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

Anonymous ACL submission

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 3,541 social media texts and traffic reports with 104K tokens, and contains 23,4K annotated entities, 15k 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 *Canceled Route* event. All relations in our dataset are n-ary, i.e. consist of more than 2 arguments, some of which are optional. Our work considerably expands the dataset of Schiersch et al. (2018) with the following contributions:

- We significantly expand the dataset with 1,973 annotated documents, more than doubling its size from 1,568 to 3,541 documents
- We add entity linking annotations to geo-linkable entity types, with references to Open Street Map[1] 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 geo-locations, entity linking of these, as well as for n-ary relation extraction. The final dataset contains $23,392$ entity, $14,931$ linking, and $2,394$ relation annotations.

[1] https://www.openstreetmap.org/

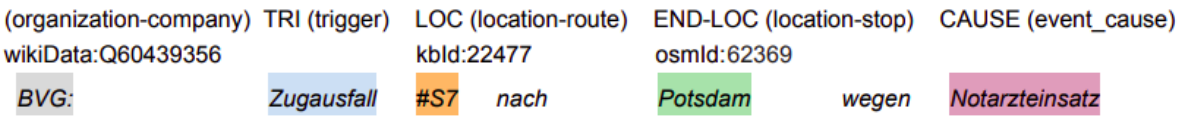

Figure 1: Traffic report annotated with entity types, entity linking and arguments of a *CanceledRoute* event

.

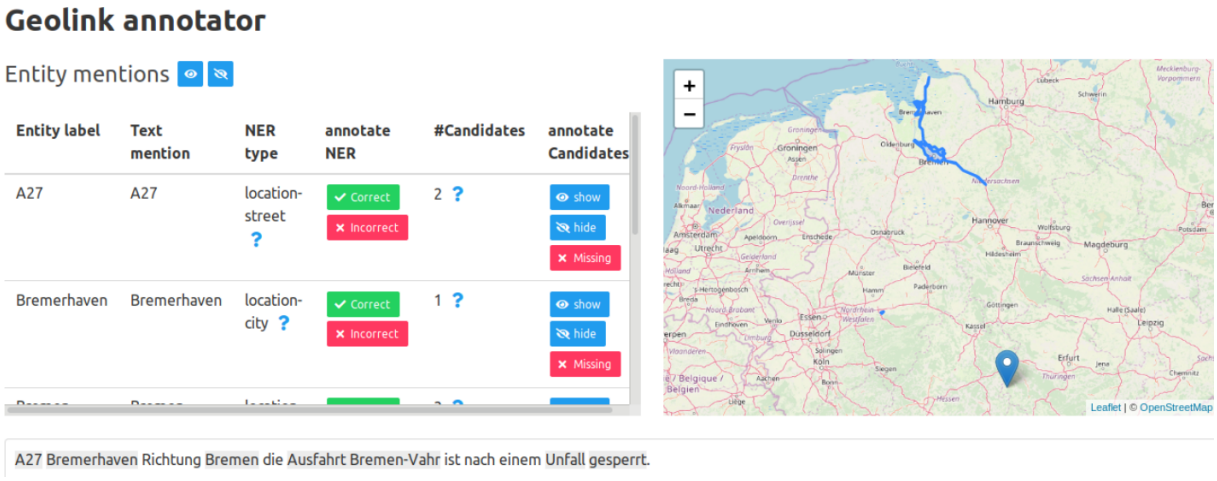

Figure 2: Geolinker: Annotation tool for entity linking

2 Data Collection and Annotation

2.1 Data Collection

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 in two formats with a fixed document schema (AVRO[2] and JSONL), but can trivially be converted to standard formats such as CONLL for NER.

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 small change to this scheme by including a new entity type label *event_cause*. That entity type serves as a replacement for entities 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.

Each document expanding the original corpus was labeled first by a single trained annotator, and then the annotations were validated by one of the authors of the paper.

2.3 Entity Linking

In contrast to the original corpus of Schiersch et al. (2018), 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 unique KB identifiers, and external identifiers (OSM, Wikidata) where possible. We also include geo-information as an additional source of ground truth whenever a location is not available in OSM[3]. Geo-information is provided as points and polygons in WKB format[4].

[2] avro.apache.org

[3] This is in particular the case for *location-route* and *location-stop* entities, which are derived from proprietary KBs. Standardized ids for these entity types like DLID/DHID were not yet available at the time of creation of this dataset.

[4] https://en.wikipedia.org/wiki/Well-known_text_representation_of_geometry

Relation	Arguments
Accident	DEFAULT-ARGS, delay
Canceled Route	DEFAULT-ARGS
Canceled Stop	DEFAULT-ARGS, route
Delay	DEFAULT-ARGS, delay
Obstruction	DEFAULT-ARGS, delay
Rail Repl. Serv.	DEFAULT-ARGS, delay
Traffic Jam	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. Each document was annotated by a single trained annotator, and then the labels were validated by one of the paper's authors.

