The creation of a benchmarking dataset for resolving ambiguous references to different mentions of people in political text

Bachelor Thesis

presented by
Angelina Basova
Matriculation Number 1653052

submitted to the
Data and Web Science Group
Prof. Dr. Simone Paolo Ponzetto
University of Mannheim

August 2021

Contents

1 Introduction **1**

2 Literature Review **4**
2.1 Sequence Labeling 4
2.2 Named entity recognition approaches 5
2.3 BERT 6
2.4 Weak supervised learning 9
2.5 Applications of Sequence Labeling 10
2.6 Research on populism communication 11

3 Methods **13**
3.1 Annotation Guidelines 13
3.2 Data Collection 14
3.3 Definition of Linguistic Rules 16
3.4 Creation of Training set 17

4 Experimental Evaluation **18**
4.1 Evaluation Metrics 18
4.2 Experiments 19
4.3 System evaluation 19
4.4 Error analysis 21
4.5 Attribute-wise evaluation 23

5 Conclusion **28**
5.1 Summary 28
5.2 Future Work 29

A Annotation guidelines **37**

B Linguistic Rules **46**

C Further Experimental Results **48**

List of Figures

2.1 BERT input representation as a sum of token-wise token embeddings, segment embeddings and position embeddings. Adapted from [11] 6
2.2 Transformer encoder architecture. Adapted from [1] 7
2.3 BERT architecture during pre-training. Adapted from [11] 8
2.4 BERT architecture during fine-tuning on named entity recognition. Adapted from [11] 9

4.1 F1 scores of BERT system based on bucket-wise evaluation on entity frequency and token frequency. The x-axis contains the attribute values of each bucket. At the top of the chart are the number of entities, that correspond to each bucket. 24
4.2 F1 scores of BERT system based on bucket-wise evaluation on entity length and sentence length. The x-axis contains the attribute values of each bucket. At the top of the chart are the number of entities, that correspond to each bucket. 26
4.3 F1 scores of BERT system based on bucket-wise evaluation on entity density. The x-axis contains the attribute values of each bucket. At the top of the chart are the number of entities, that correspond to each bucket. 27

C.1 Confusion matrix of rule-based and dictionary-based system . . . 50
C.2 Confusion matrix of BERT system 51

List of Tables

4.1 Distribution of classes in training and manually labeled set 20
4.2 Evaluation of dictionary- and rule-based system and BERT system 21

C.1 BERT class-wise evaluation based on two different evaluation schemes, partial and exact. The exact scheme considers predictions as correct if the boundaries match exactly to the gold-standard, regardless the class. The partial scheme considers predictions as correct if the boundaries match partially to the gold-standard, regardless the class. For each scheme the number of correct, incorrect, partial, missed and spurious references are shown. 49

Chapter 1

Introduction

This thesis is dedicated to the automatic recognition of references to people and groups of people in English parliamentary debates of the European Parliament. I create a big corpus that allows to identify and classify references to people that match to multiple classes. The goal is to provide the necessary data to facilitate future research on populism communication.

During the last decade populist parties have increased their representation in the European cabinet [18]. Traditionally, populism was defined as something bad and was connected to a political party family or a type of politician [10]. Nevertheless, to increase their representation, populist actors need to communicate their ideas [10]. Populist messages are characterized by a separation of the society into the 'decent people' and the 'corrupt elite' [34]. Populists appeal to give a voice to the 'decent people', whose interests are neglected by the 'corrupt elite'. Setting the focus on the way of how the 'decent people' and the 'corrupt elite' are communicated by political actors, allows to view populism as a way of political communication [10]. This allows to study a new concept of populism, namely degrees of populism. Instead of viewing populism as a binary value, either as present or not, the degree to which elements of populism communication are used, can be identified. Elements of populism communication are references to the 'decent people' and the 'corrupt elite'. There is a lot of potential in the study of populism degrees as de Vreese et al. [10] suggest. New insights could fill the gaps in populism research.

Most research on populism communication [46], [9] is based on manually analyzed data by trained coders. However, the collection of manually analyzed data is time-consuming and usually scarce. This hinders researchers at analyzing a significant number of political speeches in texts. In order to process a big amount of data, automating the task is necessary.

A suitable automatic approach is sequence labeling. Sequence labelling is a subtask of information extraction [15]. It aims to identify and classify entire phrases, namely sequences, of text into predefined categories.

Although the needed technology to identify references to people is there, automating the task was not possible due to lack of a dataset focused on populism communication. Major datasets [42], [36] only recognize references to people, that are grammatically proper nouns. For instance, references such as Mr. Johnson or Barack Obama. However, references of people such as the Germans or the Democrats are not recognized by any of the benchmarking datasets. Such references are grammatically common nouns, since they refer to a class of entities.

In order to enable the automatic recognition and classification of references to people, I develop the first benchmarking dataset. The relevant dataset for this thesis is the EuroParl dataset [27]. EuroParl consists of proceedings from the European Parliament which aligns with the focus of this thesis on political text. In this context, I create the needed dataset by achieving three key milestones. First, I define different classes of people and provide the annotation guidelines. Second, I annotate a part of the EuroParl dataset with weak supervision in a dictionary-based and rule-based approach. For each class a combination of lists of data and linguistic rules is used to automatically annotate a training set. Third, the resulting training set is used to train a BERT model. BERT is used to create the benchmarking dataset for this thesis.

The research questions that I aim to answer are the following:

1) Is it worth to use a ML-learning approach to resolve ambiguous references?

2) How good is the dataset to recognize mentions to people? How well does the dataset resolve ambiguity?

Machine-learning models have shown superior performance in comparison to other approaches of sequence labeling [24], [48]. Their ability to generalize and recognize unseen data is an important factor. For this reason, I test whether a machine learning model is able to understand the task at such a degree, that it recognizes how to correctly classify a sequence that matches to multiple classes. Through the model, I evaluate how good the resulting dataset is and which limitations arise.

The outcome of this work is a sufficiently large corpus for information extraction of categories of people.

Future researchers studying populist communication could benefit from the work of this thesis. The general classes are obtained from a study on populist communication by Wirth et al. [46]. The classes are specific person, political party or coalition, multiple persons, elite, groups of people, supranational institutions, own person and others. Although the definitions of the classes were altered, the two key actors, the 'decent people' and the 'corrupt elite', are represented by the

classes. The corpus can be used to identify and classify references to people in political speeches. This in turn allows to quickly analyze which classes a political actor references the most and to which ratio political speeches consist of references to people. Moreover, the study of the new concept in populism communication, namely the degree of populism, becomes possible.

The remainder of this thesis is organized as follows. In Chapter 2, I introduce sequence labeling and its sub-task, named entity recognition. I define their similarities and explain why the named entity recognition architecture can be used for sequence labeling. I then present common approaches to named entity recognition and explain the advantages and disadvantages of each approach. Then, I describe a powerful machine learning model called BERT. In addition, I introduce weak supervision, a method for collecting training data for machine learning models. Finally, I summarize relevant research on populism communication.

In Chapter 3, I describe the applied methods. First, I introduce the annotation guidelines. Second, I detail all the datasets that were used to create the dictionaries and rules, and explain some linguistic rules. Finally, I explain how the training set was created using dictionaries and rules.

In Chapter 4, I report the experiments and their results. I present evaluation metrics for sequence labeling. Then, I show the results of a dictionary and rule-based system and a BERT system. I then perform an error analysis and discuss which references remain ambiguous. In addition, I evaluate the BERT system using selected attributes.

Chapter 5 presents my conclusions and future work.

Chapter 2

Literature Review

In this chapter, I introduce sequence labeling and outline its subtask, named entity recognition. I describe similarities and differences between the two tasks and explain why named-entity recognition approaches can be used for sequence labeling. Then, I discuss the main approaches to named entity recognition. I emphasize machine learning as the best approach and introduce a machine learning model called BERT. Next, I present a technique for data collection called weak supervision. I also mention some applications of sequence labeling. This includes selected corpora and a tagger. Finally, I provide an overview of relevant research on populism communication.

2.1 Sequence Labeling

Sequence Labelling is a task of information extraction and involves assigning labels to words in a sequence [15].

In addition to labels an IOB tagging scheme can be adopted [37]. IOB stands for **I**nside, **B**eginning and **O**utside. **B** and **I** is used to indicate the start and end position of predefined labels. **O** is used to label words, that don't match to the predefined labels. For instance the sequence 'Wall Street' consists of two words. The word Wall is labeled as B-Street to indicate that Wall is the first word of the street. The word Street is labeled as I-Street to indicate that this word is inside a street label. Any word that is not part of a street name is labeled as O.

A popular application of sequence labeling is named entity recognition. In order to perform sequence labeling, the architecture of a sequence labeling subtask called named entity recognition [15] can be utilized.

Named entity recognition is similar to sequence tagging in that it involves the extraction of sentences and words from texts. The main difference is grammatical.

Grammatically, the extracted phrases and words consist of proper nouns. Proper nouns, also called named entities, are specific names for something concrete, such as Hanna. In contrast, common nouns are names for classes such as girl. Proper nouns are capitalized, while common nouns are not. Despite the grammatical difference, named entity recognition approaches can be applied to the task of sequence labeling.

2.2 Named entity recognition approaches

To identify named entities three major approaches exist: dictionary-based, rule-based and machine learning-based [23].

A dictionary-based approach utilizes a list of named entities also called gazetteer, dictionary or lexicon. The dictionary is used to match every token in the text against the dictionary. This approach heavily relies on quality data. New entities that are not in the dictionary can't be identified. Challenges arise when a domain uses multiple spelling variations, synonyms or new words develop. This especially applies to dynamic domains such as company names, person names and the medical domain [30]. Problems with ambiguous names remain.

A rule-based approach utilizes handcrafted rules of lexical-syntactic patterns and semantic constraints. This includes the orthography such as capitalization and part of speech tags such as nouns. It can take into account the context of named entities. Also dictionaries can be used as part of a rule to identify named entities. This approach is mostly suited for entity types that often occur in a specific context. For instance, street names often end with the word "Street" and are mentioned immediately after the word of. The advantage is that a language's syntax doesn't change often. This allows to define rules that are valid for a long period of time. The disadvantage is that it requires a huge experience and grammatical knowledge of a given language or domain. Thus, it results into named entity recognition systems that are difficult to transfer to other domains and languages and remain language or domain specific. Domain specific rules in combination with incomplete dictionaries often lead to systems with high precision and low recall [31]. The performance of named entity recognition systems decreases with noisy unstructured data sources.

