Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Extracting Political Relations from News Articles Using Transformers

HWR Berlin (Berlin School of Economics and Law)

Business Intelligence and Process Management WS2022/2023

Supervisors:	Prof. Dr. Markus Löcher
	Prof. Dr. Roland Müller
Submission:	20.01.2023
Author:	Soere Valentin Werner
	Student ID No:1892217
Word count:	16,942

Statutory Declaration

I hereby formally declare that I have written the submitted Master Thesis entirely by myself without anyone else's assistance. Wherever I have drawn on literature or other sources, either in direct quotes, or in paraphrasing such material, I have given the reference to the original author or authors and to the source where it appeared.

I am aware that the use of quotations, or of close paraphrasing, from books, magazines, newspapers, the internet or other sources, which are not marked as such, will be considered as an attempt at deception, and that the thesis will be graded with a fail. I have informed the examiners and the board of examiners in the case that I have submitted the dissertation, entirely or partly, for other purposes of examination.

Berlin, 20.01.2023

place, date signature

Abstract

Currently, political events are coded from newspaper articles worldwide using parsers based on dictionaries to generate datasets that help scientists to predict social unrest and conflicts. These parsers mostly lack flexibility when coding events, which can be improved upon using state-of-the-art natural language processing models, such as transformers. This thesis explores the use of sequence-to-sequence relation extraction transformers for political event coding. Two end-to-end relation extraction models (Hu et al., 2022; Cabot & Navigli, 2021) are proposed using both domain-specific and task-specific transfer learning.

After discussing flaws in current annotated datasets, two new datasets were created to pre-train and fine-tune the respective transformer models. The first dataset is created by a re-implementation of the currently used dictionary event coding algorithm. The second dataset comprises 2,206 annotated sentences from globally dispersed newspapers.

Several training strategies, such as entity masking and entity hinting were employed during training, finding mixed results in resulting performance. While both transformers were able to achieve excellent results on the pre-training dataset, the task-specific transfer learning model outperforms by a large margin during fine-tuning. It is expected that this result is greatly influenced by the dataset size used.

The main call for future research remains the creation of a high-quality large dataset for political relation extraction. Further directions for future transformer training are proposed. Particularly, the combination of a named entity recognition model in combination with an end-to-end relation extraction model can achieve promising results.

Keywords: Natural Language Processing, Relation Extraction, Text Generation, Political Event Coding, Deep Learning, Transformers, BERT, BART, CAMEO.

All code and documents are publicly available at:
https://github.com/valentinwerner1/Thesis_RelationExtraction_PoliticsNews

Table of Contents

Abbreviations

CAMEO	Conflict and Mediation Event Observations
FP	False positive
FN	False negative
IAA	Inter-annotator agreement
NER	Named entity recognition
NLP	Natural language processing
NLL	Negative log-likelihood
TP	True positive

List of Figures

List of Tables

1. Introduction

Every day, large amounts of news are being published by newspapers, blogs, or social media. These news are describing political interactions and opinions with high value for political and social scientists. However, due to the vast number of articles and missing structure during the search for information, such as being able to find specific political actors or specific political interactions, it is an impossible task for humans to analyse this data holistically. To overcome this issue, computerised event coding systems have been developed (e.g., Beieler & Norris, 2014; Norris, 2016; Lu & Roy, 2017) and are widely applied for these tasks already. Leveraging computers for this analysis not only creates more data in less time but can also reduce bias in news analysis, due to the global spread of data origins and opinions. Using data created by event coding systems allows to gain knowledge by creating rich visualisations, such as common interaction partners of a country or the difference in political relations reported between specific countries, while also allowing for methods such as time series forecasting to make statistically founded predictions for political climate, political instability, or risk assessment. It has been an easy task to extract political entities from a text given a list of relevant political actors or countries, however, the extraction of entities in a non-dictionary-based approach and the extraction of relations between the entities in a flexible yet reliable manner can profit from modern natural language processing (NLP) practices, such as deep learning using transformers. While the possibility to use these technologies exists, most projects currently recording political event data are still using dictionary-based pipelines (e.g., Leetaru & Schrodt, 2013; Althaus et al., 2020). These pipelines have been the target of research for a long period and although they have been improved over years of research, they have not moved on to these new technologies and have been criticised for their inefficiency (Parolin et al., 2022). The body of research looking at relation extraction for political events with transformers (e.g., Parolin et al., 2020; Hu et al., 2022; Parolin et al., 2022) has been dominated by the same researchers and is small when comparing it to the progress of research in relation extraction for domain unspecific benchmark data (e.g., Yan et al., 2021; Cabot & Navigli, 2021; Li et al., 2021). The current state of research on transformers for political relation extraction is already reporting a strong progression in performance over dictionary-based event extraction. Still, it has left many open questions and possibilities, such as new types of models, reformulating the way the task is processed and modelling techniques, which are frequently used in other domains of relation extraction.

1.1. Research Objective

This thesis aims to combine both the research in the domain of political event coding and the relation extraction modelling practices used in leading benchmark models. The potential of transfer learning with transformers will be leveraged to compare current state-of-the-art models from political relation extraction (e.g., Hu et al., 2022) and from benchmark models for domain unspecific relation extraction (e.g., Cabot & Navigli, 2021). Further, as relation extraction with sequence-to-sequence models found strong results with high flexibility for both domain-specific data in other domains (Giorgi et al., 2022) and domain-independent (Cabot & Navigli, 2021), the task of relation extraction will be conducted as a text generation task. The leading research question for this thesis is thus formulated as:

RQ: *"How can political entities and the relations between them be extracted from news article data and classified using sequence-to-sequence transformer models?"*

As transformer models have been shown to perform well when implemented with larger training corpora (e.g., Devlin et al., 2018; Cabot & Navigli, 2021) and no usable training data being publicly available, a further objective of this thesis is to create a dataset for the training of transformers in political relation extraction. To achieve this objective, first, news articles will be manually annotated to create a dataset with human-classified labels, and second, an unsupervised dictionary-based approach in combination with pre-trained transformers will be leveraged to create a large pre-training corpus.

1.2. Thesis Outline

The theoretical foundations will first provide a deeper understanding of both the task addressed in this thesis and the type of models used. The related work expands on these theoretical foundations by reflecting their implications on current research done in relation extraction and political relation extraction specifically. Building on the related work, Chapter 4 will introduce the data used in this thesis and the methodology applied for training, tuning, and evaluation of the models. Chapter 5 will describe the created artefacts, comprising the algorithm for the unsupervised dataset generation and the two transformer models trained. In Chapter 6, the model training and resulting performance will be evaluated on a multitude of metrics and an ablation study is conducted to quantify the effect of modelling measures which were taken to improve results. These results are discussed and set back into the context of the related work in Chapter 7. Finally, the thesis is concluded in Chapter 8, giving a summarisation of insights drawn from this thesis as well as an outlook for future work in political relation extraction.

2. Theoretical Foundations

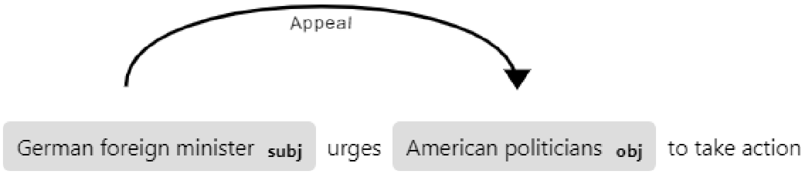
This chapter aims at giving a sufficient overview to understand decisions that have been made in the methodology, described in Section 4, and implementation of the artefacts, described in Section 5. First, the task of the problem researched, relation extraction, is described. This includes approaches that have been used in the past and current state-of-the-art methods and models. Second, the model structure of the transformer, the model type used for all recent state-of-the-art results is described in detail and its superiority over previous neural network approaches in NLP explained.

2.1. Relation Extraction

Relation extraction is the task of extracting a semantic relationship between two entities in natural language. Linguistically, these relationships require a subject, an object, and the given relation, mostly indicated by a verb connecting these entities. To visualise this, Figure 1 shows the example of “German foreign minister urges American politicians to take action” where “German foreign minister” is the leading subject and “American politicians” the object. Their relationship is an “Appeal”, indicated by the verb “urges”.

Figure 1

Example relation triplet showing entities and relation



Note. Created using spaCy 3.4.1.

Relation extraction presents a way to create structured data, such as actors, targets, and their relations, from unstructured data, such as news articles, or any other text. This data can be used to enable knowledge bases and information that is focused on interactions and co-dependency. The task of relation extraction can also be combined in other use cases such as question answering (e.g., Li et al., 2019; Zhao et al., 2020).

To accomplish this task of relation extraction, both the relevant entities must be correctly identified and the relationship between them must be correctly classified. The first part, identifying relevant entities, is referred to as named entity recognition (NER). The task of relation extraction is inseparable from NER and as result, most of the commonly used benchmark datasets for relation extraction are also used to evaluate NER performance. The challenge within the task of NER does not lie in finding candidate words, as these are mostly nouns or groups of words resolving around nouns, but in identifying which words are relevant entities. As result, this task was originally handled by utilising gazetteers and similar knowledge resources (Mikheev et al., 1999), allowing for a look-up of relevant entities. However, as the use for NER models is spread across several industrial domains, the relevant entities also differ highly and most gazetteers remain limited to specific domains. The second part of relation extraction, relation classification, is a classification task for the relation between two known named entities from a list of defined relations (Hendrickx et al., 2019). Much like the type of named entities, this list of relations is often domain related.

A related task, often incorporated with relation extraction is coreference resolution (e.g., Eberts & Ulges, 2021; Giorgi et al., 2022). Coreference resolution aims at finding entities with different names that describe the same entity. This becomes particularly critical when extracting not only relations but the knowledge implied by the relation. In a paragraph, “German foreign minister urges American politicians to take action. They have discussed with China for weeks.“, the entities “American politicians” and “They” refer both to the same group of persons. However, when splitting the text into separate sentences, this information will be lost in the second sentence and the relation to “China” can no longer be attributed to the “American politicians”. Identifying different words that refer to the same entity and replacing them in the text will allow maintaining information when splitting up text.

2.1.1. Pipelines or End-to-end

As relation extraction tasks need to accomplish both NER and relation classification, one method is to use separate models for both tasks. The necessary chaining of at least two different models into a pipeline gives this approach its name. It was commonly used with support vector machines and tree-based models (Bach & Badaskar, 2007) as these models achieved state-of-the-art results on separated tasks.

The first papers describing a joint training of both tasks used linear programming (Roth & Yih, 2007), a graph-based approach (Kate & Mooney, 2010) and a table-based approach (Miwa & Sasaki, 2014), respectively. This joint training is also referred to as end-to-end since one model performs all tasks from start to finish. A great benefit of models trained jointly is their ability to incorporate the similarities in the tasks of extracting entities and the relationships between them, making both tasks profit from one another (Yan et al., 2021). While Roth and Yih (2007) found joint training to fully outperform pipelines, the other two papers found mixed results when comparing the approaches.

The end-to-end model of Miwa and Sasaki (2014) outperforms overall scores for relation extraction but underperforms the pipeline on the NER subtask. Similarly, Kate and Mooney (2010) showed that the pipeline models performed better on specific relation types, while the end-to-end model performed better on others, indicating the effect of the structural differences in learning. End-to-end models have since been used more commonly, typically in combination with neural networks.

Several studies showed that the inherent benefit of joint training, being shared features and parameters during training can improve overall results (e.g., Yan et al., 2021, Miwa & Bansal, 2016). However, Zhong and Chen (2021) have found better results with a pipeline than with the end-to-end approach. They argue that a shared encoder for both tasks may reduce performance either due to distinctive features being important for the separate subtasks or if different input formats for the respective tasks are required. The first argument hints at the general problem that the data may require learning nuances in either task which are less focused during joint training. In recent papers, this shortcoming was addressed by utilising unshared features for named entities and relations while also using shared features separately (e.g., Crone, 2020; Yan et al., 2021), achieving new state-of-the-art results.

Although a pipeline of two or more models may allow for more specialised training, the biggest methodological drawback is the risk of error propagation. Many studies found the results on NER to be better than on relation classification (e.g., Zhong & Chen, 2021; Zhao et al., 2020), however, the models remain far from perfect. This means that several inputs for the relation classification task in a pipeline are prone to be classified incorrectly and the relation classification model can never outperform the NER model. Zhong and Chen (2021) discuss strategies to potentially reduce error propagation, such as adapted entity input markers but did not find any success using them.

Several papers suggested a third approach, using a separate NER model together with a joint relation extraction model (e.g., Wang & Lu, 2020; Han et al., 2020; Giorgi et al., 2022). This benefits from receiving entity hints from the first model which helps the joint model find relevant entities. At the same time, this method is not necessarily subjected to error propagation of a pipeline since the joint model is still able to identify different entities than given as input. The use of negative samples, i.e., samples that do not include any relevant relation, possibly helps the model to identify when these entity hints are or are not effective (Han et al., 2020).

Recent models achieving or improving state-of-the-art results on relation extraction benchmark datasets utilised mainly the end-to-end approach. However, when evaluating the approaches, a critical lack of literature exists regarding direct comparisons. While it would be generally possible to use the individual methodology of the models both on pipelines and on end-to-end models, often only one method is chosen. Table 1 shows the results of papers that incorporated both approaches, either by training them individually or by using an ablation study. Out of seven identified papers that compared pipeline models with end-to-end models, five reported better results using the end-to-end approach. This small sample size does not allow for definitive conclusions and further research is needed to compare the approaches more directly.

Table 1

Direct comparisons of end-to-end and pipeline performance on various datasets

Authors	Year	Dataset	Approach	RE
Miwa & Bansal	2016	ACE05	End-to-end	51.8
			Pipeline	51.0
Nguyen & Verspoort	2018	CoNLL04	End-to-end	66.9
			Pipeline	66.3
Han et al.	2020	MATRES	End-to-end	59.6
			Pipeline	57.2
Zhong & Chen	2021	ACE05	End-to-end	64.4
			Pipeline	64.8
Yan et al.	2021	SciERC	End-to-end	38.4
			Pipeline	36.9
Eberts & Ulges	2021	DocRED	End-to-end	59.46
			Pipeline	59.76
Giorgi et al.	2022	CDR	End-to-end	52.4
			Pipeline	34.1

Note. Yan et al. (2021) have not directly compared end-to-end to pipeline but incorporated a pipeline-like model in their ablation study using sequential encoding instead of joint encoding.

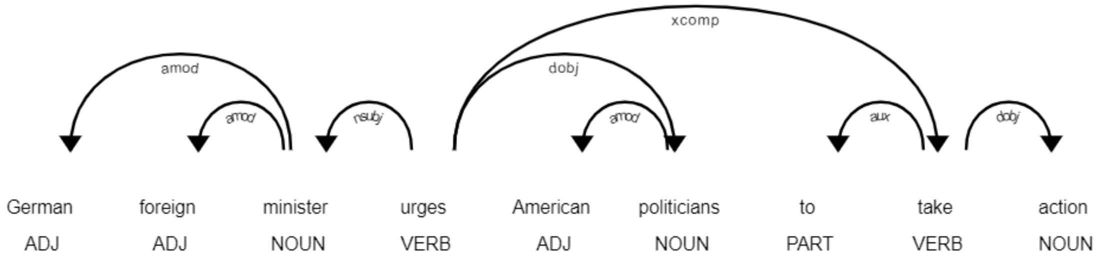
A final argument to be made for joint training is the reduction of training time (Li et al., 2021). Training time became particularly critical with large transformer models needing a longer time for training and hyperparameter tuning, only allowing the tuning on a subset of training data (e.g., Cabot & Navigli, 2021) or epochs, expecting the results to be representative for longer or full training. As such, it may be preferable to train one model for both tasks instead of training separate models on the same data.

2.1.2. Sequence-based or Dependency-based

Another differentiation of relation extraction models is separating them into sequence-based and dependency-based models (Guo et al., 2020). While sequence-based relation extraction models are purely based on text, dependency-based models also include features generated from dependencies trees, representing grammatical dependencies between tokens of a sentence. Based on a dependency tree it can be inferred which words in a sentence are connected in which manner. An example of a dependency tree is visualised in figure 2. Taking the sentence “German foreign minister urges American politicians to take action” again, “German foreign minister” and “American politicians” are compound word groups, and they are indirectly connected by the root word “urges”.

Figure 2

A dependency tree for an example sentence



Note. Created using spaCy 3.4.1. Arrows are labelled with universal dependencies (de Marneffe et al., 2014).

These trees offer further features and value when learning relation extraction as linguistic connections between words of the input are defined clearer. Dependency trees have been used successfully in the past with graph convolutional networks (Zhang et al., 2018; Guo et al., 2020) with particular success in document-level relation extraction, characterised by long sentences and paragraphs. While several pruning strategies were used to distil the information of a tree into the most relevant parts (e.g., Xu et al., 2015; Miwa & Bansal, 2016; Zhang et al., 2018), Guo et al. (2020) show that using the full unpruned input helps when extracting relations from long inputs, such as full articles. Transformer-based methods have been proven similarly successful on relation extraction tasks with long inputs both with (Zeng et al., 2020; Xu et al., 2021) and without (e.g., Tan et al., 2022) dependency components, but are currently achieving state-of-the-art results without them.

2.2. Transformers

First introduced by Vaswani et al. (2017) as a model for machine translation, transformer models revolutionised the performance of NLP models, including relation extraction models. At heart, the transformer is based on feed-forward neural networks in combination with attention mechanisms. Additionally, the transformer uses an encoder-decoder structure, which has been proven successful in previous state-of-the-art language models, such as recurrent neural networks (e.g., Cho et al., 2014). However, the attention mechanism is the key difference between the transformer model and these previous models (Vaswani et al. 2017). This section aims at explaining the transformer architecture and features, as well as how they are used for transfer learning.

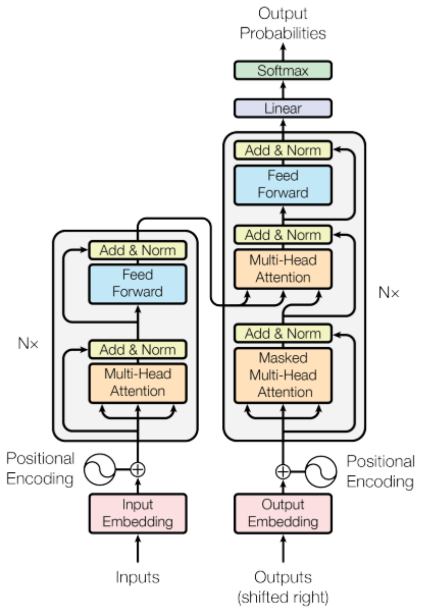
2.2.1. Architecture

The transformer can be split into three main features: Embeddings, Encoder-Decoder Structure, and Attention. As the original transformer was created for machine translation, the original goal was to create a translated output sentence given an English input sentence. The translation is generated word-by-word, each word being generated after the input went through the whole transformer architecture once.

Figure 3 shows the transformer architecture as designed by Vaswani et al. (2017). Inputs fed into the model get iteratively encoded over a set number of *N* encoders to receive a more refined word representation, based on the word embeddings, its position in the input sequence and its context to other words in the input, the latter being generated by the attention heads of the encoder. The decoder receives its input auto-regressively and creates similar word representations as the encoder for the so-far generated words of the output sequence. These inputs are also encoded based on the generated word itself, its position in the generated sentence and its context to other so far generated words. Importantly, Vaswani et al. (2017) use one Encoder-Decoder multi-head attention layer in each of the *N* decoders, setting the output representation in context to the encoded input, allowing to see which words of the generated output refer to which words of the input and to create new context vectors. These attention vectors represent the decoder output after being adjusted through a two-layer feed-forward neural network and normalisation. Based on one additional linear activation and a softmax function, the probabilities for the next output word are calculated and the next output word for the generated sequence is chosen based on the highest probability.

Figure 3

Transformer Architecture (Vaswani et al., 2017)



2.2.1.1. Multi-headed Attention

The scaled dot-product attention mechanism, in this thesis previously referred to as "context to other words", describes how close words in a sequence are related to another. Vaswani et al. (2017) calculated the attention for every word in a given sequence as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

Given *Q*, *K*, *V* being the dot-product of the input embeddings and the weight matrices W_Q, W_K, and W_V. These matrices are randomly initialised and trained during the encoder step.

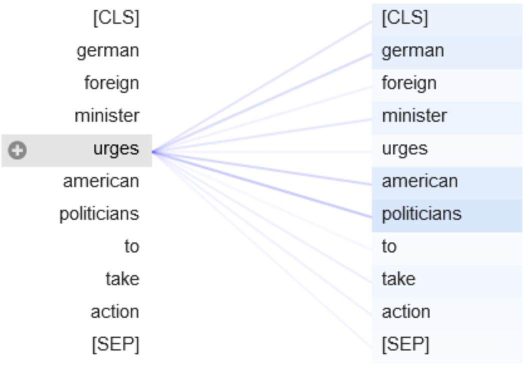
Instead of doing this scaled dot-product attention once, Vaswani et al. (2017) employ multi-headed attention over a number of heads *h* simultaneously. The weight matrices W_Q, W_K, and W_V are split into *h* matrices of the same size, being used in different heads. Finally, the attention matrices calculated are concatenated. Vaswani et al. (2017) argue that splitting this attention task into multi-head attention enables the embeddings to learn different features of the same word in its position in the sequence and of its relation to other words within the input sequence or generated sequence, potentially creating richer word encodings overall.

The multi-headed attention layer fixes the big struggle of recurrent neural networks, having to carry early token information through every following token, potentially resulting in information being lost and in exploding or vanishing gradients in case of long sequences (Le & Zuidema, 2016). Instead of gathering information from the input sequence sequentially as recurrent neural networks and long short-term memory networks do, transformers can calculate the attention matrices in parallel for the whole sequence, as the attention of a word is isolated from the attention values of a prior word (Vaswani et al., 2017). This parallelisation in encoding offers the potential of a speed-up due to the reduction of sequential steps at the cost of quadratic scaling in complexity due to the attention being calculated between every word of the sentence. To overcome this issue in the case of long sequences, it is possible to restrict the attention to a specified window size, inducing constant instead of quadratic scaling (Vaswani et al., 2017).

Figure 4 shows the attention of the token "urges" to other tokens in the sentence. The visualization only shows one self-attention attention head generated from the BERT transformer (Devlin et al., 2018). This head visualises particularly well how transformers hold high potential for relation extraction as the relevant verb pays the highest attention to relevant noun-tokens.

Figure 4

Attention of verb token to other tokens in a sequence, one head of multi-headed attention



Note. The respective head was chosen out of a selection of 144 potential heads as it visualises the relationship between the tokens particularly well. Generated with the BertViz 1.4.0 Python package, using the BERT transformer (Devlin et al., 2018).

