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 interactions between MPs and ministers. This homework attempts to programmatically identify sections that correspond to answers to written or oral questions (*"svar på skriftliga eller enkla frågor"*) using the SWERIK corpus. The output includes a list of utterance IDs labelled as answers and a graph showing their distribution over time.

II. Method

The XML corpus was parsed and processed using Python (executed in a `Conda` environment via `Jupyter Notebook`). Given the unreliable nature of surface cues—such as question marks or keywords like *"fråga"* or *"frågor"*—the approach instead relied on speaker roles and turn-taking patterns.

These surface cues appear frequently in questions and answers, and politicians often embed rhetorical or indirect questions (e.g., *"I would like the minister to reflect on this issue"*) that are hard to detect via simple keyword or punctuation-based rules. Furthermore, questions are often contextual, framed within long monologues, or phrased without traditional interrogative markers. This situation makes purely lexical or structural matching insufficient for precise annotation.

Instead, a heuristic was developed based on the formal procedure described in the task and empirical observations:

- Ministers are likely to answer when they speak after an MP within the same debate section. This pattern follows the parliamentary turn-taking structure.
- If a minister initiates a debate section, it is also likely to be a response to a previously submitted (written or oral) question. Thus, these initial statements were also marked as possible answers.
- A change in topic (new `<debateSection>`) was treated as the boundary marking the end of a Q&A sequence.
- The speaker identity and role were used to classify whether a turn was a question (MP) or an answer (minister), and only utterances by ministers were collected for the output.

While large language models (LLMs) or semantic modelling would be much more suitable for this kind of nuanced, context-dependent task—and would have been the preferred approach by the author—these methods were not feasible within the constraints of available hardware and the scope of the assignment. Such models could better interpret rhetorical intent, discourse flow, and thematic continuity, potentially improving precision and recall.

III. Results

Based on the assignment, the author produced the following:

- A CSV file containing utterance IDs labelled as answers.
- Figure 1 showing the number of detected answers per year.
- Initial testing was done on a single XML file before scaling to the full corpus, due to time constraints.

IV. Discussion and Limitations

The task proved more challenging and rewarding than expected. Despite its simplicity, the current rule-based approach yields useful results, though it struggles with older protocols due to differences in granularity.

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 solution is robust enough for full corpus processing and provides a baseline for further work. Clarifying the classification goal—whether for machine learning or qualitative analysis—is an important next step.

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.

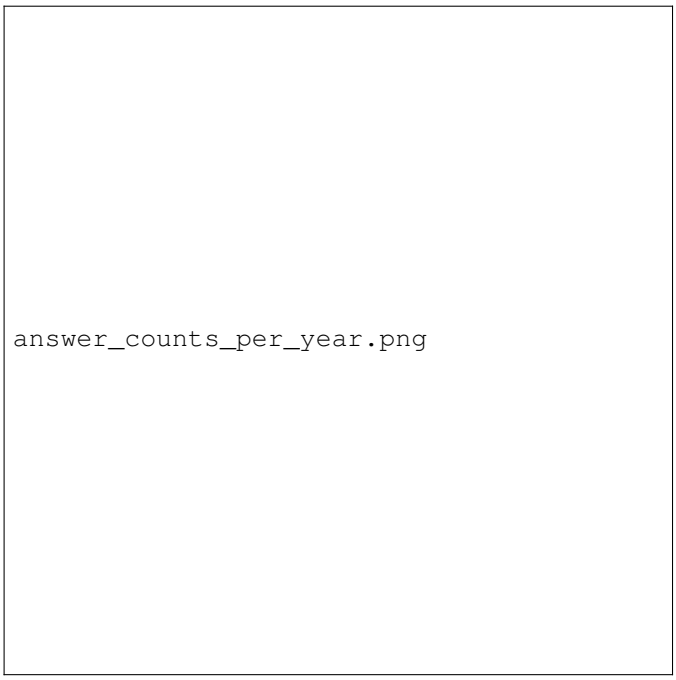

Fig. 1. Number of ministerial answers detected per year.