「論文」

How Have Political Interests of U.S. Presidents Changed?: A Diachronic Investigation of the State of the Union Addresses through Topic Modeling

Naoki KIYAMA

Abstract

In this study, I demonstrate that the topic modeling method, more specifically the latent Dirichlet allocation (LDA), is a useful method to investigate how U.S. presidents' political concerns have changed since the foundation of the United States. By applying the LDA to State of the Union Addresses, I obtain four main topics ((i) internal issues related to the federal government; (ii) international affairs tied with territorial disputes; (iii) worldwide warfare; and (iv) social welfare) and argue that each topic is a reflection of its historical background. Last, I proposed that a transition in the key topics occurred before and after World War II. In other words, the Presidents' main political concerns were influenced by whether the United States had attained the status as the world leader.

1. Introduction

There are some speeches that a U.S. president has to give. One of these is the State of the Union Address, an annual speech that the President gives to a joint session of the U.S. Congress regarding his political concerns. In the United States Constitution, the Address is described as follows:

(1) "He [The President] shall from time to time give to the Congress Information of the State of the Union, and recommend to their Consideration such Measures as he shall judge necessary and expedient." (Article II, Section 3 of the United States Constitution)

While this address has been given ever since George Washington was inaugurated, there are some characteristics to be noted. First, although most of the addresses are

given in the form of a speech, those in the 19th to the early 20th centuries were conveyed as written reports. Secondly, the U.S. Constitution directs the president to convey his message to Congress, but the spread and development of broadcasting technologies like radio and TV, seem to have changed the targeted audience. That is, before the development of these technologies, the audience was just members of congress. However, since Warren Harding's (1921−1925) address in 1923, which was the first State of the Union Address that was broadcast to the public via the radio (Kaid, 2007: 696), the targeted addressees have included not only political members, but also millions of Americans. Last, although the addresses are expected to be given once a year, in the case of the death of a president or if a president had to resign before his report of the year, the new president would give his own address. For example, William Henry Harrison (1841−1841) passed away a month after his inauguration, hence he did not give a State of the Union Address. Instead John Tyler (1841−1845), who took over the presidency, gave his own speech.

As described above, there are some difficulties in analyzing the State of the Union Addresses. Nonetheless, public speeches adhere to certain topics or themes to convey some messages to an audience. Such topics may not be made explicit, but native speakers can understand what the addresser is speaking about. This is because a "topic is the most important situational factor influencing vocabulary choice; the words used in a text are to a large extent determined by the topic of the text" (Biber and Conrad, 2009: 46). Therefore, scrutinizing how words are used in the State of the Union Addresses may reveal the main topics of the speeches or the main political concerns of the presidents. Following this assumption, I have two goals for this study: (i) what were/are the main political interests of the U.S. presidents? That is, how have the legislative concerns changed over two centuries?; (ii) what causes the changes of interest? I will argue that, with one of the topic modeling techniques called the latent Dirichlet allocation (LDA) proposed by Blei et al. (2003), the presidents' political concerns are divided into four topics, and the turning point is strongly influenced by world-wide wars.

2. Literature Review

The U.S. State of the Union Addresses have attracted a lot of attention from

researchers (Bonnefille, 2008; 2013; Crockett and Lee, 2012; Herz and Bellaachia, 2014; Tung, 2014; *interalia*). Previous literature aimed at investigating addresses from various areas of expertise, and many of them used quantitative techniques. Although there is no previous research that I have seen that investigated my specific questions, some additional studies are worth considering. Thus, this section summarizes two important previous studies and emphasizes the importance of having a diachronic perspective.

2.1 A text-mining approach to the Addresses

First, and most relevant to us, a topic analysis was done by Crockett and Lee (2012). They examined speeches from 1989 to 2011. In these 23 addresses, they found seven topics associated with "war and terror" (which included words such as *Hussein*, *Saddam*, and *Soviet*); "economics and finance" (which included terms such as *bank*, *trillion*, and *small-business*); and "inspiration of the nation and its people" (which included *hopeful*, *homeland*, and *pension*). Note that they provided no further labels nor a result of their experiments for the rest of the topics obtained.

Tung (2014) conducted another text-mining study. His research aim was two-fold: to identify the overall lexical trends in the State of the Union Addresses, and to find differences in word usage between the two major political parties, namely Republicans and Democrats. For the first research question, he mentioned that because the data collection ranges over 200 years, discovering word usage patterns in the addresses was difficult. However, when focusing on the addresses given from 1961 to 2014, some tendencies were observed. That is, Republicans used words related to external issues, such as *war, terrorists, Iraq*, and *Saddam*, inferring that these emphasize international relations. On the other hand, Democrats tended to use terms related to economics and finance (e.g., *companies, businesses, payments*, and *employment*) and related to social affairs (e.g., *education, students, wage*, and *jobs*). In other words, three major topics that the presidents tended to speak about are war and terror, economics and finance, and social affairs.

2.2 Diachronic perspective

Though the studies reviewed in the last subsection are interesting and thought-provoking, I shall point out the necessity of a diachronic transition in topics, which has

been neglected in the previous literature.

