person once. He also *liked* somebody four times. In the demo we featured in next chapter, the subject to object edge is sorted according to the attached weight. In this way, the tree graph is able to show what action an interested subject often takes.

3.2.2 Forest Graph for Semantic Role Visualization

Even though we are able to present the tree graph for semantic role labeling, once the user searches for *Trump*, he or she should see a tree with a root node that exactly matches the input role in the search bar. Every time a subject of interest is typed in, a set of its other alias should be added as well. The demo system should not only understand *Trump* but also come up with *Donald Trump* and other information, for example occupation coreference, *President Trump*.

With the help of coreference resolution from the library Allennlp [GGN18], we preprocessed the whole dataset with coreference resolution and generated local coreference clusters for each news. To obtain a global view, we use union find algorithm to merge the clusters together until neither shares a common role. To be clearer, a visualization demo for the coreference resolution is also provided.

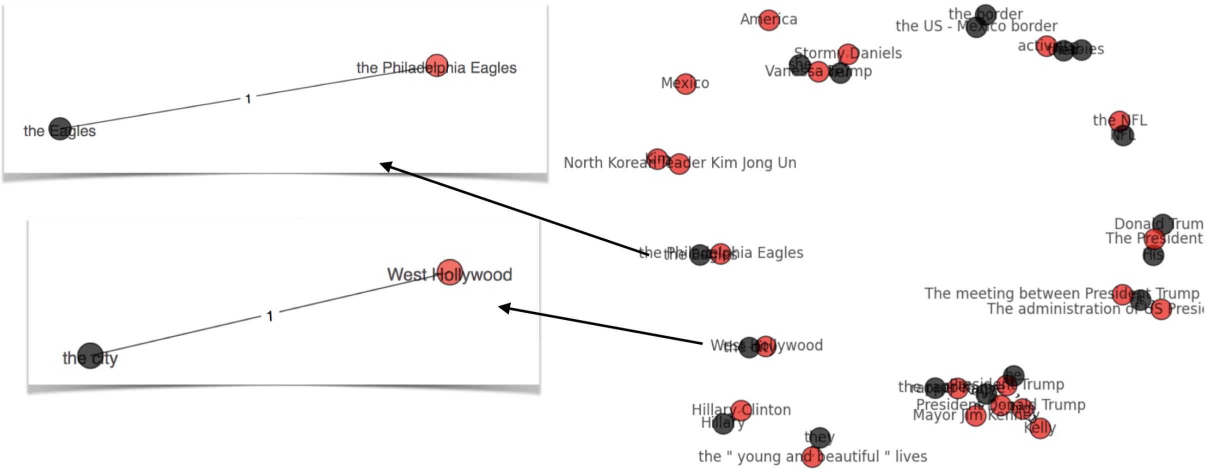


Figure 3.2: Coreference Resolution Clusters.

In the Figure 3.2, the coreference resolution successfully clusters *the Philladelphia Eagles* with *the Eagles*, and *West Hollywood* with *the city*. We define center roles in a cluster as the several most important words or phrases that can represent the whole cluster, for example, *Donald Trump* is a better representative than *the President*. The red node in the cluster indicates the center roles we should pay attention to. These are the following three ways to determine which roles among the cluster are the center roles:

- **LongestSpan** is a method which selects the role with longest length as the center role in a cluster. The idea behind the LongestSpan approach is that the longer length of the role tends to have more meanings than the shorter role.
- **WordNet** is an electronic dictionary database [Mil98]. This method marks span in WordNet as a generic role. Specific roles, however, are those which do not exist in the WordNet. The most frequent specific role in the cluster will be selected as the most important center role in the cluster. If two roles are tied in the frequency, the role with longest length will be selected.
- **NameEntity** is a method that marks specific roles as those who exist in the name entity list generated from latent dirichlet allocation. The most frequent specific role in the cluster will be selected as the most important center role. If two roles are tied in the frequency, the role with longest length will be picked. This approach highly relies on the accuracy of name entity recognition.

Figure 3.3 shows the coreference resolution clustering result using WordNet. Pink nodes indicate specific roles while red nodes indicate they are important center roles.

Equipped with the information of coreference resolution clusters, applying semantic role labeling can extract a forest graph from the news dataset to better understand the behavior of subject, verb, and object.

