**An Analysis of the German Bundestag Debates:**

**Filtering and Evaluating mentioned Entities in debates through Named Entity Recognition**

**Aust, Lisa and Tepavac, Daniel**

1. **Introduction**

It has been 74 years since the initial constituent meeting of the first German Bundestag. Over this period, 20 Bundestage came together to lead the German population through numerous troubling and tense times like the Cold War, the union of West Germany and DDR or the Covid-19 pandemic. During debates in the Bundestag politicians search for solutions regarding occurring conflicts and everyday issues which are often connected to a specific location somewhere around the world.

We present results of a project applying Named Entity Recognition (NER) on a large corpus of debates in the Bundestag. With these results statistical operations are performed and interpreted in the context of world politics at the present time. As underlying corpus serves the “**DeuPARL**”-Corpus (Walter et. al, 2021), which consists of protocols directly from the official website of the Bundestag and saves them mostly as raw, unstructured text.

We perform NER on this corpus to extract the Named Entity *location* for every protocol. An *entity* is a broad term covering everything which can be thought as a unit (Jannidis et. al, 2017). Based on this concept *Named Entities* are important words or phrases holding specific key information for a certain topic (Zitouni et. al, 2014). Typical examples are persons, organizations, or, in our case, *locations*. Those key information are highlighted through the process of *Named Entity Recognition* which then can be used for further language processing applications. We use the model “**NER for German Legal Text in Flair**” [[1]](#footnote-1) provided by the HuggingFace library, since the corpus is in German.

Our goal is to investigate and visualize the development of entity mentions throughout the decades to get a better understanding of the political discourse. Our research has implications for political science and explores quantitative and computational methods for this research area.

1. **Related Work**

Since NER is used as tool in various research, we want to give an overview about NER itself, and also in regard to research using NER for location extraction.

Furthermore, there is an insight in the research field dealing with the German Bundestag debates.

* 1. **Named Entity Recognition**

Named Entity Recognition is the first step in an Information Extraction pipeline. It is used to search through large amounts of text to extract specific words or entities. The four most common entity types are people, organization, location, and geo-political entity. Named Entity Recognition can also be framed as a sequence-labelling task with the availability of different data labelling formats.

There are three main approaches to sequence labelling in form of NER: Rule-based, feature-based (MEMM and CRF) and neural (BERT, bi-LSTM). The popularity of the approaches has changed accordingly to new developments in the broader field of Natural Language Processing.

Nowadays state-of-the-art performances are achieved using neural networks. In 2016, *Lample et al* introduced neural architectures. One using bidirectional LSTMs and conditional fields and another one deploying the construction and labelling of segments using a transition-based approach inspired by shift-reduce parsers. In 2018, *Akbik et al*. proposed contextual string embeddings for sequence labelling. Their approach achieved new state-of-the-art results in English and German named entity recognition. The authors furthermore released their pre-trained language models (see Flair embeddings). In 2021, *Wang et al.* achieved good results with their ACE + document-context model. They proposed Automated Concatenation of Embeddings (ACE), which automates the process of finding better concatenations of embeddings for structured prediction tasks.

* 1. **Uses of NER**

The method of Named Entity Recognition has many use cases. It can be applied to improve the speed and relevance of search results and is therefore valuable for search and recommendation engines or content classification among others. This functionality is also applied in health care for extracting relevant information from medical data sets. One area of application of NER is the medical data analysis. *Chen et al.* automatically reviewed medical charts for breast cancer outcome research. In 2021, *Frei and Kramer* presented their neural NLP model GERNERMED and its application for detecting medical entity types in German text data.

*Girsang et al.* applied NER to extract geolocations of types of natural disasters obtained from Twitter text. These geolocations were then classified into eight classes of natural disasters using the Support Vector Machine method.

Data mining in form of NER has also been researched in the field of political sciences. In 2019, *Imani et al.* performed a primary focus location extraction in political news reports in different languages using a variety of NER tools. Automatically detecting the geolocation of reported news events helps the contextualization and understanding of the associated events.

