Eurographics Conference on Visualization (EuroVis), Posters Track (2017)
A. Puig Puig and T. Isenberg (Editors)

A visual approach for text analysis using multiword topics

Seongmin Mun[1,2] , Guillaume Desagulier[3], Kyungwon Lee[4]

[1]Lifemedia Interdisciplinary Program, Ajou University, South Korea
[2]UMR 7114 MoDyCo - CNRS, University Paris Nanterre, France
[3]UMR 7114 MoDyCo - University Paris 8, CNRS, University Paris Nanterre
[4]Department of Digital Media, Ajou University, South Korea

Abstract

Topics in a text corpus include features and information; visualizing these topics can improve a user's understanding of the corpus. Topics can be broadly divided into two categories: those whose meaning can be described in one word and those whose meaning in expressed through a combination of words. The latter type can be described as multiword expressions and consists of a combination of different words. However, analysis of multiword topics requires systematic analysis to extract accurate topic results. Therefore, we propose a visual system that accurate extracts topic results with multiple word combinations.
For this study, we utilize the text of 957 speeches from 43 U.S. presidents (from George Washington to Barack Obama) as corpus data. Our visual system is divided into two parts: First, our system refines the database by topic, including multiword topics. Through data processing, we systematically analyze the accurate extraction of multiword topics. In the second part, users can confirm the details of this result with a word cloud and simultaneously verify the result with the raw corpus. These two parts are synchronized and the desired value of N in the N-gram model, topics, and presidents examined can be altered. In this case study of U.S. presidential speech data, we verify the effectiveness and usability of our system.

Categories and Subject Descriptors (according to ACM CCS): I.7.0 [Document And Text Processing]: General—Data Processing, H.5.2 [Information interfaces and presentation (e.g.,HCI)]: User Interfaces—Web-based Interaction

1. 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 [JZZ14, YL16, XW16, FH16, WC14, GS14, SK14]. 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 [Ram15]. Simply, multiword topics are habitual recurrent word combinations in everyday language [JR57]. 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 utility of our system.

2. Data processing

In this section, we present a data processing structure for extracting information from corpus data. Our data are taken from the Miller Center [Mil], a representative database of U.S. history and civil discourse. Figure 1 summarizes the architecture of our data processing, which is described in detail below.

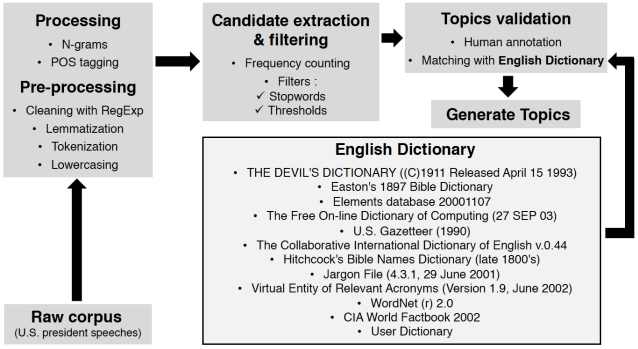

Figure 1: *Data processing structure. Framework for topic acquisition from corpus data.*

DOI: 10.2312/eurp.20171168

Processing. As 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 [Ram15].

Candidate Extraction and Filtering. Through the above procedure, we obtained unigram to trigram results with POS tagging. 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 stop-words 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 [DEV, Eas, JDi, Gaz, GNU, Hit, Jar, Wor, CIA, WSD]. 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.

We thus extracted candidate topics-45,995 from unigrams, 729,552 from bigrams, and 2,089,617 from trigrams. Of these candidates, 8910 unigram, 901 bigram, and 301 trigram topics were validated as meaningful and analyzed.