A machine learning-based approach uses a probability model. This approach can perform very good as the model can also recognize unseen entities. The ability to recognize unseen entities is called generalization. However, the model usually needs a large amount of high-quality training data [23]. Thus, this approach is best suited for domains with large amount of annotated data available. It allows better generalization and possibly good performance even when the language or the domain terminology changes.

2.3 BERT

Machine learning approaches have shown superior performance in sequence labeling tasks in comparison to rule-based and dictionary-based approaches [24], [48]. The ability to generalize is key to identify variations of sequences, especially complex sequences.

A special family of machine learning models is deep learning models [21]. Deep learning models are based on artificial neural networks [29]. They achieve state-of-the-art results on sequence labeling [31] due to their ability to automatically learn complex features of data.

A deep learning model that rapidly gained in popularity is BERT. BERT is a language representation model introduced by Devlin et al. [11]. Its name stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT achieved state-of-the-art results in a number of natural language processing tasks such as named entity recognition. The major advantages of BERT are its conceptually simple implementation and its powerful language understanding [11]. To implement BERT, only two steps are necessary: pre-training and fine-tuning.

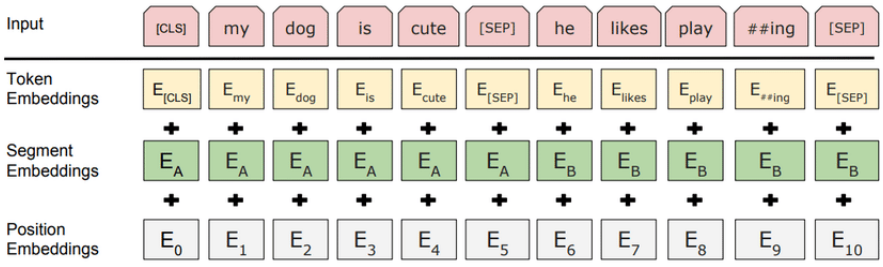


Figure 2.1: BERT input representation as a sum of token-wise token embeddings, segment embeddings and position embeddings. Adapted from [11]

For both steps, pre-training and fine-tuning, the architecture is very similar. The input is a sequence of sentences with two additional tokens as shown at the top of Figure 2.1. The first additional token is [CLS] and signals the beginning of the input. The second token [SEP] is used to separate individual sentences. The processing of the input is also the same during pre-training and fine-tuning. First words are tokenized into wordpieces [47]. Wordpieces are a limited set of common sub-word units. For instance, the word playing is split into two wordpieces, play and ##ing as shown at the top of Figure 2.1. Then, each wordpiece is converted into a word embedding E. The word embedding is the sum of the token, segment and position embeddings as illustrated in Figure 2.1. The token embedding is the

wordpiece embedding of a token. The segment embedding indicates whether a token belongs to sentence A or B. The position embedding indicates the position of the token in the input sequence. The resulting word embeddings are processed by multiple Transformer encoders, which are visualized by the blue box in Figures 2.3 and 2.4.

The Transformer encoders are stacked on each other. The architecture of a single Transformer encoder is illustrated in Figure 2.2. It consists of a self-attention and a feed forward neural network. The first component, self-attention generates an attention vectors for each word of the input. Its called an attention vector because more attention is given to relevant words of the input. The self-attention layer performs self-attention bidirectional, in other words on both directions. The output is processed by a feed forward neural network, which in turn outputs an encoded vector for each word. As the input of the Transformer encoder is an embedding for each individual wordpiece, there is no need to process the wordpieces in order. For this reason, the embeddings are processed in parallel to reduce the processing time.

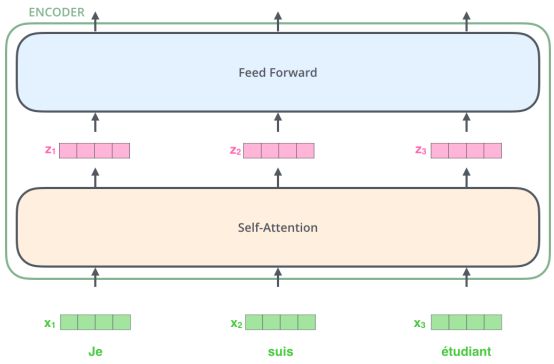


Figure 2.2: Transformer encoder architecture. Adapted from [1]

Figure 2.3 illustrates the architecture during pre-training. The input is shown at the bottom and the output is shown at the top. During pre-training, the model is trained on unlabeled data over two unsupervised tasks simultaneously. The tasks are masked language model and next sentence prediction. Masked language model randomly replaces 15% of the input tokens with a [MASK] token and predicts the masked tokens based on the context. The predictions are illustrated at the top of Figure 2.3 as Mask LM. The objective is to learn word relationships by training the model deep bidirectional on both left and right context. In other words, the model learns from the right and the left side of each token. Next sentence prediction predicts for pairs of sentences whether a sentence A follows a sentence B as shown

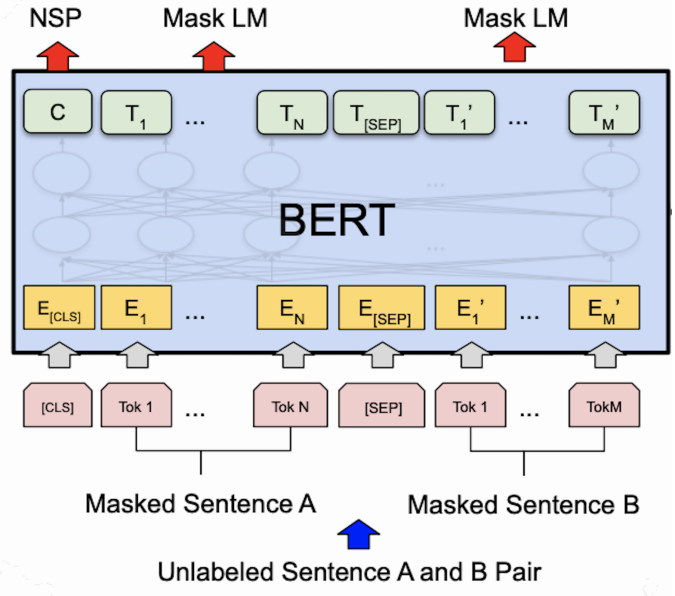


Figure 2.3: BERT architecture during pre-training. Adapted from [11]

at the bottom of Figure 2.3. The prediction is a binary value shown at the top left of the Figure as NSP. This task trains the model to understand relationships of sentences. To sum up, during pre-training the model learns to understand language by learning relationships of words and relationships of sentences.

During the fine-tuning step, BERT learns to perform a downstream task such as named entity recognition, by getting input and output data. Figure 2.4 illustrates the architecture of BERT during fine-tuning on named entity recognition. As in pre-training the input is shown at the bottom and the output is shown at the top. The input is a sequence of annotated sentences with the additional tokens [CLS] and [SEP] as shown at the bottom of Figure 2.4. Similar to pre-training, the input is processed by multiple Transformer encoders stacked on each other. The output is the prediction of each token into a pre-defined label such as O and B-PER.

The majority of model hyperparameters are the same as in pre-training except the batch size, the learning rate and the number of epochs. These three hyperparameters can be optimized by following the recommendations of the authors.

The advantage of BERT is that a number of pre-trained models are available for use. One of the pre-trained models for the English language is $BERT_{LARGE} - cased$. $BERT_{LARGE} - cased$ was trained on cased English with 24 layers, in other words Transformer encoders, 16 attention heads and 340 million parameters.

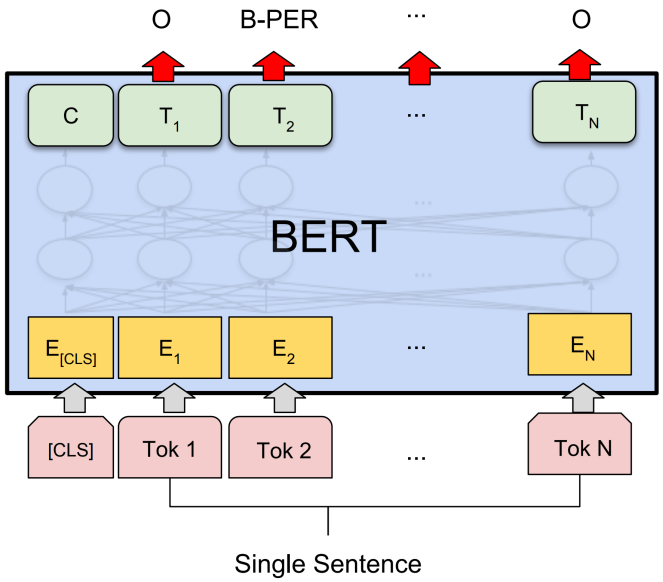


Figure 2.4: BERT architecture during fine-tuning on named entity recognition. Adapted from [11]

2.4 Weak supervised learning

When training a machine learning model in a supervised approach a large amount of annotated data is needed. Traditionally, annotated data was collected manually by domain experts [15]. This approach is time-consuming and expensive. A semi-automatic technique to annotate large amount of data is weak supervision [40].

Weak supervision allows to programmatically annotate a large amount of data without the high cost of time. Although the annotations are not as accurate as manual annotations, the high quantity results into models, that reach a reasonably high accuracy [40], [25]. A weak supervision paradigm is data programming [38].

Data programming involves creating training sets for machine learning models by using multiple labeling functions. A labeling function is a set of heuristic rules based on knowledge bases, libraries, ontologies or a combination of these types [38]. To capture all classes multiple labeling functions are implemented. This allows to collect noisy probabilistic labels.

2.5 Applications of Sequence Labeling

As machine learning models showed superior performance on sequence labeling, researchers use a plethora of methods during the implementation. This Section focuses on the implementation of sequence labeling in combination with machine learning models and the respective strengths and weaknesses of each implementation.

For the agricultural domain, Malarkodi et al. [41] create a sequence labeling system. They first define a set of 19 fine-grained classes. Then, they train a machine learning model with an annotated dataset created with weak supervision. They create 8 different training sets, with each training set consisting of a specific feature set such as part of speech tags and chunk information. Their system achieves the best performance when multiple features were used to create the training set. Despite a rich feature set, the authors perform a post-processing step in a rule-based approach. Interestingly, the added post-processing step decreases the precision by 1% and increases the recall by 4%. The hybrid approach, namely machine-learning and rule-based approach, resulted into an improved system.

