CO-Fun: A German Dataset on Company Outsourcing in Fund Prospectuses for Named Entity Recognition and Relation Extraction

Neda Foroutan and **Markus Schröder** and **Andreas Dengel**
Smart Data & Knowledge Services Dept., DFKI GmbH
Trippstadter Str. 122, 67663 Kaiserslautern, Germany
{neda.foroutan,markus.schroeder,andreas.dengel}@dfki.de

Abstract

The process of cyber mapping gives insights in relationships among financial entities and service providers. Centered around the outsourcing practices of companies within fund prospectuses in Germany, we introduce a dataset specifically designed for named entity recognition and relation extraction tasks. The labeling process on 948 sentences was carried out by three experts which yields to 5,969 annotations for four entity types (Outsourcing, Company, Location and Software) and 4,102 relation annotations (Outsourcing–Company, Company–Location). State-of-the-art deep learning models were trained to recognize entities and extract relations showing first promising results. An anonymized version of the dataset, along with guidelines and the code used for model training, are publicly available at https://doi.org/10.5281/zenodo.12745116.

Keywords— Cyber Mapping, Financial Domain, Dataset, Corpus, German, Named Entity Recognition, Relation Extraction, Conditional Random Fields, BERT, RoBERTa

1 Introduction

Cyber incidents, such as data breaches and ransomware attacks, pose potential risks to financial stability since banks and other institutes increasingly outsource processes and services to information and communication technology providers (Adelmann et al., 2020). To discover cyber risks, a conceptual method is "cyber mapping" – a process which links the financial network (e.g. banks, funds, insurance companies) with the cyber network (e.g. cloud services, datacenters, software providers) (Brauchle et al., 2020). Evidences for constructing a mapping could be descriptions of outsourced services and companies in the financial domain. A promising source to collect such hints can be found in publicly available fund prospectuses. In these documents, German Capital Management Companies (CMCs) have to state outsourcing companies and their provided services for a particular fund. To give an example, consider the following simplified sentence.

Example 1 *Die Gesellschaft hat Rechenzentrumsleistungen auf die Mercurtainment & CO KGaA ausgelagert.*
'The company has outsourced data center services to Mercurtainment & CO KGaA.'

To extract structured information, a usual step in Natural Language Processing (NLP) is the application of Named Entity Recognition (NER) to discover entities in texts. In our scenario, there are outsourced services (e.g. "data center services") and companies (e.g. "Mercurtainment & CO KGaA"). After that, Relation Extraction (RE) is commonly used to predict relationships between entities, in our case, services and companies. In order to train such NLP models, a dataset with ground truth labels is necessary.

In this paper, we present a novel dataset to support the process of cyber mapping using NLP models. Our annotated corpus consists of 948 sentences extracted from 1,054 German fund prospectuses. In total, 5,969 named entity annotations and 4,102 relation annotations were added by experts to acquire ground truth data. We conducted experiments with our dataset to evaluate the performance of trained models.

2 Related Work

NER and RE tasks are fundamental building blocks for extracting information within unstructured texts (for a recent survey see (Nasar et al., 2022)). For training models, several corpora have been built to cover specific domains, for example, the biomedical area (Khettari et al., 2023) or for clinical purpose (Báez et al., 2020). Some of them targeting specific languages, like Kazakh (Yeshpanov et al., 2022) and Italian (Paccosi and Aprosio, 2022). Regarding German language, Schiersch et al. (2018) collected data from tweets, news documents and

Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), pages 117–122
September 10-13, 2024 ©2024 Association for Computational Linguistics

RSS feeds to create a corpus with named entities such as Disasters, Triggers, Location, Organizations, Persons as well as 15 relations of the mobility and industry domain.

More related to our scenario is the business domain since the discovery of relationships between company entities is of interest. Here, Schön et al. (2018) provided an English dataset for recognizing companies, products and their relations to each other. The data was gathered from company homepages, business news portals, forums and social media channels. Instead of considering the product in the business relation, extracting the relation between two companies within unstructured texts – called Business Relation Extraction – has attracted attention in research and industry. Khaldi et al. (2021) presented a web-based English dataset for the business relation extraction between organizations. They also recommended a relation classifier using multilevel knowledge of entities to predict five types of relations between companies, i.e. Investment, Cooperation, Sale-purchase, Competition and Legal proceedings. In subsequent work, the authors provided the BIZREL dataset (Khaldi et al., 2022), a multilingual corpus in French, Spanish and Chinese in addition to their introduced English dataset. Similarly, they collected data via keyword queries using well-known search engines and the same five types of relations. Zuo et al. (2017) proposed a method of iteratively extracting asymmetric business relations like "owner-of" between two companies and indicating the relation direction between them. They evaluated their suggested method on two datasets based on *New York Times* News articles.