2.2.1.2. Encoder-Decoder-Architecture

Before using the encoder-decoder architecture with the transformer, it was used to achieve state-of-the-art performance with recurrent neural networks (Cho et al., 2014; Bahdanau et al., 2016). Encoders are used to create representations of the input sequence, while decoders create representations of the so-far-generated output sequence in an auto-regressive manner. Many tasks in NLP, such as classification tasks, however, do not require the generation of a sequence in an auto-regressive manner. Since decoders pose no values for these tasks, they are often solved by transformer encoders (e.g., Devlin et al., 2018).

Importantly, the use of an encoder-decoder attention layer allows setting both the encoder representation of the input and the decoder representation of the so far generated output in context to another, as the attention between decoded words and the whole encoded sequence can be calculated at once (Bahdanau et al., 2016). The encoder-decoder attention in recurrent neural networks uses the whole encoded sequence but is limited to the last decoded token (Bahdanau et al., 2016) while the encoder-decoder multi-head attention used in transformers can attend to every decoder input and every encoder input (Vaswani et al., 2017).

This encoder-decoder multi-head attention step is particularly important in sequence generation tasks, such as machine translation, as it helps to learn which words in the encoded language refer to which words in the decoded language. In the example of a sequence generation relation extraction where an output sequence "<triplet> subject <subj> object <obj> relation" is generated, the transformer will be able to learn special tokens and their meanings, such as the <subj> token ending the subject phrase, and after generating both subject and object, the relation indicating verb will show high attention to both the decoded subject and decoded object. With a recurrent neural network using an attention encoder-decoder architecture, the relation indicating verb could not show high attention to specific already decoded tokens, but instead only to the last decoded tokens, being <obj> and indirectly to the prior tokens which were used as input to generate the <obj> token.

2.2.2. Transfer Learning

Transfer learning describes the transfer of a model's knowledge of a domain it was trained for to a new task in the same domain or a model's knowledge of a task to a different domain with the same task (Pan & Yang, 2009). While it is possible to use transfer learning for both domain and task simultaneously, the transferred model requires a deep general understanding of knowledge. Transfer learning became relevant in the context of transformers when large language models such as GPT (Radford et al., 2018), BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2019), and DeBERTa (He et al., 2020) started using the transformer model in connection with huge training corpora. These models have further proven that using more encoder and decoder layers, resulting in larger models, performs better than the base models that try to balance performance and training time by reducing the model size.

Arguably the most widely used transformer for transfer learning is the BERT (Devlin et al., 2018), also being the basis of many following models. BERT was trained using masked language modelling, requiring the model to predict a word that was masked in the sentence, and next-sentence prediction, requiring the model to predict which sentence follows on a given input sentence. This has been shown to result in models with a strong understanding of languages, even surpassing human benchmarks in following models (He et al., 2020). While it took days to train these models (Devlin et al. 2018; He et al., 2020), it only takes minutes to access these models and transfer this understanding to new tasks or domains, achieving large performance gains over training without pre-trained transformers (Raffel et al., 2020). The models can then be used to extract the word embeddings as the basis for downstream tasks or to be fine-tuned on domain-specific and task-specific data. The results of transformer research imply that, if sufficient data size is existing, large language models can be used as basis for any NLP task, including relation extraction in a sequence-to-sequence manner. At the same time, Devlin et al. (2018) suggest, that large pre-trained models will also perform well on smaller datasets due to the larger and more expressive representations generated with more parameters from large versions of the respective models.

When fine-tuning these large language transformers, it is argued, that not all parameters should be fully adjusted. One reason may be that many features are task-independent and possibly won’t generalise as well when trained again on smaller datasets. At the same time, fine-tuning all layers will also require further training resources. Lee et al. (2019) found that freezing layers, i.e., not fine-tuning specific layers, always results in lower performance, however, they were able to maintain 90% of the quality of the original results when only training 25% of BERTs layers. Similar outcomes were achieved by other papers, showing that the full fine-tuning approach is superior to freezing all layers (Peters et al., 2019) and over gradually unfreezing layers during training (Raffel et al., 2020). As such, freezing layers should only be a measure taken if computation resources are scarce compared to available training data and model parameters.

Many of the transformer models exhibit different architectures, particularly regarding the encoder-decoder structure. For example, BERT is a transformer encoder (Devlin et al., 2018), while other models, such as GPT are considered transformer decoders, allowing for autoregressive tasks like sequence generation. To use encoder transformers for sequence generation, transferring the pre-trained domain knowledge to a new task, Rothe et al. (2020) suggest using a second model as a decoder transformer in combination with the original encoder model. Using a Bert2Bert model, utilising BERT both as encoder and decoder by randomly initialising the missing encoder-decoder layer in the decoder, showed results that were competitive with state-of-the-art results (Rothe et al., 2020). While it seems logical that the decoder model chosen for this should be a transformer with a decoder architecture, such as GPT, Rothe et al. (2020) found these combinations to perform worse. Possibly, this is due to models being trained on different vocabularies (Rothe et al., 2020). When combining two transformers to create an encoder-decoder model, due to the random initialisation of new layers totalling almost 10% of the model's parameters, it is expected that this model will require more training than a pre-trained encoder-decoder transformer. It is suggested to share parameters between the encoder and decoder when utilising the same model type both as encoder and decoder, gaining performance and reducing the model size (Rothe et al., 2020).

3. Related Work

While many relation extraction papers use the NYT benchmark dataset (e.g., Yan et al., 2021; Cabot & Navigli, 2021; Li et al., 2021), there is currently a lack of benchmark datasets for political relations from newspaper articles. Several papers have made efforts to generate event databases on political news articles (e.g., Leetaru & Schrodt, 2013; Althaus et al., 2020) using the Conflict and Mediation Event Observations (CAMEO) ontology which describes 20 main classes of interactions between political actors with overall 267 sub-classes (Gerner et al., 2002). Another commonly used classification is Pentacode, re-classifying each CAMEO interaction into material or verbal and cooperation or conflict, or a fifth, neutral class called "Make a Statement", resulting in five classes overall. While the reduction from 20 to five types of interactions results in a lot of information lost, it captures the overlying content of the interaction (Beieler, 2016).

Several large datasets using the CAMEO ontology are currently generated using PETRARCH2 (Norris, 2016). This approach utilises syntax trees, including part-of-speech tagging, finding relevant verbs within the input sentence, and then matching input words and the relevant verb with long, manually created dictionaries listing possible source actors, target actors, and verbs to recreate the triplet. One example of such a verb is "assist", being matched to the category "Engage In Material Cooperation" or more advanced verb patterns such as "assist {actor} in democracy" being matched to the category "Engage In Diplomatic Cooperation". Similarly, actors such as "Angela Merkel" are being matched to "Germany". As result, these coders create datasets that only record countries or relevant international institutions as actors and their interaction categories. These datasets are discarding the text the relation was created upon, making them less valuable as training data for NLP models.

This algorithm, basically being a text and dictionary parser, shows several weaknesses. First, to create a relation both actors and their interaction need to be identified. The common issue of having any missing values will result in a not-classified sentence (Parolin et al., 2020). Second, the list of actors, arguably also the list of verbs, needs ongoing additions of new relevant actors. The currently most updated public list does not contain commonly reported actors such as “Donald Trump” and other recently elected leaders or relevant local politicians, such as mayors of large cities. However, even when both actors can be found in the listed dictionaries, the algorithm only finds interactions in 55% of the sentences, overall only finding interactions in 15% of the sentences from newspapers (Parolin et al., 2020). Third, if a sentence has more than two source or target actors, only one will be identified in combination with the relevant verb, resulting in the further loss of information. On the other hand, the coders quickly generate a lot of interactions in an unsupervised manner.

Recent transformer research trying to substitute these algorithms tried pipeline models (Hu et al., 2022; Parolin et al., 2022) and end-to-end models (Parolin et al., 2022) that perform NER for source and target actors and classify the interaction using Pentacode. Both are built using the BERT transformer (Devlin et al., 2018) as basis, achieving state-of-the-art results which are largely outperforming earlier work, such as the PETRARCH coder. At the same time, other models based on more recent transformers, such as BART (Lewis et al., 2019) and RoBERTa (Liu et al., 2019), have been outperforming BERT by applying more advanced reconstruction tasks. This has also resulted in BART-based models achieving state-of-the-art results over BERT-based models in relation extraction tasks (e.g., Cabot & Navigli, 2021). Parolin et al. (2022) trained their model, Multi-CoPED, by utilising a multi-task approach, instantiating new task-specific layers on top of BERT. While this is considered end-to-end as only one model is used for both tasks, the parameters from the NER layers are not shared with the relation classification task. The identified named entities can then be used in combination with the CAMEO dictionaries to streamline them into the wanted event structure, recording the country or organisation code instead of the original entity, recreating the target output of PETRARCH algorithms. This approach enables the use of transformers in the CAMEO event recording task but still contains the problem of dictionaries having to be updated regularly while also reducing the recorded relations from the original 267 sub-classes and 20 main-classes from the CAMEO ontology to only five classes from Pentacode.

Hu et al. (2022) chose to pre-train BERT on a combination of large corpora of media-related data to create ConfliBERT. This pre-training was performed from pre-trained BERT and from scratch, both models yielding a similar performance on the Pentacode dataset used to evaluate it. Although BERT is likely not the highest performing choice when applying transfer learning for this use case, ConfliBERT poses a good option as its vocabulary is particularly well fit for political interactions. This is particularly valuable when working with smaller datasets, as ConfliBERT achieves the highest performance gain over BERT on tasks which were trained on the smallest datasets (Hu et al., 2022).

Parolin et al. (2022) used self-annotated training and validation data, which was annotated and validated by qualified personnel with a good inter-annotation agreement. As test set, they used gold standard records from annotated sentences from the CAMEO codebook and test sentences used to validate the PETRARCH2 (Norris, 2016) and UPETRARCH (Lu & Roy, 2017). Hu et al. (2022) only used a subset of this data to achieve their results. Critically, in the publicly available subset used by Hu et al. (2022) both the data used for training and evaluation only contains one type of event between two actors. This induces further bias, as other relations in the sentence are missed. Neither author accounted for this bias during the training of their model. Overall, Parolin et al. (2022) achieve higher scores for NER, while Hu et al. (2022) achieve higher scores on the relation classification subtask on Pentacode. There were no papers identified that evaluate transformers for CAMEO data on all 20 main-classes.

Alternatively to choosing ConfliBERT or Multi-CoPED as a specialised pre-trained option, it seems logical to choose relation extraction models, which show high performance on datasets which have overlap with the data used for CAMEO classification, such as the NYT benchmark (Riedel et al., 2010; Zeng et al., 2018). REBEL (Cabot & Navigli, 2021) is a relation extraction model based on BART (Lewis et al., 2019) which additionally to the thorough pre-training of BART, was pre-trained a second time directly on a relation extraction task with a large Wikipedia-based dataset. The REBEL dataset was created by matching entities from Wikipedia articles, which are indicated due to having their own Wikipedia article, together with their relation that is stored in WikiData. To validate whether this relationship is induced, a natural language inference model was used to predict the entailment of the relation given the sentence the entities were extracted from, reducing noise in form of false or unreliable labels in the dataset (Cabot & Navigli, 2021). This pre-training has proven to improve results on new domains (Cabot & Navigli, 2021) and, similar to the pre-training of ConfliBERT, possibly also reduces the necessary training set size as indicated by achieving the largest performance gains on the smallest datasets it was tested on.

4. Methodology

The methodology aims at explaining what decisions were made during training, why they were made and what outcomes are expected. To achieve this, first, the two datasets used during training are described, followed by the modelling methodology, describing the hyperparameters, model configuration, and training strategies.

4.1. Data

This section elaborates on all data used, the strategy in sampling the data, the purpose of the datasets, the processing of the datasets, and possible problems within and between the datasets. Further, explorative data analysis will give a deeper understanding of the structure of the data. This section first goes into the unsupervised data and then describes the annotated data.

4.1.1. Unsupervised Dataset

The unsupervised dataset is a noisy but large dataset created by a combined approach of PETRARCH2 (Norris, 2016) and entailment prediction. The algorithm used to create the dataset is described in Section 5.1 while this section focuses on the raw data used for creating it, its augmentation and analysis of the resulting dataset.

4.1.1.1. Data Sourcing and Pre-processing

As transformers require larger amounts of training data to get a better understanding of the domain they are working in, a separate dataset was created in an unsupervised manner to perform pre-training on the chosen models. The algorithm used to create the dataset is described in Section 5.1 in detail. The dataset aims at teaching the model spans for relevant entities and basic subject-verb-object patterns that are identified by current automated event coders. It is expected that training on this dataset provides a strong basis for learning nuances of political relation extraction from the annotated data.

The text for the unsupervised dataset was generated from full news articles which were pre-processed with coreference resolution using SpanBERT (Joshi et al., 2020), replacing pronouns and different names of the same entity with a uniform name of the belonging entity to increase the level of information in a sentence. This results in the text containing more relevant entities for which relations will be found, as pronouns, such as “they”, are not found as an entity by the algorithm. The articles were then split into sentences. In the case of 641 samples where the sentence splitting failed, resulting in a text length of more than 500 symbols, the full samples were removed from the dataset.

9,765 articles were retrieved from several newspapers, listed in Table A1. A portion of the articles used to generate the unsupervised dataset is identical to the articles which were used to create the texts for the annotated data. However, while the unsupervised data uses all sentences of the article, the annotated data uses the article title and description, resulting in only overlapping entities and topics, but not identical input texts. This possible overlap was accounted for in the annotated test set by using separate data.

To create a larger dataset and reduce this overlap, articles from five newspapers were further extracted from the GDELT database (Leetaru & Schrodt, 2013), which has been creating CAMEO event codes for years. These articles were chosen based on being published between January 2019 and October 2022 and based on newspapers, overlapping with the ones used for the annotated dataset. This resulted in 18,383 articles from the newspapers BBC, CBC, Eyewitness news, Sydney morning herald and TASS.

It is expected that adding articles from different periods will make the models more robust by learning further actors and relations. Because of many ongoing conflicts, the model would be able to learn correct relations purely from the entities in the sentence, such as “Russia” and “Ukraine” being mostly connected to the relation “Fight”. Adding news including these entities in different contexts may help the model to learn relevant indicators for the relations from the text instead of deducing them from the entities.

4.1.1.2. Data Augmentation

When looking at CAMEO relations, the original dataset was heavily unbalanced. To reduce this, data augmentation was conducted in three different ways. First, the heavily underrepresented relations “Engage In Unconventional Mass Violence”, “Reject”, and “Exhibit Military Posture” were fully removed from the dataset, as the number of samples was not sufficient to create a test and valuation set while leaving enough samples to train on. While this reduces the pre-training effect on CAMEO labels for these relations, the learned information for these samples would mostly be overfitting on little information, while validation and test performance would greatly vary between seeds chosen for the split.

Second, the class imbalance problem is reduced by applying random undersampling, i.e., removing data points from heavily over-represented classes. The random undersampling was employed for the relations “Make Public Statement” and “Consult”, resampling 2700 of the original samples, being approximately 50% more samples than the next most prevalent relation “Disapprove”. This resampling was set to prefer samples containing multiple relations to not further reduce the count of other relations, effectively only dropping sentences if no other than the targeted relation was present.

Third, sentences were altered by randomly swapping verbs and entities. For swapping verbs, the verbs which indicate a relationship that is prevalent in the sample are detected using the CAMEO verb dictionary as reference. Verbs were only swapped with other verbs which are recorded in the CAMEO verb dictionary for the same relationships to maintain the relation found for the sample. Further, to keep a similar meaning in the text when swapping verbs, BERT embeddings were extracted for all candidate words which indicate the same relationship to calculate the cosine similarity to the original verb. The substitution for the original verb was then drawn from the five most similar verbs in the dictionary. An example of this would be the words “report” and “describe” both indicating the relation “Make Public Statement”, while other words indicating this relation, such as “decree” would not be chosen as substitution, as it is likely does not fit the context of the sentence. Overall, this will help to represent more of the CAMEO verb dictionary inside the training data. Entities were swapped if they are included in a labelled relation. To maintain logical sentences, entities were only swapped with entities of the same type, e.g., an entity being an organisation was only substituted with another entity which is an organisation. The entities that are eligible as substitution were extracted from all identified entities in the unsupervised dataset, creating a large pool for every entity type. Entities in this pool were deduplicated to reduce the high frequency of recurrent actors. An example of this full augmentation would be the sentence “The UK also donated ventilators to India earlier this week“ being transformed to “US-backed forces also provide ventilators to Atlanta earlier this week”, forming a rephrased sentence with new entities performing the same action.

Every relation with less than 500 samples was augmented *n* times, with *n* being calculated relative to the count of samples existing for this relation. While it is recommended to augment the same sample no more than 16 times (Wei & Zou, 2019), the samples in the least prevalent class were only augmented for a maximum of nine times to reach the given threshold.

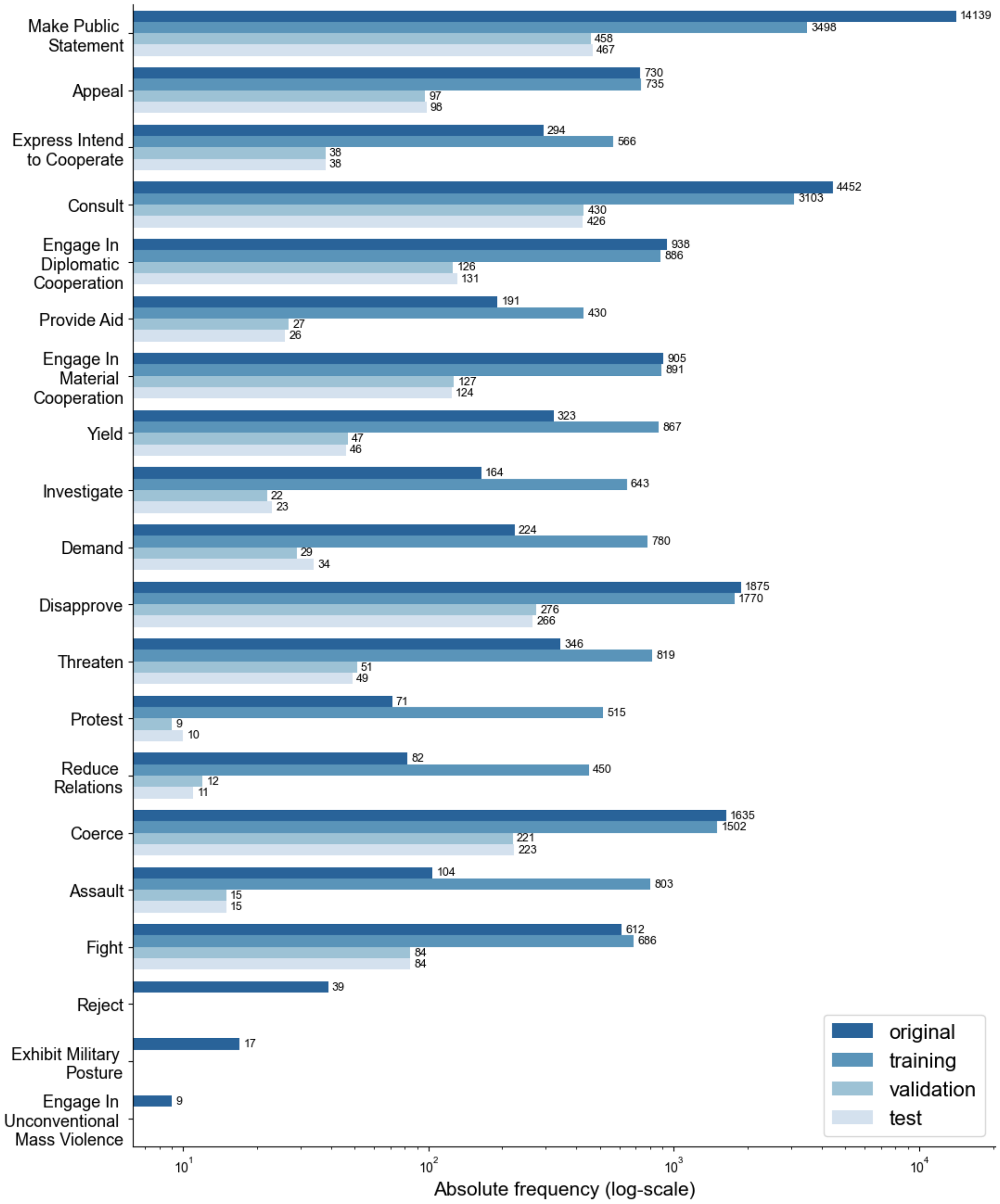
$$n_{relation} = \min(16, integer\left(\frac{500}{N_{relation}}\right)) \qquad (5)$$

4.1.1.3. Data Description

Figure 5 shows the distribution of relations within the original data, being the data before augmentations, and within the training, validation and test split for the data used during training. The augmentation resulted in a more balanced training dataset, albeit at the cost of reducing from 20 to 17 relations. In the new dataset, 18,947 samples (95.54%) contained one relation, 754 samples (3.80%) contained two relations, and 130 samples (0.66%) contained three samples. The unsupervised data was split into 70% training data, 15% validation data, and 15% test data in a stratified manner, controlling for a similar relative frequency of relations in each split. The split into the respective sets was performed before augmenting.