In Crockett and Lee (2012), the seven topics they found seem to be certain time- or period-specific issues. For example, regarding the topic of war and terror, proper nouns like *Soviet* were certainly related to the Cold War, whereas the term *Saddam Hussein* was related to Hussein's trial for genocide, which made the headlines in the mid-2000s. In other words, while it is true that both words are related to the topic of war and terror, Saddam Hussein's execution was a decade and a half after the Cold War ended. The same can be said of Tung's analysis. His results showed that words frequently used by Republicans included *Iraq, Iraqi, terrorist(s), Saddam*, and *Hussein*, which all pertained to September 11, 2001 and the Iraq War. Words used by Democrats involved *Vietnam* and *Soviet*, referring to the Cold War. As Tung himself recognized, 9/11 and the execution of Saddam Hussein took place when G. W. Bush (2001–2009) was the president. Similarly, the worst period of the Cold War, the Cuban Missile Crisis in 1962, and the first half of the Vietnam War, were overseen by Democrat presidents. Hence, the wars and terror happened at completely different times.

Considering the previous literature and the diachronic approach to the addresses, studies mentioned above emphasize the importance of investigating the changes in word and phrase usage. Put differently, in order to investigate our research questions introduced above—namely, how the presidents' political interests have changed over time and what motivates them to have such ideas—I have to analyze topics that the presidents have expressed concern about, and the topics revealed must be examined in chronological order.

With this assumption, one of the topic modeling methods, the LDA proposed by Blei et al. (2003), was applied to the collection of the State of Union Addresses.

3. Methods

This section will introduce the corpus and method this study used. Note that all the statistics and figures are proceeded by R (3.4.3) except for concordance lines and simple frequency counts.

3.1 Corpus

In this study, I collected as many speeches as possible from various web

Table 1: The information of the UA Corpus

Number of files	Tokens	Types	STTR (per 10,000)
229	11,915,915	156,860	27.28

resources. Because the manuscripts were obtained from different web sites, many symbol and spelling variations were found. Thus, non-computer friendly symbols are replaced with the HTML format and the spelling variations are standardized into the current American English orthography. For example, some manuscripts mistakenly used hyphens instead of en dashes, so I changed them into en dashes, represented as <–>. Other symbols, such as quotation marks and ampersands, were also replaced with the HTML format, which was/is enclosed in angled brackets. Next, I obtained a corpus as described in Table 1, which was calculated by CasualConc (Imao, 2015). This corpus will be called the United States presidents' State of the Union Address corpus (abbreviated as the UA Corpus).[1]

3.2 Latent Dirichlet allocation

As already explained, the aim of the current study is to investigate how the topics in the UA Corpus have changed. To achieve this goal, I used the LDA proposed by Blei et al. (2003) because its basic premise perfectly coincides with our research question. As Blei et al. (2003: 996) stated, "[t]he basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words." This idea is quite compatible with corpus linguistics research, as I quoted Biber's and Conrad's explanation.

Another important hypothesis under this method is that the LDA is based on the "bag-of-words" assumption—word orders or document orders are ignored, and all words within a document are exchangeable. This means that even though the UA Corpus is compiled with diachronic data and the current research question is to investigate transitions in topics over time, the method itself has nothing to do with a diachronic perspective.

How, then, are the topics obtained? First, stop-words should be removed. Topics are what addressers are speaking about, and thus topics are basically composed of content words, not functional counterparts. Hence, I slightly modified a stop word list used in Tabata's (2017) study and removed the words. Then, the number of topics to be

computed, which is at researcher's own discretion, must be decided (for a generous introduction to the LDA, see Kuroda (2017), Tabata (2017) and Schöch (2016)). Thus, in this study, the number of topics were calculated by changing the number ranging from ten to seventy at every five intervals. Then, the sets of words contributing to each topic were scrutinized, and the set with 35 topics seemed to capture the distribution of topics.

Note that because the aim of this study is to uncover the diachronic change in political interests of the presidents, topics that were observed only in specific periods were discarded. For example, Topic 33 was observed during William McKinley's (1897–1901) and Theodore Roosevelt's (1901–1909) administrations, but no other presidents talked about this topic. It is doubtful that this topic represents a transition in political interests in the long term. Thus, such topics were excluded from this study.

The last thing to be mentioned in relation to the LDA with the current study is that the LDA calculates the degree to which the speech belongs to the topics. In other words, all the files belong to every topic, but they differ in how high their contributions were to the topics. It would be best if all the probability ratios could be used; however, because the current research interest was to uncover the political interests of the presidents, the topics that were considered as the most likely topics of the documents were used. What I mean by "the most likely topics" are topics whose probability rates as calculated in the LDA exceed 0.3 in various or stable periods. In this study, topics that lasted over 15 years were considered as the most likely topics.

Some readers may think that 0.3 in the probability rate is too low. However, the average of the overall probability is 0.002, and topics that exceed the threshold of the probability rate, appearing in various or consecutive periods, are quite limited. Thus, I think that my standard is appropriate, albeit slightly arbitrary.

3.3 Concordance lines and n-grams

After labeling each topic obtained from the LDA, this paper goes into further detail regarding what motivated the presidents to be highly concerned about the topics. This question will be proceeded by scrutinizing how words that highly contribute to the topics are used in addresses by focusing on quad-grams recurrently used with respect to raw-frequency. The reason why I used quad-gram, and not bi-, tetra-, or hexagrams is two-fold: semantic contents and frequencies.