For the medical domain, Campillos-Llanos et al. [7] create a corpus of annotated clinical trials to help medical professionals to keep up to date with the latest advances. To accelerate the annotation process, the authors pre-annotate a corpus with rules and dictionaries from a medical lexicon and then manually annotate it. They train 3 machine learning models, one of which was BERT, on a corpus subset. The models achieve an F1 score between 79.80% and 87.03%. Although, the training set is of moderate size with 175023 tokens the performance of all machine learning models is good.

For the political domain, a named entity recognition tagger was trained by Kerkvliet et al. [28]. The tagger recognizes labels commonly found in named entity recognition datasets such as person, location and organization. However, the authors focus on political actors in Dutch parliamentary proceedings such as politicians, political groups and committees. The tagger was trained with a weak supervision technique called active learning. During active learning the tagger got a list of political actors and annotated some data. Then, an annotator corrected wrong annotations and retrained the tagger with the initial list and the corrected annotations. After a few rounds the tagger labeled all training data in a shorter amount of time compared to traditional manual annotating. The authors achieve state-of-the-art-results. The most surprising observation is that the tagger was able to identify complex references to committees and ministries consisting of several tokens.

To summarize, sequence labeling was applied with different techniques in combination with machine learning models. Although some researchers use noisy

data with weak supervision, the machine learning models achieve good performance. Some post-processing can improve the performance of a system. Also pre-processing with manual annotations shows exceptionally good results.

2.6 Research on populism communication

During the last years interest in populism communication has increased in the scientific community. De Vreese et al. [10] introduce populism as a communication style and provide a framework for research including multiple promising research ideas. One of the proposed research ideas is the investigation of degrees of populism. The authors define the degree of populism as the extent to which elements of populism communication are used, in other words references to people. Specifically, they propose to systematically analyze references to people.

Relevant to the investigation of degrees of populism is the work on populism communication by Wirth et al. [45], [46]. Wirth et al. introduce a standardized framework to investigate the flow of populism communication and its effect on people. For their research, they develop an application called Angrist-Tool to aggregate annotated data in relational databases. The data contains the analysis of political speeches in texts in three levels.

The first level focuses on the general text. This includes the content, the frame in which a political a news story is presented and the style. A frame indicates whether the author uses passive language, discusses political actions or criticizes political actors.

The second level focuses on speakers that appear in the text and their communication style. Features such as rhetoric, quotation style, a specific language style such as neologisms or emotional language are analyzed. Also the speaker's position on topics is assessed.

The third level focuses on the statements of the speaker and their value. They analyze how speakers define issues, the argumentation flow and the use of sources to justify the argumentation. Also references to other people in combination with their trait, or a link to an issue are investigated. For the third level of the analysis, the researchers identified nine classes of people, that are referenced in populist statements. The classes are specific person, political party or coalition, multiple persons, elite, groups of people, supranational institutions, country, own person and other actors. The authors systematically analyzed references to people but with multiple constraints such as ignoring historic actors and statements about the feelings of an actor. This is in contrast to the suggestion by de Vreese et al. on researching degrees of populism, as the authors don't propose any constraints. It would be interesting to follow the suggestion by de Vreese et al. [10] with classes

of people developed to investigate populism communication. The effort in this thesis is dedicated to facilitate research on the degree of populism.

Chapter 3

Methods

This chapter begins by introducing the annotation guidelines. Then, I outline the datasets that were used for the construction of rules and dictionaries. For reasons of space, Section Definition of Linguistic Rules, contains the structure of only selected rules. Finally, Section 4 outlines the creation of the training set and important considerations.

3.1 Annotation Guidelines

Wirth et al. [46] identified nine classes of people targeted by political speakers. I used the same classes of people however, I altered definitions of multiple classes. The definitions of Wirth et al. depend on speaker's metadata. They compare the political affiliation or nationality of the speaker with the targeted actor to classify the actor. Such metadata comparison is not feasible in this thesis.

The nine classes are specific person, political party or coalition, multiple persons, elite, groups of people, supranational institutions, countries, own person and other actors. The general idea in the definition of the classes was to rely on data from the context and avoid the need of speaker's metadata. Additionally, the defined classes should be able to be mapped to the two key actors of populism communication, the 'decent people' and the 'corrupt elite'. To the 'decent people' the classes groups of people and others are mapped. To the 'corrupt elite' the classes political party or coalition, elite, own person and supranational institutions are mapped. The classes specific person, multiple persons and country overlap with both key actors. This Sections contains a summary of the classes. Please refer to Appendix A for the complete annotation guidelines.

The class specific person contains names of persons. As influential people over regulations are often referenced by their job position, such as President Obama, the

job position is also annotated along with the name.

The class political party or coalition contains parties and coalitions from around the world. Additionally, political groups of the European Parliament are also considered.

The class multiple persons contains two or more references of the previously named classes in one sentence. Multiple specific persons, multiple political parties or coalitions or a combination of both are labeled as multiple persons.

The elite are the most influential people or groups of people of a nation or the world. Their influence is notable in the key sectors politics, economics, finances or media. Also groups of influential people such as committees, authorities and influential companies are annotated here. However, references by name are not labeled as elite but as specific person. For instance, Angela Merkel is annotated as specific person, but the German chancellor is annotated as elite.

Groups of people are part of the society and are identified by an explicit or implicit exclusion of other people. The group usually, but not necessarily, has a common characteristic such as a shared nationality, profession, or ideology.

Supranational institutions are entities formed by multiple countries and act on global or international level. For instance, the European Union and the World Trade Organization.

For the class country all countries of the world are considered. Additionally, country adjectives are considered such as German or French. Country adjectives define something as deriving from a particular country.

The class other actors contains references, that don't match to any of the previously mentioned classes. This includes groups of countries, groups of organizations, smaller companies and NGOs.

3.2 Data Collection

To create the training set for the BERT model a sufficient amount of annotated data was needed. Due to the diversity of references it was impossible to collect all the needed data. Therefore, the focus was to collect data that was expected to be found a lot in parliamentary debates. This includes references to members of the European Parliament and multiple governments. An additional criterion was a good coverage of the data in the European area. The EuroParl dataset [27] contains Proceedings of the European Parliament. Since every debate is about the European Union, references to actors located in the European Union are expected to occur often.

For the class specific person, version one of the Name Dataset [39] was used. There are two version of the dataset that differ in size. The first version is shorter

and was retrieved from IMDB and names databases scraped from the internet by multiple authors. For performance reasons, I chose the shorter version. Names without Latin characters such as Arabic or Chinese names were removed. This results in a total of 164410 first names and 98365 last names. Additionally, Members of the European Commission [12], [13], members of the Executive Board of the European Central Bank [3] for the years 1998-2018 were collected. Finally, political leaders of countries from the WhoGov dataset [35], [8] and REIGN dataset [6], [20], [16] were collected.

For the class political party or coalition the database Party Facts [14], [5] was used. The English names of core political parties were collected. In addition, the political groups of the European Parliament [17] were also captured for the years 1999 - 2019.

For the class multiple persons no additional data was collected. This class is defined as a combination of the previously named classes, specific person and political party or coalition. Hence, this class relies on collected data for the classes specific person and political party or coalition.

For the class elite a number of datasets was utilized. For the political elite, cabinet positions from the dataset WhoGov [35], [8] were collected. WhoGov contains members of cabinets in 177 countries from 1966-2016. For the economic elite, companies which spent on lobbying activities [33] in addition to legal forms of companies were collected [2]. For the financial elite, central banks of the European Union [4] were used.

For the class groups of people the gazetteer of the Industrial and Professional Occupation Dataset (IPOD) [32] was used. Only job titles that were annotated as a Responsibility were used to avoid job titles that consist of a location or function, such as 'Accounts' and 'Research Collaboration'. In total, 249 job titles were collected. To find nouns for groups of people, I used the Website Thesaurus [1]. Thesaurus finds synonyms for every given word and indicates with different colors how relevant each synonym is. I created two gazetteers of synonyms. The first gazetteer contains 74 synonyms related to a person's trait. The words, that I searched for synonyms were citizen, resident, person, child, woman, man, employer, worker, voter, immigrant, refugee, consumer, producer, taxpayer, reader and expert. Only very relevant synonyms were used which are indicated by a red color. The second gazetteer contains 17 synonyms related to groups of people such as 'community' and 'population'.

For the class country, I collected a list of nationalities provided by the Website of the UK government [22]. Additionally, I collected formal [44] and common names [26] of countries. Common names where collected from the ISO Online

[1] `https://www.thesaurus.com`

database of country codes.

For the class supranational institutions, I used a list of European Institutions [43] and the list of supranational institutions provided by the work on populism communication by Wirth et al. [46].

For the class others, I collected 13 synonyms from Thesaurus for the words business, enterprise,company country, SME, NGO.

3.3 Definition of Linguistic Rules

For the creation of a training set in a dictionary-based and rule-based approach I used spaCy[2]. SpaCy is an open-source software library developed for advanced natural language processing. It supports multiple linguistic features such as part-of speech tagging and lemmatization and allows to annotate in a rule-based approach. SpaCy gets as input the text that needs annotation and a set of linguistic patterns. Each linguistic rules contains a label and a pattern. SpaCy searches the input text for matches with the patterns. Whenever multiple patterns match to a sequence, SpaCy chooses the pattern matching most tokens. If multiple patterns with the same number of tokens match, then the pattern placed higher in the patterns set is chosen.

The general idea when defining linguistic rules was to define rules that resolve a number of ambiguous references. Some commonly observed ambiguities were resolved by defining rules that take more context into account and have richer dictionary lists.

For space reasons, this section contains only the definition of three linguistic rules. For more definitions of linguistic rules please refer to Appendix B.

For the class specific person a rule for different name combination was defined. The rule consists of three parts. The first part is an optional title such as Mr, Mrs and Madam. Then an optional first name followed by one obligatory last name was defined. The first and last names are list of names from the Name dataset [39].

To capture different references to cabinet members based on the cabinet position, a rule for each cabinet position was defined. The goal was to capture references such as German Minister of Agriculture or Minister of Justice. Each rule starts with an optional adjective followed by the exact position title.

To identify variations in references to influential companies, an individual rule for each company was defined. The goal was to capture references that don't contain the legal form of the company. For instance, BAYER AG can also be referenced as BAYER. For each company name the last words of the name were matched against a list of legal forms. If the company name contained a legal form

[2] `https://spacy.io`

than the legal form was defined as optional, whereas the rest of the name was mandatory.