Figure 5

Unsupervised dataset distribution for CAMEO labels

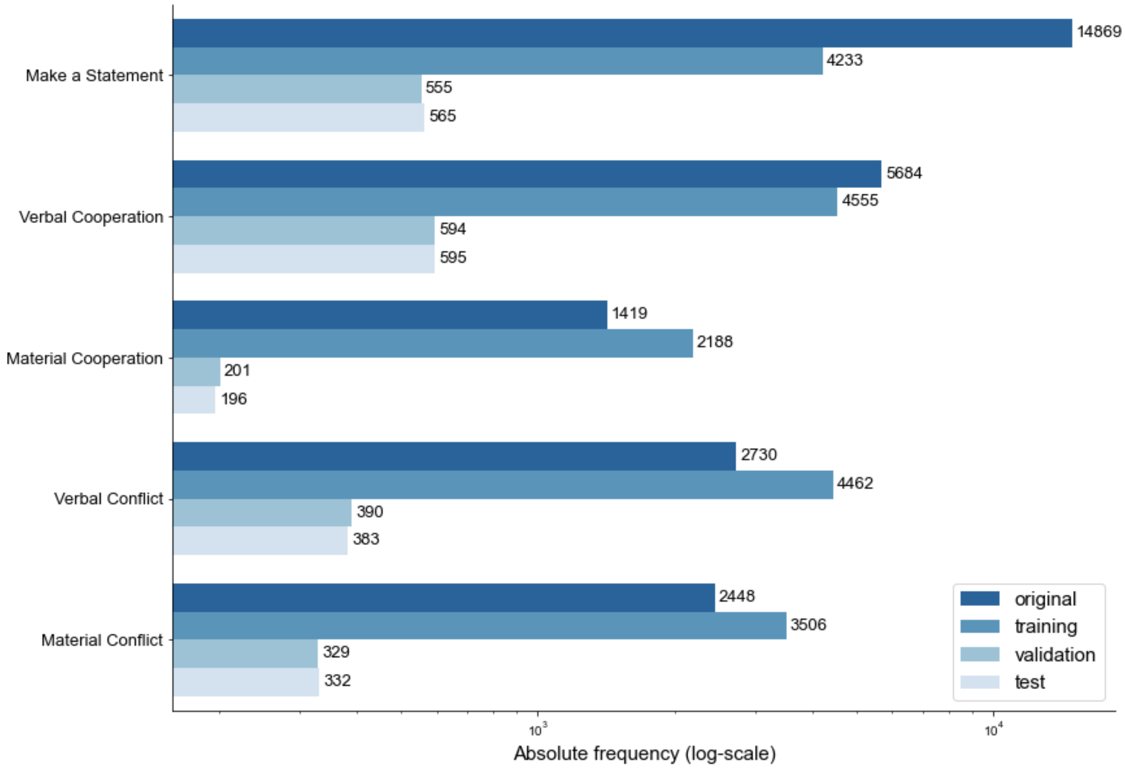


Note. “Original” refers to the original distribution of annotated data including example sentences from the CAMEO codebook. The x-axis was log scaled to help visualise relations with fewer samples. Training data is shown after random sampling and augmentation. Training, validation, and test data are shown for the first seed.

The labels were transferred into Pentacode labels as shown in Table A2, resulting in a more balanced training set for the Pentacode labels as well. Further, by using the same augmentation as for the CAMEO labels, the task becomes more comparable, allowing to evaluate the difference in training performance between both ontologies. The distribution of labels when using Pentacode is shown in Figure 6.

Figure 6

Unsupervised dataset distribution for Pentacode labels



Note. “Original” refers to the original distribution of annotated data including example sentences from the CAMEO codebook. The x-axis was log scaled to help visualise relations with fewer samples. Training data is shown after random sampling and augmentation. Training, validation, and test data are shown for the first seed.

The unsupervised dataset is not able to be used for a NER task disjoint from the relation classification task as used in the pipeline approach for relation extraction, as only entities that are included in a relation are included in the dataset. Instead, this would require all entities to be annotated to reduce error propagation within the pipeline, caused by a higher amount of false negative labels. As such, the dataset is only suitable for relation classification and end-to-end relation extraction.

4.1.2. Annotated Dataset

The annotated data is a smaller dataset aiming to create high-quality labels for political relation extraction. This section aims at explaining the remaining bias in the dataset and measures taken to reduce it. Further, the data distribution within the dataset will be reported.

4.1.2.1. Data Sourcing and Pre-processing

The annotated data comprises two parts, a manually annotated dataset from various newspaper articles and, similar to the evaluation dataset of Parolin et al. (2022) and Hu et al. (2022), the example sentences from the original CAMEO codebook. Different to these authors, the evaluation dataset was re-labelled to include all relations present in the example sentences, instead of only the example relation. This resulted in 954 relations on 372 sentences containing on average 1.75 different relations, compared to 1.0 different relations as used by Hu et al. (2022). Examples of this re-labelling can be seen in Table A6.

The main dataset was annotated by two persons between 01.11.2022 and 01.01.2023 using the spaCy annotation library Prodigy 1.11.4. The text data was sourced from 13 newspapers using their respective RSS feeds on world politics every two hours. Potential newspapers were chosen from politically or economically leading countries from every continent. Only one newspaper was chosen for each country. While this drastically reduces the number of newspaper articles, sourcing from globally dispersed newspapers reduces regional bias in the data which may be present in the way actions of certain entities, such as political enemies or allies, are described in the article. The final selection of newspapers was based on the size of the newspaper, preferring newspapers with high circulation or website visits, and political orientation, preferring newspapers that are generally viewed as central and least biased. This selection was further influenced by limiting the language of the newspaper articles to English and the necessity of a fitting RSS feed for world news being available. From the RSS feeds of the respective newspaper, the title and description of every article were collected. Sourcing of the data started on 01.11.2022 and ended on 22.12.2022. Table A1 shows selected newspapers and their origin.

The retrieved title and description of each article collected were combined, duplicates removed, and the text was pre-processed with coreference resolution using SpanBERT (Joshi et al., 2020). The data was annotated as a combination of title and description but was split into sentences after annotation. Although this combination of article title and description offers a dense proportion of relations per sample, also allowing to find relations inter-sentences, it was chosen to keep individual sentences as input. First, current research (e.g., Parolin et al., 2022; Hu et al., 2022) focuses on sentence-level relation extraction, making the results of this thesis more comparable to existing studies in this area. Second, both models trained were pre-trained on sentence-level samples, likely resulting in easier training with less fine-tuning for the attention vectors. Third, shorter samples are offering more flexibility during training, such as increased batch size.

During annotation, sentences to be annotated were shuffled several times to get articles from the full timespan in which they were retrieved. The annotation was performed with the 20 main CAMEO codes, only annotating the full subject, full object, and the CAMEO relation between them. Sentences were not annotated if the coreference resolution failed, i.e., by replacing wrong entities or identifying too many words for an entity, drastically increasing the text length. Duplicate triplets, having the same subject, object and relation, for the same text were either not annotated in the first place or removed after annotation as no additional value is gained from predicting the same relation twice. This likely will also help training as requiring text generation models to repeat output can result in endless repetitions of this relation, only limited by the maximum sequence length the model is allowed to generate.

The data was annotated without negative examples, meaning the data will only comprise texts which are connected to a relevant relation. This was chosen to reduce annotation time, as sequence generation tasks do not require negative samples (Giorgi et al., 2022). Further, only entities which are connected through a relevant relation were annotated. This also helped to reduce time and thoroughness in the annotation process, albeit at the cost of the dataset being usable for further tasks. The full codebook for entities and relations, as well as annotation guidelines used for the annotation process of this dataset, is included in Appendix D. Further, some example sentences which give further insights into the annotation process, such as skipped annotations and difficult annotations are included in Appendix B.

4.1.2.2. Data Augmentation

Several measures were taken to reduce the class imbalance in the annotated dataset and the CAMEO examples. First, to reduce the relation “Make Public Statement” in the CAMEO examples, the suffix of “, said {entity}” and the relations connected to that suffix were removed. For undersampling, 66 samples which contained at least three “Make Public Statement” relations with these making up more than 50% of all relations in the respective sample, were removed. Most importantly, the class imbalance was reduced by increasing underrepresented classes with verb and entity swapping similarly as used for the unsupervised data. Entities and the relation indicating verbs have been swapped in relative frequency to the relation appearing in the dataset. However, as the counts per label are overall smaller, the random swapping of entities and verbs was only conducted for relations which occur equal to or less than 100 times in the training set rather than 500 samples as used for the unsupervised data. This resulted in eleven out of 19 remaining relations being augmented until the threshold of 100 occurrences was reached.

4.1.2.3. Data Description

The size of the full dataset was increased from 2,206 samples to 3,052 samples using data augmentation. Compared to the unsupervised dataset, the annotated dataset finds more relations per sample. Within the augmented data, 1,549 samples (50.75%) were found to have one relation, 918 samples (30.08 %) were found to have two relations, 347 samples (11.37%) have three relations, 130 samples (4.26%) have four relations, and 108 samples (3.54%) have five or more relations, reaching up to nine relations in two cases. Overall 5,544 relations were found on 3,052 samples, averaging 1.82 relations per sample.

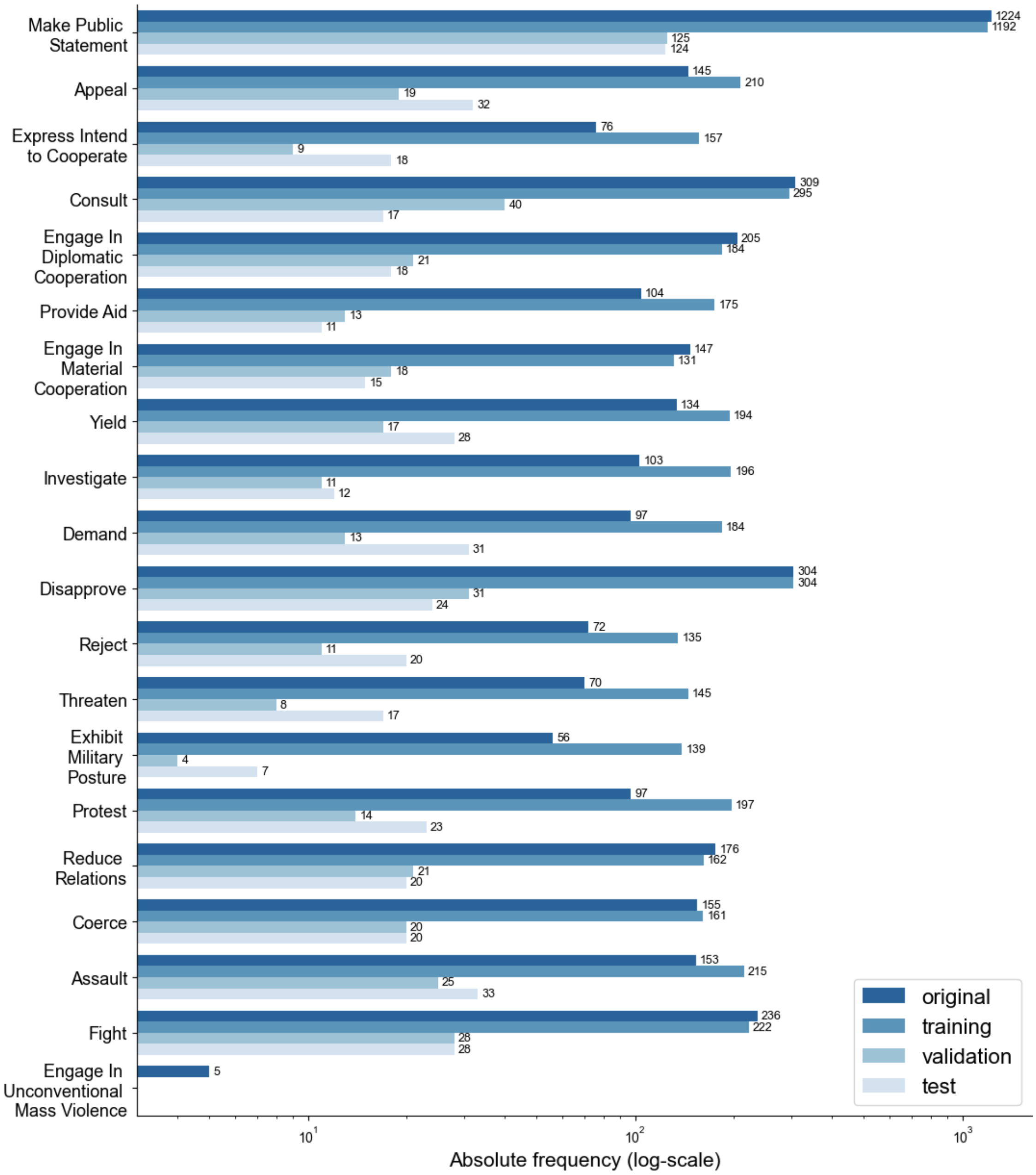
The distribution of relations in the augmented dataset is shown in Figure 7. Although the data augmentation helped to reduce the class imbalance, the data remains heavily skewed towards overrepresented relations, being “Make Public Statement”, “Consult”, and “Disapproved”. The least popular relation is “Engage In Material Cooperation”, which was not augmented.

The overall effect of this data augmentation is evaluated in the ablation study, comparing it to training on the none-augmented data. While three relations were removed in the unsupervised dataset, only one was removed from the annotated dataset. This already indicates weaknesses of syntax-tree-based event coders, as several relations seem to be found less often by the automated event coder than by human annotators.

During the split intro training, validation, and test set, it was accounted both for the count of relations per sample, as well as the class balance between splits to create homogenous datasets. Unlike for the unsupervised dataset, the relative relation frequency between sets is not the same. This is due to the CAMEO example sentences being used 20% for training, 20% for validation, and 60% for the test set to create a test set with texts that have a high similarity to Hu et al. (2022) and Parolin et al. (2022), who used 50% and 100% of the CAMEO example sentences as test set, respectively. The annotated data is used 90% for the train set and 10% for the validation set, inducing this difference between splits, with the validation split being a middle ground between the test set distribution and the training set distribution.

Figure 7

Annotated dataset distribution for CAMEO labels

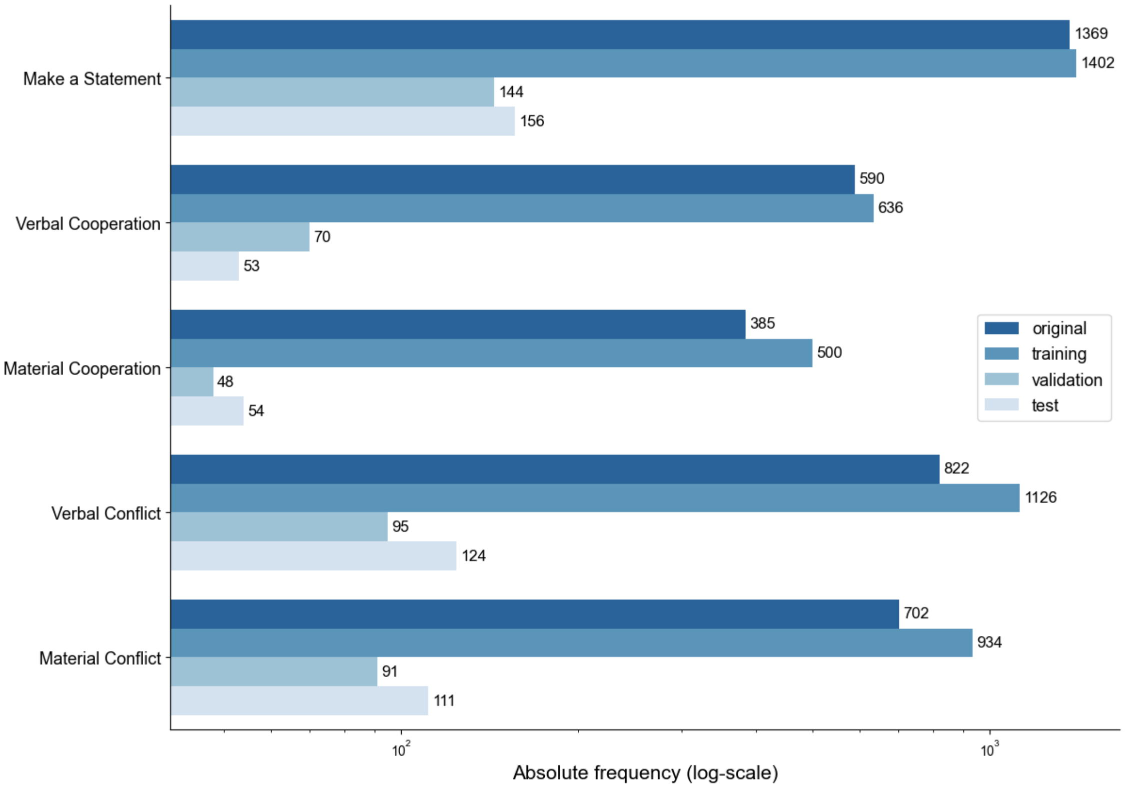


Note. "Original" refers to the original distribution of annotated data including example sentences from the CAMEO codebook. The x-axis was log scaled to help visualise relations with fewer samples. Training data is shown after random sampling and augmentation. Training, validation, and test data are shown for the first seed.

As for the unsupervised data, the annotated data was transferred into Pentacode labels. The distribution of the data on the Pentacode labels is shown in Figure 8. The relations "Material Cooperation" and "Verbal Cooperation" contain fewer samples and the relation "Make a statement" contains more samples in the train set. The former may be critical for validation and evaluation. The data was not augmented again for Pentacode labels to maintain comparability between CAMEO and Pentacode labels.

Figure 8

Annotated dataset distribution for Pentacode labels



Note. “Original” refers to the original distribution of annotated data including example sentences from the CAMEO codebook. The x-axis was log scaled to help visualise relations with fewer samples. Training data is shown after random sampling and augmentation. Training, validation, and test data are shown for the first seed.

4.1.2.4. Inter-annotator Agreement

To evaluate the quality of the annotations, a second annotator was provided with the annotation codebook and annotated 123 sentences. The inter-annotator agreement (IAA) was calculated with the F1 score since commonly used measures for IAA, such as Cohen’s Kappa, pose problems for annotations with spans like the entities in the relation extraction tasks (Hripcsak & Rothschild, 2005; Deleger et al., 2012). This is due to them requiring a true negative value, which is not present in span annotation tasks. True positives (TP) are defined as the same annotations by both annotators, while false negatives (FN) are annotations of the first annotator, which do not agree with the second annotator, and false positives (FP) are annotations of the second annotator, which do not agree with the first annotator. Every relation annotated counts as its own TP, FP, or FN, possibly creating a multitude of relations per sample.

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \qquad (4)$$

The IAA was calculated for NER in a strict and relaxed way, i.e., annotated spans are including each other but are not fully annotated like “German chancellor Olaf Scholz” and “Olaf Scholz”, relation classification, and relation extraction with strict and relaxed annotated entity spans. The results are shown in Table 2. Pentacode labels are more consistent between annotators, indicating that annotators make different decisions when annotating similar labels, which belong to the same Pentacode label.

Table 2

Inter-annotator agreement on the annotated dataset

	CAMEO			Pentacode		
Task	Precision	Recall	F1	Precision	Recall	F1
NER	66.23	71.83	68.92	66.23	71.83	68.92
NER (Relaxed)	75.76	82.16	78.83	75.76	82.16	78.83
Relation Classification	72.73	78.87	75.68	78.79	85.45	81.98
Relation Extraction	59.31	64.32	61.71	62.20	68.64	65.77
Relation Extraction (Relaxed)	66.23	71.83	68.92	71.00	77.00	73.87

Note. Relaxed Evaluation disregards exact matches for named entities, instead counting entities as correct if their spans overlap

While there are no direct guidelines to evaluate the F1 score for IAA, Deleger et al. (2012) have found the score to be close to a token-level version of Cohen’s Kappa, i.e., every token is categorised as annotated or unannotated, and evaluate it with the same scale as Cohen’s Kappa. As such, the values achieved on relation extraction can be interpreted as substantial agreement (0.61 < F1 < 0.8) between the annotators. These values are in line with evaluations from Parolin et al. (2022).

4.2. Modelling Methodology

To answer the research question, two transformer models were trained. These models were selected so they represent two different approaches, a domain-specific and a task-specific model. The domain-specific model chosen is ConfliBERT (Hu et al., 2022) in an Encoder-Decoder-Architecture. This choice was made due to the promising pretraining on several relevant corpora and public availability of the model. However, due to the model being based on BERT, it is necessary to deploy it both as an encoder and decoder to the model to generate sequences, initialising new layers. This will result in ConfliBERT needing more thorough training to train the randomly initialised encoder-decoder multi-head attention layers and linear activation layers used to decide for the most likely next token to generate. For this, it will be tested whether the model achieves better performance after pre-training on the unsupervised dataset.

The task-specific model chosen is REBEL (Cabot & Navigli, 2021). As of current benchmark tests, REBEL is among the leading models on relevant benchmarks while also having experienced a thorough universal pre-training and being publicly available. Further, the known relations of REBEL have a decent overlap with CAMEO relations, due to several political Wikipedia articles it was trained upon. Given that the training of REBEL was conducted using a specific syntax for predicting relations, the annotated labels were formed in a way to inhibit the same syntax.

Both models were trained in Python 3.8.10 using public checkpoints from the Huggingface transformers library 4.18.0 in combination with PyTorch Lightning 1.8.3 Models were trained using one Nvidia Tesla V100 SXM2 32GB GPU provided by the Business School of Economics and Law Berlin.

4.2.1. Training Methodology

Both models will be evaluated both with and without further pre-training on the unsupervised dataset. The pre-training on the unsupervised dataset does not employ masked language modelling or next sentence prediction, as used by BERT and following models, but instead pre-trains directly on the sequence generation task. This choice was made due to both models already having received a thorough basic pre-training on large corpora. After pre-training the training checkpoint with the highest validation performance is used to finetune with the annotated dataset.

The optimiser chosen for training is Adam (Kingma & Ba, 2014) with decoupled weight decay regularisation (Loshchilov & Hutter, 2019). This decision was made due to Adam being the default choice on both models trained while generally converging to lower losses faster, helping to achieve good results on large epoch training time. At the same time, weight decay regularisation has been shown to improve Adam's results in terms of generalisation (Loshchilov & Hutter, 2019).

Several strategies will be employed to reduce overfitting, given the comparatively small datasets used for fine-tuning. Due to the small size and date range of the annotated dataset, several ongoing political events keep reappearing with the same relation and entities throughout different samples. To reduce the co-occurrence of entities reappearing in the same relation type, such as "Putin" and "Russia" often appearing with "War" and "Invasion" or "Zelenskyy" and "Ukraine", entity masking is employed. This replaces random entities that are part of the label at the beginning of every epoch with a randomly initialised mask token. This can be considered an ongoing data augmentation, as new data samples are created every epoch. While it is expected for the model to easily identify the mask token as one of the relevant entities in a triplet, it will stop the model from guessing the relation based on the entity name, which is generally biased behaviour, forcing it to focus on relevant tokens as indicators for relations. As the unsupervised data contains relations from the last four years and the test set of the annotated data contains relations from over a decade ago, this will likely have a positive influence on test performance. The rate at which this masking happens is tuned for every model respectively.

Early stopping during training is used to save computation resources and reduce overfitting. The early stopping is based on the main performance metric, the macro F1 score, not improving over 4 epochs. After stopping, the checkpoint which yielded the best results on the metric is loaded and used for evaluation on the test set.

To further reduce overfitting, an exponential learning rate decay was used, reducing the learning rate *lr* relative to the number of epochs trained already by a decay factor *d*.

$$lr_{\ new} = lr_{initial} * e^{\ -d*epoch} \tag{6}$$

By slowing down the learning rate of the model, the learning of noise within data is reduced and the learning of complex patterns is improved (You et al., 2019). The decay *d* was set based on hyperparameter tuning. Similarly, warmup epochs were used to reduce early overfitting by scaling the learning rate with a factor of 0.1 for the first 2 training epochs. This has been proven to increase learning and convergence speed due to the optimisers converging to bad local optima during the first epochs of training, given the high initial learning rate (Liu et al., 2021).

