Information Extraction of Political Statements at the Passage Level

Juan-Francisco Reyes

Institute of Computer Science
Brandenburgische Technische Universität Cottbus-Senftenberg
jf.reyes@b-tu.de

Abstract

This research addresses the challenge of accurately identifying, extracting, and publishing political statements from the web. The thesis proposes a broader definition of a political statement and presents a novel information extraction system. The system leverages natural language processing techniques, a web crawler, a taxonomy of political issues, a Political Discourse Analyzer, machine learning as a service, and a content management system. The goal is to develop a theoretical model for the efficient extraction of political statements, reducing the need for human.

1 Introduction

The WWW abounds with discursive content authored by politicians but dispersed in several web resources, such as governmental websites, news websites, social media platforms, or APIs. However, finding, selecting, extracting, and making the most relevant political statements properly accessible to political scientists, journalists, linguists, citizens, and others interested in political discourse is time-consuming and involves intense human effort.

How can we adapt and enhance existing natural language processing (NLP) techniques to accurately and automatically identify, extract, and publish the most politically relevant passage-level statements with minimal human intervention while ensuring their accessibility through a suitable website?

This research aims to use NLP techniques to accurately and automatically identify, extract, and publish the most politically relevant passage-level statements with minimal human intervention. By ensuring these statements are readily accessible through a suitable website, the proposed research seeks to empower political discourse stakeholders with a streamlined and efficient means of accessing valuable political and linguistic analysis. This problem-driven research will contribute to the advancements of NLP techniques and support a more informed and transparent democratic practice in the digital age.

2 Existing solutions

Before reviewing the existing solutions, we must first answer the apparently simple and naive question, "What is a political statement?". In Linguistics, a statement is a "declarative sentence" that 1. expresses a fact, idea or opinion, 2. consists of one sentence, and 3. includes a clause composed of a doing verb and a subject, the thing or person it refers to.

In linguistics, a political statement is "a declarative sentence with political content". Nevertheless, in political science, a political statement is a verbal or non-verbal form of communication with political content that 1. expresses an intention to influence the recipient's decision, attitude or action, and 2. (in verbal form) consists of one or many sentences (a passage).

The literature review does not give a consensual definition of a political statement. Hence, rather

than being normative, we make a fundamental assumption by choosing a broader and more pragmatic way to define a political statement in this research a political statement is a coherent passage, sentence, or clause of political discourse with political content that conveys a political intention.

No existing solutions address the very same problem, but we found multiple solutions that partially solve the problem divided across the domains of computer science, political science and linguistics.

- **In computer science:** Multiple research aims to identify and extract relevant multi-sentence text (passage extraction) from an extensive collection of documents in response to a user's query (information retrieval) or in response to a user's question (question-answering) or for presenting a summary that better captures the critical information and ideas (text summarization) (Kenter et al., 2018; Xu et al., 2011).

- **In political science/Linguistics:** Multiple research aims to identify text genre/profile by analyzing linguistic features of text based on genre theory (analyzing generic constructs and the contexts in which such genres are produced, interpreted, and used), linguistic profiling by extracting lexical, grammatical and semantic features that characterize language variation, political discourse analysis (PDA) by analyzing discourse in political forums (such as debates, speeches, and hearings) and computational sociolinguistics by studying the relation between language and society from a computational perspective (Dunmire, 2012).

Most of the research in information extraction (IE) of political statements extract the embedded knowledge in a single-sentence text (not at the passage level) to populate a knowledge graph, using machine learning (ML) methods to automate the extraction task without digging much into the nuances of the political language (Bamman & Smith, 2015). Another research area in IE studies one specific aspect of the political language, such as sentiment (Bonikowski & Zhang, 2023), stance (Gambini et al., 2022), or election forecasting (Jérôme et al., 2022). More recent research analyses specific political rhetoric traits in political discourse to extract argumentation (Lapesa et al., 2020).

3 Research questions

How can we design and implement an IE system that can accurately and automatically identify and extract the most relevant political statements from the WWW by analyzing domain-specific discourse and linguistic markers while minimizing the need for human intervention?
This broad research question can be broken down into more specific sub-questions, which encompass both theoretical and practical implications for examining linguistic complexity and its computational processing, including:

1. What are the most effective computational methods for analyzing morpho-syntactic structures and patterns to automatically identify and extract coherent and cohesive statements at the passage level?

2. How can NLP techniques be adapted or combined to accurately identify and extract factually correct and politically relevant statements while minimizing the reliance on manually annotated data or human intervention?

3. What are the critical political discourse markers and linguistic features that, once systematically detected and analyzed using NLP techniques (operationalized), predict more effectively politically relevant statements?

4 Solution approach

Contrary to traditional information extraction's scope of extracting relations, entities, and facts,

extracting political statements at the passage level should identify and consolidate information from various text parts to create a more comprehensive and coherent single text. Also, a relevant political statement possesses specific linguistic markers that should be computationally analyzed before proposing statements as candidates.

Thus, we propose developing an IE system that leverages NLP techniques to automatically process texts to extract and assess passages based on their political discourse markers (via syntactic and semantic features) to propose them as relevant statements. The general approach prefers rule-based methods over ML methods to study and describe the linguistic challenges thoroughly; however, ML methods are used whenever more efficiency is required. Our solution will incorporate the following components:

4.1 Web crawler

Implement a web crawler to automatically retrieve fresh political discourse texts from the US political scene from different web resources in the WWW, such as web archives, social media outlets, news websites, Etc. The crawler recognizes political discourse content on crawled pages and classifies texts in monologic (speeches, remarks) and dialogic (interviews, conferences, debates).