First, the longer the collocations are, the more meaningful they are, and thus it is very easy to guess in what situation or context the presidents used them. On the contrary, short collocations are rarely meaningful enough to grasp the way the clusters are used. To be more specific, compare the hexagrams and bigrams of *American* attested in the UA Corpus, which are shown in Table 2. The table shows that the bigrams are too short and that hexagrams are very precise in their contexts. In this sense, longer collocations enjoy their reputation of being meaningful. On the other hand, the frequency information reverses their situations, in that the longer the collocations are, the less frequent they are. The largest number of the hexagrams in Table 2 is, at most, 6. Then, is it significant enough to conclude that the expression *of the claim of American citizen* is more important or frequent than other collocations? I would say that it is not, or at most, that it is quite a difficult question to answer. On the contrary, the bigrams show that *the American* is by far the most frequently found expression, and it should be safe to say that it is one of the most frequent collocations on the list. Thus, it is much easier to distinguish key phrases in shorter collocations than in longer ones.

These two mutually exclusive issues are resolved (but not without any issues) by assessing quadgrams. Thus, this study will focus on quadgrams, and show their concordance lines to explore the research question. Note that all the frequency counts on collocations and concordance lines presented in this and later sections are generated by CasualConc.

Table 2: Comparing frequencies of hexagrams and bigrams of *American*

No.	Hexagram	Freq.	Bigram	Freq.
1	of the claim of *american* citizen	6	the *american*	872
2	the live and property of *american*	5	of *american*	420
3	live and property of *american* citizen	4	*american* people	390
4	the bureau of the *american* republic	4	*american* citizen	195
5	the central and south *american* state	4	every *american*	122
6	the character of the *american* people	4	a(n) *american*	101
7	the claim of *american* citizen against	4	to *american*	72
8	the spirit of the *american* people	4	*american* republic	64
9	a decent home for every *american*	3	*american* family	50
10	*american* citizen against the government of	3	*american* vessel	49

4. Results and discussion

Based on the processes explained in the previous section, I obtained four topics: 6, 10, 27, and 34.[2] A smoothed diachronic figure is given in Figure 1, which illustrates that Topics 6 and 34 were the major concerns of the presidents at the beginning of the United States. However, as their importance decreases, Topics 10 and 27 gain their prominence, and the transition of Topic 27 is more radical than that of Topic 10.[3] Then, our question is, what does each topic represent? This is well-represented through a word cloud, using words whose word weight is 0.0008 or higher, and such words are plotted in Figure 2 to 5,[4] representing each topic; the following sections will deal with the topics in turn.