As some papers argue, it is easier to learn the relation classification task than to identify relevant entities suggesting the use of entity hinting (Han et al., 2020; Giorgi et al., 2022).

Similar to Giorgi et al. (2022), this was implemented in two different ways. First, utilising a pre-trained spaCy NER pipeline, adding special entities tokens in front of and behind the entities inside the text input if they are of relevant types, such as “GPE” (geopolitical entity), “NORP” (nationalities or religious or political groups), “EVENTS”, “FAC” (Buildings, airports, highways, bridges, etc.), “LAW”, “ORG” (organisations), or “PERSON”. While this uses a separate model to perform NER, Giorgi et al. (2022) argue this should not be considered a true pipeline approach to relation extraction, due to the ability of the main model to choose to ignore those extra inputs and gather entities from the text directly. Second, using gold hints, the subject and object entities are extracted directly from the relation label and marked with the special tokens inside the text input. This level of entity hinting results in the model no longer needing to identify entities by itself, but only having to identify which entities interact with one another and how to classify this interaction. This will yield unrealistic results as it reduces a huge portion of the end-to-end relation extraction task. However, this also helps to evaluate the potential of the respective model and is only used for the ablation study.

Since text generation is a classification task, due to the auto-regressive classification of the most likely next token to be generated, the negative log-likelihood (NLL) loss is widely used in end-to-end relation extraction. The loss chosen for training is an adapted version of cross-entropy, the label smoothed NLL loss (Szegedy et al., 2016). This adds a regularisation parameter ε to the predictions, aiming at reducing overconfidence during classification. If the model is very sure, that a label should not be classified, it will no longer predict 0 but instead $\frac{\varepsilon}{K}$, with *K* being the number of classes. Similarly, if the model is very sure that a label should be classified, it will no longer predict 1 but instead 1 – ε. This was found to further help training with noisy labels (Wei et al., 2022), such as the unsupervised dataset used for pre-training, and to produce better calibrated models which directly affects the beam search for sequence generation (Müller et al., 2019).

The label-smoothed NLL loss is calculated by replacing the predicted values *P(k)* for class *k* with the label-smoothed prediction $p^{LS,\varepsilon}$ within the regular NLL function with *Q(k)* being 0 or 1 depending on whether the prediction is present in the label or not.

$$p^{LS,\varepsilon} = (1-\varepsilon) * \ y + \frac{\varepsilon}{K} \tag{7}$$

$$NLL(q,p) = \ -\textstyle\sum_{k}^{K} Q(k) * \log\left(P(k)\right) \tag{8}$$

$$NLL^{LS}(q,p^{LS,\varepsilon}) = \ -\textstyle\sum_{k}^{K} Q(k) * \log\left(p^{LS,\varepsilon}\right) \tag{9}$$

4.2.2. Tuning Methodology

Several hyperparameters for the model were tuned to maximise performance. The baseline values for these hyperparameters were chosen based on the value chosen during the training of the original model. Random search has been proven to find good combinations in hyperparameters more efficiently than other search algorithms such as grid search (e.g., Bergstra et al., 2012). Trying random combinations gives several advantages, such as less none-promising combinations being evaluated and allowing meaningful inference on any number of combinations evaluated, instead of only allowing inference after all combinations were assessed. As such, random search was employed to tune several relevant hyperparameters.

Relevant hyperparameters that were tuned with random search are the batch size, learning rate, learning rate decay, ε for the label smoothed loss, weight decay, and masking rate. Apart from the baseline for the parameters, which was chosen based on the hyperparameters the original model was trained on and arbitrarily set to 0.1 for ε as used in Vaswani et al. (2017), four further values were selected. For the batch size, the values 32 and 64 were proposed, as 32 was used to train both BERT and REBEL, while 64 is the maximum batch size trainable with the available hardware. The values that were eligible for each hyperparameter are shown in Table A3. The proposed values allow for 3,750 different combinations of parameter values, however, with random search, it is possible to infer parameter relevance and better values from few runs. Performance was evaluated based on ten random trainings for ten epochs for each model. Some combinations possibly yield overall higher performance by learning slower with a longer training and require more than ten epochs, however, the training time of ten epochs was chosen to balance training time and model performance. Technically, this will favour parameters that allow for quick learning with little regularisation, such as a high learning rate in combination with a low batch size, a small learning rate decay and a small weight decay. This bias is acknowledged during the choice of final hyperparameter values.

Tuning was repeated for annotated data on new values, reducing the batch size to either 16 or 32, as fine-tuning often profits from lower batch sizes and lower learning rates (e.g., Devlin et al., 2018; Liu et al., 2019). Accordingly, the values for batch sizes were adjusted to not include the highest values and instead propose some intermediate values which were not used for the unsupervised dataset. Since the annotated dataset is smaller, the tuning was done for 15 epochs instead.

The tuning used the first seed of either dataset with CAMEO labels. Each random combination of hyperparameters took approximately 3.75 hours for REBEL and 4.5 hours for ConfliBERT for the unsupervised dataset and 2.25 hours for REBEL and 3.5 hours for ConfliBERT on the annotated dataset. The model performances with the respective hyperparameters are described in Section 5.2.1. for REBEL and Section 5.2.2. for ConfliBERT.

4.2.3. Evaluation Methodology

All evaluation is averaged across three different seeds of data to reduce bias induced by data sampling. As several strategies, such as entity masking and entity hinting are tested during training, an ablation study will evaluate the outcomes with and without these strategies. The model performance is evaluated using the metrics of micro and macro precision, recall and F1 Score. TP, FP, and FN are calculated for every relation in every sample instead of full labels, meaning one sample can create a multitude of TP, FP, and FN. This allows to also evaluate partial success of the model.

The micro metrics only look at the overall number of TP, FP, and FN, effectively disregarding class sizes. In the case of transformers, the size of data overall, as well as the size of classes, is relevant to learn connections between inputs and class labels in a robust manner. If classes are scarce in the training dataset, it is likely that the model cannot learn these connections and will fail to predict the class on unseen data used for evaluation. Intuitively, for these cases, the micro metrics will return higher performance, as underrepresented classes are also less influencing the overall count of TP, FP, and FN.

$$Precision_{Micro} = \frac{\sum TP}{\sum TP + \sum FP} \tag{10}$$

$$Recall_{Micro} = \frac{\sum TP}{\sum TP + \sum FN} \tag{11}$$

$$F1_{Micro} = 2 * \frac{Precision_{Micro} * Recall_{Micro}}{Precision_{Micro} + Recall_{Micro}} \tag{12}$$

The macro metrics average the performance of all classes. This means that underrepresented classes in the dataset will have the same influence on performance as overrepresented classes, giving more weight in the evaluation to samples from underrepresented classes. As we can expect underrepresented classes to perform worse, outliers in labels per class will also heavily reduce macro metrics. With several classes in both datasets containing significantly fewer samples, it is expected that macro scores will be lower than micro scores. Still, macro scores will describe the overall robustness of the models better than micro scores.

$$Precision_{Macro} = \frac{1}{n} \sum_x Precision_x \tag{13}$$

$$Recall_{Macro} = \frac{1}{n} \sum_x Recall_x \tag{14}$$

$$F1_{Macro} = \frac{1}{n}\sum_{x} F1_{x} \quad (15)$$

This evaluation happens both for the task of end-to-end relation extraction, meaning the model found the correct subject, object and relation, and on the task of relation classification, meaning the model found the right relations in a sample. Importantly, this also means that if a relation appears multiple times in the sample, the model needs to find it multiple times. The task of NER is not evaluated separately, as the data does not contain all entities in the input sentence, but only those which are relevant for relations. As such, the evaluation would be flawed and would not represent a realistic case in which a model is trained for NER.

5. Artefact Description

This chapter describes generated artefacts, namely the unsupervised dataset generation and the transformer models based on REBEL (Cabot & Navigli, 2021) and ConfliBERT (Hu et al., 2022). The description for the unsupervised data generation focuses on the algorithm and technologies used, while the description of the transformer models focuses on the model description and training details. All artefacts were created using Python 3.8.10.

5.1. Unsupervised Dataset Generation

The first artefact that was created has the goal of generating large amounts of training data, utilising the steady flow of news articles being published. Before generating a dataset, the articles were pre-processed by performing coreference resolution using SpanBERT (Joshi et al., 2020) with the crosslingual-coreference 0.2.9 package and splitting the articles into separate sentences with the sentence splitter from CoreNLP (Manning et al., 2014).

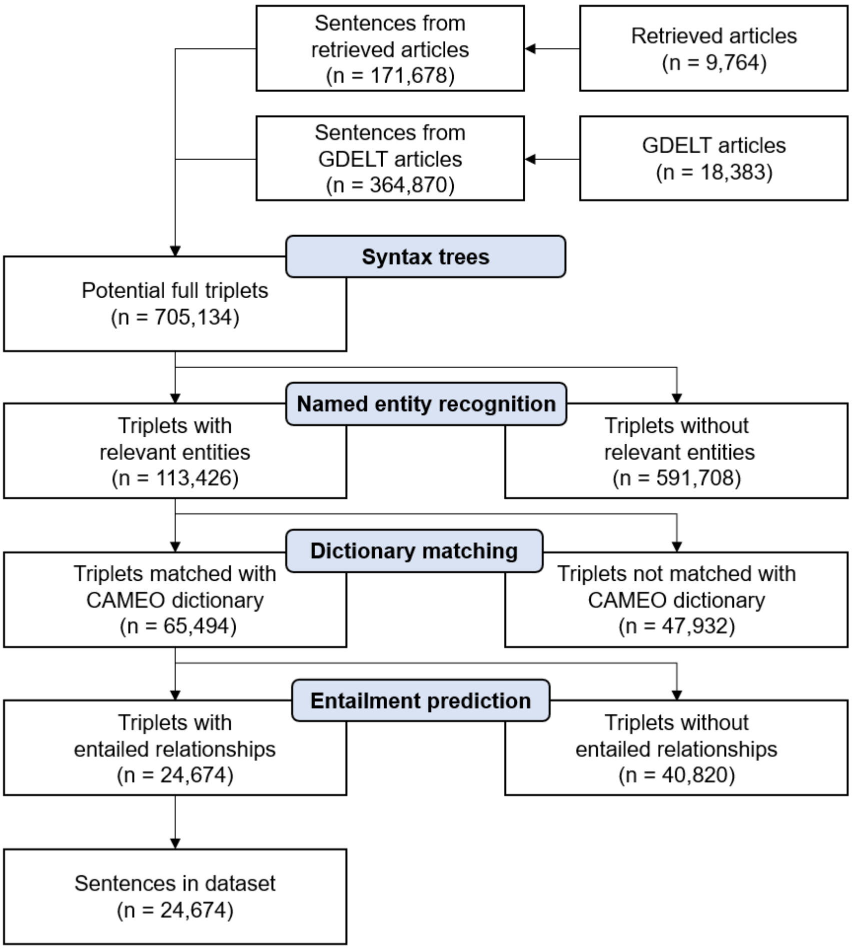
The algorithm to create the unsupervised dataset is built upon the basics of PETRARCH2 (Norris, 2016) but was adjusted both to incorporate the necessary structure for the sequence-to-sequence relation extraction task and to detach the algorithm from the outdated actor dictionaries of the CAMEO ontology. The core of the algorithm is still built on dependency trees, however, utilising spaCy 3.4.1 instead. This change was mostly made as further steps in the algorithm also utilise the spaCy pipelines. The spaCy pipeline en_core_web_trf 3.4.1 is built upon the transformer RoBERTa (Liu et al., 2019) and can be used for a multitude of tasks, including lemmatisation, dependency trees, and NER. This pipeline was chosen, as it is the highest-performing pipeline offered by spaCy. Overall, this incorporates large-scale deep neural networks with high flexibility as a substitution for the actor dictionaries, aiming at identifying more subject-object pairs overall.

Given an input sentence, a dependency tree is created and verbs in the sentence are identified. Identifying verbs is crucial, as they are indicators for possible relevant relations within the sentence while also having dependency relationships to the relevant subject and object in the sentence. When a verb is found, the algorithm looks at dependencies representing a negation, a clausal complement, or an open clausal complement. In the case of a negation of the verb, the verb becomes irrelevant, as the real relationship cannot be determined. In the case of a clausal or open clausal complement, all dependencies of the complement will further be treated as dependencies of the original verb. An example of this would be “France wants to start talks with China”, where a clausal complement exists between “wants” and “start”, while “France” is the subject to “wants” and “China” is the object to “start”. Next, all noun chunks, representing a noun and its descriptive words, such as “the german ambassador”, are identified and filtered with NER. Utilising spaCy’s NER pipeline, entities are only used in triplets if they are part of a relevant entity type, being predefined as “GPE”, “NORP”, “EVENTS”, “FAC”, “LAW”, “ORG”, or “PERSON”. The dependencies of relevant noun chunks are iterated to identify which noun chunks are related to the verb. If a noun chunk is related to the verb, the noun chunk, the dependency between the noun chunk and verb, the lemmatised related verb and the index of the related verb are appended to a list. The index of the verb is needed to identify triplets in case of duplicated verbs in the sentence. Within the list, all verbs that do not appear at least twice, as one appearance is needed for the subject and one for the object, are discarded. If a remaining verb has both a subject and an object, the verb is matched with the CAMEO verb dictionary and if the verb has a code in the dictionary, the identified triplet is recorded.

While this creates many triplets, this unsupervised data creation contains a large amount of noisy data. Similar to Cabot and Navigli (2021), a pre-trained entailment classification transformer based on BART (Lewis et al., 2019) was used to predict whether the obtained triplet is entailed by the text sentence. To increase the performance of this zero-shot model, the entailment input should be masked as a sentence. As such, a mask in the form of “{subj} does {rel} towards {obj}.” is used, as this generally forms logical and grammatically correct statements of the triplets, such as “Germany does Engage in Diplomatic Cooperation towards France”. The cut-off to classify the triplet as entailed was set to 0.65 after manually evaluating the remaining triplets. Figure 9 shows a flowchart of the dataset size at each step. Starting at 536,548 sentences, the result is drastically reduced to 24,674 sentences with relevant relations, some of which are still noisy. This conversion rate of articles to triplets is lower than the rate for PETRARCH as reported by Parolin et al. (2020), likely because of the entailment classification step. Arguably, this scaling may be sufficient, given the large number of news articles being published daily and the purpose of the dataset. While collecting more newspaper articles may help to achieve more data, they will not necessarily introduce more valuable data, due to the high overlap in topics in articles across newspapers.

Figure 9

Flowchart of the size of the unsupervised dataset at each step of generation


Sentences from retrieved articles (n = 171,678)
Retrieved articles (n = 9,764)
Sentences from GDELT articles (n = 364,870)
GDELT articles (n = 18,383)
Syntax trees
Potential full triplets (n = 705,134)
Named entity recognition
Triplets with relevant entities (n = 113,426)
Triplets without relevant entities (n = 591,708)
Dictionary matching
Triplets matched with CAMEO dictionary (n = 65,494)
Triplets not matched with CAMEO dictionary (n = 47,932)
Entailment prediction
Triplets with entailed relationships (n = 24,674)
Triplets without entailed relationships (n = 40,820)
Sentences in dataset (n = 24,674)

The resulting dataset, while generating large amounts of data in a short time, suffers several weaknesses which differentiate it from the annotated dataset. First, the algorithm is not able to find reciprocal relationships which are often present in several relations, such as “Consult”. An example of this would be “The U.S. and China met for negotiations on trade sanctions”. In this example, both relations “<triplet> The U.S. <subj> China <obj> Consult” and “<triplet> China <subj> The U.S. <obj> Consult” need to be identified. Since the algorithm is based on a subject-verb-object structure, it expects that the subject is always the source actor, and the object is always the target actor of the interaction.

Second, the algorithm is not able to find none-verb-indicated relationships. Interactions are not necessarily indicated by a verb but can instead be inferred from context or noun chunks such as “Major progress in Spain-Italy negotiations”. The dependency tree is not able to extract this relation between “Spain” and “Italy”.

Third, the algorithm will mostly only use a subject with one object, while in this domain of political news articles many entities mentioned in a sentence often relate to the same target or source. Relations such as “Putin to reassert roles in talks with Armenia, Azerbaijan” will be reduced to Putin and Armenia, missing the relation to Azerbaijan. While this does not induce FP in the relations, it will reduce the number of relations found overall.

Fourth, many relations are skipped due to verb patterns being unknown to the CAMEO ontology, which would need to be added manually to the ontology. Several types of relations are completely missed by the algorithm, inducing an imbalance in the created dataset. Parolin et al. (2020) suggest classifying unknown verb patterns into the ontology using deep learning models to reduce this problem.

5.2. Transformers

Following, the two transformer models that were trained are described in more detail regarding the model they are based on and the hyperparameter values chosen for the respective model are given and reasoned for.

5.2.1. REBEL

REBEL (Cabot & Navigli, 2021) was chosen due to it fitting the task of sequence-to-sequence relation extraction best. REBEL is a transformer based on $\text{BART}_{\text{LARGE}}$ (Lewis et al., 2019). BART is an encoder-decoder transformer that is effective in sequence-to-sequence tasks by incorporating more sophisticated pre-training tasks (Lewis et al., 2019). BART uses the transformer architecture proposed by Vaswani et al. (2017), with modified activation functions, using GeLU instead of ReLU as used in other large transformer models as well (e.g., Radford et al., 2018; Devlin et al., 2018). The large model version uses 12 encoder and decoder layers. The REBEL model used in the artefact totals 406 million parameters, none of which were frozen during training. REBEL uses an encoder layer size of 1024, turning words into embeddings of size 1024.

Since the model predicts the most likely following token during training, beam search was used to get more robust results. Beam search, instead of always just picking the word with the highest probability, picks *N* word with the highest probabilities and calculates the full sequence likelihood of the so far generated output after every generated token, continuing with the *N* most likely sequence for the next token. For the training of REBEL, *N* was set to 3 to balance performance and training time, as a higher value for *N* leads to strongly increased training time.

Table 3 shows the combinations tried during tuning. Runs with higher learning rate have achieved generally better outcomes. Only one of the three top-performing models used high regularisation in form of masking, weight decay, ε, and learning rate decay. While the other two models show higher performance, it is believed that this model will also generalise better. This is expected to have a positive effect in terms of robustness when training with more epochs, training on other seeds which contain different data, and during fine-tuning. As such, the final REBEL models for the unsupervised data were trained for a maximum of 25 epochs with a batch size of 32, a learning rate of 0.00005, a learning rate decay of 0.2, ε of 0.2, weight decay of 0.1, and a masking rate of 0.1.

Table 3

Summary of hyperparameter tuning for REBEL on unsupervised data

Batch size	Learning rate	Learning rate decay	Weight decay	ε	Masking rate	Macro F1	Micro F1
64	**0.000075**	**0.05**	**0.0075**	**0.10**	**0.0**	**97.51**	**98.54**
64	0.000125	0.10	0.0150	0.15	0.0	96.60	97.76
64	0.000025	0.15	0.0150	0.20	0.2	89.92	93.74
32	**0.000050**	**0.20**	**0.0150**	**0.20**	**0.1**	**97.33**	**98.22**†
64	0.000025	0.30	0.0075	0.15	0.2	83.24	89.72
32	0.000100	0.30	0.0150	0.20	0.2	97.07	98.12
64	0.000100	0.05	0.0100	0.15	0.0	95.64	97.51
32	0.000075	0.05	0.0100	0.05	0.2	95.80	97.38
64	0.000025	0.15	0.0050	0.05	0.0	94.63	96.96
64	**0.000125**	**0.15**	**0.0050**	**0.00**	**0.1**	**98.21**	**98.76**

Note. Performance reported on validation set. The reported result is the maximum macro F1 score reached by the model. Bold indicates highest performance. † Indicates chosen model.

The hyperparameters were tested again with lowered values for learning rate for fine-tuning. This change was made, as it is expected that the pre-training gives a thorough understanding of the language, while the fine-tuning focuses on nuances of the data that can be achieved by taking smaller steps in learning. Since learning rates were overall lower, the models were trained longer. The new maximum training duration was set to 15 epochs, while still using early stopping after 4 epochs in case the model stops improving on the macro F1 score.

Table 4 shows the tested hyperparameter combinations during tuning for the annotated data. Lowering learning rates resulted in higher results when used in combination with a low learning rate decay. Very low learning rates, however, seem to prohibit the model from learning sufficiently and yielded lower performance overall. The values chosen for fine-tuning REBEL are a batch size of 16, learning rate of 0.000025, learning rate decay of 0.10, weight decay of 0.0075, ε of 0.2, and a masking rate of 0.1.

Table 4

Summary of hyperparameter tuning for REBEL on annotated data

Batch size	Learning rate	Learning rate decay	Weight decay	ε	Masking rate	Macro F1	Micro F1
32	0.000025	0.20	0.0150	0.05	0.0	36.51	42.69
32	0.000025	0.20	0.0025	0.15	0.2	36.52	41.56
32	0.000050	0.20	0.0025	0.15	0.2	39.89	43.18
16	0.000025	0.15	0.0075	0.20	0.2	40.28	43.95
16	0.000035	0.05	0.0050	0.05	0.0	39.26	45.15
16	**0.000050**	**0.10**	**0.0075**	**0.10**	**0.1**	**41.04**	**45.80**
16	0.000005	0.05	0.0025	0.20	0.2	36.36	41.66
16	**0.000025**	**0.10**	**0.0075**	**0.20**	**0.1**	**41.22**	**45.12**†
16	0.000010	0.05	0.0100	0.10	0.2	38.42	43.79
16	**0.000025**	**0.05**	**0.0025**	**0.10**	**0.1**	**41.52**	**44.37**

Note. Performance reported on validation set. The reported result is the maximum Macro F1 score reached by the model. Bold indicates highest performance. † Indicates chosen model.

5.2.2. ConfliBERT$_{\text{ENC-DEC}}$

ConfliBERT (Hu et al., 2022) was chosen as it fits the domain of the data particularly well. ConfliBERT is a model based on BERT$_{\text{base}}$ (Devlin et al., 2018), pre-trained of several corpora of conflict-related news data using masked language modelling. Since BERT is not using an auto-regressive decoder in its architecture, it is not able to perform the sequence-to-sequence relation extraction task. Hence, the architecture of the model trained was adapted towards the encoder-decoder framework suggested by Rothe et al. (2020). While the encoder used has to be ConfliBERT, the decoder chosen could also be a decoder transformer, such as GPT, or an encoder-decoder transformer, such as BART or REBEL. However, using the same model as encoder and decoder has proven to increase performance (Rothe et al., 2020). As such, ConfliBERT was used as encoder and decoder, further sharing the parameters between encoder and decoder, significantly reducing the number of parameters that need to be trained and thus also the training time while maintaining similar results (Rothe et al., 2020). The encoder-decoder version of ConfliBERT will further be referred to as ConfliBERT$_{\text{ENC-DEC}}$.