3.4 Creation of Training set

For the creation of the training set, I used the previously outlined linguistic rules and dictionaries.

Before annotating the dataset the text was pre-processed by removing XML-tags. The XML-tags contained metadata about the speaker, such as the speaker's name, party affiliation and country affiliation. This information matches to my annotating scheme and could be annotated as well. However, only the actual speech is of interest, for this reason the XML-tags were removed.

Next, I annotated the data with SpaCy. The order of the linguistic rules didn't have any major effect on the annotations provided by SpaCy. Only the patterns of the class political party or coalition were placed higher in the set in comparison to the class groups of people. The goal was to capture the majority of references with the word group as political party, since the sequence my group usually references a political group rather than a group of people.

After annotating the training set, I removed sentences that don't contain any reference to people. The goal was to increase the ratio of references to people to improve the performance of BERT. In total, the training set consists of 86751 references to people and contains 1117311 words.

Chapter 4

Experimental Evaluation

This Chapter starts with an introduction of evaluation metrics for sequence labeling. Next, the experiments with a dictionary and rule-based system and a BERT system are described. In Section 4.3, the results of the two systems are evaluated. In the next section, common error patterns among both systems are discussed. Finally, the BERT system is evaluated based on selected attributes.

4.1 Evaluation Metrics

To evaluate a sequence labeling system, three metrics exist. The metrics compute the ratio of correctly predicted labels [15]. The correct labels are retrieved from manual annotations and are called the gold-standard. For each label, there are two errors and two ways to be correct. Errors names start with the prefix false. The first error, called false positive (FP), arises when a system incorrectly classifies a sequence that shouldn't be classified. The second error, called false negative (FN), arises when the classifier doesn't predict a label. Names of correct predictions start with the prefix true. A true positive (TP) is a correctly predicted label. A true negative (TP) arises when the system correctly doesn't predict any label.

The evaluation metrics are precision, recall and F1 score [15]. Precision defines the ratio of correctly predicted labels of all predicted labels by the model. Recall defines the ratio of correctly predicted labels given the gold standard. F1 score combines precision and recall by using the harmonic mean. When working with classes that are distributed unequally in the dataset, also called imbalanced classes, the metrics are micro-averaged [15]. Micro average is a weighted measure, that gives more weight to classes that occur often. The metrics are computes for each true positive, false positive and false negative of a class C.

The evaluation metrics are formally defined as:

$$Precision = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FP_c}$$

$$Recall = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FN_c}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.2 Experiments

For the experiments two systems were created. The two systems correspond to the introduced named entity recognition approaches from Section 2.2. The first system uses a dictionary-based and rule-based approach. The second system is based on the machine learning model BERT.

For the first system all the dictionaries from Section Data Collection and all the linguistic rules from Section Definition of Linguistic Rules were used.

For the second system a BERT model was fine-tuned. I chose the pretained-model $BERT_{LARGE} - cased$. The training data was annotated in a dictionary-based and a rule-based approach. This means that the model learned with weak supervision. The training data consists of 1117311 tokens and 86751 references to people as illustrated by Table 4.1. Before finetuning BERT, I setup the model. Based on the Devlin et al. [11], in order to finetune, three hyperparameters need to be optimized based on the task. The three parameters are batch size, learning rate and number of epochs. The paper suggests for each parameter a range of values that work well across all tasks. In order to find the optimal values, I performed an exhaustive search over all possible values. This was possible due to the fast fine-tuning. The optimal values for my experiment were 4 for batch size, 2e-5 for learning rate and 4 for number of epochs. During fine-tuning, I randomly split the input data into 90% training set and 10% validation set.

4.3 System evaluation

To evaluate the two systems a gold standard was manually annotated. The test set contains 61292 tokens and 3779 class labels. The exact distribution of the classes is shown in Table 4.1.

To evaluate the system a strict evaluation scheme [42] was used. The strict scheme evaluates an annotation as correct if the boundaries and the label match to the gold standard. In other words, the prediction and the gold standard have the

		Training set		Manually labeled set	
	Class	Total in num	Total in %	Total in num	Total in %
1	Specific person	5458	6.29	120	3.18
2	Political party or coalition	800	0.92	89	2.36
3	Multiple persons	3782	4.36	46	1.22
4	Elite	10633	12.26	674	17.84
5	Groups of people	16913	19.5	640	16.94
6	Supranational institutions	18596	21.44	656	17.36
7	Country	7373	8.5	334	8.84
8	Own person	14446	16.65	752	19.9
9	Other actors	8750	10.09	468	12.38
	Total	**86751**	**100**	**3779**	**100**

Table 4.1: Distribution of classes in training and manually labeled set

same starting token and the same subsequent tokens. Table 4.2 shows the strict evaluation results of both systems. For more detailed results please refer to the confusion matrices in Appendix C.1 and C.2. The results of different evaluation schemes are shown in Appendix C.1.

The class specific person was recognized well by the BERT model. BERT identified the vast majority of labels in the testset. The baseline achieved low recall and precision. Lengthy names and variations in names were not properly captured by the rules and the dictionaries.

The class political party or coalition achieved the lowest F1 score among both the baseline and the BERT model. The systems achieved almost identical precision, recall and F1 score values. The F1 score varies between 0.23 and 0.235. This indicates that the collected data needed some processing because the class relied solely on dictionaries of data. No linguistic rules were defined for this class.

The class multiple persons achieved less satisfactory results. A minor improvement by 16.5% achieves the BERT model as the fundamental class specific person was predicted better in comparison to the baseline. However, recall remains low in both systems.

The class elite achieved mixed results. The baseline achieved the lowest F1 score of 0.217 in this class. The baseline annotated too many sequences as elite, which result into a low precision. However, this allowed to capture almost half of the actual references. BERT achieved a high precision of 0.603 however the recall is lower in comparison to the baseline. This indicates that BERT predicted the wrong labels and was not able to ambiguously resolve references to elite people.

	Baseline			BERT		
Class	Prec	Rec	F1	Prec	Rec	F1
Specific person	0.356	0.217	0.269	0.441	0.842	**0.579**
Political party or coalition	0.320	0.180	0.230	0.340	0.800	**0.235**
Multiple persons	0.359	0.304	0.329	0.613	0.413	**0.494**
Elite	0.141	0.472	0.217	0.603	0.401	**0.481**
Groups of people	0.279	0.328	**0.301**	0.264	0.323	0.291
Supranational institutions	0.778	0.886	0.828	0.780	0.915	**0.842**
Country	0.801	0.808	**0.805**	0.791	0.814	0.802
Own person	0.916	0.983	0.948	0.920	0.989	**0.953**
Other actors	0.483	0.359	**0.412**	0.466	0.348	0.399
Micro-average	**0.433**	**0.620**	**0.510**	**0.628**	**0.633**	**0.630**

Table 4.2: Evaluation of dictionary- and rule-based system and BERT system

Overall, BERT achieves an F1 score two times higher than the baseline, with 0.481.

The class groups of people achieves a comparably low F1 score among both systems. The highest F1 score of 0.301 achieves the BERT system. The difference to BERT is only 0.017. Both baselines contain similar values in precision and recall.

The class supranational institutions achieves good results among all systems. BERT achieves the highest score, that indicates that BERT was able to identify some unseen variations in comparison to the second baseline. The F1 score varies between 0.828 and 0.842.

The class country shows comparable results among all baselines. All metrics vary between 0.791 and 0.814.

The class own person achieves the highest scores among both systems. More importantly, the systems achieve almost identical results. All metrics vary between 0.916 and 0.989.

The class other actors achieves low scores. The baseline achieves an F1 score of 0.412 and BERT 0.399. All metrics show similar results among the two systems. The highest F1 score achieves the baseline with a difference of 1.3% to BERT.

4.4 Error analysis

A strict evaluation of the systems showed that the performance varies among different classes. This section focuses on patterns that cause false predictions. For the error analysis, the results from the confusion matrices in Appendix C.1 and C.2

and the results of different evaluation schemes from Appendix C.1 were considered. Appendix C.1 shows the result of the BERT model based on exact and partial boundary matches to the gold-standard.

The class specific person showed very low precision among both systems. Both systems recognized correctly the starting word of a reference. The majority of errors are caused by references that are classified as elite. BERT wrongly predicted names followed by the word President. For instance, President Poroschenko and President Putin. BERT is more cautious when predicting specific persons which results into a higher recall of 0.842 in comparison to 0.217. When considering partial and exact evaluation scheme, Tabel C.1 highlights that BERT correctly identifies the majority of reference boundaries. In other words, the starting and ending word of each reference to a specific person.

The class political party or coalition was not recognized well among both systems. Both systems miss multiple references to a political party or coalition. The exact scheme evaluation of BERT shows similar results to the strict evaluation. This means that the main limitation of BERT is the boundary detection. For this reason, the recall is higher than the precision. The baseline struggles with both.

The class multiple persons showed mixed results. BERT achieves state-of-the-art results when evaluated on partial and exact scheme. Some errors are caused by the two previously named classes.

The class elite was not recognized well although it was the class with the biggest dictionary available. The low metrics are attributed to lengthy references. BERT is considerably rather cautious when annotating sequences as elite. The baseline wrongly annotates multiple sequences that don't reference elite people as elite. Both systems miss unseen references to committees of the European Parliament and to references to rapporteurs. Ambiguity remains on longer sequences with subsequences that could be annotated. For instance, national Member States governments contains the subsequence Member States that was wrongly annotated as others. However, the entire sequence specifies a group of governments and should be annotated as elite. BERT achieves better results on the partial and exact evaluation schemes. The results of the evaluation schemes show that BERT misses 172 references while incorrectly or partial matching 169 references.

The class groups of people was not recognized well. This class contains very diverse references, that can become very lengthy. The longest reference contained 12 tokens. Both systems missed long references with more than three tokens. This includes references such as hundreds of thousands of people. BERT missed multiple references and wrongly classified references that don't reference people. This is caused by loosely defined rules. The rules capture specifications of up to two tokens and quantification of up to three tokens. The rules are able to capture well references that include both, quantification of group members and specification.

However, the nouns that correspond to people were used as a lemma. This allowed to also capture verbs in addition to nouns. For instance, the word association allows to capture the noun associations but also the verb to associate. This resulted into wrongly classified references such as associated, reintegration of individuals and this current state. The use of lemma resulted into 192 sequences that don't reference any people to be classified as groups of people. For this reason, both systems achieved comparable results.

