Decoding the Chinese Government Work Report: A Natural Language Processing Approach

Shuyang Bian
Emory University
sbian8@emory.edu

Roberto Franzosi
Emory University
rfranzo@emory.edu

Abstract

The Chinese Government Work Report (CGWR) is delivered yearly by the presiding Premier to the National People's Congress of the People's Republic of China. Available since 1954, the published documents provide a window into the workings of the Chinese government at various levels and have been extensively studied. In this paper, we use various Natural Language Processing (NLP) tools on the English translations of the CGWR using the freeware package NLP Suite (Franzosi, 2020). We use Stanford CoreNLP, SpaCy, and Stanza for parsing. POS (Part of Speech) tags and DepRel (Dependency Relations) tags for nouns and verbs reveal the role of Presidents on government policy priorities (e.g., the changing nature of the adjectives related to the word "development," agricultural, industrial, financial, social introduced by new Presidents or the growing use of gerund tenses that make no commitment to time and nominalization which hide agency). We use a number of measures to highlight the overtime decrease of sentence complexity and vocabulary richness and the parallel increase of text readability as Premiers aim to reach wider audiences. Sentiment analysis via BERT and Stanford CoreNLP reveals an increasing optimism over the years. We geocoded via Nominatim the NER (Named Entity Recognition) location tags and map the results as pin maps via Google Earth Pro and heatmaps via Google Maps. Once again, the maps reveal distinct shifts overtime of geographic hot spots with changing Presidents.

1 Introduction

On March of every year, the presiding Premier of the People's Republic of China delivers to the National People's Congress a Government Work Report (or CGWR for Chinese) (Wang 2017). In essence, the CGWR functions not only as bully pulpit to promote the government's achievements to the Chinese public but also as a comprehensive policy plan guiding government bureaucracy.

The CWR reports have attracted considerable media and scholarly attention. Numerous universities in China have published articles on the China National Knowledge Infrastructure (CNKI, i.cnki.net). Using CiteSpace (Chen, 2006) to extract the topics discussed, preliminary research reveals that translation strategy, text analysis and governmental policies are three major clusters of CGWR analyses in past literature. Nevertheless, the methodologies employed typically encompassed closed readings on selected years' text in English.

2 Data

For the years 1954-1957, with no translation available, we used DeepL, an auto-translation tool to translate the reports for these four years. For the years 1958-1999, we use the English translations published in the Beijing Review. For the years 1961-1963 and 1965-1975 no CGWR is available, not even in Chinese, during the political upheaval of the "Great Leap Forward" of 1958-1962 and of the Cultural Revolution of 1966-1976 (Chow, 1993; Esherick et al., 2006; Landsberger, 2013). From the year 2000 onward, we used the official translations from the government domain Xinhua Net (http://www.xinhuanet.com/). Using Google Cloud Vision API for Optical character recognition, we converted the remaining reports in scanned image formats. The final corpus consisted of 51 English language CGWR documents in txt format.

3 Methods

We rely on the NLP Suite, a freeware, open-source package of Python scripts (Franzosi, 2020; https://github.com/NLP-Suite/NLP-Suite/wiki) that incorporates a wide range of tools of automatic textual analysis and visualization. In analyzing the sentiment embedded in the documents, we employed BERT (Bi-directional Encoder Representations from Transformers, Devlin et al. 2018) and spaCy (Honnibal et al. 2013). In analyzing the syntactical structure of the texts, we used Stanford CoreNLP (Manning et al., 2014) and Stanza (Qi et al., 2020). In analyzing the topics, we used Gensim (Rehurek and Sojka, 2011). Post text-parsing, our own algorithms were designed to compute corpus statistics, n-grams, sentence complexity, text readability, vocabulary richness, abstract vs. concrete, and objective vs. subjective language. We further used geographic maps to to extract geographic entities of different levels in text, and used Gephi for visualizing the Subject-Verb-Object relations as network graphs.