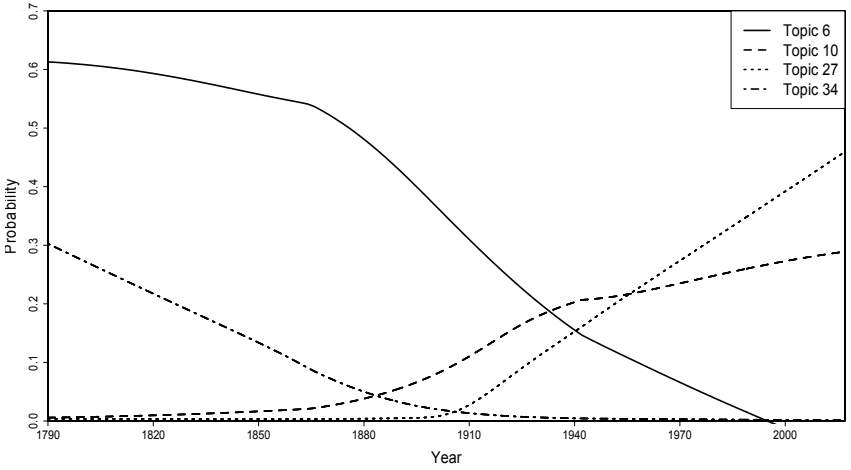

Figure 1: Diachronic transition of the topics

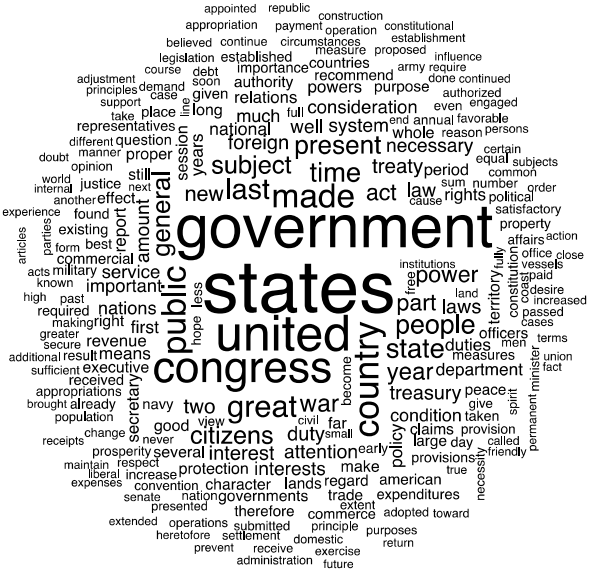

Figure 2: Word cloud on Topic 6

Figure 3: Word cloud on Topic 34

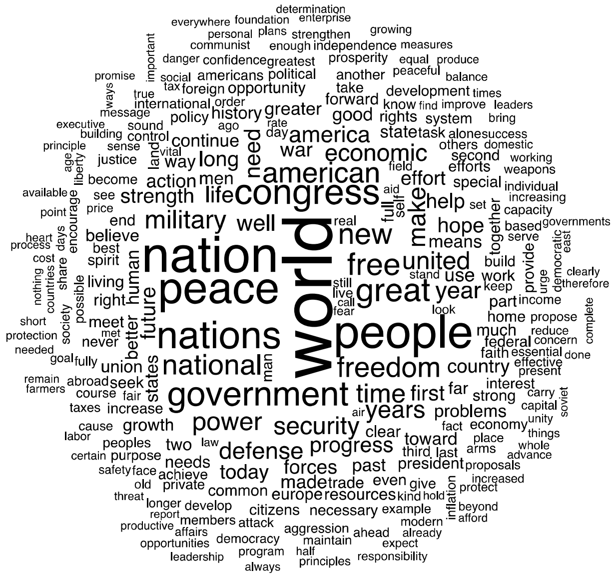

Figure 4: Word cloud on Topic 10

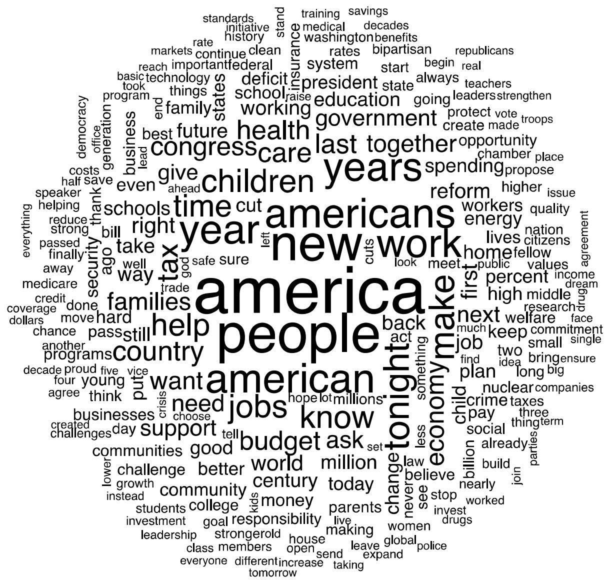

Figure 5: Word cloud on Topic 27

4.1 Topic 6 (1790−1920s)

This topic was considered to be a quite important topic at the beginning, as no other topics received more than 0.5 in Figure 1. Figure 2 shows that words related to domestic affairs, such as *congress, right(s), law(s), constitution(al), territory, duty, duties, government(s), country,* and *parties*, are frequently observed in this topic. The appearance of these words indicate that the topic is related to politics. Then, the question arises: whose politics are they, the United States' or the international community's?

In order to understand the topic more precisely, I looked at how these words were used.[5] One of the most frequently used words in the list above was *right(s)*, which appeared nearly 1,500 times out of 2,217 total frequencies in this period, and its most frequent quadgram *the right of the*, was found mostly in this period (64 times out of 69 times in total). As the concordance line in Figure 6 shows, words that follow the collocation, namely the possessor of the right, are either political organizations such as *(united) states, government*, and *party*, or those related to citizens such as *people, Indian* and *minority*. These collocational patterns indicate that the rights referred to in this period were related to domestic affairs. Expressions related to U.S. internal situations in this topic were found not only in the collocation of *right*. Let us consider two more salient examples, *the constitution of the*, the second-most commonly used quadgram of *constitution*, and *the duty of the* or *the duties of the*, the most frequent quadgram of the lemmatized form of *duty*, in turn. Figure 7 shows the linear sequence of *the constitution of the*, and it illustrates that the noun phrases following the quadgram were dominantly occupied with either *united states* or *states*, which were mostly found in the period between the 1700s and 1920 (26 times out of 28). Figure 8 also exemplifies presidents' interests, in that the presidents paid special attention to what the U.S. government or the president must work on, because the *of* phrase mostly

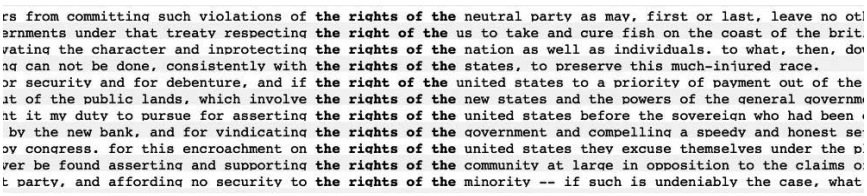

Figure 6: The concordance lines of *the rights of the*

important state of north carolina to **the constitution of the** united states (of which offi
ise the legislative power granted by **the constitution of the** united states "to lay and co
ity to preserve, protect, and defend **the constitution of the** united states", on you, gent
us depository of american happiness, **the constitution of the** united states. let them cher
the constitution of the judiciary, experimental and
mblies deriving their authority from **the constitution of the** state. each is sovereign wit
he inhabitants of the united states. **the constitution of the** united states requires that
ny would be vain and ridiculous. but **the constitution of the** united states imposes on the
by **the constitution of the** united states it is provided
ns, to which they are entitled under **the constitution of the** united states. they should h
e conformable to the requisitions of **the constitution of the** united states, i recommend t
the constitution of the united states provides that

Figure 7: The concordance lines of *the constitution of the*

rency it has been assumed that it was **the duty of the** executive not only to suppress insur
is, should be persisted in, it will be **the duty of the** united states to resist its executio
ey have taught. in the meantime it is **the duty of the** government, by all proper means with
which has so long existed and render **the duty of the** president plain in executing its pro
ect a member of congress, it shall be **the duty of the** president to cause a census of the i
is arises from an obstacle which it is **the duty of the** spanish government to remove. whilst
guidance and protection. whilst it is **the duty of the** president "from time to time to give
positions, prepare the way. i hold it **the duty of the** executive to insist upon frugality i
itution of the united states makes it **the duty of the** president to recommend to the consid

Figure 8: The concordance lines of *the duty|duties of the*

included *president, government,* or *united states*. Note that the phrase *the united states* was used in one of two ways, either referring to the country (namely, the United States) or to two or more states being in association with each other (hence, united states).