* 1. **Political Debates**

*Walter et al.* analysed the changes that happen in German parliamentary proceedings over the course of time with a focus on political and racist bias. Their research indicated the viability of analysing historical trends in a corpus consisting of German parliamentary proceedings. Political discussions of the German Bundestag have also been analysed by *Cäcilia Zirn*. The dataset consisted of transcriptions of political speeches from meetings of the German Bundestag and party manifestos. The paper provided a fine-grained analysis of speakers’ positions on political data.

1. **Research Question (preliminary)**

As the focus on this paper lays on the examination of all locations being mentioned in the German Bundestag debates since 1949, we perform statistical evaluations on the results of this examination. First, there is a general, statistical evaluation (How often is each location mentioned?, On which locations is the focus in different time periods?, Is a there valid and causal comparison between time periods or locations possible?, etc.). Second, we consider the historical context and interpret the mentioning of a location regarding this. Finally, we visualize our findings with appropriate graphs.

1. **References**

Akbik, Alan and Blythe, Duncan and Vollgraf, Roland (2018): Contextual String Embeddings for Sequence Labeling. In: *COLING 2018, 27th International Conference on Computational Linguistics*, S. 1638–1649.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. [Contextual String Embeddings for Sequence Labeling](https://aclanthology.org/C18-1139). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

M. Bahojb Imani, L. Khan and B. Thuraisingham, "Where Did the Political News Event Happen? Primary Focus Location Extraction in Different Languages," 2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC), Los Angeles, CA, USA, 2019, pp. 61-70, doi: 10.1109/CIC48465.2019.00017.

Suganda Girsang, Abba and Noveta, Bima Krisna, Location Prediction Using Named Entity Recognition for Indonesia Natural Disasters in Data Twitter. Available at SSRN: https://ssrn.com/abstract=4276345 or http://dx.doi.org/10.2139/ssrn.4276345

Frei, Johann, and Frank Kramer. "GERNERMED: An open German medical NER model." Software Impacts 11 (2022): 100212.

Hunter, J.D (2007): „Matplotlib: A 2D graphics environment. In: *Computing in Science \& Engineering*. S.90-95.

Jannidis, Fotis; Kohle, Hubertus; Rehbein, Malte. 2017. „Digital Humanities – Eine Einführung“, page 103, J.B. Metzler, Stuttgart. Springer-Verlag GmbH Deutschland.

Jurafsky, Daniel; Martin, James H. (2022): Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Leitner, Elena; Rehm, Georg; Moreno-Schneider, Julian: Fine-grained Named Entity Recognition in Legal Documents. S. 272–287. Online verfügbar unter https://link.springer.com/content/pdf/10.1007%2F978-3-030-33220-4\_20.pdf.

Walter, Tobias, et al. "Diachronic analysis of german parliamentary proceedings: Ideological shifts through the lens of political biases." 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, 2021.

Walter, Tobias, et al. DeuParl. 2021. https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2889

Lample, Guillaume, et al. "Neural architectures for named entity recognition." arXiv preprint arXiv:1603.01360 (2016).

Wang, Xinyu, et al. "Automated concatenation of embeddings for structured prediction." arXiv preprint arXiv:2010.05006 (2020).

Zirn, Cäcilia. 2014. [Analyzing Positions and Topics in Political Discussions of the German Bundestag](https://aclanthology.org/P14-3004). In *Proceedings of the ACL 2014 Student Research Workshop*, pages 26–33, Baltimore, Maryland, USA. Association for Computational Linguistics.

Zitouni, Imed (Hg.) (2014): Natural Language Processing of Semitic Languages. Berlin, Heidelberg: Springer Berlin Heidelberg; Imprint; Springer (Theory and Applications of Natural Language Processing).

1. <https://huggingface.co/flair/ner-german-legal> [↑](#footnote-ref-1)