3. Visualization Design

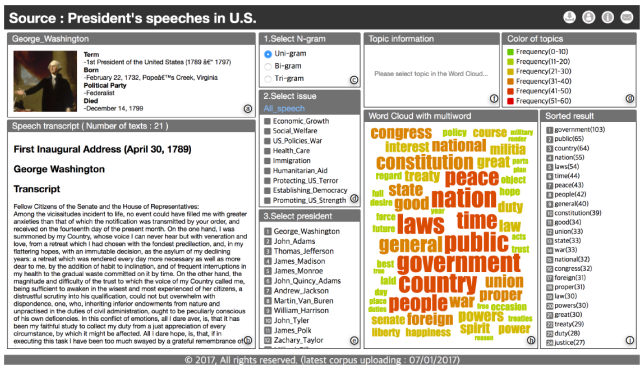

Figure 2: *Visual system interface. The interface of our visual system represents corpus data of speeches from 43 U.S. presidents-from George Washington to Barack Obama.*

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.

4. 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.

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

A serious error 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.

5. Conclusion

We have interviewed several times with domain experts who study for computational linguistic. 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. This study uses a linguistic approach to obtain accurate multiword topics and explains it via the above data processing. In future work, we plan to improve our system to show more information through combing linguistic approach and more topics with multiword without limit for the N of gram.

Acknowledgement

This work was supported by the 2017 BK21 Program, Ajou University and National Research Foundation of Korea (NRF-2015S1A5B6037107).

References

[CIA] Cia world factbook 2002. Central Intelligence Agency http://www.cia.gov/news-information/press-releases-statements/press-release-archive-2002/pr10112002.html. 2

[DEV] The devil's dictionary ((c)1911 released april 15 1993). Aloysius West http://www.alcyone.com/max/lit/devils/. 2

[Eas] Easton's 1897 bible dictionary. Matthew George Easton http://eastonsbibledictionary.org/. 2

[FH16] FLORIAN HEIMERL QI HAN S. K. T. E.: Citerivers: Visual analytics of citation patterns. In *IEEE Transactions on Visualization and Computer Graphics* (2016), vol. 22, pp. 190–199. 1

[Gaz] U.s. gazetteer (1990). U.S. Census Bureau http://ils330.wikispaces.com/file/view/US+census+bureau+us+gazetteer.pdf. 2

[GNU] Gnu collaborative international dictionary of english. C. G. Merriam Co. http://gcide.gnu.org.ua/. 2

[GS14] GUODAO SUN YINGCAI WU S. L. T.-Q. P. J. J. H. Z. R. L.: Evoriver: Visual analysis of topic coopetition on social media. In *IEEE Transactions on Visualization and Computer Graphics* (2014), vol. 20, pp. 1753–1762. 1

[Hit] Hitchcock's bible names dictionary (late 1800's). Roswell D. Hitchcock http://www.menfak.no/bibel/navn.html. 2

[Jar] Jargon file (4.3.1, 29 june 2001). Stanford http://72.9.148.189/library/Jargon_file. 2

[JDi] Jdictd. JDictd http://jdictd.sourceforge.net/JDictd/index.html. 2

[JR57] JR F.: *Papers in linguistics 1934-1951*. Oxford University Press, 1957. 1

[JZZ14] JIAN ZHAO LIANG GOU F. W., ZHOU M.: Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *IEEE Symposium on Visual Analytics Science and Technology 2014* (2014), pp. 203–212. 1

[Mil] Miller center. University of Virginia http://millercenter.org/. 1

[Ram15] RAMISCH C.: *Multiword Expressions Acquisition*. Springer, 2015. 1, 2

[SK14] 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. 1

[WC14] WEIWEI CUI SHIXIA LIU Z. W. H. W.: How hierarchical topics evolve in large text corpora. In *IEEE Transactions on Visualization and Computer Graphics* (2014), vol. 20, pp. 2281–2290. 1

[Wor] Wordnet (r) 2.0. Princeton University http://wordnet.princeton.edu/. 2

[WSD] Dictservice (wsdl). DictService (WSDL) http://services.aonaware.com/. 2

[XW16] XITING WANG SHIXIA LIU J. L. J. C. J. Z. B. G.: Topicpanorama: A full picture of relevant topics. In *IEEE Transactions on Visualization and Computer Graphics* (2016), vol. 22, pp. 2508–2521. 1

[YL16] YAFENG LU MICHAEL STEPTOE S. B. H. W. J.-Y. T. H. D. D. M. S. R. C. R. M.: Exploring evolving media discourse through event cueing. In *IEEE Transactions on Visualization and Computer Graphics* (2016), vol. 22, pp. 220–229. 1