All of these collocations lead us to conclude that Topic 6 can be labeled as the internal issues of the federal government. It is relatively straightforward as to why this topic was the main concern of the presidents; the country was founded in 1776, and the government had to develop a constitution and laws making the President's role explicit to the U.S. Congress. Therefore, the presidents frequently referred to the topics related to the federal government.

4.2 Topic 34 (1790s−1870s)

Topic 34 was salient in the first century of the United States. Figure 3 shows words that contribute to making up this topic, which show some tendencies. That is, this topic includes words pertained to regions, nationalities, and races such as *Britain, Spain, French, Florida, tribes* and *Indians*. Furthermore, we can find words related to water, such as *lake, ocean, waters*, and *navigation*. Then, does this topic consist of two different topics?

Let us take a look at a few of the words listed above. The collocational patterns of

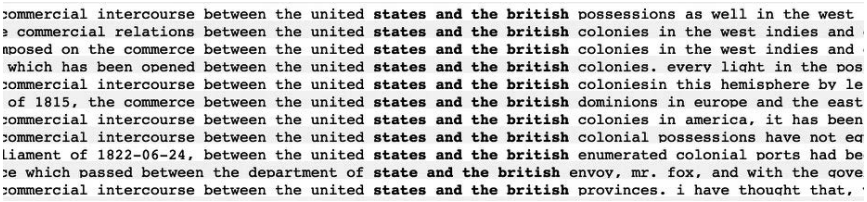

Figure 9: The concordance lines of *states and the British*

Figure 10: The concordance lines of *between Spain and*

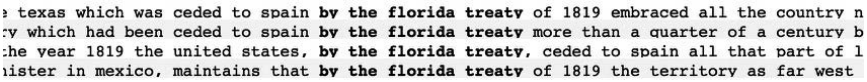

Figure 11: The concordance lines of *by the Florida treaty*

the country names reflect U.S. history. Figure 9 shows one of the frequently used collocations of *British*, and as it reads, expressions such as *colonies*, *possession*, and other related phrases follow after *United States* and *British*. Similarly, the collocations of *Spain*, demonstrated in Figure 10, indicate that there was a territorial dispute between Spain and the United States. These concordance lines suggest that there used to be many Spanish and British territories on the North American continent at that time and that the presidents were highly concerned about relationships with the two countries. This is also confirmed by the quadgram of *Florida*, which is given in Figure 11.

Another perspective is found by looking at words related to water. Consider the most frequently observed quadgram of *lake*, as given in Figure 12. As the concordance lines show, the fact that areas between the Lake of the Woods and the Rocky Mountains used to be a British territory was one of the main concerns of the presidents. In other words, there were lots of territorial disputes around the United States, and the presidents paid special attention to the issues.

As the concordance lines in Figure 12 show, *lake* is strongly tied to the territorial

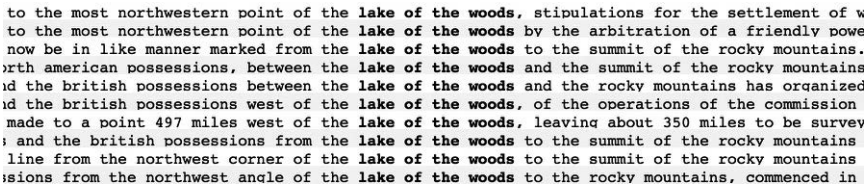

Figure 12: The concordance lines of *Lake of the Woods*

disputes with Spain and Britain. It is interesting to note that words related to water are often used to refer to a relation between the United States and other countries. For instance, many examples of *to the Pacific Ocean*, the most frequently observed quadgram of *ocean*, occur by referring to either country or racial names, as in (2) to (4):

(2) James Monroe's speech in 1818 (Democratic-Republican Party)
[I]t has been necessary during the present year to maintain, a strong naval force in the Mediterranean and in the Gulf of Mexico, and to send some public ships along the southern coast and *to the Pacific Ocean*. By these means amicable relations with the Barbary powers have been preserved, our commerce has been protected, and our rights respected.

(3) Andrew Jackson's speech in 1846 (Democratic Party)
[F]rom central America I have received assurances of the most friendly kind and a gratifying application for our good offices to remove a supposed indisposition toward that government in a neighboring state. [...] Our treaty with this republic continues to be faithfully observed, and promises a great and beneficial commerce between the two countries - a commerce of the greatest importance if the magnificent project of a ship canal through the dominions of that state from the Atlantic *to the Pacific Ocean*, now in serious contemplation, shall be executed.

(4) James Knox Polk's speech in 1846 (Democratic Party)
[O]ur laws regulating trade and intercourse with the Indian tribes east of the Rocky Mountains should be extended *to the Pacific Ocean*.

Furthermore, Figure 13 shows the most frequently observed quadgrams *navigation, of commerce and navigation*, and, as the data show, the collocations are followed by country names many times. These two observations strongly suggest that the presidents in this period were concerned about relationships with foreign countries tied with area issues.

