Methodology Backgrounder

Introduction

The Gender Gap Tracker is a unique data analytics tool that measures the ratio of men's and women's voices in influential Canadian news media. The aggregation of this rich source of information provides hard evidence that will significantly enhance knowledge regarding the scale and the impact of women's representation in public discourse. This project is a collaboration between Informed Opinions and Simon Fraser University, with in-depth research and development carried out at the Discourse Processing Lab.

Team

Our technical team at Simon Fraser University is composed of researchers, software developers and data scientists.

- Dr. Maite Taboada, Principal Investigator and Director, Discourse Processing Lab
- Dr. Fatemeh Torabi Asr, Postdoctoral Fellow and Project Manager
- Mohammad Mazraeh, Software Developer and Machine Learning Engineer
- Vasundhara Gautam, Computational Linguist
- Alexandre Lopes, Data Scientist and Database Manager
- Junette Dianne Gonzales, Language Data Annotator

Informed Opinions relies on the support of web developers at Pondstone Digital to provide the online visual interface.

Text processing pipeline

We scrape text and metadata of each news article from the daily web content of seven Canadian English-language news outlets. We then process the text using a variety of Natural Language Processing and Machine Learning techniques in order to identify quoted people and their genders. This process includes syntactic parsing, named entity recognition, quotation extraction, and gender identification, all of which happen in real time. The entire data and annotations are maintained on SFU database servers. The dashboard shows the proportion of men and women quoted daily.

Future directions

The Gender Gap Tracker is a work in progress. Since the first release of the software in February 2019, we have been continuously working to improve its coverage and accuracy. Identifying the people mentioned in news articles, unifying these mentions to accurately label multiple quotes by the same

person as one source, and tagging their gender appropriately are all complex tasks. The NLP tools sometimes miss a quotation or tag a person with the wrong gender. To identify the gender of the speakers, we rely on online services that provide databases of names. These databases have their own limitations. Nevertheless, our evaluations show that, on a sufficiently large sample of manually annotated text, the ratio of male and female sources calculated by the system is close to the actual numbers within a 5% error margin. We are continuing to work hard to minimize such errors in our new updates, but it's important to note that no media organization is likely to be unfairly affected.

We are also in the process of applying the Gender Gap Tracker to data from French-language news outlets in Canada. The challenge is that most NLP tools in French are far less accurate than their English versions, therefore a lot more in-house development is required to process French text as accurately and quickly as English text.

Finally, we are looking into different ways of separating wire copy from the original publication of each news outlet in order to provide a clearer view of the gender gap in Canadian media, produced by the news outlets themselves. In future versions of the software, we are planning to visualize more fine-grained information about who is being quoted, separating politicians, witnesses and/or victims, from experts (as informed sources of analysis, context and opinion). We are also interested in how female and male expert quotations are distributed in the articles from different topics and written by female vs. male authors.

For more information about the research methodology, please contact:

Maite Taboada mtaboada@sfu.ca
Director, Discourse Processing Lab
Simon Fraser University