4 Results

(1) Text Metrics
The CGWR is a highly formulaic discourse, with relatively stable style, schema, and topics, "a kind of 'genre'" (You et al., 2010:595). In using Yngve Depth, Frazier Depth and Frazer Sum, the three tools showed a decreased sentence complexity and an increase in readability scores, suggesting a shift towards a more accessible report.
(2) Verb Analysis
We explore on the verb form that exists in the CGWR and observe clear increase in Infinitive. A handful of verbs – improve, continue - show their continued and increasing use in CGWR language. Using Halliday's classification on modals, we found that CGWR overall shows a preference of median-value modals, with high-value modals twice as frequent as the low- value ones.With a trend in the average percentage of concrete words per sentence, normalized based on the document's length, we see a growing tendency toward greater abstractness in language. Nevertheless, nominalization increased over the years following a trend similar to Word Bank reports. As a linguistic process to obscure agency, this suggests that the nature of key messages embedded subject to future promises.
For the extract Subject-Verb-Object (SVO) triplet, we obtained 2049 nodes (as subjects and objects in total) and 4401 edges (the verb relationship) over the years. The resulting graph is a uni-mode one, with the central word "We" being a most important keyword.
(3) The Geography of the Reports: Texts as Politics
We extracted for location values using Stanford CoreNLP NER pipeline, and geocoded using Nominatim or Google. Post filtering, we obtained a total of 6,437 geocoded locations. Using a treemap and categorizing the documents based on leadership period, we reveal a steady decrease to the outside world, with shrinking references to foreign countries: 62 countries are mentioned in the reports for Mao's presidency, 77 for Deng's, 44 for Jiang's, 13 for Hu's, and 24 for Xi's. Within the country, locations within China being mentioned have been fluctuating considerably in significance. While the focus on Northern and coastal central regions remained stable from Jiang to Xi, through Hu. But with Xi, the emphasis on the South returned. Political, rather than economic, reasons, potentially underscore the shift. Using co-occurrence pipelines, we observe for a coherent Hong Kong, Macao, Taiwan trilogy over the production of reports beginning since 1990. These co-occurring tokens, which coincided historically with the principle of "one country, two systems" introduced in 1984, is observed to be repeated in nearly every report thereafter, often accompanied by high-value modals such as we will, we must (e.g., "We will unswervingly implement" Xi/Li Keqiang 2014).
(4) Looking to an Ever Brighter Future: Sentiment Analysis
The results for both BERT and CoreNLP show an increased percentage of positive emotions over time, with negative emotions remaining largely unchanged. Overall, our analysis suggests that there may be a gradual shift in the emotions expressed in governmental reports, showing growing optimism as China hurtles along its economic development plans.

Limitations

Syntactical differences in the Chinese and English language, namely, the large number of modal verbs (43 Chinese modals vs. 10 or 12 English modals) and the option of constructing a perfectly syntactically correct sentence without a subject for the verb, offer translators ample choices for their translations. Looking forward, there are at least two paths for future research: 1. apply the same set of NLP techniques in a comparative analysis of CGWR in the original Chinese language and English; 2. apply other cutting tools, such as word embeddings and word sense induction, available from both BERT and Gensim (Mikolov et al., 2013; Gupta, 2020; Lucy and Bamman, 2021).

5 References

References

2015. *It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool*. Association for Computational Linguistics.

Chaomei Chen. 2006. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *J. Am. Soc. Inf. Sci. Technol.*, 57(3):359–377.

G C Chow. 1993. Capital formation and economic growth in china. *Q. J. Econ.*, 108(3):809–842.

Joseph W Esherick, Paul G Pickowicz, and Andrew G Walder, editors. 2006. *The Chinese cultural revolution as history*. Stanford University Press.

Stefan R Landsberger. 2013. Art in turmoil: The chinese cultural revolution, 1966-76 ed. by richard king et. al. *Twent. Century China*, 38(1).

Li Lucy and David Bamman. 2021. Characterizing english variation across social media communities with BERT. *Trans. Assoc. Comput. Linguist.*, 9:538–556.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeshun You, Jianping Chen, and Zhong-Hong. 2010. Discursive construction of chinese foreign policy. *J. Lang. Polit.*, 9(4):593–614.