Having scrutinized many examples of frequently observed lexical categories, and

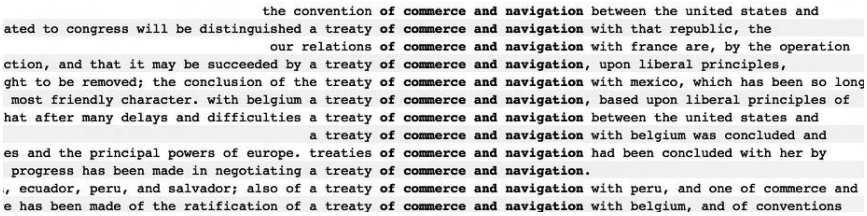

Figure 13: The concordance lines of *commerce and navigation*

having argued that the United States had territorial disputes in its early history and that presidents frequently referred to relations with others, I would like to label Topic 34 as international affairs, strongly tied with territorial disputes. Note that "international" here may be misleading, because some conflicts at that time were not strictly those between countries but were confrontations with first nations living outside of the U.S. boundary. Nonetheless, I use the term here for convenience.

4.3 Topic 10 (1880s−2010s)

The next topic, Topic 10, gradually increased its importance around the late 19th century and gained the highest score around the 1940s. Figure 4 illustrates that many military-related words, such as *peace*, *defense*, *military*, *power*, *war*, and *forces* contributed to this topic. Furthermore, the word cloud includes items referring to international relations, such as *world*, *Europe*, *Soviet*, and *countries*. These examples are enough to label this topic worldwide warfare.

In order to understand why this topic gained presidents' main political focus from the 1880s to the 2010s, let us observe some words other than the military-related words found in Figure 4. First, let us consider *communist*. The word was not very frequently used (130 times in total), because its first appearance was in Truman's speech in 1950, and not many specific collocation patterns were found. Thus, we shall take a look at words that occur within five words—the left and right windows—of *communist*. Table 3 shows the most frequently observed co-occurring words orbiting around *communist*. Figure 14 gives examples of how *communist* and *world* were used in actual manuscripts, which was mostly to evoke a negative impression or to show aggression against communism by using words such as *aggression*, *threaten*, *conspiracy*, and *painful phase*. This was apparently embodied in Ronald Reagan's expression referring

Table 3: Co-occurring words of *communist*

No.	Word	Freq.	No.	Word	Freq.	No.	Word	Freq.	No.	Word	Freq.
1	the	124	6	and	28	11	chinese	12	16	with	11
2	of	72	7	have	28	12	nation	12	17	against	10
3	be	45	8	a	24	13	world	12	18	china	10
4	to	42	9	that	17	14	aggression	11	19	threat	9
5	in	31	10	by	12	15	this	11	20	all	8

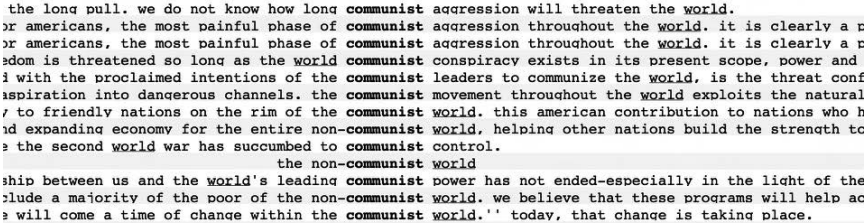

Figure 14: The concordance lines of *communist*

to the Soviet Union as the *evil empire*. Furthermore, *communist* also occurs with *Chinese* and *China* as shown in Figure 15. Given the data in Table 3 and Figure 14 and Figure 15, this topic was strongly influenced by the Cold War.

Another word that indicates that this topic was driven by the Cold War was *nuclear*. It is true that *nuclear* may refer to a nuclear plant, but most of the examples of *nuclear* occurred with military-related words such as *weapon*, *threat*, *war*, and *force*, as exemplified in Figure 16. Historically speaking, the United States competed with the Soviet Union to increase their number and the power of nuclear weapons, and in the early 1960s, the Cold War faced an urgent situation due to the Cuban Missile Crisis. Thus, it is no wonder that the presidents in this period focused on nuclear weapons.

Therefore, as the observations in the word cloud in Figure 4 and the concordance lines given in this subsection show, this international warfare topic gained the prominent focus of presidents because of the Cold War.

venth fleet no longer be employed to shield **communist** china. this order implies no aggressive intent
as a base of operations against the chinese **communist** mainland.
was required to serve as a defensive arm of **communist** china. regardless of the situation in 1950,
venth fleet no longer be employed to shield **communist** china. this order implies no aggressive intent
as a base of operations against the chinese **communist** mainland.
al methods and backward course of events in **communist** china. in these continuing efforts, the free
neral assembly, its secretary-general is in **communist** china on a mission of deepest concern to all
. in the release of our fifteen fliers from **communist** china, an essential prelude was the world
ry has continued to withhold recognition of **communist** china and to oppose vigorously the admission
ll-out bombardment of quemoy restrained the **communist** chinese from attempting to invade the off-
ill our relations with the soviet union and **communist** china. we must never be lulled into believing

Figure 15: The collocational patterns of *communist and China | Chinese*

and mobility of our present conventional and **nuclear** forces and weapons systems in the light of present
manning, and directing a truly multilateral **nuclear** force within an increasingly intimate nato allianc
ting new nations to master the black arts of **nuclear** war -- and if they are willing to turn their energ
sition of having to answer every threat with **nuclear** weapons or nothing.
diture of more than $15 billion this year on **nuclear** weapons systems alone, a sum which is about equal
reaty -- to demonstrate both the futility of **nuclear** war and the possibilities of lasting peace.
nal to launch a nuclear attack or to use its **nuclear** power as a credible threat against us or against o
is at all possible, in both conventional and **nuclear** weapons and defenses. i thought we were making som
ch an agreement that will halt the spread of **nuclear** weapons. on the basis of communications from
ast, the conflict in vietnam, the dangers of **nuclear** war, the great difficulties of dealing with the
we will maintain a **nuclear** deterrent adequate to meet any threat to the secur
rve our interests and minimize the threat of **nuclear** confrontation.
in an era where the strategic **nuclear** forces are in rough equilibrium, the risks of conf