ConfliBERT$_{\text{ENC-DEC}}$ totals 161 million parameters, making it significantly smaller than REBEL. This is partly due to shared parameters between the encoder and decoder, reducing the number of parameters from 246 million. The model uses 12 transformer layers for the encoder and 12 transformer layers for the decoder. The encoder layers have a size of 768 resulting in smaller embeddings than those created by REBEL. Beam search was employed to find the most likely sequences based on three beams.

Hyperparameter combinations tested for pre-training are shown in Table 5. It seems that the right combination of hyperparameters is more important for ConfliBERT$_{\text{ENC-DEC}}$ than for REBEL, given the higher variation in performance. The best outcome was achieved using a learning rate of 0.00005, a learning rate decay of 0.05, weight decay of 0.0075, ε of 0.1, and a masking rate of 0.2. The two best combinations only differ in weight decay and the masking rate value, indicating that these regularisations can help to find more generalisable results, achieving higher validation results during tuning and possibly higher test scores on the full training.

Although F1 values reported for the hyperparameter tuning for ConfliBERT$_{\text{ENC-DEC}}$ are lower than those for REBEL, this may be due to ConfliBERT$_{\text{ENC-DEC}}$ needing longer training due to the newly initialised layers and the new task. As the model seems to learn slower than the REBEL model, the pre-train models were trained for a maximum of 30 epochs, still using early stopping after four epochs of not improving the validation macro F1 score.

Table 5

Summary of hyperparameter tuning for ConfliBERT$_{ENC\text{-}DEC}$ on unsupervised data

Batch size	Learning rate	Learning rate decay	Weight decay	ε	Masking rate	Macro F1	Micro F1
32	0.000125	0.20	0.0025	0.15	0.0	84.91	86.04
64	0.000125	0.30	0.0050	0.10	0.1	84.71	88.01
32	0.000075	0.05	0.0100	0.20	0.1	81.88	85.59
32	0.000050	0.30	0.0025	0.00	0.2	62.96	71.05
64	0.000125	0.05	0.0100	0.20	0.0	77.06	78.66
32	**0.000050**	**0.05**	**0.0050**	**0.10**	**0.0**	**88.37**	**89.96**
64	0.000100	0.10	0.0050	0.00	0.2	73.85	78.22
32	**0.000050**	**0.05**	**0.0075**	**0.10**	**0.2**	**88.87**	**91.17**†
64	0.000050	0.05	0.0050	0.00	0.0	84.81	86.53
64	**0.000075**	**0.10**	**0.0050**	**0.15**	**0.1**	**87.49**	**89.18**

Note. Performance reported on validation set. The reported result is the maximum Macro F1 score reached by the model. Bold indicates highest performance. † Indicates chosen model.

After pre-training the model, hyperparameters were tuned again for fine-tuning. Similar to the fine-tuning hyperparameters for REBEL, the values for learning rate were overall lowered, however, leaving the range overall higher than for REBEL as ConfliBERT$_{ENC\text{-}DEC}$ seems to profit from higher learning rates. Apart from high values for learning rate, low values for learning rate decay, and ε seem to result in overall higher outcomes. The influence of the masking rate on the model performance can best be evaluated as the difference between the second and third best model. However, the slightly increased model performance may also be attributed to the lower learning rate decay. The best result was found using multiple regularisation parameters, with a high learning rate of 0.000075, masking rate of 20%, and weight decay of 0.01, while learning rate decay and ε were chosen rather low with 0.1 and 0.05, respectively. Table 6 shows the results of the hyperparameter tuning on the pre-trained model using the annotated dataset.

Table 6

Summary of hyperparameter tuning for ConfliBERT$_{ENC\text{-}DEC}$ on annotated data

Batch size	Learning rate	Learning rate decay	Weight decay	ε	Masking rate	Macro F1	Micro F1
16	**0.000075**	**0.10**	**0.0100**	**0.05**	**0.2**	**30.72**	**31.83†**
32	**0.000065**	**0.10**	**0.0100**	**0.00**	**0.1**	**29.28**	**27.76**
32	0.000050	0.10	0.0150	0.20	0.0	27.27	27.66
32	**0.000065**	**0.15**	**0.0100**	**0.00**	**0.0**	**28.21**	**27.99**
32	0.000050	0.20	0.0150	0.10	0.2	22.61	23.53
32	0.000025	0.10	0.0100	0.00	0.0	21.20	22.39
16	0.000035	0.05	0.0025	0.05	0.0	27.35	28.84
16	0.000025	0.15	0.0075	0.15	0.2	21.37	23.37
16	0.000050	0.20	0.0100	0.20	0.1	24.74	26.25
16	0.000050	0.15	0.0100	0.05	0.2	28.28	29.50

Note. Performance reported on validation set. The reported result is the maximum Macro F1 score reached by the model. Bold indicates highest performance. † Indicates chosen model.

6. Evaluation

Both models are first trained and evaluated in pre-training on the CAMEO labels and the Pentacode labels. Following, the fine-tuning on both labels is done for both models using the pre-trained checkpoints as basis. Finally, an ablation study which again fine-tunes both models from the pre-trained checkpoint helps to quantify the performance influence of further strategies used during training, being the pre-training, data augmentation, entity masking and entity hinting. All evaluation is done with micro and macro averaged precision, recall and F1 score for the tasks of relation extraction and relation classification, averaged across three seeds.

6.1. Performance on the Unsupervised Dataset

REBEL achieves the highest scores after 19 epochs, while ConfliBERT$_{ENC\text{-}DEC}$ achieved the best results after 29 epochs. The longer training of ConfliBERT$_{ENC\text{-}DEC}$ is expected, as many layers of the model were randomly initialised. It is expected that this effect is less prevalent in fine-tuning. The training of both models took on average 25 minutes per epoch.

Both models show excellent results on the task of relation extraction on the unsupervised dataset. All macro F1 scores are higher for Pentacode labels than for CAMEO labels. This is likely caused by the remaining class imbalance in the dataset. At the same time, the micro and macro F1 scores on Pentacode labels are overall more balanced than on CAMEO labels, indicating that this remaining class imbalance is less important for relation extraction on Pentacode labels than on CAMEO labels. Further, all models report both higher micro and macro precision than recall.

Overall, REBEL outperforms ConfliBERT$_{\text{ENC-DEC}}$ on both types of labels. This if further accentuated by REBEL achieving higher macro F1 scores than the micro F1 scores achieved by ConfliBERT$_{\text{ENC-DEC}}$. The full results for relation extraction are reported in Table 7.

Table 7

Model performance on unsupervised data for relation extraction

	Micro averaged			Macro averaged		
Model	Precision	Recall	F1	Precision	Recall	F1
CAMEO						
REBEL	99.47	98.99	99.23	99.16	98.49	98.81
	± 0.11	± 0.35	± 0.19	± 0.05	± 0.17	± 0.09
ConfliBERT$_{\text{ENC-DEC}}$	98.79	98.62	98.70	98.54	97.78	98.14
	± 0.31	± 0.31	± 0.31	± 0.36	± 0.62	± 0.49
Pentacode						
REBEL	99.49	98.75	99.12	99.45	98.73	99.09
	± 0.13	± 0.51	± 0.30	± 0.08	± 0.36	± 0.21
ConfliBERT$_{\text{ENC-DEC}}$	98.94	98.58	98.76	98.87	98.29	98.58
	± 0.10	± 0.22	± 0.14	± 0.16	± 0.26	± 0.19

Note. Values are averaged across three seeds. ± Indicates standard deviation.

Table 8 shows the relation classification results for both models on both CAMEO and Pentacode labels. With one exception, being REBEL on Pentacode labels, all models report higher micro F1 scores than macro F1 scores. Again, the models mostly report higher precision than recall except for ConfliBERT$_{\text{ENC-DEC}}$ reporting higher micro recall on CAMEO labels. Overall, both models perform very even on relation classification on both datasets and types of labels, with the highest difference in micro and macro F1 score being 0.18. The similar performance of both models is particularly interesting when recalling that REBEL outperforms ConfliBERT$_{\text{ENC-DEC}}$ on both types of labels at relation extraction. As all errors for relation extraction are either caused by mistakes in the subtask of NER or the subtask of relation classification, the relation extraction scores in context to the relation classification help to reason where errors exist. Although the task of NER cannot be evaluated separately, extracting relevant entities is a bigger bottleneck for ConfliBERT$_{\text{ENC-DEC}}$ as the difference between relation classification to be relation extraction is larger. The differences in scores between the tasks are lower for REBEL, explaining its overall superiority on this dataset.

Table 8

Model performance on unsupervised data for relation classification

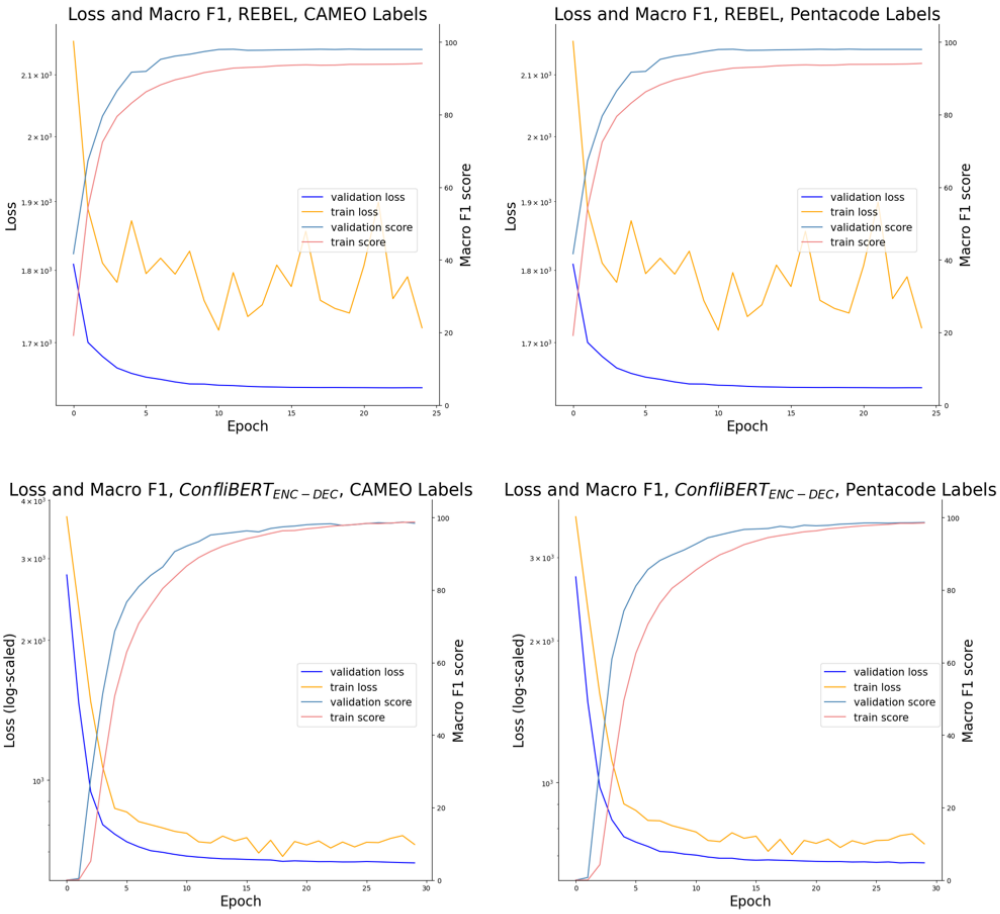
	Micro averaged			Macro averaged		
Model	Precision	Recall	F1	Precision	Recall	F1
CAMEO						
REBEL	99.68	99.19	99.43	99.47	98.80	99.12
	± 0.06	± 0.40	± 0.22	± 0.18	± 0.36	± 0.27
ConfliBERT$_{\text{ENC-DEC}}$	99.37	99.42	99.40	99.29	98.63	98.96
	± 0.08	± 0.04	± 0.06	± 0.08	± 0.28	± 0.17
Pentacode						
REBEL	99.69	98.89	99.29	99.70	98.91	99.30
	± 0.08	± 0.56	± 0.32	± 0.11	± 0.36	± 0.27
ConfliBERT$_{\text{ENC-DEC}}$	99.48	99.39	99.44	99.49	99.16	99.32
	± 0.19	± 0.18	± 0.17	± 0.20	± 0.22	± 0.20

Note. Values are averaged across three seeds. ± Indicates standard deviation

Figure 10 shows the loss and macro F1 during pre-training. The validation score increases faster than the training score for all models on all labels. Similarly, the validation loss is consistently lower than the training loss, indicating that the models are more confident in their prediction on the validation set. As both models exhibit this characteristic, it is overall more likely that this is caused by the data rather than the model architecture. Data augmentation can cause this effect in two ways. First, augmenting the data with entity and verb swapping increases the likelihood of the training set including similar data to the validation set, such as verb patterns and entities, while many verb patterns and entities inside the training set have not yet been learned. Second, the augmentation possibly increases the noise in the training set, creating samples that are illogical or confusing to the models.

Figure 10

Training and validation loss and performance on unsupervised data per epoch



Note. For plots of ConfliBERT$_{\text{ENC-DEC}}$, the y-axis of the loss (primary y-axis) is log-scaled to enhance visibility of differences in loss during later epochs. Results visualised are from training on seed 0.

The validation loss shows only slight signs of overfitting which are signalled by the validation loss increasing while the training loss keeps decreasing. This shows the effectiveness of early stopping in combination with the learning rate decay during training, allowing to learn more complex patterns in later epochs without overfitting.

6.2. Performance on the Annotated Dataset

During the evaluation of the fine-tuning performance, the pre-trained models were used as the baseline model. Table 9 shows the relation extraction results after fine-tuning the pre-trained models. The baseline models show very low scores on the new dataset compared to prior results on the unsupervised dataset. This is partly due to two CAMEO labels being trained during fine-tuning which were removed during pre-training, causing a large penalty, particularly on the macro F1 scores. The overall low performance of both models indicates that only little overlap exists between the annotated dataset and the information learned from the unsupervised. This opens questions about whether the pre-training on unsupervised data helps model performance and will be further examined in the ablation study.

Both the fine-tuned REBEL and fine-tuned ConfliBERT$_{\text{ENC-DEC}}$ show higher relation extraction results on Pentacode labels. However, the difference in performance is higher for ConfliBERT$_{\text{ENC-DEC}}$, indicating that the model struggles to differentiate the CAMEO classes more than REBEL. Overall, ConfliBERT$_{\text{ENC-DEC}}$ performs worse than REBEL on every metric on either type of labels. This margin between the models is unexpected, as both models are pre-trained on the same data and showed similar scores during pre-training. One reason for this difference is possibly the difference in model sizes, hindering the smaller model to account for the high complexity in the limited number of layers. As overall scores are drastically lower than on the unsupervised dataset, this further argues for increased complexity in the annotated dataset, which affects ConfliBERT$_{\text{ENC-DEC}}$ stronger than REBEL.

Interestingly, REBEL achieves higher micro F1 scores on CAMEO labels than on Pentacode labels while macro F1 scores are higher on Pentacode labels. The high standard deviations of most metrics further indicate a large difference between the training data, as all seeds use the same test data. This argues for data augmentation, being the main cause of differences inside the training data, having a high influence on the resulting performance. The ablation study will further indicate whether this influence is positive or negative.

Table 9

Model performance on annotated data for relation extraction

	Micro averaged			Macro averaged		
Model	Precision	Recall	F1	Precision	Recall	F1
CAMEO						
REBEL^{BL}	18.14	8.31	11.40	16.71	7.39	10.25
ConfliBERT$_{ENC-DEC}$^{BL}	8.85	4.02	5.52	11.35	4.53	6.48
REBEL	51.95 ± 1.87	44.29 ± 1.57	47.79 ± 1.35	49.47 ± 2.52	42.17 ± 1.10	44.17 ± 1.16
ConfliBERT$_{ENC-DEC}$	29.20 ± 1.21	24.56 ± 1.19	26.65 ± 0.50	28.40 ± 3.19	23.02 ± 1.15	24.56 ± 1.63
Pentacode						
REBEL^{BL}	25.00	11.79	16.02	24.78	12.54	16.65
ConfliBERT$_{ENC-DEC}$^{BL}	14.62	6.67	9.16	14.91	6.80	9.34
REBEL	52.83 ± 1.34	48.55 ± 2.30	50.58 ± 1.69	51.99 ± 0.59	47.07 ± 2.39	49.13 ± 1.40
ConfliBERT$_{ENC-DEC}$	36.80 ± 1.17	29.52 ± 1.50	32.75 ± 1.32	35.57 ± 0.61	28.45 ± 1.01	31.50 ± 0.75

Note. Values are averaged across three seeds. Baseline values are reported for models trained on first seed. ± Indicates standard deviation. ^{BL} indicates baseline.

Table 10 shows the results of relation classification on the annotated dataset. Like previous results, all models achieve higher micro F1 scores than macro F1 scores. REBEL still finds stronger performance than ConfliBERT$_{ENC-DEC}$, however, the margin between the models, particularly on Pentacode labels is reduced when compared to relation extraction. As for the results on the unsupervised dataset, ConfliBERT$_{ENC-DEC}$ struggles less with relation classification than with full relation extraction, further indicating problems in the NER subtask.

Table 10

Model performance on annotated data for relation classification

	Micro averaged			Macro averaged		
Model	Precision	Recall	F1	Precision	Recall	F1
CAMEO						
REBEL^{BL}	42.48	19.47	26.70	42.39	17.97	25.24
ConfliBERT$_{ENC\text{-}DEC}^{BL}$	42.48	19.28	26.52	50.76	19.17	27.83
REBEL	70.25 ± 1.82	59.97 ± 1.50	64.67 ± 0.78	68.44 ± 3.49	59.13 ± 0.88	61.66 ± 1.08
ConfliBERT$_{ENC\text{-}DEC}$	61.16 ± 4.38	52.41 ± 0.87	56.34 ± 1.43	60.51 ± 5.12	49.37 ± 1.19	52.26 ± 1.70
Pentacode						
REBEL^{BL}	58.48	27.58	37.48	56.76	26.80	36.41
ConfliBERT$_{ENC\text{-}DEC}^{BL}$	56.11	25.83	35.38	56.16	26.70	36.19
REBEL	72.76 ± 1.42	67.55 ± 1.84	70.02 ± 0.66	71.67 ± 1.53	65.36 ± 2.58	68.03 ± 1.27
ConfliBERT$_{ENC\text{-}DEC}$	71.10 ± 1.86	60.04 ± 0.87	65.08 ± 0.27	68.75 ± 1.72	57.24 ± 0.44	62.20 ± 0.56

Note. Values are averaged across three seeds. Baseline models are pre-trained models without fine-tuning. ± Indicates standard deviation. ^{BL} indicates baseline.

6.3. Ablation

The ablation was conducted for both models to see how much models can improve using entity hints, or decrease performance when reducing training strategies, such as data augmentation, pre-training and entity masking.

Table 11 shows the ablation results for REBEL. Both the pre-training on the unsupervised dataset and entity masking during training seem to have no positive effect on REBEL when trained on CAMEO labels. Interestingly, removing the pretraining results in a reduction of the relation classification scores, while the relation extraction scores rise. Further, removing data augmentation resulted in a large decrease in performance on both tasks. While the use of spaCy entity hints resulted in a decrease in performance, the use of gold entity hints resulted in large gains at relation extraction and small improvements at relation classification.

Table 11

Performance of REBEL when removing or adding training strategies

	Relation Classification			Relation Extraction			
	Precision	Recall	F1	Precision	Recall	F1	Δ
Baseline	68.44 ± 3.49	59.13 ± 0.88	61.66 ± 1.08	49.47 ± 2.52	42.17 ± 1.10	44.17 ± 1.16	
- Pre-training	62.04 ± 2.19	58.0 ± 1.32	58.34 ± 1.25	48.57 ± 3.37	44.32 ± 1.82	45.04 ± 2.38	+ 0.87
- Augmentation	68.62 ± 3.08	54.67 ± 3.01	60.01 ± 1.98	46.55 ± 2.06	37.39 ± 1.31	40.46 ± 1.23	- 3.71
- Ent. masking	69.20 ± 3.36	59.86 ± 1.89	61.90 ± 1.92	50.44 ± 2.34	42.38 ± 0.18	44.50 ± 0.92	+ 0.33
+ Gold hints	68.84 ± 2.03	60.39 ± 1.30	61.91 ± 0.96	62.15 ± 3.32	54.46 ± 1.37	55.92 ± 1.68	+ 11.75
+ spaCy hints	64.15 ± 1.64	55.76 ± 0.63	57.61 ± 0.09	47.25 ± 3.67	40.56 ± 1.87	42.04 ± 1.73	- 2.13

Note. Baseline model is fine-tuned REBEL. All metrics are macro averaged. Performance is reported averaged over all three seeds. Difference is only reported on relation extraction score. ± Indicates standard deviation.

Table 12 shows the same ablation study for ConfliBERT$_{\text{ENC-DEC}}$. Not using the pre-training data resulted in two of the three models returning a macro F1 score of 0, not being able to learn the task at all. To counteract this effect, the models without pre-training were instead trained for 50 epochs without early stopping, however achieving the same result on these models. Not using entity masking during training increased the macro F1 score for both relation classification and relation extraction. The data augmentation has a stronger effect on ConfliBERT$_{\text{ENC-DEC}}$. On one hand, this effect may be stronger than for

REBEL due to the lower performance overall. On the other hand, this possibly indicates that ConfliBERT$_{\text{ENC-DEC}}$ requires more training data overall to generalise the information learned. Unlike for REBEL, spaCy hints improve ConfliBERT$_{\text{ENC-DEC}}$ results on relation classification and relation extraction. The performance gain from gold hints was smaller than for REBEL but made a large difference on both tasks. Even with gold hints, ConfliBERT$_{\text{ENC-DEC}}$ is behind on the results achieved by REBEL without these hints.

