A visual approach for text analysis using multiword topics

[*] Seongmin Mun , [†] Guillaume Desagulier , [‡] Kyungwon Lee

[*] Lifemedia Interdisciplinary Program, Ajou University [*] UMR 7114 MoDyCo - CNRS, University Paris Nanterre

[†] UMR 7114 MoDyCo - University Paris 8, CNRS, University Paris Nanterre

[‡] Department of Digital Media, Ajou University

[*] stat34@ajou.ac.kr, [†] guillaume.desagulier@univ-paris8.fr, [‡] kwlee@ajou.ac.kr

Introduction

Visual analysis of text data can support users in acquiring a general understanding of information about corpus without actually reading it. This can be very helpful when the task involves large volumes of text. Research in extracting topics is very common for the visual analysis of corpora. These topics can be categorized as those that have a meaning that can be expressed in one word and those whose meaning must be described using a combination of words. This latter type is called a multiword topic. Simply, multiword topics are habitual recurrent word combinations in everyday language. For example, if people say that Barack Obama sets the bar high, we understand it as a metaphor that President Obama's competitors will have a hard time trying to beat him. However, analysis of multiword topics requires a system based on systematic analysis and verification with a raw corpus. Therefore, we have created a visual system that covers necessary parts for exploring more information in a corpus using multiword topics. This work provides the following contributions: (1) We present the two topic types in corpus data to explore more information and find accurate results. (2) We present a systematic analysis for extracting accurate topic results. (3) We assess our system via case studies using U.S. Presidential Addresses to verify the assets of our system.

Data Processing

We present a data processing structure for extracting information from corpus data. Figure 1 summarizes the architecture of our data processing, which is described.

Preprocessing. We conducted cleaning with RegExp, lemmatization, tokenization, and lowercasing. We then conducted an N-gram analysis and part-of-speech (POS) tagging on the extracted topic candidates in the processing stage.

Candidate Extraction and Filtering. We counted these results by their frequency value and filtered the data by applying a threshold (frequency value greater than or equal to 10). In addition, topic candidates were extracted without stopwords by each gram. For instance, for bigrams, "house i," "power we," etc. are stop-words and removed from the candidate topics.

Topic Validation. We verified the filtered candidate topics with computational linguistics and several English dictionaries. The output of candidate topic filtering must be verified. For this verification, we developed a working algorithm that automatically compares the results with several English dictionaries; if the candidate topic is defined in dictionaries, the algorithm returns this candidate topic as an available result. Additionally, the primary validated candidate topics are manually verified by computational linguistics researchers. If candidate topics not in the dictionary are determined by the researchers to be meaningful, they are stored in a user dictionary and utilized in later analysis.

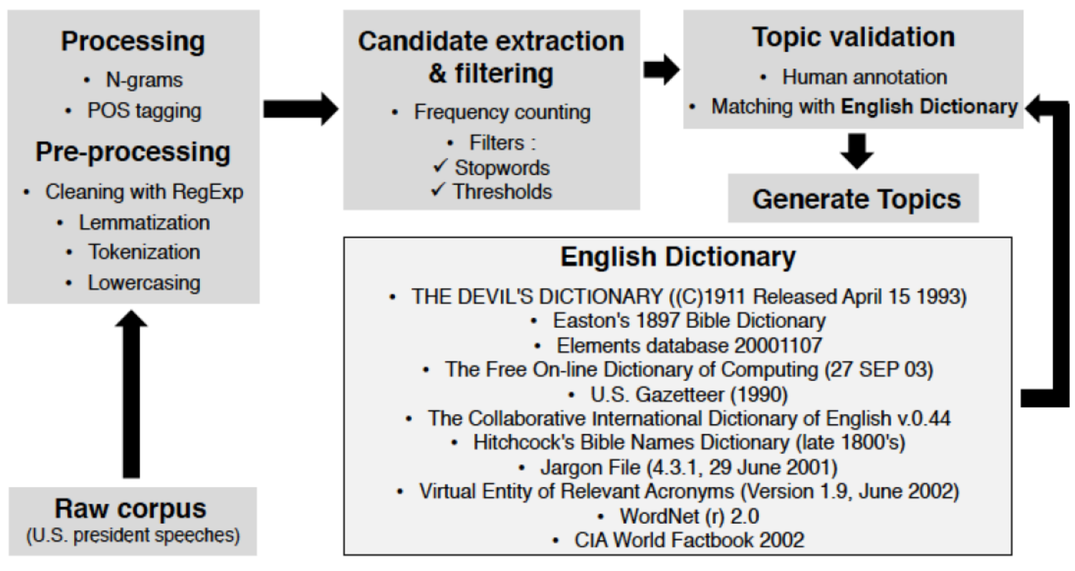

Figure 1: Data processing structure. Framework for topic acquisition from corpus data.

Visualization Design

Figure 2 depicts the main workspace of our visual system after loading all the presidents' speeches. Three buttons in the middle of layer headers (Figure 2 (c)) provide options for changing topic word combinations by the value of N in each N-gram. Additionally, users can change visual result by selecting options in the middle(Figure 2 (d), (e)), making our system very flexible because different visual results of a president's speech can be viewed easily.