4.2 Taxonomy of political issues

Implement a taxonomy of political issues that classifies all political issues (persons, organizations, places, concepts, Etc.) linked to their respective representation in a knowledge base (Wikidata). Each entity has a lexicon with various ways to refer to itself ("aliases").

4.3 NLP pipeline

Implement an NLP pipeline using spaCy with the following components/tasks:

1. Named-entities recognition (NER), rule-and-lexicon-based and linked to Wikidata.
2. Named-entity disambiguation and linking (NED/NEL): used in case of ambiguous concepts or entities (i.e., Columbus [PERSON] and Columbus [PLACE]). ML model trained using automatically retrieved-context sentences from the WWW.
3. Coreference resolution: Identifying and linking different textual mentions that refer to the same entity or concept within a given text to improve the understanding of relationships between words, phrases, and sentences and to provide a more coherent representation of the text's meaning.
4. Relation extraction (RE): Using an open relation extraction (ORE) approach, which extracts relations and their arguments without a predefined schema (ClauseIE). More meaning may be inferred while extracting more relations.
5. Triple extraction: Knowledge in the form of triples is extracted using dependency parsing and matching algorithms to ensure a correct representation of facts in the real world.
6. Political Discourse Analyzer (PDA): Using multiple algorithms and matching rules, assessing political discourse markers in the statements (via syntactic and semantic features) classifies them as valid as a relevant candidate.

4.4 Machine Learning as a Service (MLaaS)

Implement ML models deployed on a cloud computing service (Google Cloud), accessed by the IE system via APIs.

4.5 Content Management System (CMS)

Implement a CMS to allow editors to promote candidate statements to be published on an observatory website.

4.6 ObPolDis –Observatory of Political Discourse

Implement a CMS to allow editors to promote candidate statements to be published on an observatory website, https://obpoldis.netlify.app/.

5 Evaluation methods

This research aims to comprehend the linguistic intricacies of addressing the problem and its computational implementation utilizing NLP techniques. As a result, the primary endeavor involves systematically exploring novel insights related to the studied artifacts' linguistic principles, methodologies, and performance. Substantial advancements in IE can be realized by understanding the NLP pipeline components or techniques employed to tackle the issue.

This research is problem-oriented, meaning that the research problem itself is on focus rather than the methods and tools to solve it. By identifying the existing knowledge base and its gaps through literature reviews and conducting preliminary experiments on an NLP pipeline prototype within the IE system, the core nature of the problem is determined. Once linguistic features that contribute to the relevance of political statements are defined (i.e., cohesion, coherence, correctness, accuracy, readability, complexity, and other syntactic and semantic features found in political discourse), fundamental experiments can be performed to establish relationships between variables along the NLP pipeline components.

Throughout the research, if existing theories cannot explain a phenomenon, multiple experiments are conducted to verify the accuracy of the proposed model. The focus is on achieving a strong qualitative correlation (through observation) rather than quantitative agreement. If verification fails, the model must be refined, and new observations may be required. Upon achieving a verified model, large-scale extractions can be conducted to gather information about the IE system's characteristics and performance.

Experiments involve manipulating variables along the NLP pipeline to improve the prediction of relevant political statements and evaluating individual components through experiments. For example, an essential aspect of the study is testing the Political Discourse Analyzer (PDA) component ("algorithm") with a dataset of political and non-political statements to determine which political discourse markers (or "index") better predict relevant political statements. After numerous iterations, the model's efficiency could be improved by defining linguistic attributes and political discourse markers that predict relevant political statements. The newly acquired knowledge can be framed as design considerations, which can be incorporated by modifying the initial product or developing a new design. When implemented effectively, the new product addresses the original problem.

6 Research objectives

The overall purpose of this work is to achieve a fundamental understanding of how a passage-level political statement can be extracted automatically from the WWW. This thesis aims to develop a theoretical model that describes the most efficient way to extract political statements automatically in mathematical terms.

References

Bamman, D., & Smith, N. A. (2015). Open Extraction of Fine-Grained Political Statements. Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/d15-1008

Bonikowski, B., & Zhang, Y. (2023). Populism as Dog-Whistle Politics: Anti-Elite Discourse and Sentiments Toward Minority Groups. Social Forces. https://doi.org/10.1093/sf/soac147

Dunmire, P. (2012). Political Discourse Analysis: Exploring the Language of Politics and the Politics of Language. Language & Linguistics Compass. https://doi.org/10.1002/lnc3.365

Gambini, M., Fagni, T., Senette C. & Tesconi, M. (2022). Tweets2Stance: Users stance detection

exploiting Zero-Shot Learning Algorithms on Tweets. arXiv. https://doi.org/10.48550/arXiv.2204.10710

Jérôme, B., Mongrain, P., & Nadeau, R. (2022). Forecasting the 2022 French Presidential Election: From a Left–Right Logic to the Quadripolarization of Politics. PS Political Science & Politics. https://doi.org/10.1017/s1049096522000488

Kenter, T., Borisov, A., Van Gysel, C., Dehghani, M., de Rijke, M., & Mitra, B. (2018). Neural Networks for Information Retrieval. arXiv preprint arXiv:1801.021782 .

Lapesa, G., Blessing, A., Blokker, N., Dayanik, E., Haunss, S., Kuhn, J., & Pado, S. (2022). Analysis of Political Debates through Newspaper Reports: Methods and Outcomes. Datenbank-Spektrum. https://doi.org/10.1007/s13222-020-00344-w

Xu, W., Grishman, R., & Zhao, L. (2011). Passage Retrieval for Information Extraction using Distant Supervision. In H.

Wang, & D. Yarowsky (Eds.), IJCNLP 2011 - Proceedings of the 5th International Joint Conference on Natural Language Processing (pp. 1046-1054). Association for Computational Linguistics (ACL)