The class supranational institutions was recognized very good. Institutions are known from the dictionary and are expressed with a small number of variations. The beginning and the entire sequences were predicted mostly correct. Some errors are due to references that were not present in the dictionaries such as the European Community and this present body.

The class country was also recognized very well. As country names and country adjectives are known from the dictionaries and are expressed with a small number of variations. Only a few abbreviations and variations not present in the dictionary were not recognized such as US and America.

The class own person was recognized the best. Almost all references were recognized because there is a small number of possible references. Some errors were caused by the word my. When the speaker talks about his thoughts and actions then the word my is annotated as own person. However, when the speaker uses the same word to talk about other people, then the reference should not be annotated as own person. For instance, the reference my group often references the own political group and should be annotated as political party or coalition. Both systems caused this error.

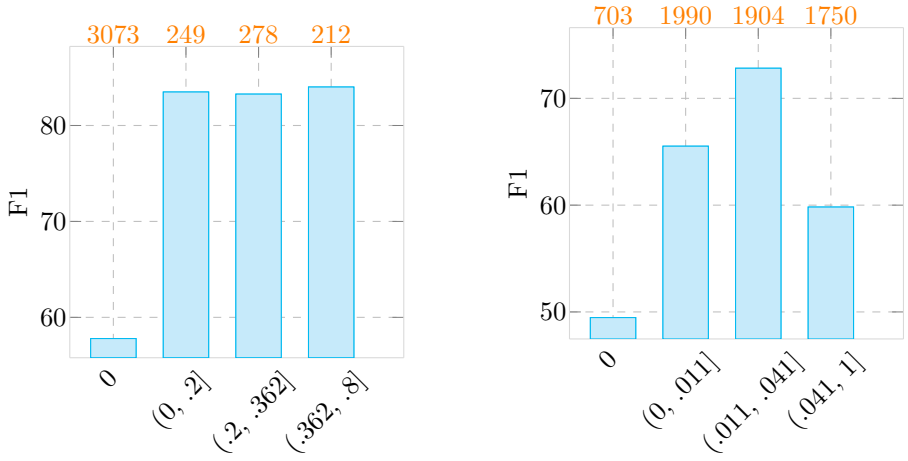
The class other actors was not recognized well. Both systems didn't recognize correctly boundaries of references. For instance, the reference Member States were always recognized, but several of our Member States was never recognized. Both systems were unable to recognize longer references. Also diverse references were not recognized. This class contained a relatively small dictionary, but it was defined very broadly. For this reason, BERT missed 80 references completely.

4.5 Attribute-wise evaluation

As stated in the Introduction, the first research question was whether a machine learning machine learning approach is worth it to resolve ambiguous references. The micro-averaged metrics in Section System evaluation show that the machine learning system outperformed the rule-based and dictionary-based system despite the noisy training data that was collected with weak supervision. The precision was 20% higher, recall was 13% higher and the F1 score was 12% higher. The

lower cost of data collection with weak supervision shows that a machine learning model is definitely worth to resolve ambiguous references to people.

To better evaluate the strengths and weaknesses of the system, a new evaluation approach was used. Fu et al. [19] introduced an approach based on bucketing. The authors split the entities from the test set into four sets, also called buckets, based on an attribute. For each bucket the F1 score is computed individually.



(a) Entity Frequency

(b) Token Frequency

Figure 4.1: F1 scores of BERT system based on bucket-wise evaluation on entity frequency and token frequency. The x-axis contains the attribute values of each bucket. At the top of the chart are the number of entities, that correspond to each bucket.

The first attribute, that was considered to evaluate the dataset of this thesis is entity frequency. Entity frequency is the ratio of seen entities in the training set. In this context, the terms entity and reference to people are used interchangeably. The chart 4.1a shows that unseen entities are correctly classified to a satisfactory degree, as the F1 score is 0.58. The F1 score among seen entities varies between 0.83 and 0.84. This shows that a higher entity frequency influences minimally the performance of the system. It should be highlighted that the majority of entities in the test set were unseen. In total, 3073 entities were unseen and only 739 were previously seen. The high number of unseen entities reduced the micro-average metrics of the system to a significant degree. Nevertheless, the chart shows that the system is able to generalize, but achieves better results on seen data.

Similar results are observed from the chart 4.1b on token frequency. Token frequency is the ratio of entity tokens in the training set. Entity tokens are the indi-

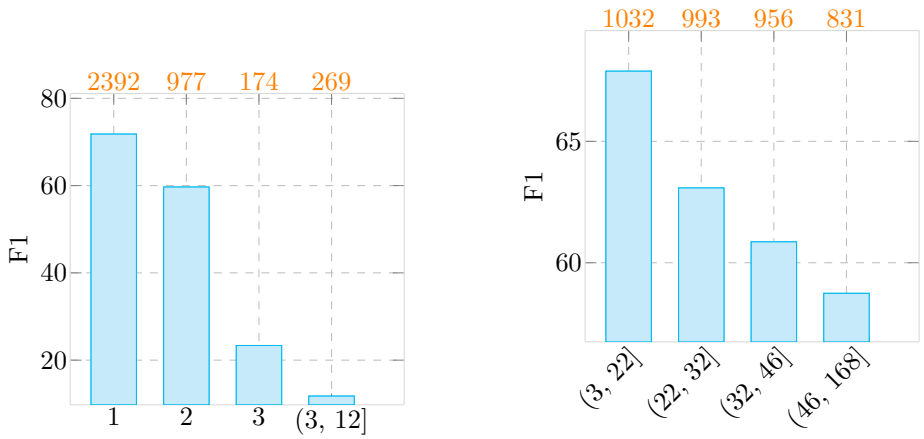
vidual words of an entity such as Angela and Merkel in the entity Angela Merkel. Entity tokens that were unseen achieve an F1 score of 0.49. Futhermore, the F1 score increases as the token frequency increases. Interestingly, a high token frequency of over 0.041 decreases the F1 score by 17.8%. The decrease is possibly caused by tokens that appear in multiple class entities, in other words by ambiguous references. This would explain the higher frequency in the training set in combination with a lower F1 score. In contrast, entities already seen in the training set that are less ambiguous are easier to classify correctly as the F1 score of 0.72 shows. To sum up, considering token frequency the model is able to correctly classify unseen tokens but achieves the best results on tokens that account for up to 0.041 of the training set. Higher frequency tokens are more difficult to correctly classify, but still achieve an acceptable F1 score of 0.59.

The majority of ambiguous references consist of multiple tokens and are found in long sentences. This characteristic is observed on the charts 4.2 on sentence length and entity length. Both charts show that an increase in tokens leads to a decrease in F1 score. This relationship was expected regarding the entity length. As the majority of the linguistic rules was defined for sequences of up to three tokens, the majority of the training set consisted of entities up to three tokens. The chart on entity length shows that the longest entity consisted of 12 tokens. Such a long sequence corresponds to entities with a lengthy specifications and quantification of a collection of people. Lengthy references possibly occur in long sentences. This would explain the similar structure to the chart on entity length. It should be highlighted that the buckets regarding sentence length are almost balanced. The bucket size ranges between 1032 and 831 entities. This indicates that longer sentences are no exception in parliamentary proceedings.

Regarding sentences, a sentence can consist of multiple or just a few entities. Entity density computes the ratio of entities within a sentence. The chart 4.3a shows that an increase of entity density over 0.147 results into a reduction of F1 score of up to 0.12. Some long references were found in high density sentences.

To summarize, the attribute-wise evaluation revealed some strengths and weaknesses of the BERT system. The majority of references were unseen which resulted into the low F1 score of 0.63. The system achieves state-of-the-art-results on seen entities despite the frequency of the entity in the training set. Entities were classified correctly when the entity tokens account for up to 0.042 of the training set. References with more than two tokens were rarely classified correctly and remain ambiguous. Long references are possibly found in longer sentences with over 46 tokens or in sentences that consist of over 0.214 of entities.

Overall, the system achieved the best results on short entities of up to two tokens. On a sentence level, best results were achieved on sentences of up to 22 tokens and on sentences that consist of up to 14.7% of entities.



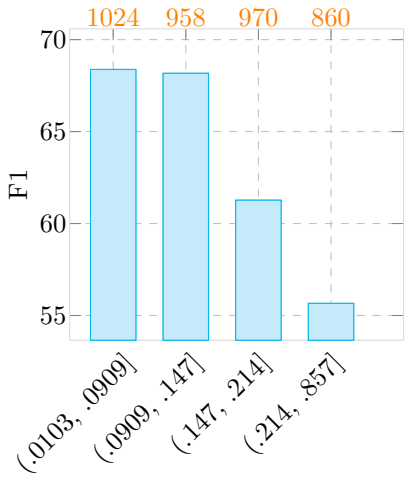
(a) Entity Length

(b) Sentence Length

Figure 4.2: F1 scores of BERT system based on bucket-wise evaluation on entity length and sentence length. The x-axis contains the attribute values of each bucket. At the top of the chart are the number of entities, that correspond to each bucket.

It should be highlighted that the results match with the two observations of the authors of the attribute-wise analysis [19]. The authors evaluated multiple datasets and observed that entity length consistently influences the performance. BERT shows the same characteristic as the difference among the highest and lowest performance is 60%. The second observation of the authors is that frequency and sentence length matter but to a smaller degree. BERT showed smaller performance variations in these attributes. The performance variation in sentence length is 9%, in token frequency 23% and in entity frequency 26%.

The observations of the authors indicate that the limitations of the dataset of this thesis are common limitations in sequence labeling tasks. While there is always some room for improvement, the limitations won't be completely eliminated.



(a) Entity Density

Figure 4.3: F1 scores of BERT system based on bucket-wise evaluation on entity density. The x-axis contains the attribute values of each bucket. At the top of the chart are the number of entities, that correspond to each bucket.

Chapter 5

Conclusion

5.1 Summary

In this thesis I created the first benchmarking dataset to recognize mentions to people. The EuroParl dataset, that contains English parliamentary proceedings was annotated. Data from multiple datasets was collected to either match it against the text or to integrate it in linguistic rules. I created a training set with weak supervision which was used to train a BERT model. The baseline was a rule-based and dictionary-based system. I showed that the BERT model achieves better results in comparison to the baseline system. BERT outperforms in six out of nine classes and achieves a micro-average F1 score of 0.63. It also achieves state-of-the-art results on the classes supranational institutions, country and own person. This confirms that noisy annotations with weak supervision can train good machine learning models if they are available in a big quantity. For this reason, I positively answer the first research question, whether machine learning models are worth using to resolve ambiguous references.