Table 12

Performance of ConfliBERT$_{\text{ENC-DEC}}$ when removing or adding training strategies

	Relation Classification			Relation Extraction			
	Precision	Recall	F1	Precision	Recall	F1	Δ
Baseline	60.51 ± 5.12	49.37 ± 1.19	54.24 ± 1.72	28.40 ± 3.19	23.02 ± 1.15	25.36 ± 1.76	
- Pre-training	40.28 ± 3.11	17.47 ± 13.22	21.42 ± 12.37	1.33 ± 1.88	0.74 ± 1.05	0.94 ± 1.33	- 24.42
- Augmentation	59.93 ± 2.43	45.63 ± 2.19	49.53 ± 0.60	18.18 ± 1.23	13.61 ± 0.33	15.08 ± 0.56	- 10.28
- Ent. masking	62.10 ± 2.36	50.41 ± 0.93	53.57 ± 0.86	31.48 ± 1.10	24.93 ± 0.82	26.71 ± 0.18	+ 1.35
+ Gold hints	63.01 ± 1.22	51.99 ± 0.68	55.22 ± 0.20	38.82 ± 1.66	31.84 ± 1.04	34.04 ± 1.31	+ 8.68
+ spaCy hints	60.28 ± 2.03	51.80 ± 1.71	54.07 ± 1.91	29.42 ± 2.22	25.19 ± 1.89	26.54 ± 2.06	+ 1.18

Note. Baseline model is fine-tuned ConfliBERT$_{\text{ENC-DEC}}$. All metrics are macro averaged. Performance is reported averaged over all three seeds. Difference is only reported on relation extraction score. ± Indicates standard deviation.

7. Discussion

Both trained models were able to fully learn the algorithm used for the unsupervised dataset generation, as they achieve results close to a perfect F1 score on both CAMEO and Pentacode labels. However, neither of the models was able to match this performance on the annotated dataset. While this hints at a higher complexity of the annotated data, present in the way triplet patterns were labelled, several other factors may influence the score as well. Although the IAA found for the annotation data is decent, the IAA only evaluates reliability between annotators but not the validity of the dataset, as both annotators may give the same wrong label to the data. The data was not annotated by domain experts, possibly resulting in less consistent or more incorrect samples. It is expected that this produces an unquantified noise in the dataset. This does not only affect the model negatively in the way it learns to find triplets independent of the relation but also reduces the samples which the model can learn generalisable information for the respective relation from. As data augmentation has a strong impact on model performance it can be argued that more training data will improve results. Having a larger dataset with less noise, labelled by domain experts, would help to find the real upper limit of these models on real data.

Overall, REBEL matched or outperformed ConfliBERT$_{\text{ENC-DEC}}$ on every metric on either label, while also training for shorter periods. This can have a multitude of reasons. First, the model and embedding size of ConfliBERT$_{\text{ENC-DEC}}$ is not large enough to learn high-complexity data, as the model is based on BERT$_{\text{BASE}}$, while REBEL is based on a larger model and uses 1024 instead of 768 dimensions for embeddings. This would also fit the excellent performance on the unsupervised dataset in combination with the weak performance on the annotated dataset. Second, the training data possibly is not sufficient. On the unsupervised data, ConfliBERT$_{\text{ENC-DEC}}$ had a large dataset to learn from, while the annotated dataset is small. This goes in hand with the finding that the pre-trained baseline for ConfliBERT$_{\text{ENC-DEC}}$ finds higher performance on the annotated test set than ConfliBERT$_{\text{ENC-DEC}}$ without pre-training. This is further supported by the model taking several epochs before returning any predictions in the correct format, which requires fewer epochs on the unsupervised dataset. Third, it is likely that the task-specific knowledge of REBEL, in combination with some overlap in domains it was trained on, gives a better basis for learning sequence-to-sequence relation extraction compared to the domain-specific transfer learning method applied on ConfliBERT$_{\text{ENC-DEC}}$. This was unexpected but is further facilitated by the random initialisation of new layers which could not profit from ConfliBERT's thorough pre-training on domain-specific knowledge.

Two measures taken during the training of the models resulted in lower scores. The first measure decreasing performance is the pre-training on the unsupervised dataset. REBEL achieved higher relation extraction results when skipping the pre-training on the unsupervised dataset, while the relation classification score decreased. The weaknesses mentioned in Section 5.1 likely influence the suitability of the unsupervised dataset as the pre-training dataset. Furthermore, it cannot be ruled out that the pre-training on the unsupervised dataset reduced the performance of REBEL due to REBEL having already experienced a thorough relation extraction pre-training on a large dataset. As such, training on the unsupervised data possibly results in less generalisable token embeddings and parameters. Interestingly, the relation classification being lower when not pre-training the model suggest that particularly the entities in the pre-training are misleading the model, while the verb patterns used to create relations in the unsupervised dataset are helping to identify more relations. When reconnecting this information with the way the attention mechanism works for the transformers used it can be deducted that the model picks entities due to paying high attention towards relation-indicating words. This could mean that a pre-training on a wider repertoire of relation-indicating words, such as the CAMEO verb dictionary, can help the model. If this is the case, the way the entities and subjects are structured in the unsupervised dataset is too different from the structure in the annotated dataset and the dataset is thus not fit for thorough pre-training. A new pre-training dataset should use a different NER pipeline to identify actors, which is overall more coherent with the way that entities are labelled on the annotated dataset.

Importantly, the heavy decrease of performance for ConfliBERT$_{\text{ENC-DEC}}$ when not pre-training the model is mostly accounted for by the models not being able to learn the task of relation extraction in a text generation manner on the small dataset. As such, the pre-training on a larger unsupervised dataset poses great value when only having a small annotated dataset and training a model with a new task or new randomly initialised weights, as in the case of ConfliBERT$_{\text{ENC-DEC}}$. However, in such a situation, task-specific transfer learning should be preferred and a more fitting option, such as REBEL should be used instead.

The second measure taken to cause a reduction in performance is entity masking. Entity masking decreased performance on both tasks for both models. This comes highly unexpected, given its recurring appearance in high-performing hyperparameter combinations during tuning, and must be caused by the newly initialized masking token. It was expected that the masking token can be easily identified as part of a relation while not being able to deduce relations from specific biased entities, appearing often in specific relations. It is unclear whether the drop in performance is due to the model not being able to classify relations from these biased entities or whether the masking needs a larger dataset to be more effective. Overall, entity masking cannot be recommended for training on this data.

The performance increase in both tasks when using gold entity hints shows high potential for complimenting the end-to-end relation extraction model with a fine-tuned NER model.

These findings replicate the findings of prior papers using gold-level entity hints (e.g., Giorgi et al., 2022). Further, the gold entity hints improved the performance on the relation classification task. This indicates that relation extraction can indeed profit from joint training, as a model that finds more correct entities seems to also find more correct relations. However, using a pre-trained NER model like the spaCy transformer pipeline does not work as well. This might be due to the model being pre-trained on a dataset which was annotated using spaCy NER, resulting in the model already learning the patterns the spaCy model uses to perform NER, gaining little extra value from these new hints. This fits the decrease of performance on REBEL, which's baseline generalises better after pre-training, while the performance is increased on ConfliBERT$_{\text{ENC-DEC}}$, which seems to not have fully learned the entity patterns, instead further profiting from the spaCy entity hints. The performance decrease on REBEL further argues for a large difference between the spaCy entities and the annotated entities, resulting in confusing the model with wrong entities rather than giving actual hints for the entities it should use to form relations. This would agree with the arguments made towards the ineffectiveness of the employed pre-training due to misleading entities. It is not expected that the results increase when using entity dictionaries, such as those used for PETRARCH. Instead, a separate NER model needs to be fine-tuned on a fitting NER dataset. Such a dataset does not yet exist to the degree where it would be usable with the annotation strategy used in this thesis, requiring the annotation of all relations in a sentence. As mentioned during the description of the data, the annotated dataset created in this thesis is not fit for such training, as the NER model would need to identify all relevant entities in the sentence, not only those included in relations. Although error propagation is less relevant with an end-to-end model as relation extraction part, it is expected that the way the relation extraction models learn heavily differs when being used in such a pipeline.

Even though the results achieved in the ablation study are mostly conclusive, it is expected that all performances reported can be increased upon, as the models were not tuned again for the respective ablation. As such, better hyperparameter combinations can likely be found to further improve the results of model training without pre-training and entity masking, while possibly reducing the loss in performance without augmentation or when using spaCy hints.

The outcomes of REBEL, even with gold entity hints, are still far behind the results reported by previous studies (Parolin et al., 2022), which reported a macro F1 score of up to 64.5 for relation extraction and 78.4 for relation classification. However, due to the use of Pentacode labels instead of CAMEO labels and particularly due to the structural differences in data, the results cannot be directly compared. While it is likely, that increasing the relations that must be found within the data reduces the overall score, the performance of Parolin et al. (2022) might also be affected by the model guessing relations which are present in the text but are not annotated in the label.

While Parolin et al. (2022) found higher recall than precision on the overall score, all models trained in this thesis found equal or higher precision than recall on either type of label. It is expected that this high recall is due to the models needing to only find on relation within the data. Recall values are lower, when many FN predictions are made, indicating that the models rather miss relations than assign wrong labels for a sample. The latter case would result in lower precision of the models instead. While this indicates that the model is not able to find all relations, this outcome should be preferred over a high recall with lower precision. As the data from which the model should extract relations exist in vast quantities in form of blogs, news articles or social media opinions, it is a wanted outcome to find fewer but correct relations within this data. This will result in the creation of decent amounts of correct data per sample, rather than finding larger volumes on relations of which many are not entailed by the text.

8. Conclusion

This thesis created two datasets for the training of political relation extraction in an end-to-end manner. First, an unsupervised dataset was created by an algorithm similar to current event coding algorithms, which has proven to be easily learnable by transformers using sequence-to-sequence relation extraction. Second, an annotated dataset with substantial IAA, which differs from existing datasets in the complexity of labelling, requiring the models to find multiple political relations per sample.

This thesis further contributes to the existing research in political relation extraction and event coding using transformers in three ways. First, no other paper was identified to evaluate performance both on CAMEO and Pentacode labels, being able to draw insights in terms of the trade-off between the higher level of information provided by CAMEO labels and the higher accuracy of the model provided by Pentacode labels. Second, no other papers were identified to measure the effect of entity hinting and entity masking on these labels. It is unknown, whether other papers implemented other regularisation techniques, such as the label-smoothed NLL loss. Third, no other papers were identified to solve the task of political event coding on either CAMEO or Pentacode labels using sequence-to-sequence relation extraction.

To answer the research question leading this thesis, transformers can fully learn the algorithms currently applied in automatic event encoders. Further, sequence-to-sequence text generation transformers showed heavily reduced performance on the annotated data, indicating that there is a long way to go to achieve human-level event extraction. Several measures which were taken to improve the generalisability of the model, such as entity masking and pre-training on the large unsupervised dataset resulted in a decreased performance. However, the data augmentation method used shows promising results. Further, end-to-end models in combination with an additional NER step beforehand may find promising results in the future, given the strong performance increase found with entity hinting.

When training transformers to this end, using task-specific transfer learning outperformed domain-specific transfer learning. It remains unclear, which effect a larger and less noisy dataset can have on the models used. Further, testing is needed whether ConfliBERT$_{\text{ENC-DEC}}$ performed worse due to the model size, being too small for high-complexity data, or whether REBEL is superior due to being a task-specific model.

To implement one of the proposed transformer models as a substitution for current event coding algorithms such as PETRARCH, another step must be built atop the transformer model. Matching the result of current event coding algorithms, it would be necessary, to map the subject and object in created triplets back to relevant actors, being state actors, organisations and similar. However, in newspaper articles, several actors do not necessarily indicate their country, e.g., “20 victims”. For this, it may be helpful to separately construct a knowledge graph for the text at hand, profiting from further information given in a news article, possibly stated in earlier or later sentences, e.g., “Attack in Kherson”. Using knowledge graphs instead of actor mappings will create more events with higher flexibility regarding the actors which are recorded, allowing to map “20 victims” to “Kherson”. Also instead of matching this knowledge with a dictionary, it is likely less restrictive to match it with huge knowledge graph databases, such as WikiData, allowing to further map “Kherson” to “Ukraine”.

While it is likely, that transformers will replace current event coding algorithms at some point, research efforts in this area would profit from having a high-quality public benchmark dataset. This would facilitate future research both by increasing the comparability of results, as well as removing the barrier of data labelling to perform research in this area.

References

Althaus, S., Bajjalieh, J., Carter, J. F., Peyton, B., & Shalmon, D. A. (2020). *Cline Center Historical Phoenix Event Data*. Cline Center for Advanced Social Research. v.1.3.0. University of Illinois Urbana-Champaign.

Bach, N., & Badaskar, S. (2007). A Review of Relation Extraction. *Literature review for Language and Statistics II*, *2*, 1-15.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016). End-to-End Attention-based Large Vocabulary Speech Recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4945-4949.

Beieler, J. and Norris, C. (2014). *Petrarch: Python Engine for Text Resolution And Related Coding Hierarchy*. Available at https://github.com/openeventdata/petrarch

Beieler, J. (2016). Generating Politically-Relevant Event Data. *arXiv preprint arXiv:1609.06239*.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*(2), 281-305.

Cabot, P. L. H., & Navigli, R. (2021). REBEL: Relation Extraction by End-to-End Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2370-2381.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint arXiv:1406.1078*.

Crone, P. (2020). Deeper Task-Specificity Improves Joint Entity and Relation Extraction. *arXiv preprint arXiv:2002.06424*.

De Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 4585-4592.

Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., Stoutenborough, L., Kouril, M., Marsolo, K., & Solti, I. (2012). Building Gold Standard Corpora for Medical Natural Language Processing Tasks. In *AMIA Annual Symposium Proceedings* (Vol. 2012), 144-153.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Eberts, M., & Ulges, A. (2021). An End-to-end Model for Entity-level Relation Extraction using Multi-instance Learning. *arXiv preprint arXiv:2102.05980*.

Gerner, D. J., Schrodt, P. A., Yilmaz, O., & Abu-Jabr, R. (2002). Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions. *International Studies Association, New Orleans*.

Giorgi, J., Bader, G. D., & Wang, B. (2022). A sequence-to-sequence approach for document-level relation extraction. *arXiv preprint arXiv:2204.01098*.

Guo, Z., Zhang, Y., & Lu, W. (2020). Attention Guided Graph Convolutional Networks for Relation Extraction. *arXiv preprint arXiv:1906.07510*.

Han, R., Ning, Q., & Peng, N. (2020). Joint Event and Temporal Relation Extraction with Shared Representations and Structured Prediction. *arXiv preprint arXiv:1909.05360*.

He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint arXiv:2006.03654*.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Séaghdha, D. O., Padó, S., Pennacchiottim M., Romano, L., & Szpakowicz, S. (2019). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American medical informatics association*, *12*(3), 296-298.

Hu, Y., Hosseini, M., Parolin, E. S., Osorio, J., Khan, L., Brandt, P., & D’Orazio, V. (2022). ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5469-5482.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, *8*, 64-77.

Kate, R., & Mooney, R. (2010). Joint Entity and Relation Extraction using Card-Pyramid Parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning,* 203-212.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Le, P., & Zuidema, W. (2016). Quantifying the vanishing gradient and long distance dependency problem in recursive neural networks and recursive LSTMs. *arXiv preprint arXiv:1603.00423*.

Lee, J., Tang, R., & Lin, J. (2019). What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning. *arXiv preprint arXiv:1911.03090*.

Leetaru, K., & Schrodt, P. A. (2013). Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, 2(4), 1-49.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv preprint arXiv:1910.13461*.

Li, X., Luo, X., Dong, C., Yang, D., Luan, B., & He, Z. (2021). TDEER: An Efficient Translating Decoding Schema for Joint Extraction of Entities and Relations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8055-8064.

Li, X., Yin, F., Sun, Z., Li, X., Yuan, A., Chai, D., Zhou, M., & Li, J. (2019). Entity-Relation Extraction as Multi-Turn Question Answering. *arXiv preprint arXiv:1905.05529*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2021). On the Variance of the Adaptive Learning Rate and Beyond. *arXiv preprint arXiv:1908.03265*.

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*.

Lu, J. and Roy, J. (2017). *Universal Petrarch: Language-agnostic political event coding using universal dependencies.* Available at https://github.com/openeventdata/UniversalPetrarch

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings* *of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55-60.

Mikheev, A., Moens, M., & Grover, C. (1999). Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, 1-8.

Miwa, M., & Sasaki, Y. (2014). Modeling Joint Entity and Relation Extraction with Table Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*,1858-1869.

Miwa, M., & Bansal, M. (2016). End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. *arXiv preprint arXiv:1601.00770*.

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? *Advances in neural information processing systems*, *32*.

Norris, C. (2016). Petrarch 2: Petrarcher. *arXiv preprint arXiv:1602.07236*.

Pan, S. J., & Yang, Q. (2009). A Survey on Transfer Learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.

Parolin, E. S., Khan, L., Osorio, J., D’Orazio, V., Brandt, P. T., & Holmes, J. (2020). HANKE: Hierarchical Attention Networks for Knowledge Extraction in political science domain. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics*, 410-419.

Parolin, E. S., Hosseini, M., Hu, Y., Khan, L., Brandt, P. T., Osorio, J., & D'Orazio, V. (2022). Multi-CoPED: A Multilingual Multi-Task Approach for Coding Political Event Data on Conflict and Mediation Domain. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 700-711.

Peters, M. E., Ruder, S., & Smith, N. A. (2019). To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. *arXiv preprint arXiv:1903.05987*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*(140), 1-67.

Riedel, S., Yao, L., & McCallum, A. (2010, September). Modeling Relations and Their Mentions without Labeled Text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 148-163.

Roth, D., & Yih, W. T. (2007). Global Inference for Entity and Relation Identification via a Linear Programming Formulation. *Introduction to statistical relational learning*, 553-580.

Rothe, S., Narayan, S., & Severyn, A. (2020). Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics*, *8*, 264-280.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818-2826.

Tan, Q., He, R., Bing, L., & Ng, H. T. (2022). Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation. *arXiv preprint arXiv:2203.10900*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in neural information processing systems*, *30*.

Wang, J., & Lu, W. (2020). Two are Better than One: Joint Entity and Relation Extraction with Table-Sequence Encoders. *arXiv preprint arXiv:2010.03851*

Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *arXiv preprint arXiv:1901.11196*.

Wei, J., Liu, H., Liu, T., Niu, G., Sugiyama, M., & Liu, Y. (2022). To Smooth or Not? When Label Smoothing Meets Noisy Labels. *arXiv preprint arXiv:2106.04149.*

Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., & Jin, Z. (2015). Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Path. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1785-1794.

Xu, W., Chen, K., & Zhao, T. (2021). Discriminative reasoning for document-level relation extraction. *arXiv preprint arXiv:2106.01562*.

Yan, Z., Zhang, C., Fu, J., Zhang, Q., & Wei, Z. (2021). A partition filter network for joint entity and relation extraction. *arXiv preprint arXiv:2108.12202*.

You, K., Long, M., Wang, J., & Jordan, M. I. (2019). How Does Learning Rate Decay Help Modern Neural Networks?. *arXiv preprint arXiv:1908.01878*.

Zeng, X., Zeng, D., He, S., Liu, K., & Zhao, J. (2018). Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 506-514.

Zeng, S., Xu, R., Chang, B., & Li, L. (2020). Double Graph Based Reasoning for Document-level Relation Extraction. *arXiv preprint arXiv:2009.13752*.

Zhang, Y., Qi, P., & Manning, C. D. (2018). Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. *arXiv preprint arXiv:1809.10185*.

Zhao, T., Yan, Z., Cao, Y., & Li, Z. (2020). Asking Effective and Diverse Questions: A Machine Reading Comprehension based Framework for Joint Entity-Relation Extraction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20),* 3948-3954.

Zhong, Z., & Chen, D. (2021). A Frustratingly Easy Approach for Entity and Relation Extraction. *arXiv preprint arXiv:2010.12812*.

Appendices

Appendix A – Additional Tables

Table A1

Newspapers used to retrieve data and their origin

Newspaper	Continent	Country
New York Times	North America	United States
CBC	North America	Canada
Folha de S. Paulo	South America	Brazil
Buenos Aires Times	South America	Argentina
BBC	Europe	United Kingdom
Der Spiegel	Europe	Germany
France24	Europe	France
TASS	Europe / Asia	Russia
Japan Times	Asia	Japan
Times of India	Asia	India
Egypt Independent	Africa	Egypt
Eyewitness News	Africa	South Africa
Sydney Morning Herald	Australia	Australia

Table A2

Label mapping from CAMEO labels to Pentacode labels

Make a statement	Verbal Cooperation	Material Cooperation	Verbal Conflict	Material Conflict
Make a public statement	Express Intend to Cooperate	Engage in Material Cooperation	Reduce Relations	Exhibit Military Posture
Appeal	Engage in Diplomatic Cooperation	Provide Aid	Investigate	Engage in Unconventional Mass Violence
	Consult	Yield	Disapprove	Coerce
			Reject	Assault
			Threaten	Fight
			Demand	Protest

Table A3

Hyperparameters and values searched during tuning

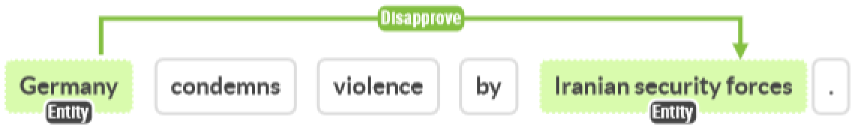
Batch size	Learning rate	Learning rate decay	ε	Weight decay	Masking rate
16 ^{R,C}	0.000005 ^{R}	0.05	0.00	0.0025	0.0
32	0.000010 ^{R}	0.10	0.05	0.0050	0.1
64*	0.000025	0.15	0.10	0.0075	0.2
	0.000035 ^{R,C}	0.20	0.15	0.0100	
	0.000050	0.30*	0.20	0.0150	
	0.000065 ^{C}				
	0.000075				
	0.000100*				
	0.000125*				

Note. * Value was only used during tuning on unsupervised dataset. ^{R} Value was only used during tuning of REBEL on annotated dataset. ^{C} Value was only used during tuning of ConfliBERT$_{\text{ENC-DEC}}$ on annotated dataset. Hyperparameters for unsupervised dataset were set identically for both models.

Appendix B - Annotation Examples

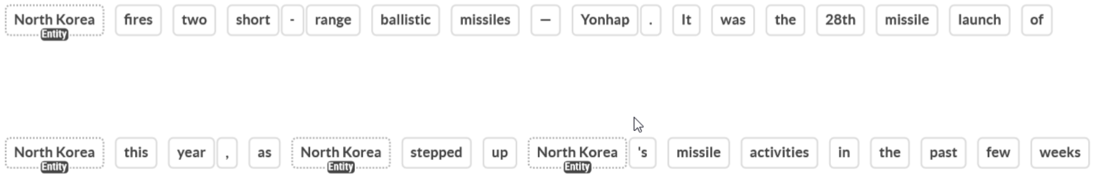
All examples are provided as screenshot directly from the annotation tool used, Prodigy. The examples are picked due to being special cases that were very notable in their characteristics during annotation. Screenshots were taken during annotation before training model or pre-processing the dataset.

