

MobIE: A German Dataset for Named Entity Recognition, Entity Linking and Relation Extraction in the Mobility Domain

Leonhard Hennig Phuc Tran Truong Aleksandra Gabryszak

German Research Center for Artificial Intelligence (DFKI)
Speech and Language Technology Lab

{leonhard.hennig, phuc_tran.truong, aleksandra.gabryszak}@dfki.de

Abstract

We present `MobIE`, a German-language dataset, which is human-annotated with 20 coarse- and fine-grained entity types and entity linking information for geographically linkable entities. The dataset consists of 3,232 social media texts and traffic reports with 91K tokens, and contains 20.5K annotated entities, 13.1K of which are linked to a knowledge base. A subset of the dataset is human-annotated with seven mobility-related, n-ary relation types, while the remaining documents are annotated using a weakly-supervised labeling approach implemented with the Snorkel framework. To the best of our knowledge, this is the first German-language dataset that combines annotations for NER, EL and RE, and thus can be used for joint and multi-task learning of these fundamental information extraction tasks. We make `MobIE` public at <https://github.com/dfki-nlp/mobie>.

1 Introduction

Named entity recognition (NER), entity linking (EL) and relation extraction (RE) are fundamental tasks in information extraction, and a key component in numerous downstream applications, such as question answering (Yu et al., 2017) and knowledge base population (Ji and Grishman, 2011). Recent neural approaches based on pre-trained language models (e.g., BERT (Devlin et al., 2019)) have shown impressive results for these tasks when fine-tuned on supervised datasets (Akbik et al., 2018; De Cao et al., 2021; Alt et al., 2019). However, annotated datasets for fine-tuning information extraction models are still scarce, even in a comparatively well-resourced language such as German (Benikova et al., 2014), and generally only contain annotations for a single task (e.g., for NER CoNLL’03 German (Tjong Kim Sang and De Meulder, 2003), GermEval 2014 (Benikova et al., 2014);

entity linking GerNED (Ploch et al., 2012)). In addition, research in multi-task (Ruder, 2017) and joint learning (Sui et al., 2020) has shown that models can benefit from exploiting training signals of related tasks. To the best of our knowledge, the work of Schiersch et al. (2018) is the only dataset for German that includes two of the three tasks, namely NER and RE, in a single dataset.

In this work, we present `MobIE`, a German-language information extraction dataset which has been fully annotated for NER, EL, and n-ary RE. The dataset is based upon a subset of documents provided by Schiersch et al. (2018), but focuses on the domain of mobility-related events, such as traffic obstructions and public transport issues. Figure 1 displays an example traffic report with a *Cancelled Route* event. All relations in our dataset are n-ary, i.e. consist of two or more arguments, some of which are optional. Our work expands the dataset of Schiersch et al. (2018) with the following contributions:

- We significantly extend the dataset with 1,686 annotated documents, more than doubling its size from 1,546 to 3,232 documents
- We add entity linking annotations to geolinkable entity types, with references to Open Street Map¹ identifiers, as well as geo-shapes
- We implement an automatic labeling approach using the Snorkel framework (Ratner et al., 2017) to obtain additional high quality, but weakly-supervised relation annotations



The dataset setup allows for training and evaluating algorithms that aim for fine-grained typing of geolocations, entity linking of these, as well as for n-ary relation extraction. The final dataset contains 20,484 entity, 13,104 linking, and 2,036 relation annotations.

¹<https://www.openstreetmap.org/>

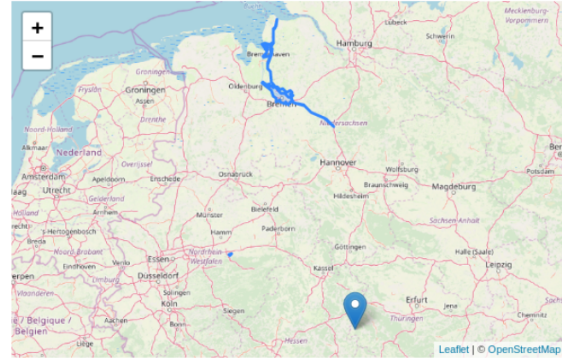
(organization-company) TRI (trigger) LOC (location-route) END-LOC (location-stop) CAUSE (event_cause)
 wikiData:Q60439356 kbld:22477 osmId:62369
 BVG: Zugausfall #S7 nach Potsdam wegen Notarzteinsatz