Figure 16: The concordance lines of *nuclear*

4.4 Topic 27 (1950s–2010s)

Lastly, consider Topic 27. The word cloud in Figure 5 shows that this topic includes words that can be clustered in various ways, such as family-related issues (*child(ren), education, schools, college, family*), labor (*businesses, jobs, income,* and *companies*), and health care (*medical*, *insurance, health*, and *Medicare*). With these observations, I would like to label this topic social welfare, and consequently, more lexical items are found to be related to this topic, such as *tax(es), woman, women, social, budget, vote, safe, energy*, and *right*.

Let us take a look at some words that seem to be unrelated to social welfare. The word *security* is found in the word cloud, and this may evoke national security related to warfare. It is true that *security* is often used to refer to national defense, as its most frequently observed quadgrams are *the security of the, to the security of*, and *for the security of*, but a collocation that is strongly tied to this topic, namely *the social security system*, is found to be the fifth-most frequent quadgram. Furthermore, if we restrict ourselves to the bigram and do not use the quadgram, *social security* occurs only in this period (199 times).

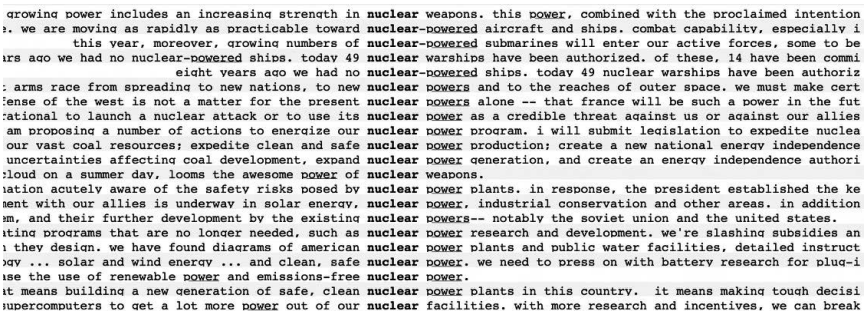

Figure 17: The concordance lines of *nuclear*

Another controversial word is *nuclear*, in that it is also a keyword in the last topic, namely worldwide warfare. One of the frequently used bigrams, *nuclear power*, presents an interesting change in the context of the bigram (see Figure 17). The first half (from 1955 to 1977) of the collocations shows that the presidents used *nuclear power* to refer to a source of weapons, whereas the later presidents (1981 to 2011) tended to use *nuclear power* to refer to electricity. This change reflects differences in the two topics described above and here, namely, worldwide warfare and social welfare.

Why did social welfare gain the primary attention of presidents? First, the mid-1960s was a period of civil rights campaigns in the United States, as various facts show (e.g., in 1963, Martin Luther King Jr. gave his famous “I have a dream” speech; Michael Marrington published *The Other America*, which discussed the existence of economically handicapped people such as the elderly and minorities; and Medicaid was legislated). Furthermore, the growing feminism campaign motivated a change in the way women were treated. That is, inequality in wages between males and females was legally eliminated in 1963. This is probably why *woman* or *women* is made explicit in the context of serving the country in one of the most frequent quadgrams, *men and women who*, as exemplified in Figure 18. Thus, it is no wonder that the topic of social welfare starts to gain presidents’ attention at this time. Second, environmental issues became known to the world. As various leading industrial areas are concentrated in the United States, the consumption of energy that causes destruction of the environment must be dealt with. Hence, the topic also covers energy affairs during this time.

Summarizing this section, I have shown that the presidents’ main political

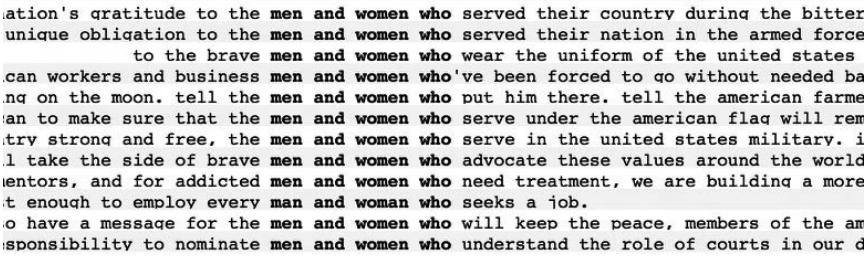

Figure 18: The concordance lines of *men and women who* and *man and woman who*

concerns have changed among four topics. The Presidents of the first century who were involved in foundation of a nation, were mainly concerned about internal or domestic affairs. At the same time, the presidents of the early days of the United States also focused on international relationships, especially emphasizing territorial issues. After World War II, the political concerns of presidents gradually changed, and worldwide warfare and social welfare became the presidents' primary considerations.