The second research question was: how well does the dataset detect mentions of people and how well is ambiguity resolved? The error analysis showed that short references with up to two words and references that were previously seen in the training set were classified exceptionally well. Ambiguity was resolved the best among references with words that don't appear a lot in other entities. Ambiguity remains in long references with more than three words and is mostly found in sentences with more than 46 words or in sentences that consist of more than 14.7% of entities.

Additionally, specifications and quantification of people were only considered in the beginning of the sequence. Rules that consider specifications at the end of a sequence could improve the dataset. However, only a handful of such rules was

defined.

Ambiguous references could be resolved if tokens that appear in multiple entities were collected and precise rules for these tokens were defined. For instance, the token group can be labeled with three different classes depending on the previous and following tokens. The reference to the Group of the Greens/European Free Alliance is labeled as political party or coalition, a reference to a an advisory group is labeled as groups of people and a group of NGOs is labeled as others. A more exhaustive list for the class other actors could also be used as it was too small and simple.

Interestingly, the results of the experiments indicate the characteristics of named entity recognition approaches as introduced in Section 2.2. A dictionary-based approach achieves the best results when the list is exhaustive and a small number of spelling variations exist. The classes political party or coalition, supranational institutions, country and own person rely only on dictionary data and are characterized by a limited number variations. All classes except the class political party or coalition achieve state-of-the-art results among both systems. A rule-based approach with domain specific rules in combination with incomplete dictionaries often lead to systems with high precision and low recall. The classes that used rules are specific person, elite, groups of people and other actors. Although the majority of the classes don't achieve a high precision, the precision is in most cases higher than the recall. Finally, a machine-learning approach performs better than a rule-based and dictionary-based approach due to the model's ability to generalize. The BERT model achieves higher F1 scores in six out of the nine classes.

It is interesting to note that BERT identified well the relationship of entities within a sentence. The class multiple persons relies on the identification of multiple specific persons and political parties or coalitions within a sentence. BERT was able to capture well this relationship as it achieved an F1 score of 0.49. This indicates that the masked language model task, on which BERT was trained during pre-training, did result into a better understanding of word relationships within a sentence.

5.2 Future Work

The dataset of this thesis is the first focusing only on references to people. As this dataset is mostly appropriate for short references to people, it is a good starting point in automating the task. Two promising directions to advance populism communication research were inspired by the methodology of other authors as introduced in Section 2.5.

The first direction inspired by Kerkvliet et al. [28] is to leverage the dataset

of this thesis with active learning. The authors showed great performance by actively training a tagger for political actors. This approach needs some upfront data to start learning. The dataset of the thesis can be used as upfront data for the classifier. Kerkvliet et al. also used parliamentary proceedings as me and noticed that complex references consisting of several tokens were recognized well. This observation is interesting, as the main limitation of my dataset is the recognition of lengthy references. Due to the similarities of the political domain and the input data, that consists of parliamentary proceedings, using active learning could provide some improvements.

The second direction inspired by Malarkodi et al. [41] involves post-processing the predictions of a machine learning model. The authors applied linguistic and heuristic rules to improve the accuracy of their system. Overall, they used a hybrid approach to sequence labeling as they first implemented a machine learning approach followed by a rule-based approach. Adding a rule-based approach could potentially resolve moderately long references. Longer references will likely remain ambiguous.

At this point, research on degrees of populism communication becomes feasible.

Bibliography

[1] Jay Alammar. The Illustrated Transformer. Accessed on: August 2, 2021. [Blog post]. Available: https://jalammar.github.io/illustrated-transformer/, 2018.

[2] European Central Bank. List of legal forms. Accessed on: July 15, 2021. [Online]. Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjWjImAvZzyAhUN2qQKHZtfCWIQFnoECAMQAw&url=https%3A%2F%2Fwww.ecb.europa.eu%2Fstats%2Fmoney%2Faggregates%2Fanacredit%2Fshared%2Fpdf%.

[3] European Central Bank. Executive Board members – terms of office. Accessed on: June 7, 2021. [Online]. Available: https://www.ecb.europa.eu/ecb/orga/decisions/eb/html/ebtimeline.en.html, 2021.

[4] European Central Bank. Monetary Financial Institutions (MFIs): Download area. Accessed on: June 7, 2018. [Online]. Available: https://www.ecb.europa.eu/stats/money/mfi/general/html/dla/mfi_MID/mfi_csv_210709.csv, 2021.

[5] Paul Bederke, Holger Döring, and Sven Regel. "Party Facts – Version 2020b". Accessed on: June 7, 2021. [Online]. Available: https://dataverse.harvard.edu/file.xhtml?fileId=4274158&version=1.1.

[6] Curtis Bell, Clayton Besaw, and Matthew Frank. The Rulers, Elections, and Irregular Governance (REIGN) Dataset. Accessed on: June 7, 2021. [Online]. Available: https://oefdatascience.github.io/REIGN.github.io/, 2021.

[7] Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC Medical Informatics and Decision Making*, 21(1):69, dec 2021.

[8] Nuffield Politics Research Centre. WhoGov Dataset - Nuffield Politics Research Centre. Accessed on: June 5, 2021. [Online]. Available: https://politicscentre.nuffield.ox.ac.uk/whogov-dataset/.

[9] Iskander De Bruycker and Matthijs Rooduijn. The People's Champions? Populist Communication as a Contextually Dependent Political Strategy. *Journalism & Mass Communication Quarterly*, page 107769902199864, apr 2021.

[10] Claes H. de Vreese, Frank Esser, Toril Aalberg, Carsten Reinemann, and James Stanyer. Populism as an Expression of Political Communication Content and Style: A New Perspective. *International Journal of Press/Politics*, 23(4):423–438, oct 2018.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, oct 2019.

[12] Holger Döring. The Composition of the College of Commissioners: Patterns of Delegation. *European Union Politics*, 8(2):207–228, jun 2007.

[13] Holger Döring. Composition of the College of Commissioners – Version 2014. UNF:6:fdbGsOLBvPq1Ptl4lHbsCA== [fileUNF]. Accessed on: July 15, 2021. [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/A5ECON, 2016.

[14] Holger Döring and Sven Regel. Party Facts: A database of political parties worldwide. *Party Politics*, 25(2):97–109, mar 2019.

[15] Jacob Eisenstein. *Natural Language Processing*. Adaptive Computation and Machine Learning series. MIT Press, 2019.

[16] Cali Mortenson Ellis, Michael C. Horowitz, and Allan C. Stam. Introducing the LEAD Data Set. *International Interactions*, 41(4):718–741, jul 2015.

[17] European Parliament. Seats by political group and country. Accessed on: June 19, 2021. [Online]. Available:https://www.europarl.europa.eu/election-results-2019/en/seats-political-group-country/1999-2004/constitutive-session/, 2021.

[18] Stefano Fella, Elise Uberoi, and Richard Cracknell. European Parliament Elections 2019: results and analysis. Technical report, 2019.

[19] Jinlan Fu, Pengfei Liu, and Graham Neubig. Interpretable Multi-dataset Evaluation for Named Entity Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, Stroudsburg, PA, USA, nov 2020. Association for Computational Linguistics.

[20] Henk E. Goemans, Kristian Skrede Gleditsch, and Giacomo Chiozza. Introducing archigos: A dataset of political leaders. *Journal of Peace Research*, 46(2):269–283, mar 2009.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. 2016.

[22] UK government. List of nationalities CSV file. Accessed on: June 16, 2021. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/664133/CH_Nationality_List_20171130_v1.csv/preview, 2020.

[23] Venkat N. Gudivada, Dhana Rao, and Vijay V. Raghavan. Big Data Driven Natural Language Processing Research and Applications. In *Handbook of Statistics*, volume 33, pages 203–238. Elsevier, jan 2015.

[24] Zhiyong He, Zanbo Wang, Wei Wei, Shanshan Feng, Xianling Mao, and Sheng Jiang. A Survey on Recent Advances in Sequence Labeling from Deep Learning Models. *ArXix*, abs/2011.0, nov 2020.

[25] Stefan Helmstetter and Heiko Paulheim. Weakly Supervised Learning for Fake News Detection on Twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 274–277, Barcelona, Spain, aug 2018. IEEE.

[26] ISO - International Organization for Standardization. Online Browsing Platform (OBP). Accessed on: June 4, 2018. [Online]. Available: https://www.iso.org.

[27] Alina Karakanta, Mihaela Vela, and Elke Teich. EuroParl-UdS: Preserving and Extending Metadata in Parliamentary Debates. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japan, may 2018.

[28] Lennart Kerkvliet, Jaap Kamps, and Maarten Marx. Who Mentions Whom? Recognizing Political Actors in Proceedings. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 35–39, Marseille, France, may 2020. European Language Resources Association.

[29] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015.

[30] Ulf Leser and Jörg Hakenberg. What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4):357–369, jan 2005.

[31] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, dec 2020.

[32] Junhua Liu, Yung Chuen Ng, Kristin L. Wood, and Kwan Hui Lim. IPOD: A Large-scale Industrial and Professional Occupation Dataset. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, volume 2, pages 323–328, New York, NY, USA, oct 2020. ACM.

[33] LobbyFacts. LobbyFacts Database. Accessed on: June 7, 2021. [Online]. Available: https://lobbyfacts.eu/.

[34] Cas Mudde. The Populist Zeitgeist. *Government and Opposition*, 39(4):541–563, mar 2004.

[35] Jacob Nyrup and Stuart Bramwell. Who Governs? A New Global Dataset on Members of Cabinets. *American Political Science Review*, 114(4):1366–1374, nov 2020.

[36] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards Robust Linguistic Analysis using OntoNotes. pages 143–152, 2013.

[37] Lance A Ramshaw and Mitchell P Marcus. Text Chunking using Transformation-Based Learning. *Third ACL Workshop on Very Large Corpora*, pages 82–94, 1995.

[38] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data Programming: Creating Large Training Sets, Quickly. *Advances in neural information processing systems*, 29:3567–3575, 2016.

[39] Philippe Remy. Name Dataset. Accessed on: June 12, 2021. [Online]. Available:https://github.com/philipperemy/name-dataset, 2021.

[40] Yuji Roh, Geon Heo, and Steven Euijong Whang. A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347, apr 2021.