In the financial domain, Jabbari et al. (2020) created a French corpus including 26 entity types and 12 relation types gathered from French financial newspapers. They trained a BERT-based (Devlin et al., 2018) NER model on five types of entities (Person, Location, Organization, Role and Currency) and investigated a rule-based RE method for the relationship around the Role entity (i.e. "hasRole"). Moreover, Hillebrand et al. (2022) recommended a BERT-based architecture that employs a Gated Recurrent Units tagger coupled with conditional label masking to jointly predict entities tags sequentially and links the predicted entities. Additionally, they built a dataset from real-world German financial documents. The main entity type is Key Performance Indicators (KPI), such as revenue or interest expenses. Generally, entity classes include KPI, change of it, its monetary value and their sub-types. Linked relations are considered between KPI and sub-types or their values.

Still, there seems to be no dataset which meets our requirements. To train NLP models for performing cyber mapping, we need realistic sentences in German language explicitly mentioning outsourced services in the financial domain. Therefore, we built our own dataset from fund prospectuses which is covered in detail in the next section.

3 Corpus Creation

The corpus was created in a collaborative research lab of Deutsche Bundesbank[1] (the central bank of the Federal Republic of Germany) and the German Research Center for Artificial Intelligence[2] (DFKI). In this project a set of 1,054 publicly available fund prospectuses (PDFs) were collected from websites of 37 well-known Capital Management Companies (CMCs) in Germany.

Our corpus is built upon these documents by first converting the PDFs into plain texts using Apache's PDFBox[3] text stripper routine. The fund prospectuses consist of 92 pages on average, however, only a certain section in the document, usually no longer than a full or half page, mentions outsourced services. Conveniently, independent of the CMC, such a section is commonly named 'Auslagerung [Outsourcing]' followed by a section labeled 'Interessenkonflikte [Conflicts of Interest]' with some minor variations. Therefore, with a proper regular expression we were able to identify the beginning and end of these sections in our plain texts. For sentence splitting, Apache's OpenNLP[4] sentence detector loaded with a German pre-trained model[5] was applied. To turn words with hyphens such as 'Dienst–leistung [ser–vice]' in their hyphenless form, string matching and string manipulation was performed with regular expressions. Finally, 1,267 sentences could be collected of comparable shape as Example 1. However, roughly half of them assemble bullet point lists.

3.1 Annotation Process

Three subject-matter experts of the Deutsche Bundesbank annotated the corpus with named entities and relations. For this, a Graphical User Interface

[1] https://www.bundesbank.de/en

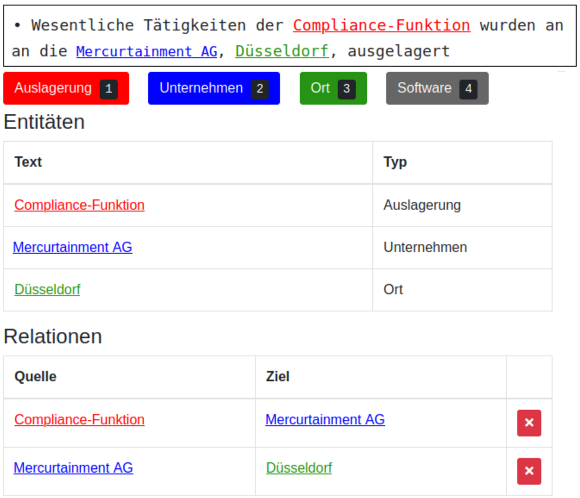
[2] https://www.dfki.de/

[3] https://pdfbox.apache.org/

[4] https://opennlp.apache.org/

[5] opennlp-de-ud-gsd-sentence-1.0-1.9.3.bin

(GUI) was provided which is depicted in Figure 1. Sentences are randomly distributed to the annotators who independently worked on them. The experts sporadically annotated sentences during their working hours and were done after around one month. Because of limited time available, same texts were not sent to multiple annotators, therefore, inter-annotator agreement is not considered. The following named entity types could be annotated: ‘Auslagerung’ [Outsourcing], ‘Unternehmen’ [Company], ‘Ort’ [Location] and Software. Additionally, we allow users to declare the following two relationships: Outsourcing–Company and Company–Location. Annotators could mark sentences as ignorable if they recognize that no entities are present (this happened 85 times). To reduce annotation efforts, our system pre-annotates sentences with already collected named entities once they exactly match in the text. A three-page annotation guideline was provided to give a brief tutorial and to cover special cases during the annotation process.