Note that the presidents were concerned about more than one topic in the same period. As Figure 1 shows, the moves in Topics 6 and 34, and those in Topics 10 and 27, are quite similar, in that the time the first two topics falls increases the other two topics. Put differently, Topics 6 and 34 are negatively correlated to Topics 10 and 27. There is no clear-cut boundary of exactly when the relations changed, but the mid-1930s, namely the time when the United States attained the world's supremacy, seems to be the threshold during which Topics 10 and 27 took the primary considerations of the presidents.

5. Conclusion

In this study, I have demonstrated how the presidents' primary concerns have changed over two centuries by applying the LDA to the State of the Union Addresses and showed that four main topics are obtained: internal issues related to the federal government, international affairs tied with territorial disputes, worldwide warfare, and social welfare. Lastly, I proposed that the transition in the topics occurred around the mid-1930s. In other words, the presidents' main political concerns were influenced by whether or not the United States had attained the status of a world leader.

There are a few remaining issues to be addressed. First, as was stated earlier, not all of the speeches were given as oral presentations; some were merely submitted written reports. Assuming that register variants do not affect the topics, I intentionally ignored the difference in register. Nonetheless, it may have some influence on our conclusion, and thus, register variations should be given further concern. Secondly, there are some criteria that were arbitrarily made, such as selecting the number of topics and which topics were to be scrutinized. Thus, it is necessary to find less ad-hoc methods (a heat map may help us resolve this issue). Lastly, since 1923, the targeted audience has changed, and this may also have affected the transitions of topics. More specifically, the social welfare topics may be associated with the change in target audience, though I do not have enough confidence to add this explanation for Topic 27.

Acknowledgment

An earlier version of this paper was presented at JAECS 43, which was held at Kwansei Gakuin University. I would like to thank the insightful comments from the audience. I am also grateful to Tomoji Tabata for providing me with his stop word list and to Roger Prior for his invaluable comments. Lastly, I would like to thank Yoshiyuki Nakao, the editor of the English Corpus Studies 25, and three anonymous reviewers for their insightful comments. Of course, any remaining errors are my own. This study was supported by a research grant from University of Kitakyushu.

References

Biber, D. and S. Conrad (2009) *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Blei, D., A. Ng, and M. Jordan (2003) "Latent Dirichlet Allocation." *Journal of Machine Learning Research 3*: 993–1022.

Bonnefille, S. (2008) "When Green Rhetoric and Cognitive Linguistics Meet: President G. W. Bush's Environmental Discourse in his State of the Union Addresses." *Metaphorik.de 15*: 27–61.

Bonnefille, S. (2013) "Energy Independence: President Obama's Rhetoric of a Success Story." *Research in Language 11*: 189–212.

Crockett, S. and C. Lee (2012) "Does It Matter Qhat They Said? A Text Mining Analysis of the State of the Union Addresses of USA Presidents." *13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*: 77–82.

Herz, J. and A. Bellaachia (2014) "The Authorship of Audacity: Data Mining and Stylometric Analysis of Barack Obama Speeches." In Stahlbock, R., G. M. Weiss, M. Abou-nasr and H. R. Arabnia (eds.), *Data Mining: Proceedings of the 2014 International Conference on Data Mining (Worldcomp Iinternational Conference Proceedings 2014)* Nevada: CSREA Press, pp. 148–154.

Imao, Y. (2015) CasualConc (Version 2.0.5) [Computer Software] URL: https://sites.google.com/site/casualconc/

Kaid, L. L. (2007) "Radio, Politics and." In Kaid, L. L. and C. Holtz-Bacha (eds.), *Encyclopedia of Political Communication.* California: SAGE Publications, pp. 696–697.

Kuroda, A. (2017) "Quantitative Analysis of Literary Works: Novels of Sir Arthur Conan Doyle." In the Institute of Statistical Mathematics (ed.). *Text-mining and Digital Humanities*. Tokyo: the Institute of Statistical Mathematics, pp. 55–70.

Schöch, C. (2016) "Topic Modeling Genre: An Explanation of French Classical and Enlightenment Drama." *Digital Humanities Quarterly 11*, URL http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html

Tabata, T. (2017) "The Semantic Structure of the FLOB Corpus: Using Probabilistic Models for Corpus Description." In the Institute of Statistical Mathematics (ed.). *Text-mining and Digital Humanities*. Tokyo: the Institute of Statistical Mathematics, pp. 1–18.

Tung, J. (2014) Text Mining Analysis of State of the Union Addresses: With a Focus on Republicans and Democrats Between 1961 and 2014 URL: https://statoftheart.files.wordpress.com/2014/05/text-mining-analysis-of-state-of-the-union-addresses.pdf

Notes

1. The dataset is available at <https://drive.google.com/file/d/0B27VXLzIM-qhNmxNVnNsTk84bEk/view>
2. The topic numbers are computed through the LDA, their orders have nothing to do with diachronic counterparts.
3. The movement of the topics indicates that the U.S. presidents do not have any everlasting political philosophy that is consistent through 240 years.
4. Readers may think that 0.0008 as a cutting point looks too low. However, the average of the overall word weight is 0.00004, and the maximal value is 0.04. I did trials-and-errors many times to find a value to plot visible word clouds, and 0.0008 was the lowest relatively visible value. Thus, I would like to say that the value is appropriate.
5. It should be natural to examine words represented in the center of Figure 2, such as *united*, *states*, and *government*, however the quadgram of such words show that they are mostly used as *government of united states of America*, and no insightful quadgrams were found. Thus, being certainly quite important to this topic, such words do not provide any answers to the current question, and will not be further investigated.

（木山　直毅　北九州市立大学　Email: n-kiyama@kitakyu-u.ac.jp）