1. Sentence example which is easy to annotate. These sentences only appeared rarely and were often at least two sentences long.



Germany condemns violence by Iranian security forces .

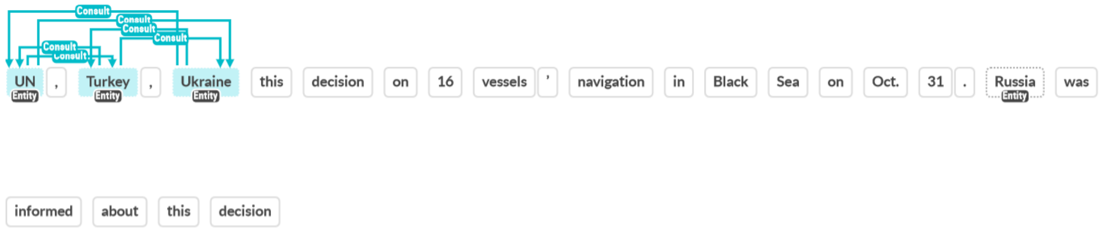
2. Sentence example with relevant event but only actor entity. These relations appear frequently and if the whole article was being used as input as once, some remark about South Korea would possibly induce an object for this relation of “Exhibit Military Posture”.



North Korea fires two short - range ballistic missiles – Yonhap . It was the 28th missile launch of

North Korea this year , as North Korea stepped up North Korea 's missile activities in the past few weeks

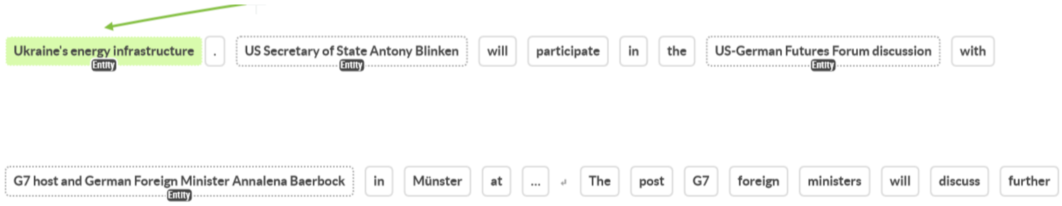
3. Sentence example for extreme reciprocal relationships. In most cases, reciprocal relationships are only existing between two entities. This also poses the risk of a quickly scaling number of relations, when the list of actors is longer.



UN , Turkey , Ukraine this decision on 16 vessels ’ navigation in Black Sea on Oct. 31 . Russia was

informed about this decision

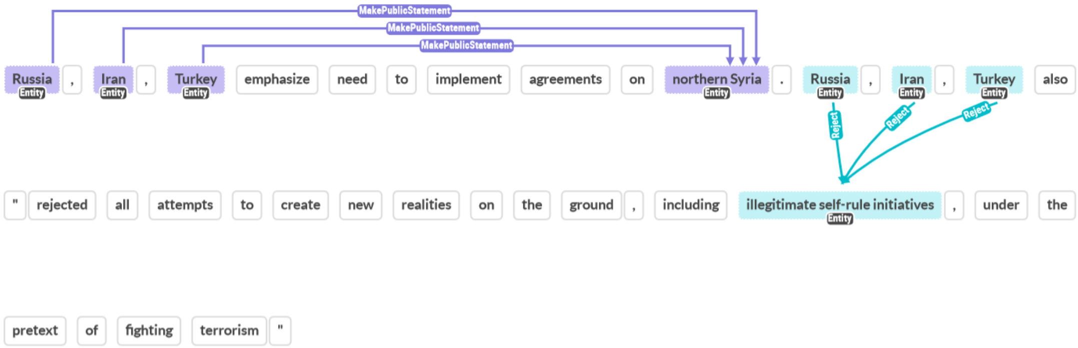
4. Sentence example where entities are unclear, is “G7 host” and “German Foreign Minister Annalena Baerbock” the same or different entities in this sentence. A model would have to know that Münster is in Germany to have a human-level conclusion on this.



Ukraine's energy infrastructure . US Secretary of State Antony Blinken will participate in the US-German Futures Forum discussion with

G7 host and German Foreign Minister Annalena Baerbock in Münster at ... ↵ The post G7 foreign ministers will discuss further

5. Sentence example where coreference resolution allows for the recording of more relationships. In the original text, all three entities in the second sentence would have been “they”, which holds less value for relation extraction.

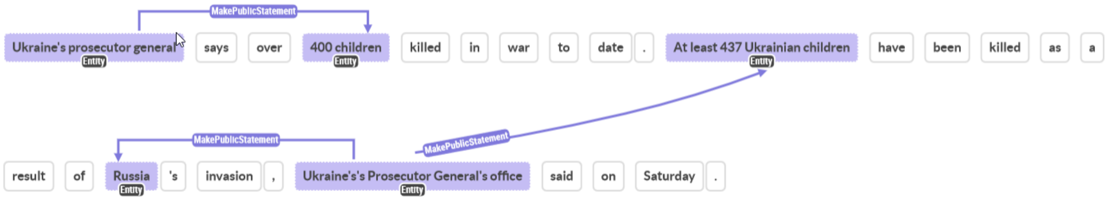


Russia , Iran , Turkey emphasize need to implement agreements on northern Syria . Russia , Iran , Turkey also

" rejected all attempts to create new realities on the ground , including illegitimate self-rule initiatives , under the

pretext of fighting terrorism "

6. Sentence example where duplicate relationship are recorded due to paraphrasing one of the entities. Identical duplicates were removed, but often duplicates remain due to changes in entities. This happens regularly, as for example “Putin” and “President Putin”, or other leaders, are commonly used within the same text.

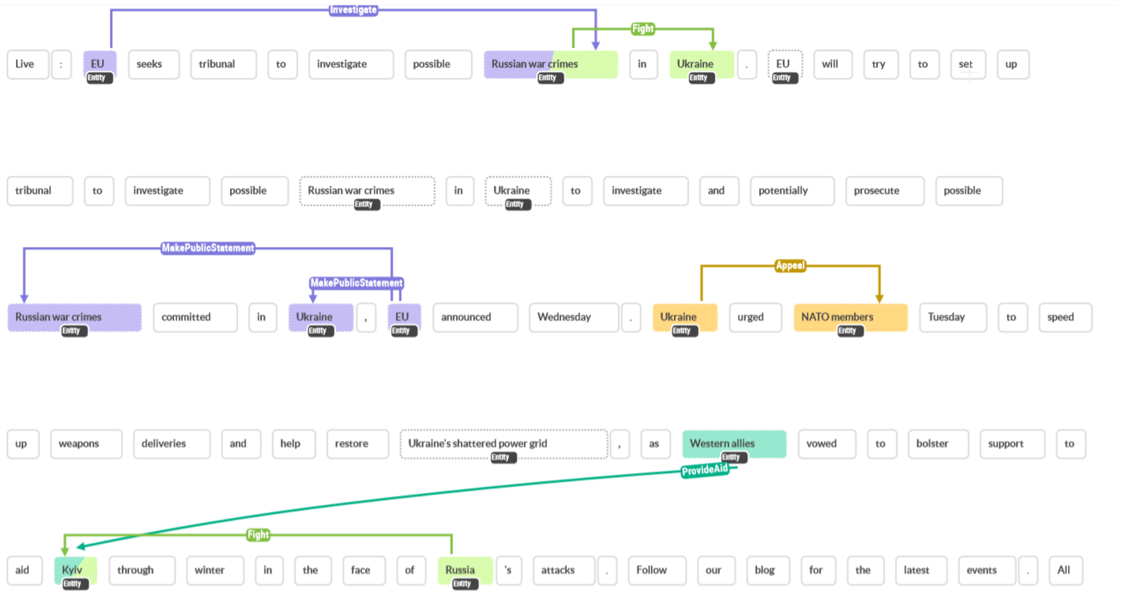


Ukraine's prosecutor general says over 400 children killed in war to date . At least 437 Ukrainian children have been killed as a

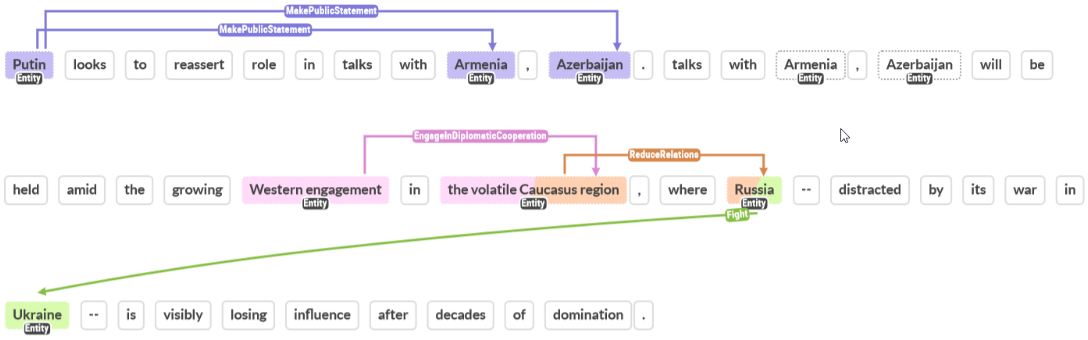
result of Russia 's invasion , Ukraine's's Prosecutor General's office said on Saturday .

7. Sentence examples where many relations are contained in one article title and description combination.

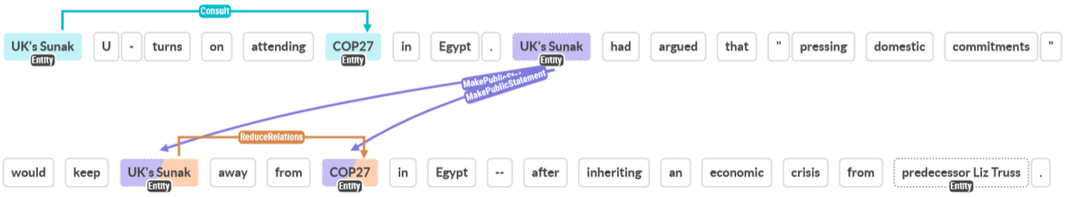
7A. Annotating these samples with high number of different entities and relations takes longer and increases the risk of missing relations. It is expected that these examples would be easier to annotate if split into sentences in the first place.


Live : EU seeks tribunal to investigate possible Russian war crimes in Ukraine . EU will try to set up tribunal to investigate possible Russian war crimes in Ukraine to investigate and potentially prosecute possible Russian war crimes committed in Ukraine , EU announced Wednesday . Ukraine urged NATO members Tuesday to speed up weapons deliveries and help restore Ukraine's shattered power grid , as Western allies vowed to bolster support to aid Kyiv through winter in the face of Russia 's attacks . Follow our blog for the latest events . All

7B. Often, these longer samples also change the topic and mention several ongoing events which are not directly connected to another.

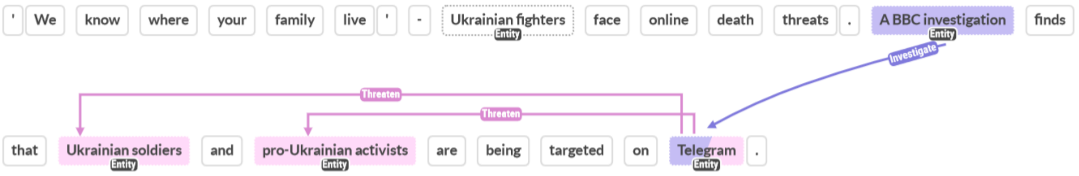

Putin looks to reassert role in talks with Armenia , Azerbaijan . talks with Armenia , Azerbaijan will be held amid the growing Western engagement in the volatile Caucasus region , where Russia -- distracted by its war in Ukraine -- is visibly losing influence after decades of domination .

8. Sentence example where information in the text is contradicting due to timeliness of some information that has changed. Timeliness of information is a problem that was encountered multiple times, also in reports about World War 2 and similar.



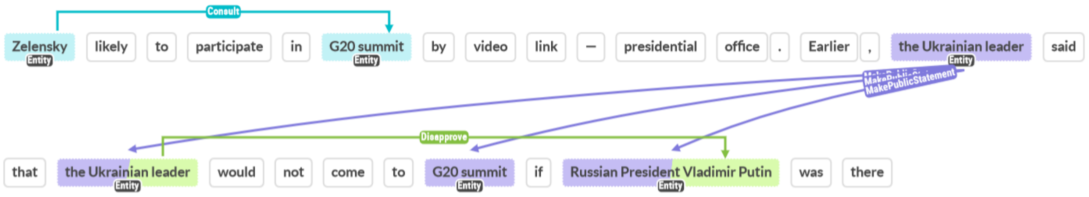
UK's Sunak U - turns on attending COP27 in Egypt . UK's Sunak had argued that " pressing domestic commitments " would keep UK's Sunak away from COP27 in Egypt -- after inheriting an economic crisis from predecessor Liz Truss .

9. Sentence where the relation type is only induced by first sentence, information being lost when splitting back to sentence. Targeting something itself is no indication of a threat, while the threat in the first sentence is.



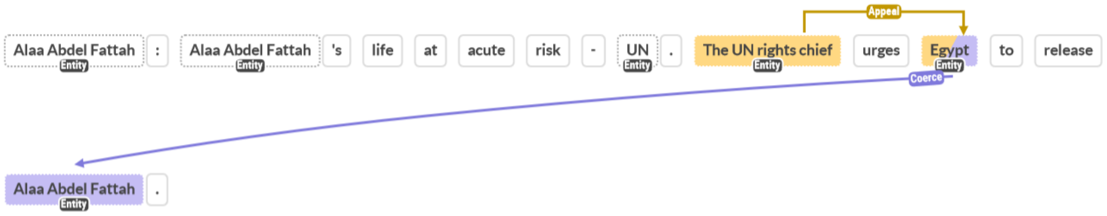
' We know where your family live ' - Ukrainian fighters face online death threats . A BBC investigation finds that Ukrainian soldiers and pro-Ukrainian activists are being targeted on Telegram .

10. Sentence where simple text results in many relationships, these texts were quite common throughout the annotation process but usually 3-4 sentences long.



Zelensky likely to participate in G20 summit by video link — presidential office . Earlier , the Ukrainian leader said that the Ukrainian leader would not come to G20 summit if Russian President Vladimir Putin was there

11. Sentence where the complement clause (urge – release) changes the meaning of the relation indicating verb. Without the urge, the relation would be “Yield” instead. However, knowing that Egypt is being urged implicates that they are currently coercing the entity.



Alaa Abdel Fattah : Alaa Abdel Fattah 's life at acute risk - UN . The UN rights chief urges Egypt to release Alaa Abdel Fattah .

Appendix C - Data Examples

This appendix includes ten random samples from all splits of both datasets used during training. The full datasets are available in the files which were handed in with this thesis or on GitHub: github.com/valentinwerner1/Thesis_RelationExtraction_PoliticsNews

Table A4

Samples of the annotated training set (augmented)

Text	Label
Clips on social media show hundreds protesting in Tibet.	<triplet> hundreds <subj> Tibet <obj> Protest
“Premeditated”: an Israeli peace movement attack an Israeli peace movement's tax inspector.	<triplet> an Israeli peace movement <subj> an Israeli peace movement's tax inspector <obj> Assault
Footballer describes Italy stabbing attack.	<triplet> Footballer <subj> Italy stabbing attack <obj> Make Public Statement
one U.S. soldier rejects President Yoweri Museveni for Minsk, tell offensive will continue.	<triplet> one U.S. soldier <subj> President Yoweri Museveni <obj> Reject <triplet> President Yoweri Museveni <subj> Minsk <obj> Make Public Statement
special police troops in Kosovo enable only vulnerable people to communicate a migrant ship stuck for two weeks.	<triplet> special police troops in Kosovo <subj> vulnerable people <obj> Provide Aid
President El-Sisi’s's Speeches “In the name of Allah the Most Benevolent, the Most Merciful” Today, we celebrate a very dear and cherished occasion to the heart of every Egyptian; the 50th anniversary of the Egyptian-Emirati relations."	<triplet> Egyptian <subj> Emirati <obj> Engage In Diplomatic Cooperation
Demonstrations over Zambia that erupted at the weekend appear to have engage down for now.	<triplet> Demonstrations <subj> Zambia <obj> Protest
ASEAN was founded on August 8, 1967, in Bangkok by five Southeast Asian countries	<triplet> five Southeast Asian countries <subj> ASEAN <obj> Engage In Diplomatic Cooperation

"response to South Korea crowd crush is now obliged to announce A senior British minister how Missile incident in Ukraine's's energy network happened, why the territory of Ukraine's's energy network was thump, who plotted Missile incident in Ukraine's's energy network, why that missile, allegedly assault eastwards, suddenly made a Uturn to fly west, and why that missile, allegedly assault eastwards was launched at a moment when there were no Russian missiles in the air," Quatar's emir, Sheikh Hamad bin Khalifa alThani see	<triplet> Quatar's emir, Sheikh Hamad bin Khalifa al-Thani <subj> response to South Korea crowd crush <obj> Demand <triplet> Quatar's emir, Sheikh Hamad bin Khalifa al-Thani <subj> A senior British minister <obj> Make Public Statement <triplet> Quatar's emir, Sheikh Hamad bin Khalifa al-Thani <subj> Ukraine's's energy network <obj> Make Public Statement
Iran captain "supports" anti-government protests ahead of match against England.	<triplet> Iran captain <subj> anti-government protests <obj> Engage In Diplomatic Cooperation

Table A5

Samples of the annotated validation set

Text	Label
Kidnapped Franco-Australian conservationist released by captors in Chad.	<triplet> captors <subj> Kidnapped Franco-Australian conservationist <obj> Yield
French President Emmanuel Macron, during a phone call with Ukrainian President Volodymyr Zelensky	<triplet> French President Emmanuel Macron <subj> Ukrainian President Volodymyr Zelensky <obj> Consult <triplet> Ukrainian President Volodymyr Zelensky <subj> French President Emmanuel Macron <obj> Consult
Zelesnkyy rejects claims Polish missile strike launched from Ukraine.	<triplet> Zelesnkyy <subj> claims Polish missile strike launched from Ukraine <obj> Reject
In UN council's first ever-resolution on the South-east Asian nation, UN council also urged the release of all "arbitrarily detained" prisoners including former leader Aung San Suu Kyi.	<triplet> UN council <subj> release of all "arbitrarily detained" prisoners <obj> Appeal

US ready to work with Russia on preparation of treaty to replace New START - Pentagon.	<triplet> US <subj> Russia <obj> Express Intend to Cooperate
Taliban ruled Afghanistan has been sharing expertise with the Liberation Tigers of Tamil Eelam according to a special report submitted to the Canadian Security Intelligence Service	<triplet> a special report submitted to the Canadian Security Intelligence Service <subj> Taliban ruled Afghanistan <obj> Make Public Statement <triplet> a special report submitted to the Canadian Security Intelligence Service <subj> the Liberation Tigers of Tamil Eelam <obj> Make Public Statement <triplet> Taliban ruled Afghanistan <subj> the Liberation Tigers of Tamil Eelam <obj> Engage In Material Cooperation <triplet> the Liberation Tigers of Tamil Eelam <subj> Taliban ruled Afghanistan <obj> Engage In Material Cooperation
The European Community may halt mediation efforts among Yugoslavia's feuding republics if cooperation by all parties founders	<triplet> The European Community <subj> Yugoslavia's feuding republics <obj> Threaten
CSTO Secretary-General Stanislav Zas made a report on the work of the CSTO mission sent to Armenia in connection with the aggravation of the situation on the Armenian-Azerbaijani border	<triplet> CSTO Secretary-General Stanislav Zas <subj> CSTO <obj> Make Public Statement <triplet> CSTO <subj> Armenia <obj> Investigate
A prominent anti-Syria journalist has been killed in a car bomb explosion in a residential sector of mostly Christian eastern Beirut	<triplet> a car bomb explosion <subj> a residential sector of mostly Christian eastern Beirut <obj> Assault <triplet> a car bomb explosion <subj> A prominent anti-Syria journalist <obj> Assault
The secretary general of NATO said that a Ukrainian air defense missile most likely caused Deadly Blast.	<triplet> The secretary general of NATO <subj> Ukrainian air defense missile <obj> Make Public Statement

Table A6

Samples of the annotated test set

Text	Label
Israeli Prime Minister Ehud Barak has agreed to US mediation in the final status talks with the Palestinians, a senior Israeli official said.	<triplet> Israeli Prime Minister Ehud Barak <subj> US mediation <obj> Engage In Diplomatic Cooperation <triplet> a senior Israeli official <subj> Israeli Prime Minister Ehud Barak <obj> Make Public Statement <triplet> a senior Israeli official <subj> US mediation <obj> Make Public Statement <triplet> a senior Israeli official <subj> the Palestinians <obj> Make Public Statement <triplet> Israeli Prime Minister Ehud Barak <subj> the Palestinians <obj> Express Intend to Cooperate
With the signing of a Memorandum of Understanding (MOU) on Cessation of Hostilities, the Sudanese government and SPLM/A have agreed to allow “unimpeded humanitarian access to all areas and for people in need.”	<triplet> the Sudanese government <subj> humanitarian access <obj> Engage In Material Cooperation <triplet> SPLM/A <subj> humanitarian access <obj> Engage In Material Cooperation
Arab League Secretary General Chadli Klibi undertook mediation mission between Syria and Palestinian leader Yasser Arafat.	<triplet> Arab League Secretary General Chadli Klibi <subj> Syria <obj> Consult <triplet> Arab League Secretary General Chadli Klibi <subj> Palestinian leader Yasser Arafat <obj> Consult
Zimbabwean Prime Minister Robert Mugabe today accused the United States of restoring the blackmail in the negotiations on independence for Namibia.	<triplet> Zimbabwean Prime Minister Robert Mugabe <subj> Namibia <obj> Make Public Statement <triplet> Zimbabwean Prime Minister Robert Mugabe <subj> the United States <obj> Disapprove
A member of the Syrian parliament, Mohammed Mamoun, started a hunger strike yesterday to protest President Assad’s failure to usher in meaningful political reforms	<triplet> A member of the Syrian parliament, Mohammed Mamoun <subj> President Assad <obj> Protest