3.3 Challenges

For now, searching for *Trump* will create a forest graph result whose roots are *Trump*, *Donald Trump*, and *The President*. However, inside the second verb layer of each tree, there may

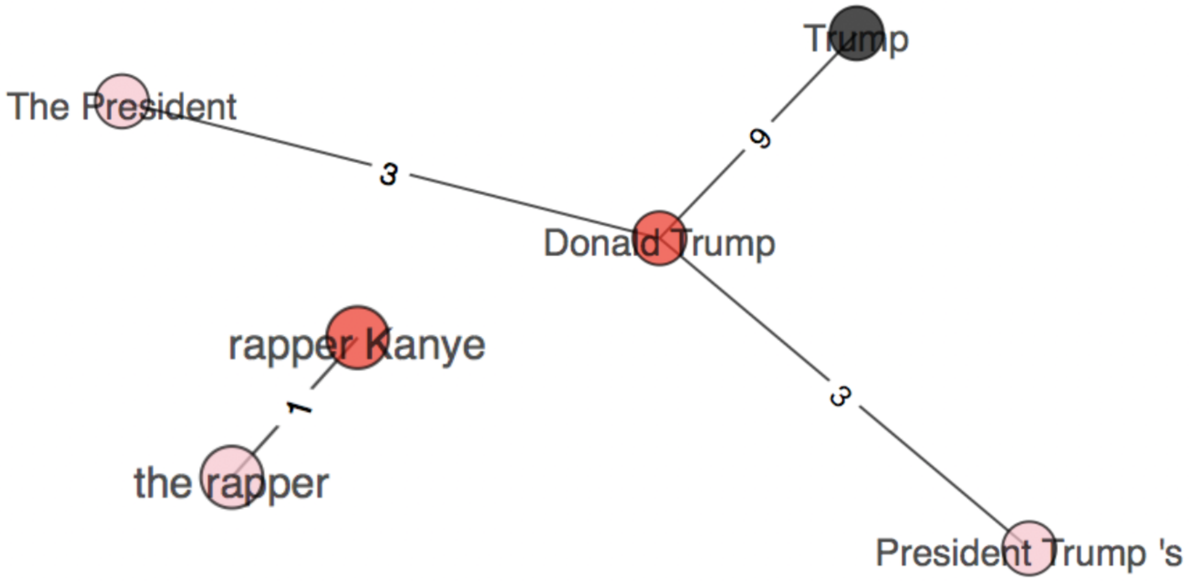


Figure 3.3: Center Roles for Coreference Resolution Clusters using WordNet Method

exist multiple objects sharing the same or similar meaning. There are often cases when two verbs under a certain subject share similar meanings as well. Our challenge is how can we merge verbs and objects reasonably.

3.3.1 Merging Objects under A Certain Verb

There are cases when under a certain verb, say *escalated*, there are two similar objects *his war on football players* and *his war of words with the Philadelphia Eagles* sharing the same meaning (Figure 3.4). It would be overwhelming to analyze what happened in the news dataset if we keep all of them inside the forest graph. In this section, we merge objects under a certain verb and then sum up their frequency weight on the edges to form a new node and a new edge respectively (Figure 3.5).

The way we merge object arguments is by $TF-IDF$, a measurement of how important a term is in reflecting its document in the corpus. From the equation 3.1, we denote $O(t,d)$ as the number of times term t occurs in document d. Term frequency $TF(t,d)$ can be generated

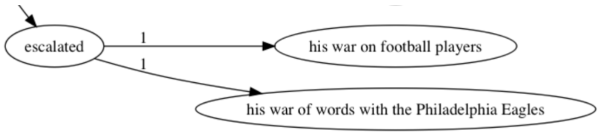

escalated
1
his war on football players
1
his war of words with the Philadelphia Eagles

Figure 3.4: Result **before** Merging Objects

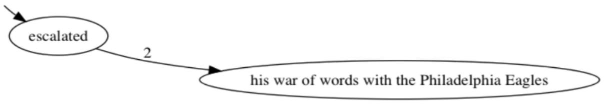


Figure 3.5: Result **after** Merging Objects

from $O(t,d)$ divided by the overall number of words in document d. Intuitively, the more representative the term t is to the document, the higher $TF(t,d)$ score it can achieve.

$$TF(t,d) = \frac{O(t,d)}{\sum_i O(i,d)} \tag{3.1}$$

Inverse document frequency for term t in corpus C can be computed through equation (3.2). $|C|$ denotes the number of documents in corpus C, and $|d \in C : t \in d|$ is the number of documents in corpus that has term t in it. In order to avoid the error of dividing by zero, we add +1 to the denominator. The basic idea behind this is that inverse document frequency $IDF(t,C)$ measures how well a term t can represent a topic. If there are a fewer number of documents containing this term t, this term can achieve higher $IDF(t,C)$.

$$IDF(t,C) = log(\frac{|C|}{|d \in C : t \in d| + 1}) \tag{3.2}$$

Then $TF - IDF$ score can be computed by the multiplicaiton between term frequency and inverse document frequency.

$$TF - IDF(t,d,C) = TF(t,d) \cdot IDF(t,C) \tag{3.3}$$

After calculating $TF - IDF$ score for each object role under the same verb, we set up a threshold to capture the importance between roles and merge them together if the difference between their $TF - IDF$ score falls in the threshold. Intuitively, the term t with highest $TF - IDF$ score will be chosen to represent the terms with similar meanings. When they are tied in $TF - IDF$ score, the one with the longest length will be selected as longest length tends to be more detailed in meaning empirically.

