A New Data Set for Speaker Attribution in German Parliamentary Debates

I. Motivation For Data Set Creation

A. Why was the data set created?

The annotated dataset has been created to train machine learning systems for the task of speaker attribution, to enable text analyses in the political domain.

The goal of this task is the identification of speakers in political debates and in newswire, and the attribution of speech events to their respective speakers. Being able to identify this information automatically, i.e., identifying who says what to whom, is a necessary prerequisite for a deep semantic analysis of unstructured text.

B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

The raw data includes transcripts of parliamentary debates from the German Bundestag. This data is freely available from `https://www.bundestag.de/services/opendata` and has already been used in different projects and publications, mostly by political scientists.

C. What (other) tasks could the dataset be used for?

Beyond serving as training data for speaker attribution systems, we expect that the dataset will also be interesting for corpus-linguistic investigations or discourse studies of parliamentary debates and, more general, of argumentative text and political communication.

D. Who funded the creation dataset?

The annotation work has been supported by the German Society for Computational Linguistics & Language Technology (GSCL) and by the Baden-Württemberg Ministry for Science, Research and the Arts.

II. Dataset Composition

A. What are the instances?

Our data is a text corpus of political speeches by members of the German Bundestag. We provide the data in json format. Each document includes the text for a speech held in the German Bundestag on a specific agenda item. In addition to the raw text, the json files include the annotations, i.e., the cue word(s) that trigger a speech event and the corresponding role spans (Source, Addressee, Message, Topic, Medium, Evidence).

B. How many instances are there in total?

The dataset includes text from 267 speeches held in the German Bundestag by 196 different speakers (213,617 tokens). The time frame covers the 19th legislative term (2017–2021). The data has been split into train, development and test data. More detailed information on the number of annotations will be added once the adjudication process is completed.

C. What data does each instance consist of?

Each instance consists of the text of one paragraph where all words that trigger a speech event (also including writing and thought) have been annotated as cues. Each cue is linked to its roles, encoding the Source of the speech event, its Addressee, Topic, Evidence, Message, the Medium used to convey the message and obligatory particles (for example, separated verb prefixes).

D. Is there a label or target associated with each instance? If so, please provide a description.

For more information on the annotation scheme, please refer to the annotation guidelines available from our anonymous github repository.

E. Is any information missing from individual instances?

The dataset has been created from transcripts of the parliamentary debates and should thus be considered as a normalised version of the original speech data. We do not include the audio files in the corpus (those are, however, accessible at `https://www.bundestag.de/dokumente/textarchiv/`. We also removed all comments from the speeches so that the documents (speeches) only include speech events produced by the politician who gave the speech.

F. Are relationships between individual instances made explicit?

The relation between individual speakers can be inferred through the meta-information provided in the file names (e.g., the speaker names and party affiliation of the different speakers). The information on date and agenda item is also included and allows to reconstruct which speeches have been given on the same day and topic (however, the topic itself is only specified on an abstract level, e.g., "Agenda item 1").

G. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset is a sample of the speeches from the 19th legislative term (2017–2021) of the German Bundestag. The distribution of topics in the data is not representative of the larger data but has been sampled to cover a more diverse range of topics, with contributions from all parties distributed over the whole legislative term. Below, we describe the sampling procedure in more detail.

a) Sampling procedure: We extracted a dataset of parliamentary debates from the German Bundestag, covering a time period from the 19th legislative term (2017 to 2021).[1] The corpus includes speeches by 807 different speakers, with over 900,000 sentences and over 16 mio tokens. From this corpus, we selected individual speeches for annotation as follows. Our goal was to create a gold standard, controlled for topic and including speeches for each of the political parties. In addition, we wanted the texts to be evenly distributed over the time span of the legislative term (2017–2021). To achieve this goal, we selected specific agenda items that covered a range of topics, and then sampled all speeches that belong to this specific agenda item, to increase the comparability of the contributions made by speakers from different parties.

b) CAP topics: We based our topic selection on the coding scheme developed in the Comparative Agendas Project (CAP) [1]. The coding scheme includes 21 major topics (see Table I) and more than 200 fine-grained subtopics. The topics we selected have been annotated as major CAP topics, which allowed us to use the annotated CAP data to train a topic classifier.

c) Training a CAP topic classifier: For training data, we used the Parliamentary Question Database[2], a data set with more than 10,000 major and minor interpellations posed by parliamentarians to the government. The data set ranges over the 8th to the 15th legislative periods (1976–2005). Each interpellation has been assigned to a major and a minor topic, according to the CAP coding scheme.

Before training, we did some standard preprocessing and clean-up of the data where we lower-cased the text and used a number of regular expressions to remove non-ascii characters, listings of politicians' names, header and footer information and so on. We also removed stopwords and punctuation and extracted a tokenised and lemmatised version of the speeches.[3] This resulted in a training set with 10,033 interpellations, with an average length of 388 tokens per interpellation. We then trained a feature-based classifier, based on tf-idf weighted bag-of-words (BOW) features. We experimented with different classifiers provided by the scikit-learn library[4] and found that the linear SVM gave us best results for predicting topics on the interpellations. For the 21 major topics, our classifier achieves a micro F1 of 72.9% on the indomain interpellation data.