Figure 1: Traffic report annotated with entity types, entity linking and arguments of a *Canceled Route* event

Geolink annotator

Entity mentions  

Entity label	Text mention	NER type	annotate NER	#Candidates	annotate Candidates
A27	A27	location-street ?	<input checked="" type="checkbox"/> Correct <input type="checkbox"/> Incorrect	2 ?	<input type="button" value="show"/> <input type="button" value="hide"/> <input type="button" value="Missing"/>
Bremerhaven	Bremerhaven	location-city ?	<input checked="" type="checkbox"/> Correct <input type="checkbox"/> Incorrect	1 ?	<input type="button" value="show"/> <input type="button" value="hide"/> <input type="button" value="Missing"/>



A27 Bremerhaven Richtung Bremen die Ausfahrt Bremen-Vahr ist nach einem Unfall gesperrt.

Figure 2: Geolinker: Annotation tool for entity linking

2 Data Collection and Annotation

2.1 Annotation Process

We collected German Twitter messages and RSS feeds based on a set of predefined search keywords and channels (radio stations, police and public transport providers) continuously from June 2015 to April 2019 using the crawlers and configurations provided by Schiersch et al. (2018), and randomly sampled documents from this set for annotation. The documents, including metadata, raw source texts, and annotations, are stored with a fixed document schema as AVRO² and JSONL files, but can be trivially converted to standard formats such as CONLL. Each document was labeled iteratively, first for named entities and concepts, then for entity linking information, and finally for relations. For all manual annotations, documents are first annotated by a single trained annotator, and then the annotations are validated by a second expert. All annotations are labeled with their source, which e.g. allows to distinguish manual from weakly supervised relation annotations (see Section 2.4).

2.2 Entities

Table 3 lists entity types of the mobility domain that are annotated in our corpus. All entity types except for *event_cause* originate from the corpus of Schiersch et al. (2018). The main characteristics of the

original annotation scheme are the usage of coarse- and fine-grained entity types (e.g., *organization*, *organization-company*, *location*, *location-street*), as well as trigger entities for phrases which indicate annotated relations, e.g., “*Stau*” (“*traffic jam*”). We introduce a minor change by adding a new entity type label *event_cause*, which serves as a label for concepts that do not explicitly trigger an event, but indicate its potential cause, e.g., “*technische Störung*” (“*technical problem*”) as a cause for a *Delay* event.

2.3 Entity Linking

In contrast to the original corpus, our dataset includes entity linking information. We use Open Street Map (OSM) as our main knowledge base (KB), since many of the geo-entities, such as streets and public transport routes, are not listed in standard KBs like Wikidata. We link all geo-locatable entities, i.e. *organizations* and *locations*, to their KB identifiers, and external identifiers (Wikidata) where possible. We include geo-information as an additional source of ground truth whenever a location is not available in OSM³. Geo-information is provided as points and polygons in WKB format⁴.

³This is mainly the case for *location-route* and *location-stop* entities, which are derived from proprietary KBs of Deutsche Bahn and Rhein-Main-Verkehrsverbund. Standardized ids for these entity types, e.g. DLID/DHID, were not yet available at the time of creation of this dataset.

⁴<https://www.ogc.org/standards/sfa>

²avro.apache.org

Relation	Arguments
<i>Accident</i>	DEFAULT-ARGS, delay
<i>Canceled Route</i>	DEFAULT-ARGS
<i>Canceled Stop</i>	DEFAULT-ARGS, route
<i>Delay</i>	DEFAULT-ARGS, delay
<i>Obstruction</i>	DEFAULT-ARGS, delay
<i>Rail Repl. Serv.</i>	DEFAULT-ARGS, delay
<i>Traffic Jam</i>	DEFAULT-ARGS, delay, jam-length

Table 1: Relation definitions of the MOBIE dataset. DEFAULT-ARGS for all relations are: location, trigger, direction, start-loc, end-loc, start-date, end-date, cause. Location and trigger are essential arguments for all relations, other arguments are optional.

Figure 2 shows the annotation tool used for entity linking. The tool displays the document’s text, lists all annotated geo-location entities along with their types, and a list of KB candidates retrieved. The annotator first checks the quality of the entity type annotation, and may label the entity as *incorrect* if applicable. Then, for each valid entity the annotator either labels one of the candidates shown on the map as correct, or they select *missing* if none of the candidates is correct.