3.3.2 Merging Verbs under A Certain Subject

There are verbs like *believe*, *say*, *think* that have similar meanings in Figure 3.6. Even though they differ in spelling, all of the three verbs convey the meaning of the subject holding a certain view. If we can merge such verbs together by similarity, we can possibly clean up the mess and focus more on the verbs with different and important meanings. In this section, we merge verbs under certain subjects by word embedding, a method which can show the similarity between every two verbs under a certain subject.

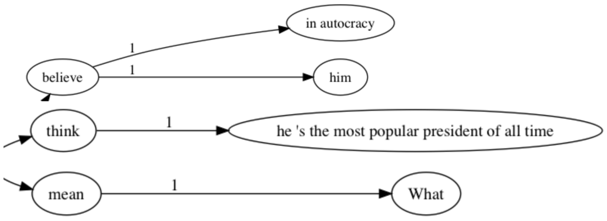


Figure 3.6: Result **before** Merging Verbs

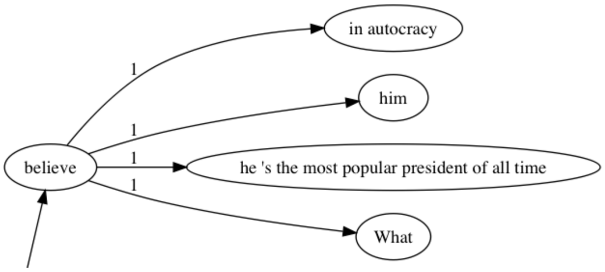


Figure 3.7: Result **after** Merging Verbs

Word embedding can be considered as a word to vector mapping. Here we are using the word embedding produced by word2vec [MSC13]. In the high dimension, word2vec puts similar words nearby in space. Using cosine distance, we can get the similarity between words under certain subject in a convenient manner.

After calculating similarities among every verb under a certain subject, a distance threshold is set up to capture the similarity and to merge them together if the cosine distance falls in this threshold. This merging process is also conducted by union find algorithm. The figures 3.6 and 3.7 present the result of merging verbs $\{believe, think, mean\}$ under the certain subject *Trump*. After merging, the result in Figure 3.6 will be gathered together as Figure 3.7 shows.

3.3.3 More than Subject, Object and Verb

In this previous sections, we mainly focus on semantic arguments like subject, object and verb. However, other parts including modifier, negative, location also play important roles