• Wesentliche Tätigkeiten der Compliance-Funktion wurden an an die Mercurtainment AG, Düsseldorf, ausgelagert

Entitäten

Text	Typ
Compliance-Funktion	Auslagerung
Mercurtainment AG	Unternehmen
Düsseldorf	Ort

Relationen

Quelle	Ziel	
Compliance-Funktion	Mercurtainment AG	×
Mercurtainment AG	Düsseldorf	×

Figure 1: A graphical user interface in German to annotate a sentence (top) with named entities (center) and relations (bottom). Entity types are ‘Auslagerung [Outsourcing]’, ‘Unternehmen [Company]’, ‘Ort [Location]’ and Software.

3.2 Resulting Dataset

Our Company Outsourcing in Fund Prospectuses (CO-Fun) dataset consists of 948 sentences (900 textually unique) with 5,969 named entity annotations, including 2,340 Outsourced Services, 2,024 Companies, 1,594 Locations and only 11 Software annotations. While the sentences have an average length of 314.8 ± 393.7[6] characters (w/o markup tags) and 44.9 ± 53.5 tokens, they contain 6.3 ± 9 annotations on average. All sentences contain at least one named entity, while 193 of them do not include any relations. Without considering duplicates, our corpus mentions 270 outsourced services, 323 companies, 84 locations and one software. Although, the software type is very limited with only one unique entity, we still decided to leave it in the dataset. Regarding relations, we have found 2,573 Outsourcing-Company relationships and 1,529 links between Companies and Locations (in total 4,102). On average, 4.3 ± 8.6 relations can be found in the sentences.

The raw data of CO-Fun consists of records formatted in JavaScript Object Notation[7] (JSON) which are sequenced in a JSON-line file (UTF-8 encoding). Each entry has the following properties: In the `text` property, the annotated text is present in form of HyperText Markup Language (HTML). We use `span` tags[8] to annotate named entities in text. Annotations are uniquely identified with a Universally Unique Identifier (UUID) (`id`). The entity’s type is given in the `type` attribute which can be one of the following options: ‘Auslagerung’ [Outsourcing], ‘Unternehmen’ [Company], ‘Ort’ [Location] or Software. Additionally, named entities are listed in a JSON array called `entities`, again with their ID, type and covered text. Another JSON array (`relations`) defines the relationships between a source entity (`src`) and a target entity (`trg`).

Gathering a dataset about outsourced services to German companies naturally raises concerns of potential misuse. We therefore performed an anonymization of all companies by randomly swapping their names with other companies with the same postfix e.g. GmbH. Replacing name of companies by their postfix helps to maintain legal and business consistency. For this, we make use of OffeneRegister[9] – a database dump of the German commercial register.

The anonymized CO-Fun dataset is publicly available[10] under MIT license together with other related materials such as the annotation guideline, derived data and source code. In the next section,

[6] using ± notation for standard deviation

[7] https://www.json.org/

[8] https://html.spec.whatwg.org/#the-span-element

[9] https://offeneregister.de/

[10] https://doi.org/10.5281/zenodo.12745116

initial experiments with our dataset are presented.

4 Experiments

In our study, we investigated extracting two types of structured information from our corpus. Firstly, we recognized entities within our sentences by applying Named Entity Recognition (NER) methods. Secondly, we detected relations between entities using a Relation Extraction (RE) model.

4.1 NER and RE Methods

We employed two NER models: Conditional Random Fields (CRF) and BERT (Bidirectional Encoder Representations from Transformer) (Devlin et al., 2018). For applying CRF, we utilized CRFsuit toolkit (Okazaki, 2007) and derived the features related to the token itself and its neighborhood information. The token features include the word itself, its part-of-speech tag, whether the word is capitalized, starts with a capital letter or is a digit. In order to extract tokens and their part-of-speech tags, the SpaCy library and the "de_core_news_sm" German language model was ustilized. Additionally, we considered the bigram and trigram characters the word ends with, and each token was assigned the same bias feature. Furthermore, we captured neighborhood information from the two words to the left and right of the token, checking their part-of-speech tags and if they start with a capital letter or are entirely in uppercase. If a token is at the beginning or end of the sentence, we provided BOS or EOS as the left or right neighbor to CRF, respectively.

In order to apply the pre-trained BERT model, we fine-tuned the German language version of it on our data using the SpaCy 3 library[11]. As a result, the model with about 110 million parameters is capable of predicting our four entity types.