[41] Malarkodi C. S., Elisabeth Lex, and Sobha Lalitha Devi. Named Entity Recognition for the Agricultural Domain. *Research in Computing Science*, 117(1):121–132, dec 2016.

[42] Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, volume 4, pages 142–147, Morristown, NJ, USA, jul 2003. Association for Computational Linguistics.

[43] European Union. Institutions and bodies. Accessed on: June 7, 2021. [Online]. Available: https://europa.eu/european-union/about-eu/institutions-bodies_en, 2021.

[44] United Nations. Official Names of the United Nations Membership. https://www.un.int/protocol/sites/www.un.int/files/Protocol%20and%20 Liaison%20Service/officialnamesofcountries.pdf, Accessed on 01 July 2021.

[45] Werner Wirth, Frank Esser, Martin Wettstein, Sven Engesser, Dominique Wirz, Anne Schulz, Nicole Ernst, Florin Büchel, Daniele Caramani, Luca Manucci, Marco R Steenbergen, Laurent Bernhard, Edward Weber, Regula Hänggli, Caroline Dalmus, Christian Schemer, and Philipp Müller. The appeal of populist ideas, strategies and styles: A theoretical model and research design for analyzing populist political communication. 2016.

[46] Werner Wirth, Martin Wettstein, Dominique Wirz, Nicole Ernst, Florin Büchel, Anne Schulz, Frank Esser, Edward Weber, Caroline Dalmus, Sven Engesser, and Luca Manucci. NCCR Democracy Module II: The Appeal of Populist Ideas and Messages. Technical report, 2019.

[47] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa,

Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv*, abs/1609.0, sep 2016.

[48] Vikas Yadav and Steven Bethard. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA, oct 2019. Association for Computational Linguistics.

Appendix A

Annotation guidelines

A.1 Introduction

This chapter contains the annotation guidelines for the EuroParl corpus. The first Section introduces general considerations regarding the annotation.
The Sections 2 to 10 contain the definitions of the classes of people and multiple examples. The nine classes of people are Specific person, Political party or coalition , Multiple persons, Groups of people, The elite, Supranational institutions, Country, Own person and Other actors.

A.1.1 References to a target actor

References to a target actor by a common noun (1) and a proper noun (2) are annotated. References by an adjective are considered only for countries and supranational institutions (3), (4). References by a pronoun (5) are considered only for the class own person. Any other references to a target actor are not considered (6).

(1) These are points which are being raised by people all over the EU.

(2) Fahma Mohamed from Bristol in my constituency at the age of 14, has led a campaign against FGM.

(3) We should be careful not to fall into the American trap.

(4) Lithuania is firmly rooted in the European economy.

(5) I am pleased to make my first speech.

(6) There are many *long-term unemployed* with no real prospect of finding jobs.

A.1.2 Sequences with nested labels

Subsets of a sequence can be labeled individually. Whenever the entire sequence specifies an actor, then subsets of the sequence are not labeled individually. Examples are not treated as a specification and therefore can be annotated individually.

To avoid uncontrollably long sequences, specifications defined with a question word or a definite article are ignored as in the examples (7) and (8). Question words are words used to ask a question, such as who, which, and what. Definite articles are used to indicate someone or something specific such as that and the.

(7) I had hundreds of messages from constituents, trade unions, campaign groups and NGOs in Wales *who were gravely concerned about its effects*.

(8) We have 60 million citizens *who are living under the poverty line*.

In sentence (7) the Wales is used to further specify the group of NGOs. This specification changes the target actor as it implies that NGOs from other countries are excluded. For this reason, the entire phrase is labeled. NGOs and Wales shouldn't be labeled individually although they match to the definitions of the respective classes.

A.1.3 Quantification of actors

Any quantification of members within a group should also be annotated. A quantification is expressed with a numeral word such as four, a determiner such as several, or a noun such as thousands. This is relevant for the classes groups of people, elite and others because they include collections of people.

(9) ...since the three Baltic countries will now become fully integrated into the European structures...

(10) ...Europe is dependent on Russian gas for winter warmth and that most of the constituents I represent are already struggling to pay their energy bills each month...

A.2 Specific person

A specific person is a phrase that contains the name of a specific person. The person needs to be mentioned by the speaker and can't be the speaker itself. If the job position is mentioned in addition to the name, then both are annotated (13).

(11) Fahma Mohamed $_{\text{specific person}}$ from Bristol in my constituency at the age of 14, has led a campaign against FGM.

(12) Dr Martin Luther King specific person wanted to see, in his dream, people treated on the merit of their character...

(13) We have the pleasure of Commissioner Anna Diamantopoulou's specific person

(14) Mr Draghi specific person stated that the ECB was aware that the Lithuanian economy has extreme variabilities. presence.

A.3 Political party or coalition

Phrases that contain the name or abbreviation of a political party or a political coalition should be annotated. General references to political parties and members of political parties are considered if the names of the members are not stated (17). Additionally, names and abbreviations of political groups of the European Parliament are also considered.

(15) The union pour francaise political party or coalition has had a huge impact on ...

(16) All these three parties political party or coalition are pro-euro.

(17) The agreement of the leaders of the political groups political party or coalition ...

(18) On behalf of my group political party or coalition I would like to propose...

(19) The conclusion will not be that the PPE Group political party or coalition must be dissolved...

A.4 Multiple persons

Multiple persons are phrases that contain multiple specific persons, multiple political parties or coalitions or a combination of both within a sentence.

(20) ...what I have seen in this Chamber is large parties, such as the PPE multiple persons and the Socialists multiple persons, voting en bloc, sometimes together but always en bloc.

(21) Congratulations to the rapporteur, Mr Tarabella multiple persons, and thanks to Commissioner Jourová multiple persons for her statement.

This sentence contains two labels. PPE and the Socialists are two individual political parties. As both appear within the sentence boundary they will be annotated as multiple persons.

A.5 Groups of people

Groups of people are parts of the society or the whole society. The main characteristic is the explicit or implicit exclusion of other people. The speaker can include or exclude himself, as long as he is referencing to a group of people, then it should be annotated as such. In populist statements these groups are considered to be the 'true people'. Group members can have a common characteristic such as belong to a nation, an ethos, a function within the society or an occupation. The common characteristics differ with each other by the excluded group of people.

If multiple groups of people are named, then each group is annotated individually. In contrast to the description of the Section Sequences with nested labels from the Introduction, the overall target actor does not change. For instance, the phrase "workers, producers and consumers" does not specify a new target actor. The phrase enlarges the target actor as the group of workers, the group of producers and the group of consumers are merged. The increase of members in the target actor indicates that each group is labeled individually.

A.5.1 Nation

Phrases that refer to the people as a nation indicate the nationality of the group members. The excluded people in this sub-label are foreign people such as companies and other countries.

(22) He has not thought about the young people of Lithuania$_{\text{groups of people}}$ who are the future of Lithuania...

(23) ...we are not as rich as the Germans$_{\text{groups of people}}$.

A.5.2 Ethnos

Groups of people as an ethnos are characterized by the exclusion of people within the own country that have another origin.

(24) Innocent civilians$_{\text{groups of people}}$, including women and children, are caught in the crossfire. We must do all we can to protect and support them. Each civilian killed is one too many. The increase in the numbers of displaced people is also of concern.

A.5.3 Function

Groups of people as a function contain groups with a specific function within the society or the nation. The people that don't belong to the function are implicitly excluded.

(25) Mr President, the Transatlantic Trade and Investment Partnership has the potential to damage our environment and devastate the rights of workers groups of people, producers groups of people and consumers groups of people.

A.5.4 Occupation

Groups of people with a common occupation or an occupation in a common industry should be annotated here.

(26) We should rely on the opinions of our own scientists in EFSA groups of people and not on ideology.

(27) Irish farmers groups of people are rightly demanding that their livelihoods and their communities are protected.

(28) I take note of your views and will pass them on to my colleagues groups of people for further reflection. I am a medical doctor and I say very clearly that, first and foremost, health is number one.

(29) ...which measures would the new Commission envisage undertaking in order to avoid a further deterioration of the European hauliers groups of people situation.

A.5.5 Other groups

This sub-category contains groups of people with a characteristic that doesn't fall under any of the previously mentioned characteristics. This can include groups of people within a regional area, or general references to the people without a specific characteristic.

(30) These are points which are being raised by people all over the EU groups of people, and I think it is good to have an active debate.

(31) The results of a recent Eurobarometer show that nearly all Europeans groups of people agree that equality between women groups of people and men groups of people is a fundamental right, and a large majority of citizens groups of people believe that tackling inequality between women groups of people and men groups of people should be a priority for the EU.

(32) And I want to tell these people in these two big groups, and maybe in other groups, that by doing so what they are doing is intending to side with the anti-Europeans groups of people.

A.6 The elite

The elite are the most influential people in various sectors of society. The four key sectors are politics, finance, business and media. Influential people from other sectors of society are also considered if they have the ability to influence other elite actors. References to elite people by their name should be annotated as specific persons.

A.6.1 Political

The political elite consists of the elite in the three branches of power executive, legislative and judicative. This includes governments, national authorities, politicians, judges, members of national authorities and members of supranational institutions.

(33) ...the way that they inform their Members of Parliament$_{\text{elite}}$ is by inviting them to a reading room.

(34) ...to lock in the current UK Government's$_{\text{elite}}$ back-door privatisation of our National Health Service.

(35) Mr President$_{\text{elite}}$, I have listened closely to this very interesting debate.

(36) Authorities of all countries$_{\text{elite}}$ should also continue their efforts to solve the problem of domestic violence.

(37) ...and Mrs Merkel of course crushed him because what the German Chancellor$_{\text{elite}}$ says goes in the modern Europe.

(38) Some Members of the European Parliament$_{\text{elite}}$ have advocated an autonomous monetary policy which can help to solve some social and economic problems...

A.6.2 Financial

The financial elite contains the most influential financial institutions such as central banks and stock exchanges. Also presidents and board members of financial institutions are considered.

(39) Mr Draghi stated that the ECB$_{\text{elite}}$ was aware that the Lithuanian economy has extreme variabilities.

A.6.3 Economic

The economic elite comprises of influential companies, enterprises, corporations and employers. This elite is often mentioned in discussions about lobbying or with statements saying that this elite poses a threat to other companies and employees.

Influential members of elite companies such as owners, CEOs and board members of are also annotated as economic elite.

(40) Others do not want to attack Google$_{\text{elite}}$ openly because they fear retaliation measures such as demotion or exclusion by Google$_{\text{elite}}$.