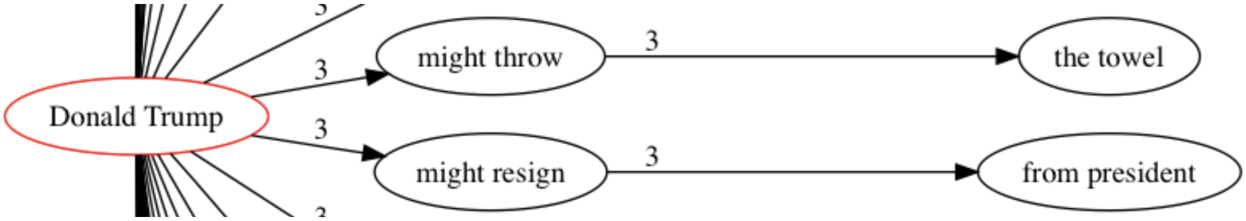


Figure 3.8: Result after Adding Modifier to Verbs

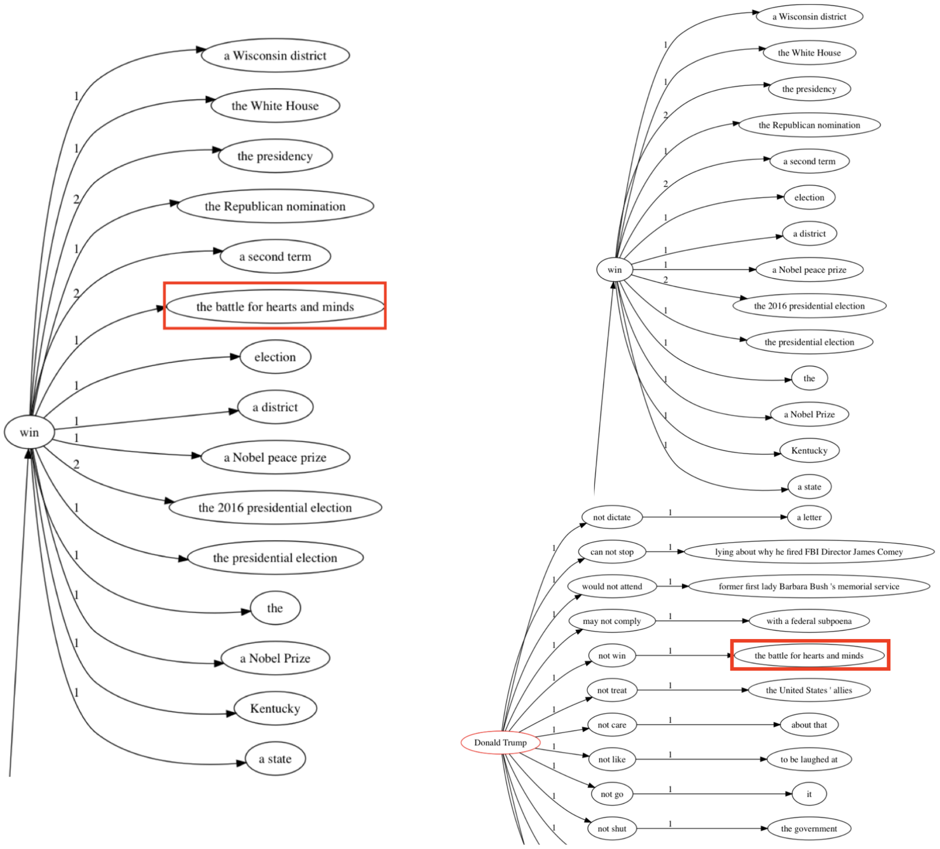
in understanding the news. Some of them can even change the meaning of the news. How to deal with these vital roles remains a question in the forest graph generated by semantic role labeling and coreference resolution.

3.3.3.1 Modifier Changes the Meaning of Verbs

Previously before taking modifier into consideration, the only known information through subject and verb is that a certain subject performs certain action. However, there might have several ways of taking this action. They may intend to take the steps, have already conducted the behavior or just claim there is no possibility of doing that. This uncertainty is measured by modifier. Take Figure 3.8 as an example, by adding modifier, instead of the old news *Donald Trump resign from president*, which is not true for sure in year 2019 even with grammar correction, a completely different news saying *Donald Trump might resign from president* becomes more authentic for a news in 2019.

3.3.3.2 Negative Preserves Sentiment of the News

To preserve the correctness of the sentiment the news convey, we add negative to the verb as an extra sentiment information. According to Figure 3.9, before adding negative roles, it is misleading to see *Donald Trump win the battle for hearts and minds.* The real key sentence in the news is *Donald Trump not win the battle for hearts and minds.*



(a): Before Adding Negative

(b): After Adding Negative

Figure 3.9: Result before and After Adding Negative Role

3.3.4 An option: Lemmatized Verbs

Lemmatization focuses on verbs with tense like *winning* and *won*, which they share the same stem. Merging them together using the lemmatized result can merge the duplicate information together to some extent. Here by $TF - IDF$ approach referred in section 3.3.1, we can later on merge the objects *presidential election* and *the 2016 presidential election* together in Fugure 3.10.

However, one shortback of lemmatization is that lemmatized verbs lose information of tense. According to the lemmatized result (b) in Figure 3.10, we lost the information of whether the subject is winning or has already won the election. Therefore whether lemmatize a verb or not remains an option in the configuration for users to decide according to different circumstances.