As a basis for relation extraction, we used SpaCy's tutorial for a relation extraction component on GitHub[12]. In this project, the pre-trained RoBERTa model (Liu et al., 2019) is fine-tuned to extract relations.

4.2 NER and RE Datasets

Before applying the CRF model, each sentence was tokenized and an Inside-Outside-Beginning (IOB) format label was assigned to each token. The IOB scheme gives each token one of the following labels: B-ent, I-ent or O. If the token is the beginning of an entity, it is labeled as B-ent (begin-of-entity) but if the token is part of the entity but not its beginning, I-ent (inside-of-entity) is assigned to the token. If the token does not belong to any of the entity types, it is tagged as 'O'. After IOB tagging the tokens of each sentence, we randomly split the data with the proportion of 80%, 10% and 10% to create the training, development and test sets, respectively.

The same sentences that were used in each set (training, development and test) for the CRF model were also considered for the BERT model. In other words, the sentences in each set from split data for CRF model were labeled in the format required for training the BERT model. Each set includes a list of sentences with the list of tuples containing their entities and labels specified with the location of the entity in the sentence (start and end character position as well as entity label). For later reuse, the training, development and test sets were converted into SpaCy binary files.

The dataset split from the NER case is the same for the RE datasets. For each sentence, a list of entities and relations were prepared. A structure is provided for each entity to record an entity's text and label as well as its character and token position in text. Each relation entry has a label and refers to a child and head entity using their token positions. Ultimately, dataset text files were converted to binary files in SpaCy format.

The training, development and test sets in text format for the CRF model, as well as in text and SpaCy formats for the BERT and RoBERTa models are publicly available in the anonymized CO-Fun dataset[13].

4.3 NER and RE Results

The CRF model was run for 100 iterations using the L-BFGS training algorithm. The L1 and L2 regularization terms tuned by using cross validation are 0.05 and 0.01, respectively. Default values were used for the remaining hyperparameters provided by the CRFsuite toolkit. The BERT model was fine-tuned on the German training set for unlimited number of epoches with the early-stopping of 1600 and batch size of 128 (default values of Spacy library). The initial learning rate and warm-up step were set to $5 * 10^{-5}$ and 250, respectively.

[11] https://spacy.io/

[12] https://github.com/explosion/projects/tree/v3/tutorials/rel_component

[13] https://doi.org/10.5281/zenodo.12745116

Also, the L2 weight decay rate with value of 0.01 was applied. Similarly, the RoBERTa model was fine-tuned for 52 epochs with a $1{,}000$ batch size. Moreover, there is a max-length parameter representing the furthest distance at which existing relation is sought between any two entities. We discovered in tests that the model performed best with a max-length of 20. Remaining parameters were configured the same as in BERT's configuration. Both models were trained on a NVIDIA RTX A6000 GPU which took 40 minutes (NER) and 9 minutes (RE).

We evaluated the performance of our models in terms of exact match using precision, recall and F1-score (Nadeau and Sekine, 2007). Table 1 demonstrates the performance of the NER and RE models on the training and test sets of CO-Fun, measured by micro-averaging. Both models of CRF and BERT face overfitting as test F1-scores show lower scores than their training values. However, CRF performs better than BERT on the test set with F1-score of 94%. Furthermore, RoBERTa could classify 86.35% of the relations that exist between entities in the test set.

		Train			Test	
Models	P	R	F1	P	R	F1
CRF	96.7	95.1	95.9	95.7	93.0	94.3
BERT	99.8	94.2	97.0	92.9	91.5	92.2
RoBERTa	89.4	81.7	85.3	86.5	86.1	86.3

Table 1: Precision (P), Recall (R) and F1-score results of the NER models (CRF and BERT) and RE model (RoBERTa) on the training and test sets of CO-Fun.

5 Conclusion and Future Work

In this paper, we introduced an annotated German dataset called CO-Fun which is a NER and RE dataset on company outsourcing in fund prospectuses. Our dataset contains 948 sentences with $5{,}969$ named entity annotations (including Outsourced Services, Companies, Location and Software) and $4{,}102$ annotated relations (Outsourcing–Company and Company–Location). Applying state-of-the-art NER and RE models showed promising performances on CO-Fun.

In the future, we aim to extend this dataset with similar data and improve the performance of applied models by using additional knowledge, for example, by incorporating knowledge graphs in the training process. Additionally, we will investigate the impact of bullet points within sentences, examining whether their presence facilitates processing for NER and RE tasks.

