CSCI699 Introduction to Information Extraction: Project mid-term report

Akira Matsui
Student ID: 5421284665
amatsui@usc.edu

1 Introduction

In this mid-term project report, I will describe the problem setting, the data, the methods, and the preliminary results. Also I will conduct error analysis. This project aims to find the importance of social science terminologies in policy decision making.

I hypothesize that the policymakers utilize economics concepts in the complicated way when they make difficult decisions such as policy rule changes. In other words, I assume that the way of using those terminologies in policy making is complicated and in a frequent manner.

To test this hypothesize, I use FOMC (The Federal Open Market Committee) minute data from 1967 to 2018 and the economics terminology dictionary to make the text dataset with economics terminologies. Also, I use the historical data of Federal Funds Rate (FFR) and Monetary Base (MB) to see the usages of the terminologies reflect the policy changes.

To extract the information of the usages of the terminologies, I use dependency parser, focusing on the dependency of terminology in each sentence. Also I compare these dependencies with the policy indicators.

Although the dependency parsers return the robust results, these parsers do not correctly extract the dependency of compound words such as fiscal policy. I plan to develop the model to address this drawback in the rest of the project.

2 Data

In this section, I will give a brief explanation of data collection and describe the data. I crawl the FOMC minute from 1967 to 2018 and the economics terminology dictionary.

FOMC minute is the minute of the FOMC meeting. In most cases, at each time after a committee, the monetary policy decision become open to the public. FOMC minute plays the pivotal role in monetary policy. FOMC minute contains not only the detail of policy decision making but also the perspectives and the outlook of the economic situation.

To have the FOMC minute datasets, I use the existing datasets that contains the minute from 1967 to 2007. In addition, I wrote a crawler to get the minute from 2008 to 2018. Also, I collect the economic terminologies from the website that gives brief explanations of basic economics concepts, The Economics Classroom. This website is run by the professor in Economics.

The Figure 1 shows that text length of minutes over the entire period. This plot describes the length of minute increases over time. The striking point is between 1992 and 1993 periods. The FRB changed the style of the minute and they started to publish digital version minutes from 1993. This fact must be considered in the data analysis since the length of text and the minute style might affect the usage of terminologies.

To find the relation between the economic terminology usages and the monetary policy changes, I use the historical data of Federal Funds Rate (FFR) and Monetary Base (MB). FFR and MB are the important monetary policy operation tools and the FRB change them to conduct their policy packages. Therefore, we can assume that the changes in these policy tools can be interpreted as the changes in monetary policy. The Figure 2

1https://www.federalreserve.gov/monetarypolicy/fomc_historical_year.htm
2I will discuss this in the related research
3https://stanford.edu/~rezab/useful/fomc_minutes.html
4https://econclassroom.com/
5The detail about monetary policy operation and its tool are points discussed in the review paper.
and 3 shows the percent changes of FFR and MB. The clear change of the monetary policy was made after the financial crisis in 2008.

3 Methods

I firstly extract the sentences with economics terminologies in FOMC minutes. Any sentences with at least one terminology in the dictionary are extracted in this process. Next, I use dependency parsers to extract the dependency information of each terminology in the sentences. To check the robustness of the parsers, I use two well-known dependency parsers, Stanford CoreNLP and spaCy. To see how each terminology depends on other words in the sentence, I collect the children of the terminologies in dependency. To see the relationship between the policy indicators and terminology usages, I compute the correlation among them.

4 Results

Firstly, I plot the economic terminology counts in the minutes overtime. To take into account the fact that the text length of minute varies among the minutes, I divide the terminology counts by the text length of the minute. Fig 4 shows that the frequency in use economic terminology increases over time. Not only, frequency, Figure 5 describes that the variety of economic terminology increases over time. This is consistent with the fact that the transparency of monetary policy has been emphasized and the monetary policy became complicated recently.

Next, as we hypothesized, if the policymakers let the economic terminology to explain the economic condition and their decision, the usage of terminology become complicated. To check this hypothesis, I plot the total number of terminologies’ children in dependency parsing over time. Figure 6 shows that the number of dependencies increases over time. I use two dependency parsers, Stanford CoreNLP and spaCy, and both of them return the robust results.

Lastly, I compute the correlation between the monetary policy indicators and the number of terminology dependencies. Fig 7 shows the correlation heatmap among variables and shows that there are weak correlations between the number of dependency and the monetary policy indicators.

5 Remaining errors

There are two rooms to be improved in the rest of project: 1) the methods to evaluate the results with policy indicators and 2) the dependency parsing. In this report, I do not use the statistical model for time series data even though the data used in this project is time series data. To consider the typical tendency in time series data might improve the result. Secondly, I will modify the dependency parsing so that I can extract how the compound word are used in the dictionary. For example, Figure 8 shows an example that a compound terminology are not considered as one word. Also, I will modify the model to extract the terminology usage other than counting the number of children in dependency parsing.

6 Other

I changed the direction of my project. In the project proposal, I planned to use the relation extraction to extract how the terminologies are used in the minute. However, as I process the data, I found that there are not many relation expressions of entities in the data set. Therefore, I changed to use the dependency parser to extract the usage of the terminologies.

7 Conclusion

In this report, I describe the problem settings of my project, the data, the methods. I hypothesized that the way of using economics concepts in minutes reflects the policy changes. I used the dataset of FOMC minute and the economic terminology dictionary. To find how the policymakers use the economic concepts in FOMC, I utilize the dependency parsers. Also, I give preliminary results with the existing model and error analysis of the results. The preliminary results show that the frequency and variety of economic terminologies in the minute increases over time, which is consistent with the fact I discussed in the review paper. In the rest of the project, I will 1) sophisticate the evaluation method to see the terminology usage reflect the policy change and 2) augment model so that the model so that it can extract the more information of terminology usages.
Figure 1: The number of words in each minute.

Figure 2: Effective Federal Funds Rate (percent change).

Figure 3: St. Louis Adjusted Monetary Base (percent change).

Figure 4: Economics terminology per words. (the number of economic terminology divided by the total number of words of the minute)

Figure 5: The variety of economics terminology. The number of different terminologies in each minute.

Figure 6: The total number of children dependencies of the economic terminology in dependent parser (divided by the total number of words in the minute).

Figure 7: The heatmap of correlations among the monetary policy indicators and the number of terminologies’ dependency.

Figure 8: The error example for a compound terminology word. fiscal policy