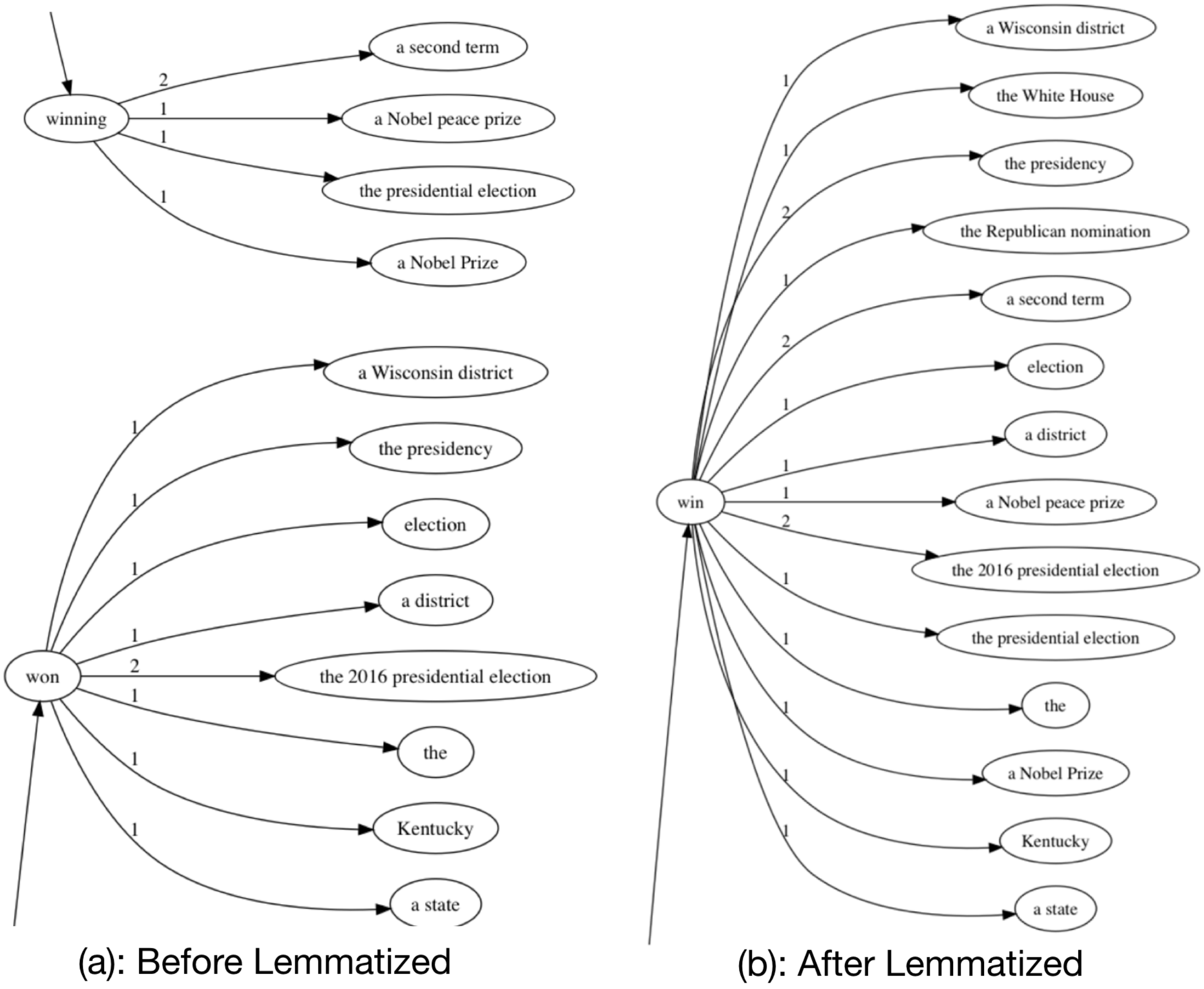


Figure 3.10: Result before and After Lemmatized Verbs

3.4 Chapter Summary

In this chapter, we described the format of two dataset we used, a topic specified Trump dataset and a general 6 months dataset. We also presented methodology applying semantic role labeling to extract key entities information including subject, verb and object from news. With the help of coreference resolution, a tree based graph demo evolved into a forest graph visualization. This greatly improves the understanding of the structure and the key point of a piece of news and provides a global view towards the entities from the whole dataset.

To deal with the duplicates meaning within objects and verbs, we adopted $TF-IDF$ and word2vec approaches to objects and verbs respectively as criteria of merging algorithm as well as the trick of union find for merging verbs. However, semantic roles are more than just subject, verb and object. Among those important semantic roles, we focused on modifier, negative for verbs and discussed on the problem whether we need lemmatized verb or not.

CHAPTER 4

Experiment Analysis

This Chapter presents the experiment findings and analysis on extracting semantic roles information from the news dataset.

4.1 Action Tracking on Verbs

The forest graph generated by previous chapter provided a unique prospective on tracking the actions of a certain subject. In this section, we apply the approach to 6_months dataset from November 2018 to March 2019 on taxonomy (category): */sports/basketball*. The subject we are interested in is one of the basketball stars *LeBron James*.

For each month, we generate top ten frequent verbs by sorting the label weights on the edges between subject *LeBron James* and its verbs. Through the 5 months, there are following verbs appearing at least 4 times in the tree graph. $Top\ frequent\ verbs = \{Leave, Take, Score, Make, Have, Miss\}$, among which *take*, *make* and *have* are abstract verbs that tend to have different meanings in different scenarios. Take *have* as an example, *LeBron James* ***have*** *nothing to prove to anyone*, *LeBron James* ***have*** *a rough debut season* and *LeBron James* ***have*** *30 points, 12 assists and 10 rebounds* are major key information we observed under the verb *have*.

