


Annotation of Fine-Grained Geographical Entities in German Texts

Julián Moreno-Schneider ✉ 

DFKI GmbH, Berlin, Germany

Melina Plakidis ✉

DFKI GmbH, Berlin, Germany

Georg Rehm ✉ 

DFKI GmbH, Berlin, Germany

Abstract

We work on the creation of a corpus, crawled from the internet, on the Berlin district of Moabit, primarily meant for training NER systems in German and English. Typical NER corpora and corresponding systems distinguish persons, organisations and locations, but do not distinguish different types of location entities. For our tourism-inspired use case, we need fine-grained annotations for toponyms. In this paper, we outline the fine-grained classification of geographical entities, the resulting annotations and we present preliminary results on automatically tagging toponyms in a small, bootstrapped gold corpus.

2012 ACM Subject Classification Information systems → Entity resolution; Information systems → Information extraction

Keywords and phrases Named Entity Recognition, Geographical Entities, Annotation

Digital Object Identifier 10.4230/OASICS.LDK.2021.11

Supplementary Material *Collection (Collection of documents about Moabit district annotated with Geographical Entities):* <https://gitlab.com/jmschnei/Moabit-Collection>

Funding The research presented in this paper is funded by the German Federal Ministry of Education and Research (BMBF) through the project QURATOR (<http://qurator.ai>) (Unternehmen Region, Wachstumskern, grant no. 03WKDA1A).

1 Introduction

The amount of information available in digital form is continuously growing, and a significant portion of it is accessed through mobile devices [2]. The fact that such devices are mobile in the first place, and usually also equipped with geolocation functionality, enables providing localised information to their users. A use case that can benefit from customised information, i. e., content tuned to the particular location, is tourism, e. g., an interactive travel guide, bringing points of interest in the vicinity to the user’s attention [14, 15].

While exploiting a user’s geographical location is relatively straightforward (privacy issues aside), combining this with information in textual form is less trivial. Typical corpora annotated for named entities, of which location is usually one of a small number of classes, do not distinguish between more detailed location-type entities, and, consequently, NER taggers do not make this distinction. We argue that for our use case of a travel guide, a more detailed distinction for toponyms, differentiating between, for example, (train/bus) stations, parks, streets and squares, is needed, allowing for more relevant, tailored recommendations. We explore this by defining a semantic classification of fine-grained geographical entities and by annotating a collection of documents accordingly. Our envisioned use case is to semi-automatically create a route or a guided tour, along which the user can explore the Berlin district of Moabit. This use case is to be understood in the context of the project



© Julián Moreno-Schneider, Melina Plakidis, and Georg Rehm;
licensed under Creative Commons License CC-BY 4.0

3rd Conference on Language, Data and Knowledge (LDK 2021).

Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch; Article No. 11; pp. 11:1–11:8

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

QURATOR¹, dealing with *digital curation technologies* [13]; promising results for a similar approach, but for a different domain, have been reported in [8, 9]. In QURATOR, we process large and multi-media document collections and analyse, re-arrange, summarise and visualise information contained in the collections, to generate stories – including the guided tours in our tourism use case – through *semantic storytelling*.

The rest of this paper is structured as follows. Section 2 describes related work focusing on the identification of geographical entities. Section 3 further motivates our work and explains the annotation guidelines. Section 4 describes the data we work with and provides results from first annotation efforts. Section 5 sketches a usage scenario for the annotated collection and reports on preliminary classification results. Finally, Section 6 sums up the main findings and provides pointers to future work.

2 Related Work

Specifically focusing on historical corpora, Won et al. [19] investigate the performance of five different NER systems for toponyms and spatial information. They find that using an ensemble method based on voting performs best. Similar to our envisioned approach, they experiment with combining NER modules with gazetteers. The use case of Alex et al. [1] is similar to ours. They use the Edinburgh Geoparser to tag and resolve fine-grained geographical locations in and around Edinburgh (both historical and contemporary). Also using the Edinburgh Geoparser, Gritta et al. [5] use two corpora (WikToR and LGL [4]) to evaluate five geo-parsers: CLAVIN², Yahoo!PlaceSpotter³, The Edinburgh Parser [6], Topocluster [3] and GeoTxt [7] and conclude that the Edinburgh parser works best for them.

We adopt the idea of combining gazetteers (and also simple pattern-matching based components) with NER modules from earlier approaches, but the key feature that sets our use case apart from the ones mentioned above, is the fact that we include German as a language to detect more fine-grained geographical entities.