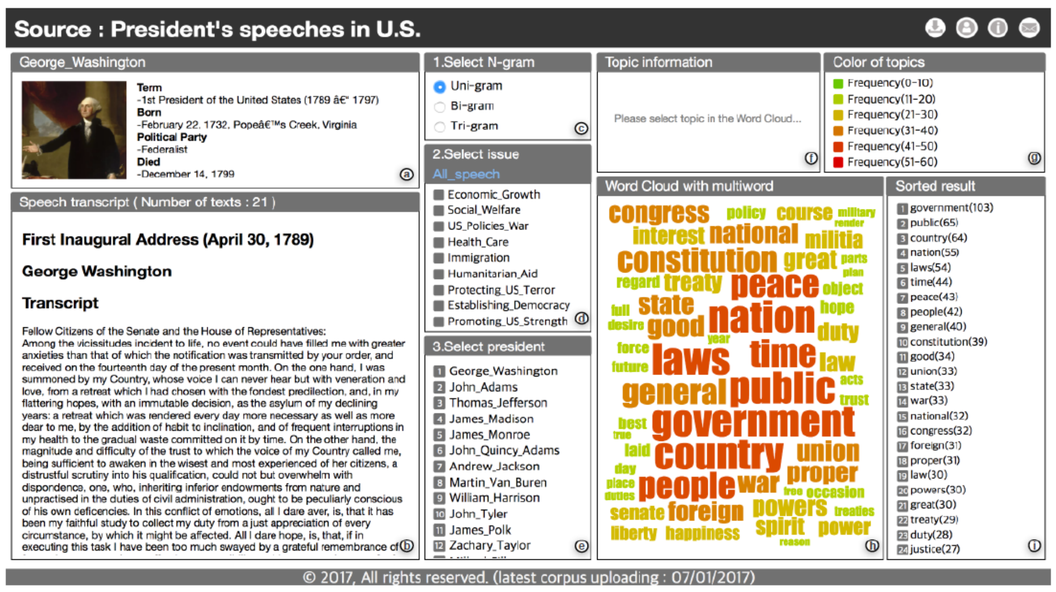

Figure 2: The interface of our visual system represents corpus data of speeches from 43 U.S. presidents from George Washington to Barack Obama.

Case Studies

We conducted case studies to evaluate the effectiveness and usability of our system. We worked with computational linguistics researchers who study multiword topic analysis and have expert knowledge of it. They used our system to find information about their research questions. A serious problem will occur in the analysis result if the researcher used a topic that has a meaning in one word only. For example, the topic "United States" frequently appears in the speech. However, if we do not use multiword analysis, the words "United" and "States" will account for a large proportion of the analysis results. Our visual system has addressed this problem, as shown in Figure 3.

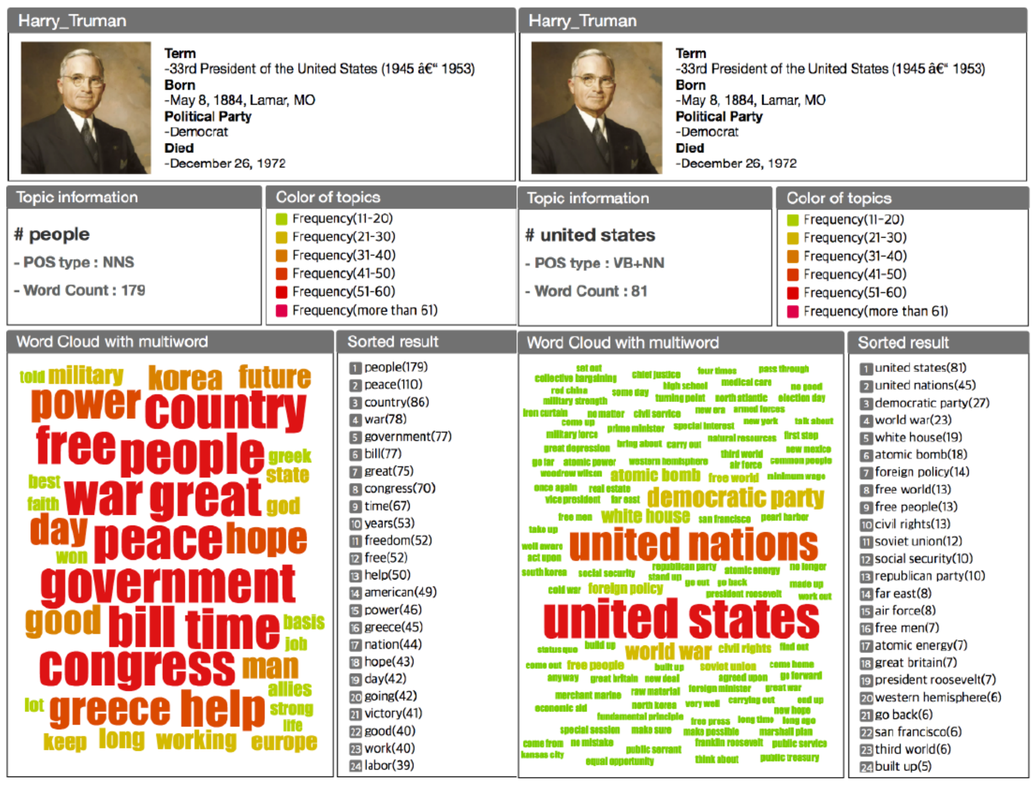

Figure 3: Analysis result of Harry Truman's speech by (a) unigram and (b) bigram.

Conclusion

We have interviewed several times with computational linguists. And they agreed that the exploration of multiword topics by N-gram is a major strength of our system. Further, this system can facilitate quick exploration of the information in a corpus and get accurate results, as shown in the above case studies. This study reveals the data processing required to acquire accurate topic results from corpus data by N-gram and uses a linguistic approach to obtain accurate multiword topics and explains it via the above data processing.

References

[1] RAMISCH C.: Multiword Expressions Acquisition. Springer, 2015.

[2] STEFFEN KOCH MARKUS JOHN M. W. A. M. T. E.: Varifocalreader-in-depth visual analysis of large text documents. In IEEE Transactions on Visualization and Computer Graphics (2014), vol. 20, pp. 1723–1732.