For the rest $Specific\,Top\,Frequent\,Verbs = \{Leave,\ Score,\ Miss\}$, they all have fixed major objects. *LeBron James* ***leave*** *the Cleveland Cavaliers*, *LeBron James* ***score*** *points* and *LeBron James* ***miss*** *throws* in November while *LeBron James* ***miss*** *games* in the rest of months. A figure is drawn below to show the changing of actions through time.

According to Figure 4.1, y axis represents the reverse rank of each verb. The three

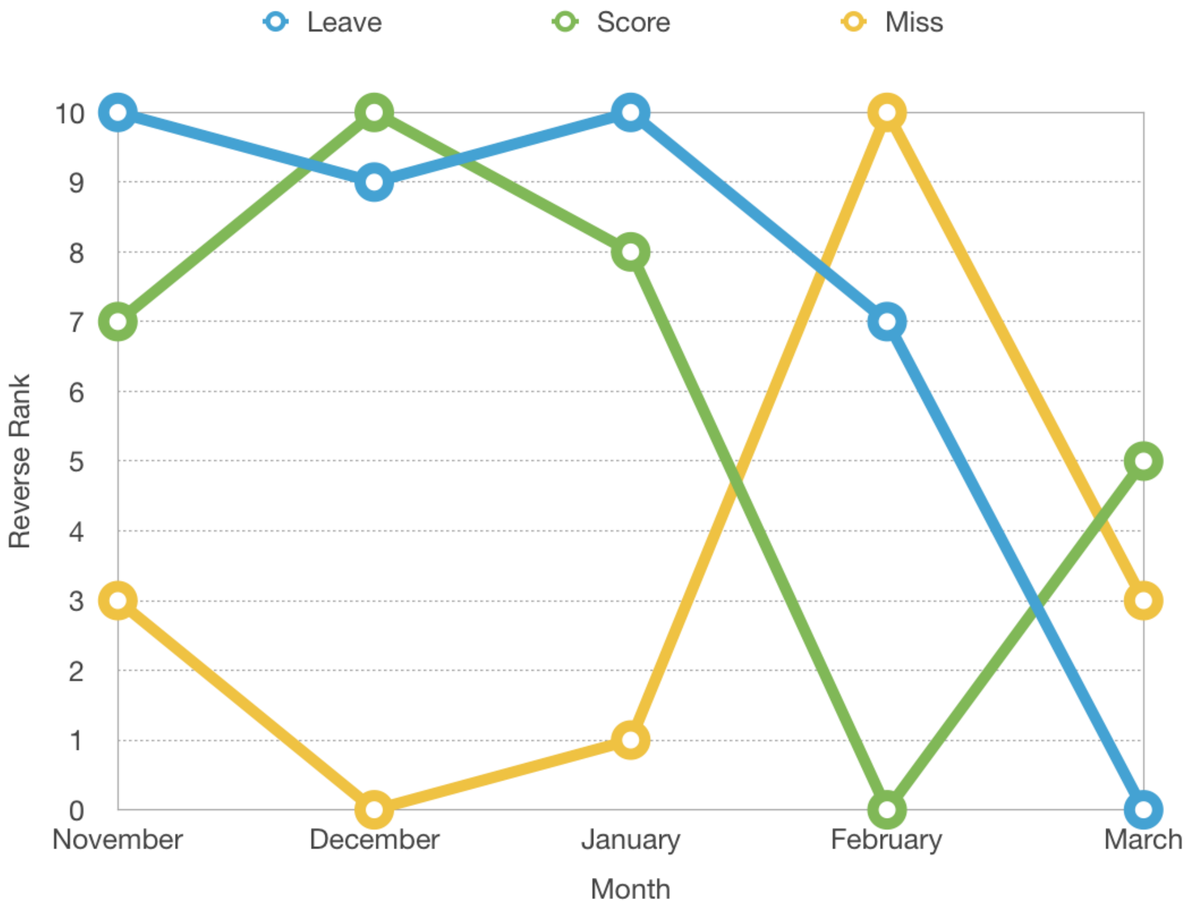


Figure 4.1: Action Tracking for *LeBron James*

specific top frequent verbs have 1-based ranking in the top frequent ten verbs per month. For visualization purpose, we assume if the verb does not exist in the top ten frequent verbs list in a month, its ranking for that month will be set as 11. In the Figure 4.1, we plot $11 - ranking$ as the reverse rank for the three verbs every month.