A roadside bombing near the town of Samarra on Sunday killed one U.S. soldier and wounded two others, the military said.	<triplet> the military <subj> one U.S. soldier <obj> Make Public Statement <triplet> the military <subj> two others <obj> Make Public Statement <triplet> the military <subj> A roadside bombing near the town of Samarra <obj> Make Public Statement <triplet> A roadside bombing near the town of Samarra <subj> one U.S. soldier <obj> Assault <triplet> A roadside bombing near the town of Samarra <subj> two others <obj> Assault
Red Cross officials suspended aid deliveries in Baghdad on Wednesday after two of their vehicles were hit by gunfire and a staff member went missing	<triplet> Red Cross officials <subj> Baghdad <obj> Reduce Relations <triplet> Baghdad <subj> two of their vehicles <obj> Assault <triplet> Baghdad <subj> a staff member <obj> Assault
Angola has asked Portugal for military aid, especially instructors for its Soviet- and Cuban-trained armed forces	<triplet> Angola <subj> Portugal <obj> Appeal
Sri Lanka has established diplomatic ties with and opened an embassy in Tehran	<triplet> Sri Lanka <subj> Tehran <obj> Engage In Diplomatic Cooperation
King Hassan of Morocco was quoted today as saying he would be ready to host a meeting between an Israeli peace movement and the Palestine Liberation Organization (PLO).	<triplet> King Hassan of Morocco <subj> an Israeli peace movement <obj> Express Intend to Cooperate <triplet> King Hassan of Morocco <subj> the Palestine Liberation Organization (PLO) <obj> Express Intend to Cooperate

Table A7

Samples of the unsupervised training set (augmented)

Text	Label
At the same time, Michel is seeking to improve EU's economic standing in the world's second largest economy, which has experienced significantly slower growth partly as a result of rigid anti-COVID-19 restrictions that also limited Michel's visit to just one day.	<triplet> Michel <subj> EU's economic standing <obj> Engage In Material Cooperation

Colombian Olga Lucía Álvarez Benjumea comes from a very devout Catholic family but had the support of Colombian Olga Lucía Álvarez Benjumea's mother, a former nun.	<triplet> Colombian Olga Lucía Álvarez Benjumea <subj> a very devout Catholic family <obj> Consult
Mohammad Ahmadzadeh, who played for Iran from 1988 to 1990 and coached Malavan F. C. from 2018 to 2020 want to ask Mr. Donyamali, who considers Bandar Anzali’s member of parliament Ahmad Donyamali a representative of Bandar Anzali city, northern Iran“ why are Bandar Anzali’s member of parliament Ahmad Donyamali silent?	<triplet> Mohammad Ahmadzadeh <subj> Mr. Donyamali <obj> Appeal
NASA ministers disturb to Mar-a-Lago to deliver a memorandum of demands, NASA ministers also demonstrated outside the Johannesburg Central Police Station earlier.	<triplet> NASA ministers <subj> Mar-a-Lago <obj> Protest
Nigerian President Muhammadu Buhari also congratulated Biden "on Biden's election at a time of uncertainty and fear in world affairs".	<triplet> Nigerian President Muhammadu Buhari <subj> Biden's election <obj> Engage In Diplomatic Cooperation
Picture: AFPBrexitEuropean UnionTheresa MayJean-Claude Juncker Email PrintTweetShareAFP | 24 May 2019 13:53BRUSSELS - EU said Friday that May's resignation does nothing to change EU's position on the Brexit withdrawal deal that EU's members agreed with Britain.	<triplet> EU's members <subj> Britain <obj> Engage In Diplomatic Cooperation
Washington: The top two Republicans in the US Congress have broken The top two Republicans in the US Congress's silence about former president Donald Trump’s dinner last week with white supremacist Nick Fuentes, saying the The top two Republicans in the US Congress's party has no place for anti-Semitism or white supremacy.	<triplet> The top two Republicans <subj> The top two Republicans <obj> Fight
ANC President Cyril Ramaphosa also decree Indian-administered Kashmir's special status, be Indian-administered Kashmir's special status, be they Mapuche or not Mapuche or	<triplet> ANC President Cyril Ramaphosa <subj> Indian-administered Kashmir's special status <obj> Demand

not, to have easier access to traditional healers, known as Machi.	
the Saudi-led coalition leading the House committees investigating Joemat-Pettersson and Joemat-Pettersson's aides over Guinea and his aides over their pressure campaign on Ukraine's pressure campaign on Ukraine would like to discuss from Gordhan's High Court review application, but it's not clear if Gordhan's High Court review application'll ever appear to give testimony.	<triplet> Joemat-Pettersson <subj> Guinea <obj> Investigate <triplet> the Saudi-led coalition <subj> Gordhan's High Court review application <obj> Consult
Leading economist Bismarck Rewane tells CNN that the changes to the currency naira look adds nothing to naira's value and is insignificant in curbing counterfeiting.	<triplet> Leading economist Bismarck Rewane <subj> CNN <obj> Make Public Statement

Table A8

Samples of the unsupervised validation set

Text	Label
In a statement following a bill to stave off a strike by railway workers that could potentially devastate the US economy's passage in the House of Representatives, Biden urged the Senate to "act urgently".	<triplet> Biden <subj> the Senate <obj> Appeal
WATCH | A look inside one of the state's three abortion clinics:the only abortion clinic serving northern Louisiana months agoDuration 5:32The last abortion clinic in northern Louisiana is managing a surge in demand after Texas tightened abortion restrictions, while waiting to hear if the Supreme Court will overturn the landmark Roe v. Wade ruling, which has upheld federal abortion rights for nearly 50 years.	<triplet> the Supreme Court <subj> the landmark Roe v. Wade ruling <obj> Coerce

Masuku over the weekend told the Sunday Times that Masuku had still not been furnished with any charges by Gauteng ANC.	<triplet> Masuku <subj> the Sunday Times <obj> Make Public Statement <subj> Gauteng ANC <obj> Make Public Statement
a vote next week, if successful, would pass a D. C. statehood bill for the first time in the U. S. House of Representatives, but a D. C. statehood bill faces insurmountable opposition in the Republican-controlled Senate.	<triplet> a D. C. statehood bill <subj> the Republican-controlled Senate <obj> Consult
The the Group of Seven (G7) nations price cap will allow non-EU countries to continue importing seaborne Russian crude oil, but The price cap, an idea of the Group of Seven (G7) nations, will prohibit shipping, insurance and re-insurance companies from handling cargoes of Russian crude around the globe, unless Russian crude is sold for less than the cap.	<triplet> non-EU countries <subj> seaborne Russian crude oil <obj> Engage In Material Cooperation
Experts believe Kim, who is enjoying warmer relations with North Korea's and the easing of pressure from Russia and Mark Chinoy, senior fellow at U. S. -China Institute at the University of Southern California,, will seek a U. S. commitment for improved bilateral relations and partial sanctions relief while trying to minimize any concessions on North Korea's Kim Jong-un's nuclear facilities and weapons.	<triplet> Kim <subj> a U. S. commitment <obj> Appeal
Kurdish-led forces in Syria have called on Russia to dissuade Turkey from launching a ground offensive against Kurdish-led forces in northern Syria, Kurdish-led forces in northern Syria's commander said on Tuesday.	<triplet> Kurdish-led forces <subj> Russia <obj> Consult <subj> Turkey <obj> Make Public Statement <subj> Russia <obj> Make Public Statement
But it is reported that A Japanese man disagrees with A Japanese man's lawyers.	<triplet> A Japanese man <subj> A Japanese man's lawyers <obj> Disapprove
Some Republicans are already accusing Democrats of using a shooting at a primary	<triplet> Some Republicans <subj> Democrats <obj> Disapprove

school in south Texas to cynically further Democrats's own political objectives.	
Picture: AFPMozambiqueManuel ChangCredit Suisse Group Email PrintTweetShareBonga Dlulane | 04 January 2019 19:19JOHANNESBURG - London authorities have confirmed that London authorities have the arrests Former Credit Suisse bankers arrested in London over Mozambique loans in connection with a fraud scheme allegedly involving AFPMozambiqueManuel ChangCredit.	<triplet> Former Credit Suisse bankers <subj> London <obj> Coerce

Table A9

Samples of the unsupervised test set

Text	Label
Picture: MOHAMMED ABED/AFPIsraelPalestineGazaIsrael strikesIslamist militant Email PrintTweetShareAFP | 02 July 2021 15:12GAZA CITY, PALESTINIAN TERRITORIES - Israel hit Islamist militant sites in Gaza with air strikes on Friday in retaliation for incendiary balloon launches from Gaza, in the latest unrest since a ceasefire ended May's conflict.	<triplet> Israel <subj> Islamist militant sites <obj> Fight
Wray, who last week drew criticism from U. S. President Donald Trump for FBI Director Christopher Wray's description of Russian election interference and the threat posed by the anti-fascist movement known as Antifa, said in the Senate testimony Thursday that the U. S. has only experienced occasional voter fraud on a local level.	<triplet> Wray <subj> the Senate testimony <obj> Make Public Statement
Picture: EWNCity of EkurhuleniEkurhuleni Metro Police Department EMPD trainees Email PrintTweetShareMia Lindeque | 12 October 2021 06:30JOHANNESBURG -	<triplet> Angry EMPD trainees <subj> City <obj> Disapprove

Angry EMPD trainees have accused City of Ekhuruleni of deliberately the hold up Angry EMPD trainees's training in order to cut costs.	
Separately, the White House announced U. S. would welcome 100,000 Ukrainian refugees and provide an additional $1 billion US in food, medicine, water and other supplies.	<triplet> U. S. <subj> 100,000 Ukrainian refugees <obj> Express Intend to Cooperate
The European Union's leaders have decided to extend the anti-Russian economic sanctions expiring on July 31 for six more months, an The European Union's spokesman said on Thursday.	<triplet> an The European Union's spokesman <subj> the anti-Russian economic sanctions <obj> Make Public Statement
AdvertisingRead moreHobbs, who is Arizona’s secretary of state,'s victory adds further evidence that Trump is weighing down Trump's allies in Arizona’s as Trump gears up for an announcement of a 2024 presidential run.	<triplet> Trump <subj> Trump's allies <obj> Consult
And so the purpose of the visit is to coordinate with nine other heads of state, what are in Washington interest, and President Joe Biden believe in Israel's interest as well," President Joe Biden stressed.	<triplet> President Joe Biden <subj> Israel's interest <obj> Express Intend to Cooperate
But after sustained pressure from the US and Russia, a compromise was struck and a pan-Afghan delegation that held many rounds of dialogue with the country's hardline Islamist former rulers, alongside the months of peace talks the US held agreed to talk to an unofficial Afghan delegation.	<triplet> a pan-Afghan delegation <subj> an unofficial Afghan delegation <obj> Consult
US President Donald Trump's national security adviser arrived in Jerusalem on Saturday to take part in the trilateral meeting with Russian Security Council Secretary Nikolai Patrushev and Israeli National Security Adviser Meir Ben-Shabbat on June 24-25.	<triplet> US President Donald Trump's national security adviser <subj> Jerusalem <obj> Consult

"Matt Binder, a journalist for Mashable and one of those suspended've been very critical of Elon Musk in Matt Binder, a journalist for Mashable and one of those suspended's reporting," Matt Binder, a journalist for Mashable and one of those suspended told the BBC.	<triplet> Matt Binder <subj> the BBC <obj> Make Public Statement

Appendix D - Annotation Code Book

The Codebook used for annotation was mostly transferred from the original CAMEO Codebook[1].

Entity annotations:

- Sentences where coreference resolution failed and entities became mixed up are fully ignored and skipped.
- ‘s is not part of the entity if the entity ends on it (e.g., Putin’s speech at G20 …; Putin = entity, G20 = entity)
- Articles are part of entities (e.g., “the”, “an”, “a”)
- All descriptive words connected to an entity (e.g. “the former president Donald Trump”)

Types of entities to be annotated:

Table A10

Entity types to annotate with examples

Entity	Example
Official institutions, mostly governmental	“The White House”, “The Pentagon”
Persons, can be political leader or any other named or unnamed person	“President Joe Biden”, “Alla”, “An Ukrainian Soldier”
Political Organizations	“NATO”, “EU”
Corporations	“Adidas”, “Microsoft”
Geographical Locations	“Germany”, “Donesk Region”, “Madrid”
Political Events	“Protests”, “2nd confidence vote”
Laws, Policies and Rights	“Qatari Laws”, “Local policies”

Relation annotations:

- In case of bidirectional relations (e.g. Marcon-Schulz meetup in Paris; Consult), mark both directions as separate relations (Macron, consult, Schulz & Schulz, consult, Macron)

[1] Retrievable on https://parusanalytics.com/eventdata/data.dir/cameo.html

- In case a relation contains the same entities twice (exactly same relation), it is sufficient to mark the relevant entity once

Types of relations to be annotated:

Table A11

Relation types to annotate with examples

Type of Relation	Meaning	Examples
Make Public Statement	All public statements not falling in other categories Includes but is not limited to: - Decline comment - Make pessimistic or optimistic comment - Consider policy option - Claim or deny responsibility or acknowledgement - Reject accusation - Engage in symbolic act - Express accord	1) “President Reagan and Egyptian President Hosni Mubarak agreed today there was an urgent need for progress towards a Middle East settlement“ 2) “Secretary-General Boutros Boutros-Ghali on Saturday expressed condolences to the United States for the death of three American diplomats.“ 3) “The government of Liberia denied on Thursday charges by Ivory Coast that Monrovia is committing genocide“
Appeal	All requests, proposals, suggestions and appeals. Includes but is not limited to: - Appeals for cooperation (economic, military, judicial, …) - Appeals for aid	1) “Kenyan President Daniel Arap Moi on Monday urged Uganda to to repatriate “all Kenyan criminals hiding there” to face trial, accusing them of killing Kenyan policemen in cross-border raids recently.” 2) “U.S. President George W. Bush said Friday that he will tell Japanese Prime Minister Junichiro Koizumi that Japan

		needs to enact significant economic reforms.”
Express intend to Cooperate	Offer, promise, agree to, or otherwise indicate willingness or commitment to cooperate. Includes but is not limited to: - Economical cooperation & aid - Military cooperation & aid - Judicial cooperation & aid - Diplomatic cooperation & aid - Negotiations	1) “Senior Hungarian and Romanian officials agreed on Wednesday that their countries should cooperate to encourage Romanian refugees in Hungary to return home.” 2) “Portugal will support Turkey’s efforts to become a full member of the European Community, Portuguese President Mario Soares said”
Consult	Consultations and meetings. Includes but is not limited to: - Discussion by phone - Make a visit - Host a visit - Meet at a “third” location - Engage in mediation - Engage in negotiation	1) “U.S. Secretary of State Warren Christopher telephoned Russian Foreign Minister Andrei Kozyrev on Tuesday to discuss efforts to forge a peace settlement in former Yugoslavia, Itar-Tass news agency said” 2) “Israel and Lebanon renewed negotiations today on an Israeli troop pullback from Lebanon and their future relations.”
Engage in Diplomatic Cooperation	Initiate, resume, improve or expand diplomatic, non-material cooperation or exchange. Includes but is not limited to:	1) “A top U.S. official today praised Haiti’s efforts to improve its record on human rights and said it was an important partner for the United States.”

	- Praise or Endorse - Defend verbally - Rally support on behalf of - Grant diplomatic recognition - Apologize - Forgive - Sign formal agreements	2) “Saudi Arabia has mobilized pressure groups in the United States to help support the rights of Palestinians in their struggle against Israel.” 3) “Argentina has apologized to Brazil for one of its gunboats intercepting a Brazilian ship in the Beagle Channel, disputed by Argentina and Chile.”
Engage in Material Cooperation	Initiate, resume, improve or expand material cooperation or exchange. Includes but is not limited to: - Cooperate economically - Cooperate militarily - Judicial cooperation - Share intelligence or information	1) “European foreign direct investment flows in Latin America and the Caribbean rose more than eightfold“ 2) “Zambia extradited suspected British militant Haroon Rashid Aswad to Britain on Sunday, a senior Zambian government official said.“
Provide Aid	Provisions, extensions of material aid. Includes but is not limited to: - Monetary aid and financial guidance - Military aid - Humanitarian aid - Military protection & asylum	1) ”The United States continued to send arms to Pakistan last year, a State Department Spokesman said Wednesday.” 2) “U.N. helicopters evacuated the wounded from the besieged Bosnian town of Gorazde on Friday”
Yield	Yieldings and concessions Includes but is not limited to: - Ease sanctions and restrictions - Ease political freedoms - Ease bans on political entities	1) “The Nigerian Union of Teachers (NUT), the umbrella union for primary school teachers, announced Thursday that it has called off a four-day strike after deliberations with the Nigerian government.”

	- Ease political dissent - Accede to demands - Return, release of persons or property - Allow for humanitarian actions - Demobilize armed forces - Retreat or surrender militarily	2) “The Rwandan government on Thursday accepted demands from Hutu rebels that it initiate political reforms.“ 3) “Georgian President Eduard Shevardnadze resigned Sunday as the opposition threatened to storm his residence in Tbilisi.“
Investigate	Investigations Includes but is not limited to: - Crime & corruption investigations - human rights abuses investigations - investigate military action & war crimes - Monitoring or surveillance	1) “Israel’s high court opened a landmark hearing Wednesday into the legality of secret interrogation techniques used against Palestinian detainees.” 2) “A US national has been put under investigation in Italy for her possible role in rioting during a G8 summit in Genoa last month, Ansa news agency reported.”
Demand	Demands or Requires Includes but is not limited to: - Demand any kind of cooperation - Demand any kind of aid - Demand political reform - Demand yields - Demand negotiations	1) “Some 800,000 Iraqi Kurds sought refuge in Germany last month.” 2) “Opposition groups in Zimbabwe are demanding that President Mugabe abandon his controversial policy of land confiscations.“ 3) “Russia said on Tuesday that Sudan must return a Mi-26 helicopter that was captured by the Sudanese authorities last week.”
Disapprove	Disapproves, Criticize, Denounce, Accuse, Complain, Lawsuits	1) “Lebanon complained to the United Nations on Tuesday over two Israeli air raids last Friday in

	Includes but is not limited to: - Accuse of aggression or abuse - Accuse of crime, corruption - Rally opposition against - Find guilty or liable	which it said 20 people were killed or wounded” 2) “Archbishop Desmond Tutu on Sunday called for sanctions against Nigeria in the wake of the execution of Ken Saro-Wiwa.“
Reject	Reject or refuse appeals and demands Includes but is not limited to: - Defy norms, laws - Veto - Reject requests for rights, changes - Reject aid or cooperation - Refuse to ease dissent	1) “The U.S. said it would not meet hostage-takers demands to release prisoners in Iraq, including a number of females.” 2) “Israel is opposed to French mediation in peace negotiations with Syria“ 3) “The United States on Wednesday vetoed a Security Council resolution censuring as a violation of international law its military sweep of the Nicaraguan ambassador’s home in Panama on December 29.”
Threaten	All threats, coercive or forceful warnings with serious potential repercussions Includes but is not limited to: - Non-force threats - Threaten to reduce or stop aid - Threaten to boycott, embargo or sanction - Threaten to reduce or break relations - Threaten with restrictions - Threaten to halt negotiations or mediation - Threaten with military force	1) “Iran on Tuesday threatened to cut off electricity to the autonomous Azerbaijani republic of Nakhichevan over non-payment of bills” 2) “The Hamas threatened Monday to resume terrorist activities in Israel in an escalation of the intifada“ 3) “Iraq’s interim government announced that it is prepared to impose martial law as street battles raged in central Baghdad between insurgents and security forces“

Exhibit military posture	Increase of alert status, mobilize power, however, not yet attacking Includes but is not limited to: - Mobilizing army or police - Increase cyber forces - Weapon tests	1) “government of Sindh province has ordered patrols by police and paramilitary soldiers after violent protests by Muslim groups.“ 2) “North Korea has trained more than 500 computer hackers capable of launching cyber warfare against the United States“
Protest	Civilian demonstrations and other collective actions carried out as protest Includes but is not limited to: - Demonstrate or rally for changes or rights - Conduct strike or boycott - Obstruct passage, block - Protest violently, riot	1) “Up to 100 ethnic Albanians demonstrated on Tuesday in the Yugoslav province of Kosovo, where 24 people were killed in nationalist riots last March” 2) “Palestinian youths resorted to throwing stones during demonstrations against the alleged human rights violations by the Israeli military, officials said on Thursday.“
Reduce relations	All reductions in normal, routine, or cooperative relations Includes but is not limited to: - Expel or withdraw cooperation or aid - Halt mediations or negotiations - Impose embargo, sanctions, boycott - Reduce or stop assistance - Reduce or break diplomatic relations	1) “Switzerland said today it had expelled two Soviet diplomats based in Geneva for spying, adding to a long series of espionage scares.“ 2) “Japan said on Tuesday it had halted economic aid to Yugoslavia in line with Western efforts to end the fighting there” 3) “President Bill Clinton has imposed sanctions on the Taliban religious faction that controls Afghanistan“ 4) US has been violating the UN resolution on Iran since 2018

	- Violating resolutions / agreements	
Coerce	Repression, Violence against civilians or their rights or properties Includes but is not limited to: - Confiscate property - Impose sanctions against civilians - Arrest, detain - Attack cybernetically - Discrimination	1) “The British government on Monday outlawed the largest Protestant extremist organization in Northern Ireland because of what it called its direct involvement in killing in the strife-torn province.” 2) “Israeli soldiers arrested more than 100 Palestinians on Saturday in a security sweep of the Hebron area of the occupied West Bank“ 3) “North Korea has tried to hack into the computers of South Korean army officers, officials said Tuesday“
Assault	Use of unconvential violence Includes but is not limited to: - Torture - Hostage taking - Physical assault - Sexual assault - Any type of bombing - (attempted) Assassination	1) “The Sri Lankan army has been holding thousands of Tamil civilian refugees as human shields in the battle zones of the southern sector of the Jaffna peninsula“ 2) „Three US servicemen were killed by an improvised explosive device outside of the Iraqi city of Basra“
Fight	All uses of conventional military force Includes but is not limited to: - Blockades - Occupy territory - Use small arms & light weapons - Use artillery & tanks	1) “Both the Phillippines military and the Moro Islamic Liberation Front are guilty of violating the ceasefire agreement signed in March 2001” 2) “British aircraft using precision guided missiles killed 4 Iraqis in an attack on a suspected weapons supply in Basra.“

	- Employ aerial weapons - Violate ceasefire	
Engage in unconventional mass violence	All uses of unconventional force that are meant to cause mass destruction, casualities and suffering Includes but is not limited to: - Mass expulsion - Mass killings - Ethnic cleansing - Use weapons of mass destruction - Use chemical, biological or radiological weapons - Detonate nuclear weapons	1) “The Israeli army forced out on Wednesday more than 1,000 Palestinian refugees from their homes in a West Bank refugee camp“ 2) “Serb forces were engaged in ethnic cleansing in Kosovo against the majority Albanian population of the province, according to the US government.“ 3) “Sudan’s government is responsible for mass killings and other atrocities in the Darfur region, according to a United Nations report.“