2.4 Relations and Events

Table 1 lists all annotated relation types, together with their definitions and arguments. The relation set focuses on events that may negatively impact traffic flow, such as e.g. *Traffic Jam*s and *Accident*s. 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 *TrafficJams*, which are often reported together.

Human annotation. The annotation in Schiersch 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 the usefulness of automatic, weakly supervised labeling approaches. Our intuition is that due to the formulaic nature of texts in the traffic report domain, weak heuristics that exploit the contextual combination of trigger key phrases and specific location types, provide a good

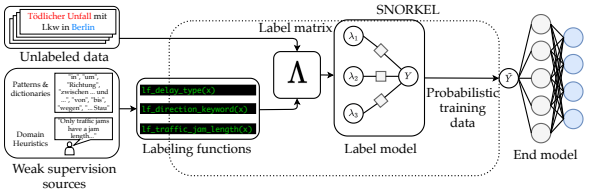

Figure 3: Snorkel applies user-defined, 'weak' labeling functions (LF) to unlabeled data and learns a model to reweigh and combine the LFs' outputs into probabilistic labels.

signal for relation labeling. For example, *"A2 Dortmund Richtung Hannover 2 km Stau"* is easily identified as a *TrafficJam* relation mention due to the occurrence of the *"Stau"* trigger in combination with the road name *"A2"*.

We used Snorkel (Ratner et al., 2017) as our weak labeling framework. 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 the learned label model combines the votes of all labeling functions weighted by their estimated accuracies and outputs a set of probabilistic labels (see Figure 3). We implemented a set of labeling functions for the relation classification of trigger concepts, and role classification of trigger-argument concept pairs. The output of these labeling functions can then be used to reconstruct n-ary relation annotations.

Trigger classification LFs include keyword list checks as well as examining contextual entity types (e.g. *TrafficJam* triggers often co-occur with *distance* entities). 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, among others. In particular, we looked at the entity type to see if it matched an argument role class, e.g. *duration*s for role *delay*, or *distance*s for role *jam-length*. For some of the role classes we looked for typical context words, e.g. *"Richtung"* for the *direction* role, *"von"* and *"bis"* for *start-loc,end-loc*, respectively, and *"wegen"* for *cause*. We also applied regular expression based patterns as in Grusdt et al. (2018), to output the relevant role label given a matching pattern.

We trained the Snorkel label model on all unlabeled documents in the dataset that contained at

	Twitter	RSS	Total
# docs	2,825	716	3,541
# sentences	5,983	2,028	8,011
# tokens	69,188	34,630	103,818
# entities	15,407	8,154	23,561
# linked	9,909	5,034	14,943
# events	1,719	651	2,370

Table 2: Dataset statistics per source.

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, and typically contain more annotations (e.g., 11.3 entities/doc vs 5.4 entities/doc for Twitter). The annotated corpus is provided with a standardized Train/Dev/Test split, in two formats (AVRO[5] and JSONL). To ensure a high data quality for evaluating event extraction, we do not include automatically annotated events in the provided *Test* split of the dataset.

Table 3 list the distribution of entity annotations in the dataset, Table 4 the distribution of linked entities. Of the 23,561 annotated entities covering 20 entity types, 14,943 *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 43.3% 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 our KB. In addition, candidate retrieval for *organization*s often returned no viable candidates, especially for non-canonical name variants used in tweets.

[5] avro.apache.org

	Twitter	RSS	Total
date	487	724	1,211
disaster-type	78	19	97
distance	39	179	218
duration	439	196	635
event-cause	1,062	146	1,208
location	964	1,107	2,071
location-city	967	1,220	2,187
location-route	2,667	431	3,098
location-stop	2,195	1,466	3,661
location-street	699	641	1,340
money	17	3	20
number	564	246	810
org-position	4	0	4
organization	315	121	436
organization-company	2,102	48	2,150
percent	1	0	1
person	135	0	135
set	19	64	83
time	773	505	1,278
trigger	1,880	1,038	2,918

Table 3: Distribution of entity annotations.

	# entities	# KB	# NIL
location	2,071	728	1,343
location-city	2,187	1,646	541
location-route	3,098	2,533	565
location-stop	3,661	2,245	1,416
location-street	1,340	1,116	224
organization	436	0	436
organization-company	2,150	201	1,949

Table 4: Distribution of entity linking annotations.

	Twitter	RSS	Total
Accident	359	80	439
CanceledRoute	334	95	429
CanceledStop	33	56	89
Delay	376	58	434
Obstruction	443	159	602
RailReplacementService	104	38	142
TrafficJam	70	165	235

Table 5: Distribution of relation annotations.

The dataset contains 2,370 annotated traffic events, 992 manually annotated and 1,378 obtained via weak supervision (Table 5). We can see that *CanceledStop* and *RailReplacementService* relations occur less frequently 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 Ger-

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

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.

Britta Grusdt, Jan Nehring, and Philippe Thomas. 2018. Bootstrapping patterns for the detection of mobility related events. In *14th Conference on Natural Language Processing*, pages 50–59. Verlag der Österreichischen Akademie der Wissenschaften.

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 Techologies*, 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, Xiangrong 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.