Figure 4.1 indicates that, the media has been reporting on *LeBron James*'s leaving from a basketball team called Cleveland Cavaliers since November. However, the frequency began to fall in February 2019. In the meaning time, instead of a few news on missing throws in November, news on *LeBron James* ***miss*** *games* ranked first in February. Besides, *LeBron James* ***score*** *points* was not even reported in February. The lines in the figure depicts how the leaving action disappeared and how the missing games action emerged.

To explain what happened to the missing game action in February, the top 5 frequent verbs are listed below. Verbs that appear only in January and February is marked in red.

ranking	**verbs for *LeBron James***	**fixed main objects**
1	miss	games
2	suffer	a groin strain injury
3	make	no fixed main objects
4	leave	Cleveland Cavaliers
5	lead	the team

Table 4.1: Verb Rankings for *LeBron James* in Feburary

From Table 4.1, one biggest event happened to *LeBron James* in Feburary was suffering the groin strain injury. This can well explain why his action transfered from scoring to missing the game. Therefore, extracting semantic roles from news is able to track actions on verbs for certain subjects along time.

4.2 Breaking News Tracking on Objects

Moving from verbs to objects, breaking news could also be tracked. For 6_months dataset, we run our semantic role labeling on Janurary to April, 2019 on taxonomy (category): */sports/basketball*, which has overall 75,827 peices of news. The input of those news is title description. For every month, we search the subject for *Lakers*, a basketball team at Los Angeles. And we sum up all the label weights on the edges between verb and object. We denote $W(V, o|S = s)$ as the sum of weight on edges between all the verbs $v \in V$ and a specific object o under certain subject s.

$$W(V, o|S = s) = \sum_{v \in V} W(v, o|S = s) \tag{4.1}$$

For every interested object i in interested objects $IO = \{Davis, James, Game, Ariza, Others\}$ which have top most $W(V, i|S = s)$, we draw the pie chart according to the percentage P computed as follows.

$$P(i) = \frac{\sum W(V, i|S = Lakers)}{\sum_{o \in O} W(V, o|S = Lakers)} \tag{4.2}$$

In the equation (4.2), top interested Objects IO is a subset of all objects O under subject $Lakers$.