6 Acknowledgements

This work was funded by the TransferLab Cybermapping which is a collaborative transfer lab of Deutsche Bundesbank and DFKI. We would like to thank Christoph Fricke, Ezgi Delikanli and Jacqueline Krüger at Deutsche Bundesbank for their support and contributions to the dataset.

7 Ethical Considerations

Gathering a dataset about outsourced services to German companies naturally raises concerns of potential misuse. Although the raw information is publicly available, but not in a digested and enriched version, the cooperating official authority raised concerns that the real information could be misused by malicious players. To address these concerns, an anonymization strategy was chosen.

8 Limitations

One limitation is the number of annotators and a missing agreement. To annotate the sentences, we had only three experts sporadically annotated sentences during their working hours. As the experts had limited time, same texts were not sent to multiple annotators, therefore, inter-annotator agreement was not considered.

Another limitation for our corpus was the small set of $1{,}054$ documents provided to us: in fact, only few pages contain some sentences about outsourcing statements. Reasons for that are our special language, domain and selection constraints, thus, it was not possible for us to compile a larger dataset. Moreover, the collaborating partner could not provide more documents since other regulatory data is usually confidential and we are not allowed to get access to them. As a result, the size of the Co-Fun dataset is rather small.

References

Frank Adelmann, Jennifer A. Elliott, Ibrahim Ergen, Tamas Gaidosch, Nigel Jenkinson, Tanai Khiaonarong, Anastasiia Morozova, Nadine Schwarz, and Christopher J. Wilson. 2020. Cyber risk and financial stability : It's a small world after all. *IMF Staff Discussion Notes*.

Pablo Báez, Fabián Villena, Matías Rojas, Manuel Durán, and Jocelyn Dunstan. 2020. The chilean waiting list corpus: a new resource for clinical named entity recognition in spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop, ClinicalNLP@EMNLP 2020, Online, November 19, 2020*, pages 291–300. Association for Computational Linguistics.

Jan-Philipp Brauchle, Matthias Göbel, Jens Seiler, and Christoph von Busekist. 2020. Cyber mapping the financial system. Technical report, Carnegie Endowment for International Peace. `http://www.jstor.org/stable/resrep24291`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Lars Patrick Hillebrand, Tobias Deußer, Tim Dilmaghani Khameneh, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. 2022. KPI-BERT: A joint named entity recognition and relation extraction model for financial reports. In *26th International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25, 2022*, pages 606–612. IEEE.

Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. 2020. A french corpus and annotation schema for named entity recognition and relation extraction of financial news. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2293–2299. European Language Resources Association.

Hadjer Khaldi, Farah Benamara, Amine Abdaoui, Nathalie Aussenac-Gilles, and EunBee Kang. 2021. Multilevel entity-informed business relation extraction. In *Natural Language Processing and Information Systems - 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23-25, 2021, Proceedings*, volume 12801 of *Lecture Notes in Computer Science*, pages 105–118. Springer.

Hadjer Khaldi, Farah Benamara, Grégoire Siegel, Camille Pradel, and Nathalie Aussenac-Gilles. 2022. How's business going worldwide? a multilingual annotated corpus for business relation extraction. In *13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3696–3705. European Language Resources Association (ELRA).

Oumaima El Khettari, Solen Quiniou, and Samuel Chaffron. 2023. Building a corpus for biomedical relation extraction of species mentions. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, BioNLP@ACL 2023, Toronto, Canada, 13 July 2023*, pages 248–254. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30:3–26.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2022. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv.*, 54(1):20:1–20:39.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs). `https://www.chokkan.org/software/crfsuite/`.

Teresa Paccosi and Alessio Palmero Aprosio. 2022. KIND: an italian multi-domain dataset for named entity recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 501–507. European Language Resources Association.

Martin Schiersch, Veselina Mironova, Maximilian Schmitt, Philippe Thomas, Aleksandra Gabryszak, and Leonhard Hennig. 2018. A german corpus for fine-grained named entity recognition and relation extraction of traffic and industry events. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Saskia Schön, Veselina Mironova, Aleksandra Gabryszak, and Leonhard Hennig. 2018. A corpus study and annotation schema for named entity recognition and relation extraction of business products. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. Kaznerd: Kazakh named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 417–426. European Language Resources Association.

Zhe Zuo, Michael Loster, Ralf Krestel, and Felix Naumann. 2017. Uncovering business relationships: Context-sensitive relationship extraction for difficult relationship types. In *Lernen, Wissen, Daten, Analysen (LWDA) Conference Proceedings, Rostock, Germany, September 11-13, 2017*, volume 1917 of *CEUR Workshop Proceedings*, page 271. CEUR-WS.org.