(41) In America there are other companies like Yahoo$_{\text{elite}}$ which can compete

A.6.4 Media

The media elite contains influential journalists, TV channels, opinion leaders and other media. They are considered to be elite if they influence other elite people.

(42) ...rather than commenting on an article in The Guardian$_{\text{elite}}$ which refers to some so-called EU paper, on which I cannot respond.

A.7 Supranational institutions

Supranational institutions are entities formed by multiple countries and act on global or international level. Phrases that contains the entire name, abbreviation or any other words that clearly refer to a supranational institution are annotated here.

(43) ...they are still fighting with each other about who in fact invented this proposal in the Convention to elect the Commission$_{\text{sup. institutions}}$ by the European Parliament$_{\text{sup. institutions}}$.

(44) Lithuania is firmly rooted in the European$_{\text{sup. institutions}}$ economy.

(45) But in the time-honoured EU$_{\text{sup. institutions}}$ tradition of ignoring the will of the people...

A.8 Country

Any country of the world, either dependent or independent, is annotated here. Direct references to a country by the country's common name, formal name or a well known abbreviation of its name are considered. For instance, Germany's common name is Germany, its formal name is Federal Republic of Germany and a well known abbreviation is DE. Additionally, country adjectives are considered. Country adjectives describe something originating from a country as in example (48).

(46) In my view, Lithuania's $_{\text{country}}$ successful participation in the euro area will depend on five factors.

(47) We welcome the diplomatic efforts launched in Normandy on 6 June by Ukraine $_{\text{country}}$, the Russian Federation $_{\text{country}}$, France $_{\text{country}}$ and Germany $_{\text{country}}$, and we hope that this process – as well as the joint Berlin Declaration of 2 July ...

(48) We should be careful not to fall into the American $_{\text{country}}$ trap.

A.9 Own person

This label consists of references to the speaker's self. The references are grammatically pronouns such as I, me and myself.

(49) I $_{\text{own person}}$ am pleased to make my $_{\text{own person}}$ first speech since being elected as Coordinator Spokesperson for the Socialist and Democrat Group on Foreign Affairs, to renew my $_{\text{own person}}$ warm.

A.10 Other actors

Other actors are people not covered by any of the previously mentioned labels.

A.10.1 Countries

This category contains references to a collection of countries. References can be based on a common characteristic such as the geographic location of the countries, or be unspecified.

(50) ...those Member States $_{\text{others}}$ who are losing their youngest and brightest talents to other countries $_{\text{others}}$.

(51) ...since the three Baltic countries $_{\text{others}}$ will now become fully integrated into the European structures,

(52) ...workers who wish to work and contribute to their new host countries $_{\text{others}}$, as well as in controlled migration...

A.10.2 Organizations, NGOs

This category contains references to organizations and companies that are not considered to be the most influential over regulations. This can include a general statement such as organizations, a specification of the region or the business industry.

(53) This would cut the cost of raising capital, mainly for small and medium-sized enterprises $_{\text{others}}$.

Appendix B

Linguistic Rules

Prefix definition

The annotation guidelines define that any specification and quantification of a collection of people should be annotated as well. In order to capture specifications and quantifications, a set of five prefixes was defined. The prefixes were used for the labels elite, groups of people and others.

The first prefix starts with an optional determiner, followed by an obligatory adjective. The adjective acounts for both quantification and specification. Examples include unemployed people, many people, Israeli women, civil society and these people.

The second prefix is an adjective followed by multiple nouns are used to captures references such as several minority people.

The third prefix starts with an optional determiner, followed by a noun and then by an adposition. Common references include sections of society and thousands of people.

The fourth prefix starts with a possessive pronoun to account for examples such as his family.

The fourth prefix starts with a noun followed by adposition a lemma for a reference to people and finally a noun. References include family members.

The last prefix starts with an optional determiner followed by a numeral and an adposition for references such as five people.

Linguistic rules

For the class specific person multiple rules were defined. The first rules contains general name combinations. A name can start with an optional title Then leader

names were integrated into rules. Each name of a leader starts with an optional titles such as Mr, Mrs, Madam etc. Then the first half of the name is optional, the second half is obligatory. The goal was to define capture with a simple rule with first names as optional and the last names as obligatory. For this reason, the name was split into half. Rules for the European Parliament members and the European Central Bank board members are similar. Each rule contains a member name. Each rule starts with an optional title such as Mr, Mrs, Madam. Then the first names of the members are optional. Only the last word of the name was obligatory. Rules for Commission members start with an optional word "Commissioner". Then the first words of the name are optional. Finally the last name word is obligatory.

For the class political party or coalition no linguistic rules were defined. The annotations rely on the data of the dictionaries.

For the cabinet positions one rule per cabinet position was created. Each rule accepts an optional adjective followed by the entire cabinet position. For the economic elite a rule for each company with lobbying spendings was defined. Each rule contains the words of the company name. In case a company names contains the legal form as part of the name, then the legal form was optional. For each committee of the European Parliament a rule was defined. Each rule starts with the optional words "Committee on" followed by the complete committee name. Each rule end with an optional "Committee" word.

For financial elite only dictionaries were used and no linguistic rules.

For the class groups of people a number of rules were defined. All rules contain one of the prefixes defined to capture a specification or quantification of an actor followed by a list of group actors.

For the class own person no rules were identified as it solely relies on dictionary data. For the class supranational institutions only dictionaries were used.

For the class country no linguistic rules were defined. This classes is recognized only with a list of possible words.

For the class others five rules were defined. Each rule contains one of the prefixes defined to capture a specification or quantification of an actor followed by a list of actors.

Appendix C

Further Experimental Results

		Correct	Incorrect	Partial	Missed	Spurious	Precision	Recall	F1
Specific person	Partial	118	0	2	0	48	0.708	0.992	0.826
	Exact	118	2	0	0	48	0.702	0.983	0.819
Political party or coalition	Partial	27	0	41	22	12	0.594	0.528	0.559
	Exact	27	41	0	22	12	0.338	0.3	0.318
Multiple persons	Partial	38	0	5	3	3	0.88	0.88	0.88
	Exact	38	5	0	3	3	0.826	0.826	0.826
Elite	Partial	339	0	169	172	41	0.680	0.623	0.689
	Exact	339	169	0	172	41	0.680	0.499	0.552
Groups of people	Partial	216	0	311	122	192	0.517	0.572	0.543
	Exact	216	311	0	122	192	0.3	0.333	0.316
Supranational institution	Partial	629	0	26	13	66	0.89	0.961	0.924
	Exact	629	26	0	13	66	0.872	0.942	0.906
Country	Partial	275	0	3	56	42	0.864	0.828	0.846
	Exact	275	3	0	56	42	0.859	0.823	0.841
Own person	Partial	746	0	7	0	50	0.933	0.995	0.963
	Exact	746	7	0	0	50	0.929	0.991	0.959
Others	Partial	190	0	202	80	48	0.661	0.617	0.638
	Exact	190	202	0	80	48	0.432	0.403	0.417

Table C.1: BERT class-wise evaluation based on two different evaluation schemes, partial and exact. The exact scheme considers predictions as correct if the boundaries match exactly to the gold-standard, regardless the class. The partial scheme considers predictions as correct if the boundaries match partially to the gold-standard, regardless the class. For each scheme the number of correct, incorrect, partial, missed and spurious references are shown.

Figure C.1: Confusion matrix of rule-based and dictionary-based system

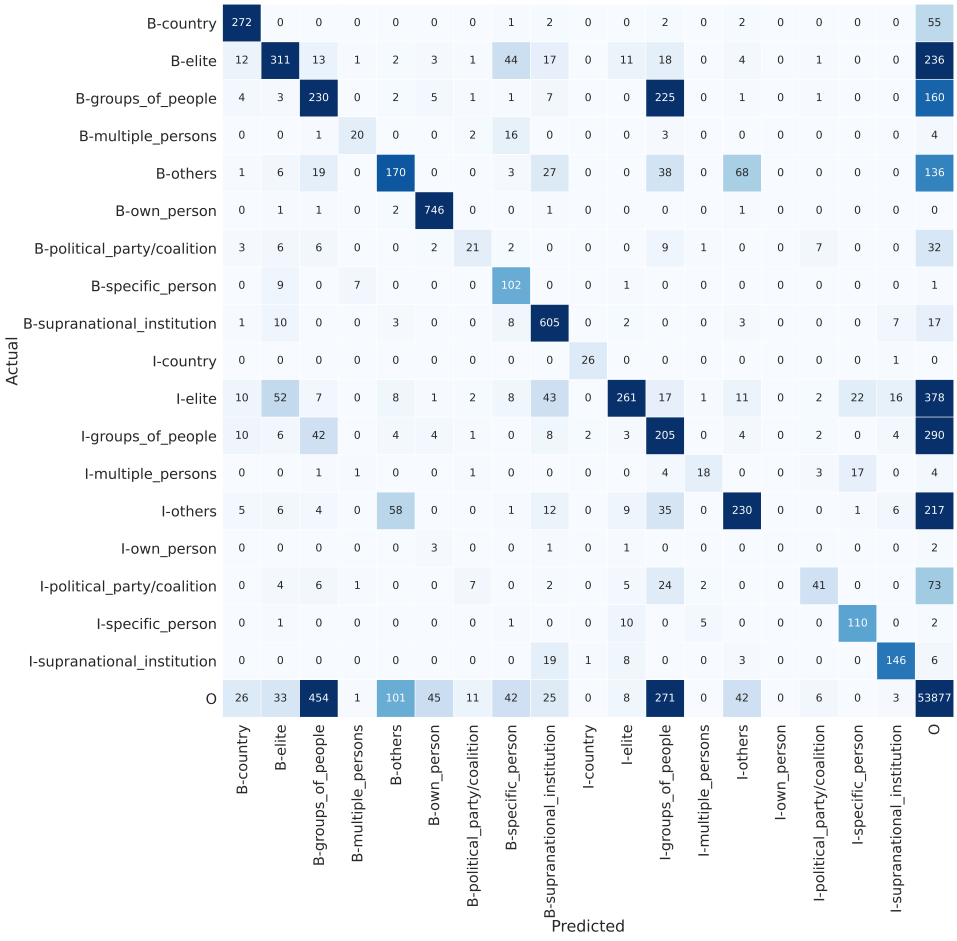


Figure C.2: Confusion matrix of BERT system

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Bachelorarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Mannheim, den 5.08.2021 Unterschrift