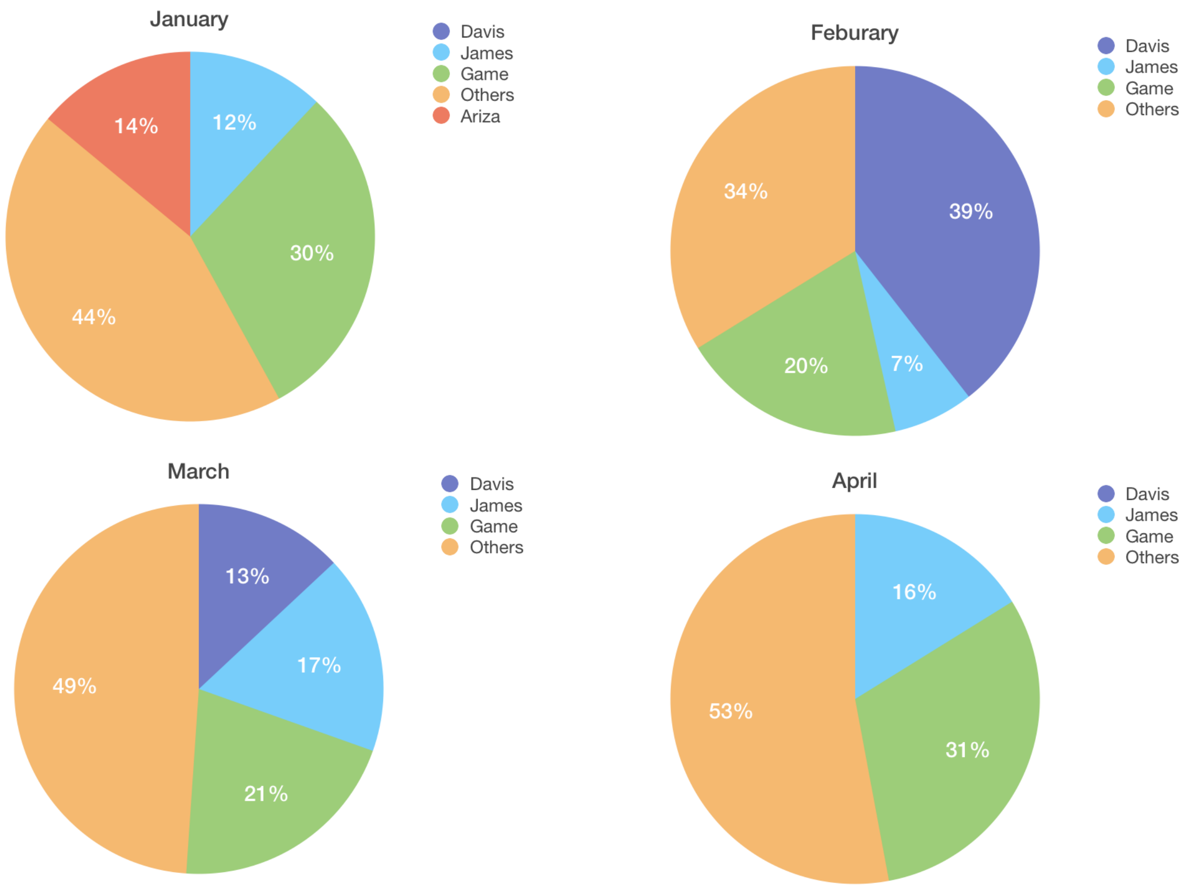


Figure 4.2: Breaking News Tracking on Trade Rumors

From the figure 4.1,we could find out that *LeBron James* has always been a role related to *Lakers* as he is a super star in the team. However, according to Figure, we can also observe that *Anthony Davis* in purple has played an important part in the news related to *Lakers* since Feburary. He was not even appear in Janurary but the number of news mentioned him and *Lakers* emerged and even beat *James* and *Lakers* in Feburary but gradually decreased in March. His breaking news disappeared completely in April. The event happened in Feburary and March was the trade rumors on *Anthony Davis*. He annouced intention of being on the trading market and rumors said that *Lakers* wanted him and would possibly trade him with some players. This affected the performance of Lakers' young players and *Lakers* lost some matches. After the trade deadlines, the topic eventually disappeared.

Intuitively, the weight on the edges of forest graph represents the frequency of how often this news is reported. Therefore, semantic role labeling is capable of visualizing the trend in news articles from the level of semantic roles.

4.3 Chapter Summary

In this chapter, we have discussed about the experiment we run on the dataset. We can track actions for individuals and the breaking news by the frequency weight on edges between subject and verb and edges between verb and object respectively. Basketball is just an example and this can be widely applied to non-specific domain of news.

CHAPTER 5

Conclusion and Future Directions

5.1 Conclusion

In this thesis, we extracted global entities information from news dataset using natural language processing technique. This research proposed a demo application to visualize the key information of news data including semantic roles like subject, verb and object. This method provided a new perspective towards news data and was able to observe action tracking for fixed subject, breaking news tracking and specific characters of semantic roles. By extracting global entities information from news in this way, accessing information from news can be more efficient.

5.2 Future Directions

Due to the limitation of time, this work mainly focused on semantic roles like subject, verb and object. Despite the fact that we solved the challenges of modifier and negative, there are still other important roles in sentences like adjectives, adverbs, location and time. One future direction could lie in adding more semantic roles and layers in the forest graph to extract more information from news. How to select those semantic roles could also be a question.

Currently we added subject roles by coreference resolution. However, the best coreference resolution result nowadays is still not satisfying enough to do the preprocessing for the whole dataset for semantic role labeling. Another future directions could be improving the coreference resolution result and applying it directly to the news dataset as preprocessing to remove all the pronouns in the news data. This can also replace the pronouns with specific entities for semantic role labeling result. Or another possible thought is that we can utilize other knowledge base by applying entity linking to merge subject roles together.

Another problem is that in this work, semantic role labeling was also capable of labeling roles with long length. This is because there is no limitation on length during the process of semantic role labeling. In this thesis, we provide a cofigurable length for user to prune. However, there are some possible ways to deal with such long roles like extracting the head words for object arguments. This could be another future direction as well.

REFERENCES

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation." *Journal of machine Learning research*, **3**(Jan):993–1022, 2003.

[GGN18] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. "AllenNLP: A deep semantic natural language processing platform." *arXiv preprint arXiv:1803.07640*, 2018.

[HLL17] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. "Deep semantic role labeling: What works and whats next." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 473–483, 2017.

[LHL17] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. "End-to-end neural coreference resolution." *arXiv preprint arXiv:1707.07045*, 2017.

[Mil98] George Miller. *WordNet: An electronic lexical database.* MIT press, 1998.

[MSC13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[Sin18] Sonit Singh. "Natural Language Processing for Information Extraction." *CoRR*, **abs/1807.02383**, 2018.

[ZSY18] Rui Zhang, Cicero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. "Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering." *arXiv preprint arXiv:1805.04893*, 2018.