Task for Junior SWERIK Research Engineer

Erik Mandel
Homework Solutions for SWERIK research team
erik.mandel@outlook.com

I. Introduction and Data

The Riksdag's chamber protocols contain diverse utterances, including question-answer exchanges between MPs and ministers. This project programmatically identifies utterance sections corresponding to ministerial answers to written or oral questions (*"svar på skriftliga eller enkla frågor"*) using the SWERIK XML corpus. The output includes a list of answer-labeled utterance IDs and a temporal distribution plot.

II. Method

The XML corpus was parsed and processed using Python within a `Conda` environment and `Jupyter Notebook`. Due to the unreliable nature of surface cues such as question marks or keywords like *"fråga"* or *"frågor"*, the method focused instead on speaker roles and turn-taking patterns to identify ministerial answers.

Such surface cues frequently appear in both questions and answers, and politicians often use rhetorical or indirect formulations (e.g., *"I would like the minister to reflect on this issue"*) that simple keyword or punctuation rules fail to capture. Additionally, questions are often embedded within extended monologues or lack traditional interrogative markers, making lexical or structural matching insufficient for precise annotation.

Therefore, a heuristic approach was developed, based on formal parliamentary procedures and empirical observations:

- Ministers' utterances immediately following those of MPs within the same debate section were considered likely answers, reflecting parliamentary turn-taking conventions.
- If a minister initiated a new debate section, this was also marked as a potential answer, as it likely responds to a previously submitted question.
- Changes in topic, marked by new `<debateSection>` tags, delineated boundaries of Q&A sequences.
- Speaker identity and role were used to classify turns as questions (MP) or answers (minister), and only minister utterances were extracted as answers.

While advanced approaches such as large language models or semantic modeling would better capture the nuanced, context-dependent nature of this task—and were the preferred option—they were not feasible due to hardware constraints and the assignment's scope. Such models could improve precision and recall by interpreting rhetorical intent, discourse flow, and thematic continuity.

III. Results

The outcome of the assignment includes:

- A CSV file listing utterance IDs identified as ministerial answers.
- A time-series plot (Figure V) illustrating the yearly distribution of detected answers.
- Initial development and testing were conducted on a single XML file, with subsequent scaling to process the full corpus within project constraints.

IV. Discussion and Limitations

The task proved more challenging than expected. Although the rule-based approach provides useful results, it struggles with older protocols due to inconsistent formatting and granularity. In particular, the very low number of detected answers throughout much of the 1900s likely reflects these limitations rather than actual absence of data.

Key challenges included:

- **Ambiguity:** Many utterances contain rhetorical or implicit questions, making rule-based labeling difficult.
- **Formatting inconsistencies:** Older protocols lack clear structure, complicating segmentation.
- **Modelling limits:** Data size and time constraints prevented use of advanced methods like LLMs, which would likely improve results.

The method offers a baseline for full corpus processing but requires clearer definition of classification goals—e.g., for machine learning or qualitative analysis.

Introducing more detailed categories such as *speech*, *answer*, *question/remark*, and *description/context* could significantly improve precision and recall by better capturing the nuances of parliamentary dialogue.

Beyond enhancing rule-based heuristics, other methods to consider include:

- Training supervised classifiers on hand-labelled data.
- Applying discourse segmentation to identify topic boundaries.
- Using transformer models for deeper semantic and contextual understanding.
- Combining lexical, structural, and contextual features for refined classification.

Future directions also include expanding rule-based cues and improving historical alignment through better segmentation.

This work lays groundwork for analyzing how political communication evolves—potentially revealing trends toward more indirect or formal speech.

I hereby grant permission to the SWERIK research team to use the results, code, and files I have submitted

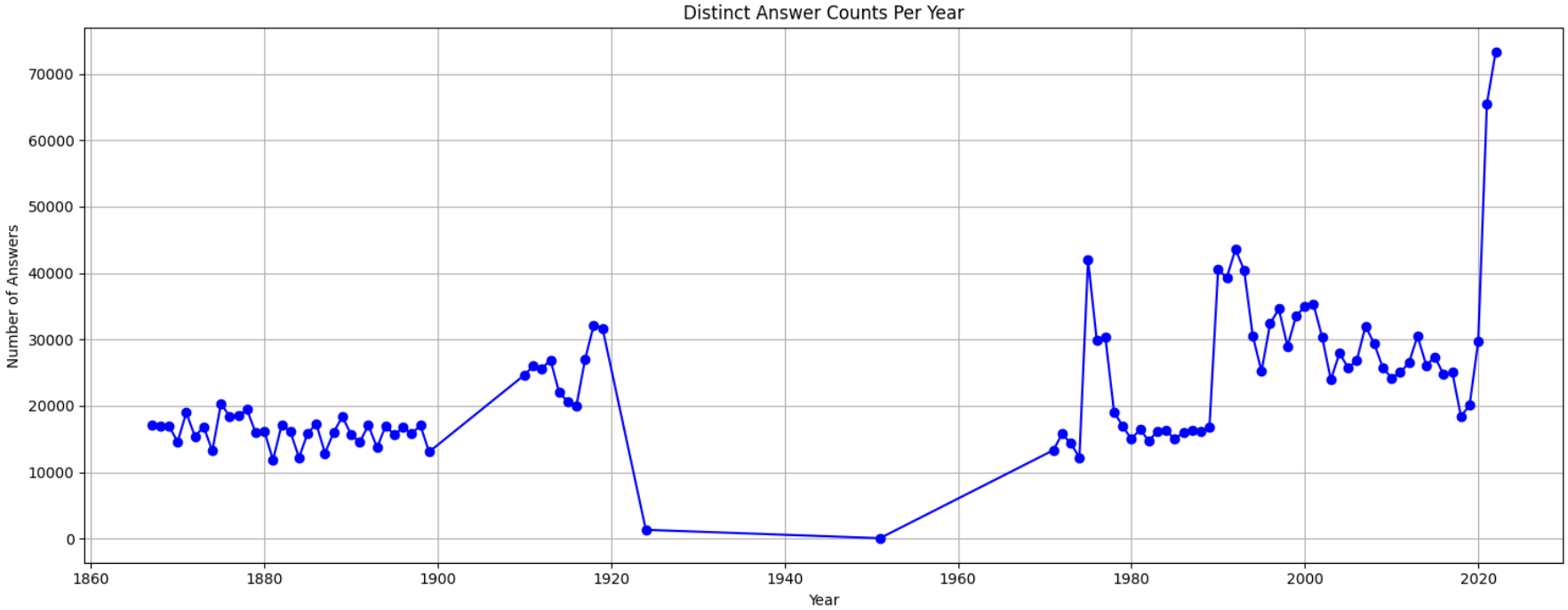

Figure A.1: Number of ministerial answers detected per year.