2.4 Relations

Table 1 lists relation types and their arguments. The relation set focuses on events that may negatively impact traffic flow, such as *Traffic Jams* and *Accidents*. All relations have a set of required and optional arguments, and are labeled with their annotation source, i.e., human or weakly-supervised. Different relations may co-occur in a single sentence, e.g. *Accidents* may cause *Traffic Jams*, which are often reported together.

Human annotation. The annotation in Schierich et al. (2018) is performed manually. Annotators labeled only explicitly expressed relations where all arguments occurred within a single sentence. The authors report an inter-annotator agreement of 0.51 (Cohen’s κ) for relations.

Automatic annotation with Snorkel. To reduce the amount of labor required for relation annotation, we explored an automatic, weakly supervised labeling approach. Our intuition is that due to the formulaic nature of texts in the traffic report domain, weak heuristics that exploit the combination of trigger key phrases and specific location types provide a good signal for relation labeling. For example, “A2 Dortmund Richtung Hannover 2 km Stau” is easily identified as a *Traffic Jam* relation mention due to the occurrence of the “Stau” trigger

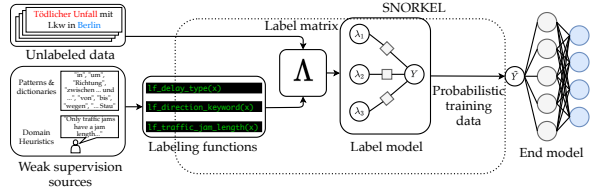


Figure 3: Snorkel applies user-defined, ‘weak’ labeling functions (LF) to unlabeled data and learns a model to reweigh and combine the LFs’ outputs into probabilistic labels.

in combination with the road name “A2”.

We use the Snorkel weak labeling framework (Ratner et al., 2017). Snorkel unifies multiple weak supervision sources by modeling their correlations and dependencies, with the goal of reducing label noise (Ratner et al., 2016). Weak supervision sources are expressed as labeling functions (LFs), and a label model combines the votes of all LFs weighted by their estimated accuracies and outputs a set of probabilistic labels (see Figure 3).

We implement LFs for the relation classification of trigger concepts and role classification of trigger-argument concept pairs. The output is used to reconstruct n-ary relation annotations. Trigger classification LFs include keyword list checks as well as examining contextual entity types. Argument role classification LFs are inspired by Chen and Ji (2009), and include distance heuristics, entity type of the argument, event type output of the trigger labeling functions, context words of the argument candidate, and relative position of the entity to trigger. We trained the Snorkel label model on all unlabeled documents in the dataset that contained at least a *trigger* entity (690 documents). The probabilistic relation type and argument role labels were then combined into n-ary relation annotations.

We verified the performance of the Snorkel model using a randomly selected development subset of 55 documents with human-annotated relations. On this dev set, Snorkel-assigned trigger class labels achieved a F1-score of 80.6 (Accuracy: 93.0), and role labeling of trigger-argument pairs had a F1-score of 72.6 (Accuracy: 83.1). This confirms our intuition that for the traffic report domain, weak labeling functions can provide useful supervision signals.

3 Dataset Statistics

We report the statistics of the MOBIE dataset in Table 2. The majority of documents originate from Twitter, but RSS messages are longer on average,

	Twitter	RSS	Total
# docs	2,562	670	3,232
# sentences	5,409	1,668	7,077
# tokens	62,330	28,641	90,971
# entities	13,573	6,911	20,484
# linked	8,715	4,389	13,104
# events	1,461	575	2,036

Table 2: Dataset statistics per source

and typically contain more annotations (e.g., 10.3 entities/doc versus 5.3 entities/doc for Twitter). The annotated corpus is provided with a standardized *Train/Dev/Test* split. To ensure a high data quality for evaluating event extraction, we include only documents with manually annotated events in the *Test* split.

Table 3 lists the distribution of entity annotations in the dataset, Table 4 the distribution of linked entities. Of the 20,484 annotated entities covering 20 entity types, 13,104 *organization** and *location** entities are linked, either to a KB reference id, or marked as NIL. The remaining entities are non-linkable types, such as time and date expressions. The fraction of NILs among linkable entities is 43.1% overall, but varies significantly with entity type. *Locations* that could not be assigned to a specific subtype are more often resolved as NIL. A large fraction of these are highway exits (e.g. “Pforzheim-Ost”) and non-German locations, which were not included in the subset of OSM integrated in our KB. In addition, candidate retrieval for *organizations* often returned no viable candidates, especially for non-canonical name variants used in tweets.