1	Cultural Policy Issues
2	Defense
3	Domestic Macroeconomic Issues
4	Education
5	Environment
6	Health
7	Immigration and Refugee Issues
8	Law, Crime, and Family Issues

TABLE I
MAJOR TOPICS FROM THE COMPARATIVE AGENDAS PROJECT THAT WE SAMPLED TO BE INCLUDED IN OUR DATA SET.

d) Sampling based on predicted CAP topics: We then used the classifier to predict topics for each speech in the parliamentary debates, after applying the same preprocessing steps to the data. This gives us topic predictions for each individual speech. To guide our sampling process, we aggregated the predictions for all speeches belonging to the same agenda item. We call the topic based on a "majority vote" for each agenda item the *major topic* of the agenda. Our assumption is that all speeches given on the same agenda item should belong to the same major topic. As a result, we obtained a distribution of topics over all speeches for each respective agenda item. We sorted the predictions and *manually selected and validated* agenda items for each of the CAP topics in Table I, where the majority of the speeches for this agenda item have been predicted as belonging to this topic.

We only selected agenda items where each of the political parties participated in the debate, and also aimed at selecting items that are roughly evenly distributed over the time period of the legislative term, to *ensure that our dataset is covering a range of different topics, is distributed evenly over the whole legislative term and includes speeches from all different parties on the same set of topics.*

H. Are there recommended data splits (e.g., training, development/validation, testing)?

We provide the train/dev/test splits used in the experiments described in our paper. Please note that we assured that none

[1] The data is freely available from https://www.bundestag.de/services/opendata, the Open Data service of the German Bundestag.

[2] https://www.comparativeagendas.net/datasets_codebooks

[3] For lemmatisation, we used the spaCy library: https://spacy.io with the de_core_news_sm model.

[4] https://scikit-learn.org

of the agenda items in the test set are included in the training set. This results in a more realistic setting as compared to distributing speeches from the same agenda item into training and test sets.

I. Are there any errors, sources of noise, or redundancies in the dataset?

While we removed comments from the speeches to avoid including speech events that have been produced by persons other than the speaker, the speeches might include some interposed questions or closing remarks not properly marked in the XML version of the data.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained and does not rely on other external resources. But note that the audio and video data for the speeches can also be accessed at https://www.bundestag.de/.

The raw data is in the public domain. The annotated version of the data will be made available under the Creative Commons BY-SA 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/).

III. Collection Process

A. How was the data collected?

The data has been downloaded from the open data service of the German Bundestag who provide the transcripts of all recent debates in XML format: https://www.bundestag.de/services/opendata.

B. If the dataset is a sample from a larger set, what was the sampling strategy?

See Section II, G.

C. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Three of the organisers of the shared task from Mannheim university and IDS Mannheim and four student assistants were involved in the data creation process. Students were paid in accordance with the applicable collective bargaining agreements at Mannheim University.

D. Over what timeframe was the data collected?

The data was collected in January 2021.

IV. Data Preprocessing

A. Was any preprocessing/cleaning/labeling of the data done?

We removed comments and tokenized the data.

B. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The raw data is available from https://www.bundestag.de/services/opendata in XML format.

C. Is the software used to preprocess/clean/label the instances available?

We used the German spaCy model de_core_news_sm for tokenization (https://spacy.io/).

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?

To be answered in July 2023.

V. Dataset Distribution

A. How will the dataset be distributed?

We will make the data available via our university's GitHub account.

B. When will the dataset be released/first distributed?

The data has been released in an anonymous github with the submission of our paper to COLING-LREC 2024.

C. Are there any copyrights on the data?

No.

D. Are there any fees or access/export restrictions?

No.

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

The dataset will be distributed via the GitHub account of our university.

B. Will the dataset be updated?

No.

C. If the dataset becomes obsolete how will this be communicated?

We do not foresee a scenario where the dataset will become obsolete.

D. Is there a repository to link to any/all papers/systems that use this dataset?

No.

E. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?

The data is available via the Creative Commons BY-SA 4.0 license, so others may extend/augment/build on this dataset, given that they also make the new resource available under the same license.

VII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., by an institutional review board)?

No.

B. Does the dataset contain data that might be considered confidential?

No.

C. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

The data might include racist and discriminating statements by politicians that might be considered as offensive.

D. Does the dataset relate to people?

Yes.

E. Does the dataset identify any subpopulations (e.g., by age, gender)?

The dataset includes speeches by members of the German parliament, held in the Bundestag. The data collection was conducted by the Bundestag itself and all speakers were aware of the data collection and consented to it.

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

Yes, all speakers are known.

G. Were the individuals in question notified about the data collection?

The data collection was conducted by the German Bundestag and all speakers were aware of the data collection and consented to it. In addition, the recordings of all debates are freely available on the Bundestag website.

H. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

No.

References

[1] Shaun Bevan. Gone Fishing: The Creation of the Comparative Agendas Project Master Codebook. In Frank R. Baumgartner, Christian Breunig, and Emiliano Grossman, editors, *Comparative Policy Agendas: Theory, Tools, Data*. Oxford: Oxford University Press, 2019.