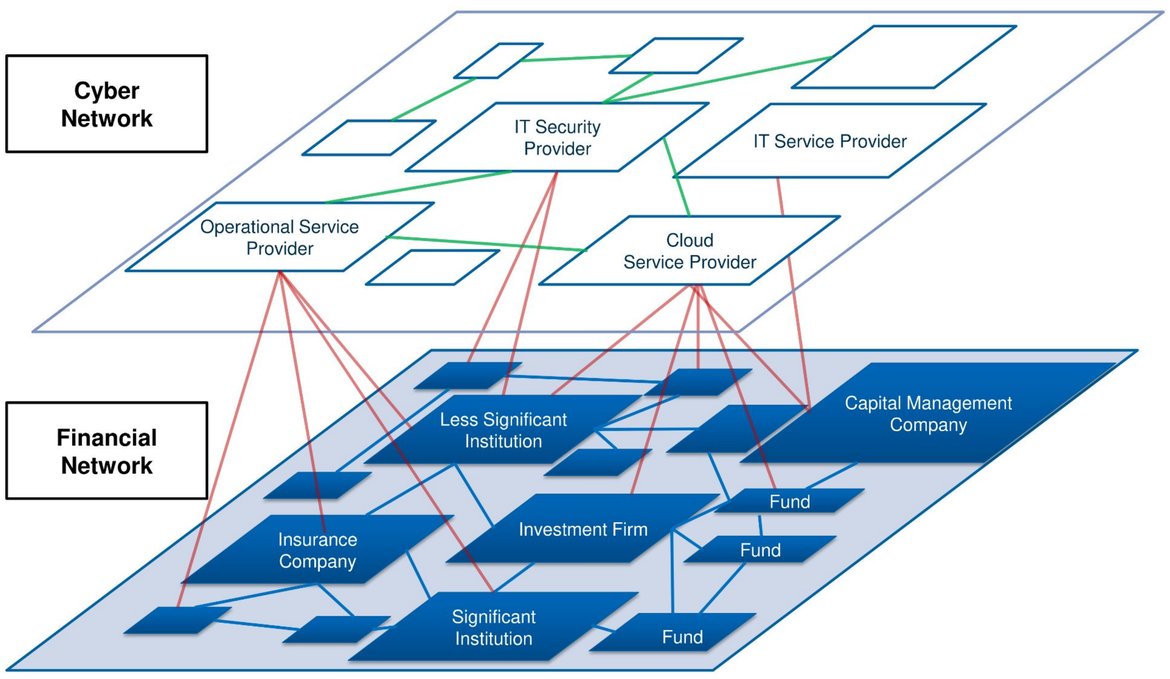
dfki ai Deutsches Forschungszentrum für Künstliche Intelligenz
*German Research Center for Artificial Intelligence*

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

***Neda Foroutan, Markus Schröder, Andreas Dengel***

## I. Motivation

- Banks and other institutes are increasingly outsourcing critical processes and services to third-party information and communication technology (ICT) providers.
- This outsourcing trend introduces potential risks to cyber incidents and financial stability.
- Discovering cyber risks requires “cyber mapping” giving insights in relationships among financial entities and service providers.
- Accurate recognition of entities and their relationships in financial documents necessitates the use of advanced NLP models trained on a dataset with ground truth labels.
- Lack of annotated datasets focused on outsourced services in the financial domain.
- No dataset links companies with the services they outsource.




## II. Data Collection

- 1,054 publicly available fund prospectuses (PDFs) were collected from websites of 37 well-known Capital Management Companies (CMCs) in Germany.
- Converting PDFs into plain texts.
- Sentences were splitted and preprocessed.
- 1,267 sentences could be collected.
- Roughly half of them assemble bullet point lists.
- Three subject-matter experts of the Deutsche Bundesbank annotated the corpus with named entities and relations.
- Named entity types: *‘Auslagerung’ [Outsourcing], ‘Unternehmen’ [Company], ‘Ort’ [Location]* and *Software*.
- Relationships types: *Outsourcing–Company* and *Company–Location*.

**Example:**

***‘Die Gesellschaft hat Rechenzentrumsleistungen auf die Mercurtainment & CO KGaA ausgelagert.’***

***‘The company has outsourced data center services to Mercurtainment & CO KGaA.’-***

**Named Entities:**

**Outsourcing: “data center services” and Company: “Mercurtainment & CO KGaA”**

**Relations:**

**Outsourcing-Company**

## III. CO-Fun Dataset

- The raw data of CO-Fun consists of records formatted in JSON.
- Each entry has the following properties:
  - “entities”: Lists of Named Entities with their type.
  - “text”: the annotated text is present in HTML Format.
  - “relations”: Lists of source and target entities with their relation type.

***** Anonymization of all companies by randomly swapping their names with other companies with the same postfix e.g. GmbH.

**CO-Fun Dataset**
https://zenodo.org/records/12745116


```
"entities":[
  {
    "text":"Fondsadministration",
    "type":"Auslagerung"
  },
  …
],
"text":"<html><head><\/head><body>Die Gesellschaft hat die folgend…
"relations":[
  {
    "src":{
      "text":"Fondsadministration",
      "type":"Auslagerung"
    },
    "trg":{
      "text":"OPEX Corporation",
      "type":"Unternehmen"
    },
    "type":"Auslagerung-Unternehmen"
  },
```


| | | | |
|---|---|---|---|
| **# Sentences** | 984 (900 unique) | | |
| **Sentences average length** | 314.8 ± 393.76 characters (w/o markup tags) | | |
| | 44.9 ± 53.5 tokens | | |
| **Sentences contain on average** | 6.3 ± 9 annotations | | |
| **# Named Entity annotations** | 5,969 | | |
| | # Outsourcing Services | 2,340 (W) | 270 (O) |
| | # Companies | 2,024 (W) | 323 (O) |
| | # Locations | 1,594 (W) | 84 (O) |
| | # Software | 11 (W) | 1 (O) |
| **# Relations annotations** | 4,102 | | |
| | # Outsourcing-Company | 2,573 | |
| | # Companies-Locations | 1,529 | |

(W): With duplicates (O): Without duplicates

## IV. Results

- Data was randomly split to training, development and test sets with 80%, 10% and 10% ratio.
- Applying Named Entity Recognition (NER) and Relation Extraction (RE) models on CO-Fun.
- CO-Fun is a NER and RE dataset on company outsourcing in fund prospectuses.

| Models | | Train | | | Test | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 |
| NER | 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 |
| RE | RoBERTa | 89.4 | 81.7 | 85.3 | 86.5 | 86.1 | 86.3 |

## 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.

## Limitations

- No inter-annotator agreement: As the expert annotators had limited time.
- Small size of the Co-Fun dataset:
  - The collaborating partner could not provide more documents since other regulatory data is usually confidential.
  - Our corpus was the small set of the provided documents and only a few pages contain some sentences about outsourcing statements due to our special language, domain and selection constraints.

**Towards Cyber Mapping the German Financial System with Knowledge Graphs**
https://link.springer.com/chapter/10.1007/978-3-031-60626-7_15