With regard to NER in general, we use and evaluate SpaCy⁴, Stanford CRF NER⁵ and an approach based on BERT⁶, to recognise location-type entities. Out of the box and without specific re-training on a corpus annotated for more fine-grained types, they only output (among other types) locations, without any sub-classes.

3 Annotation of Fine-grained Geographical Entities

Many datasets for NER distinguish between persons, organisations and locations, and reserve a miscellaneous/other category for the remaining entities not captured by the former three categories [17, 18, 11]. In order to identify more specific (fine-grained) geographic entities, we need, first, a classification of the types of entities that we want to recognize, and, second, a collection of documents in which the entities are annotated based on that classification, that can be used as training data for new NLP models.

¹ <https://qurator.ai>

² <https://clavin.bericotechnologies.com>

³ <http://boss.yahoo.com>, retrieved 2016-07-31.

⁴ <https://spacy.io>

⁵ <https://nlp.stanford.edu/software/CRF-NER.shtml>

⁶ <https://github.com/kamalkraj/BERT-NER>

3.1 Semantic Classification

To annotate a corpus that further sub-classifies locations, we follow the NoSta-D guidelines [16]. We keep the four main categories of the NoSta-D-TagSet (PER, LOC, ORG, OTH; for persons, organisations, locations and other entities, respectively) and use the subcategories of the LOC category listed in the NoSta-D-TagSet as the foundation of our own fine-grained classification. As a result, a total of 14 fine-grained classes were developed for the LOC class (see Table 1). The other core NEs are kept (PER, ORG, OTH).

Furthermore, we follow the NoSta-D guidelines in that we do not annotate *dates, religions, names of animals, dynasties, cardinal directions, technical terms, salutations or political tendencies*. Additionally, as defined in the NoSta-D guidelines, we annotate categories such as *languages, websites, book/movie titles, wars or currencies* as belonging to the class OTH.

■ **Table 1** Classification of subcategories for named entities, specifically for LOCATION entities.

Category	Label	Description	Examples
City	CTY	Capital cities, major and minor cities and smaller towns.	Berlin, Leipzig
Country	CNY	Based on the United Nations Member States ⁷ .	Deutschland
State	STA	The 16 states (Bundesländer) of Germany.	Bayern, Brandenburg
Address	ADD	Toponym consisting of at least a street name and a number.	Unter den Linden 6, 10117 Berlin
Continent	CON	The seven continents.	Afrika, Australien
Building	BUILD	All buildings that are not considered as particular points-of-interest.	Anne-Frank-Grundschule, Johann von Neumann-Haus
Sight	SIGHT	Particular points-of-interest (i. e., popular sight-seeing locations) ⁸ .	Brandenburger Tor, Siegestssäule
Waters	WTR	Bodies of water such as lakes, rivers and channels.	Spree, Müggelsee
Address-Sub	ADD-SUB	Part of an address, not matching the address description.	Friedrichstraße, Potsdamer Platz
District	DST	Administratively recognised district or area of a city.	Mitte, Neukölln
Station	STN	Any train, bus or subway station.	U Turmstraße, S Friedrichstraße
Park	PARK	Parks and recreational areas.	Tiergarten, Britzer Garten
Shop	SHOP	Restaurants, cafés, bars and shops.	Sapori di Casa, Barcomi's Deli, George R
LOC-oth	LOC-OTH	Places that do not match any of the above.	Moltkebrücke
Org.	ORG	Companies, agencies, institutions, etc.	Apple, Samsung, Google
Person	PER	People, including fictional.	Angela Merkel
Other	OTH	All derived named entities (see NoSta-D NEDeriv), websites, book or movie titles, currencies, eras, languages, wars, etc.	www.google.de, Deutsch, Erster Weltkrieg

We deviate from the NoSta-D guidelines in that we do not use the *part* attribute for entities part of longer, complex tokens and we disregard the category *VLOC* (virtual location). We annotate derived entities (NEderiv in NoSta-D) as *OTH*, and finally we tag *restaurants*, *bars*, *cafes* and *hotels* as belonging to (sub-types of) locations, rather than organisations, because this better suits our tourism-inspired use case.

3.2 Integration of Linked Open Data

Annotated collections of documents are additionally improved through the inclusion of information from Linked Open Data (LOD) sources, to make them suitable for use with Linked Data approaches. For this we use a semi-automatic approach composed of two steps: first, we request two sources of information from which we automatically obtain URLs (looking up the entity on its (language-specific) DBpedia page using DBpedia spotlight [10]) and latitude and longitude (we use a SPARQL query against the Geonames⁹ ontology) of the entities. Second, we extend this automatic process by validating the information obtained (because the correct information is not retrieved in all cases) and we complete manually the information in those cases in which it is missing (either URL or latitude/longitude).