The dataset contains 2,036 annotated traffic events, 1,280 manually annotated and 756 obtained via weak supervision. Table 5 shows the distribution of relation types. *Canceled Stop* and *Rail Replacement Service* relations occur less frequently in our data than the other relation types, and *Obstruction* is the most frequent class.

4 Conclusion

We presented a dataset for named entity recognition, entity linking and relation extraction in German mobility-related social media texts and traffic reports. Although not as large as some popular task-specific German datasets, the dataset is, to the best of our knowledge, the first German-language dataset that combines annotations for NER, EL and RE, and thus can be used for joint and multi-task learning of these fundamental in-

	Twitter	RSS	Total
date	434	549	983
disaster-type	78	18	96
distance	37	175	212
duration	413	157	570
event-cause	898	116	1,014
location	887	1,074	1,961
location-city	844	1,098	1,942
location-route	2,298	324	2,622
location-stop	1,913	1,114	3,027
location-street	634	612	1,246
money	16	3	19
number	527	198	725
org-position	4	0	4
organization	296	121	417
organization-company	1,843	46	1,889
percent	1	0	1
person	135	0	135
set	18	37	55
time	683	410	1,093
trigger	1,614	859	2,473

Table 3: Distribution of entity annotations

	# entities	# KB	# NIL
location	1,961	703	1,258
location-city	1,942	1,486	456
location-route	2,622	2,138	484
location-stop	3,027	1,898	1,129
location-street	1,246	1,036	210
organization	417	0	417
organization-company	1,889	192	1,697

Table 4: Distribution of entity linking annotations

	Twitter	RSS	Total
Accident	316	80	396
Canceled Route	259	75	334
Canceled Stop	25	42	67
Delay	337	48	385
Obstruction	386	140	526
Rail Replacement Service	71	27	98
Traffic Jam	67	163	230

Table 5: Distribution of relation annotations

formation extraction tasks. The dataset is freely available under a CC-BY 4.0 license at <https://github.com/dfki-nlp/mobie>.

Acknowledgments

We would like to thank Elif Kara, Ursula Strohriegel and Tatjana Zeen for the annotation of the dataset. This work has been supported by the German Federal Ministry of Transport and Digital Infrastructure as part of the project DAYSTREAM (01MD19003E), and by the German Federal Ministry of Education and Research as part of the project CORA4NLP (01IW20010).

References

- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. [Contextual String Embeddings for Sequence Labeling](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. [Improving Relation Extraction by Pre-trained Language Representations](#). In *Proceedings of AKBC 2019*, pages 1–18, Amherst, Massachusetts.
- Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. [NoSta-D Named Entity Annotation for German: Guidelines and Dataset](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1251.
- Zheng Chen and Heng Ji. 2009. [Language specific issue and feature exploration in Chinese event extraction](#). In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 209–212, Boulder, Colorado. Association for Computational Linguistics.
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. [Autoregressive Entity Retrieval](#). In *Proceedings of ICLR 2021*. ArXiv: 2010.00904.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Heng Ji and Ralph Grishman. 2011. [Knowledge Base Population: Successful Approaches and Challenges](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA. Association for Computational Linguistics.
- Danuta Ploch, Leonhard Hennig, Angelina Duka, Ernesto William De Luca, and Sahin Albayrak. 2012. [GerNED: A German corpus for named entity disambiguation](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3886–3893, Istanbul, Turkey. European Language Resources Association (ELRA).
- Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. [Snorkel: Rapid Training Data Creation with Weak Supervision](#). *Proceedings of the VLDB Endowment*, 11(3):269–282. ArXiv: 1711.10160.
- Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. [Data programming: Creating large training sets, quickly](#). In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Sebastian Ruder. 2017. [An Overview of Multi-Task Learning in Deep Neural Networks](#). *arXiv:1706.05098 [cs, stat]*. ArXiv: 1706.05098.
- 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. European Language Resources Association (ELRA).
- Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, Xianrong Zeng, and Shengping Liu. 2020. [Joint Entity and Relation Extraction with Set Prediction Networks](#). *arXiv:2011.01675 [cs]*. ArXiv: 2011.01675 version: 2.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the conll-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2017. [Improved Neural Relation Detection for Knowledge Base Question Answering](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–581, Vancouver, Canada. Association for Computational Linguistics.