4 Data and Annotation

Since our envisioned use case is in tourism, namely the generation of travel guides or guided tours, we decided to annotate a particular collection of documents, instead of taking a benchmark NER corpus and annotating our fine-grained location classes. The document set we used was collected through focused web crawling. We used Spidey¹⁰ in combination with a list of manually generated seed terms, such as *Moabit*, *Kleiner Tiergarten* (a particular park in Moabit), *Kulturfabrik Moabit* (an event location) and *Kurt Tucholsky* (a German-Jewish author born in Moabit). In total, the complete list contains 28 items – places, buildings or persons – related to Moabit. This returned a list of URLs, which we crawled and boilerplated to extract the content and metadata using Newspaper3k.¹¹

The resulting collection consists of 380 documents in German and 92 documents in English. We first tagged these documents with SpaCy and Stanford CRF-NER and proceeded with the Stanford CRF-NER output. After a manual revision, the documents contained 2682 (for German) and 777 (for English) LOC entities which we use as gold annotations. As a next step, we analysed the manually corrected LOC entities again and annotated them for the sub-classes listed in Section 3.1. The results are included in Table 2.

Examples 1 to 3 show three sentences extracted from the collection where several fine-grained geographical entities are annotated.

► **Example 1.** The site is openly accessible and you can stroll along the river Spree WTR .

► **Example 2.** To get to the AEG turbine factory BUILD from Hauptbahnhof STN , take the TXL OTH bus going to Tegel airport LOC-OTH and get off at Beusselstraße STN .

⁷ <https://www.un.org/en/member-states/>

⁸ Sights are a subcategory of buildings that are considered to be famous by the general population.

⁹ <http://www.geonames.org>

¹⁰ <https://github.com/vikrambajaj22/Spidey-Focused-Web-Crawler>

¹¹ <https://github.com/codelucas/newspaper>

■ **Table 2** Absolute and relative frequency of each sub-class.

LOC sub-class	German		English	
	#	%	#	%
City	778	17.68	246	32
Country	678	15.41	14	2
State	90	2.04	2	<1
Address	295	6.70	13	2
Continent	32	0.73	0	0
Building	40	0.91	57	7
Sight	182	4.14	19	2
Waters	88	2.00	10	1
Address-Sub	251	5.7	37	5
District	441	10.02	32	4
Station	119	2.70	209	27
Park	111	2.52	2	<1
Shop	162	3.68	22	3
LOC-oth	1134	25.77	114	15
Total	2682	100	777	100

► **Example 3.** [...] of the author **Hans Magnus Enzensberger** **PER**, in **Fregestraße 19** **ADD**, as well as in the studio apartment of the author **Uwe Johnson** **PER**, who was staying in the **United States** **CNY**, at **Niedstraße 14** **ADD** in the **Berlin** **CTY** district of **Friedenau** **DST**.

As can be seen in Examples 4 and 5 the annotations still contain ambiguities. Example 4 has “Berlin” annotated as a city, although it could also be considered as incomplete, because the annotation could also include “North-West”. If there are nested entities, we only annotate the longest entity which contains the nested ones.

► **Example 4.** Designed for **Allgemeine Elektrizitäts-Gesellschaft** **ORG** in 1908 and constructed in 1910, it is located in North-West **Berlin** **CTY**, in the district of **Moabit** **DST**, around 3 Kilometers far from **Reichstag** **SIGHT**.

► **Example 5.** Since 1987, a memorial on the **Putlitzbrücke** **LOC-OTH**, which connects the districts of **Moabit** **DST** and **Wedding** **DST**, has commemorated the 30 **Berlin** **CTY** Jews who were deported from the nearby **Moabit** **DST** freight depot.

The annotated collection of documents is stored in a repository,¹² which is private to avoid any licensing issues (access can be granted upon request for research purposes). At the time of writing this paper, we are including the Linked Open Data information, which will be made available in the repository as an extended version of the collection.

¹²<https://gitlab.com/jmschnei/Moabit-Collection>

5 Geographical Entity Analyser

In order to demonstrate the functionality of the annotated document collection, we trained various named entity recognition modules using the collection. The current number of samples per class is too low to train a model on, but through our manual evaluation of the automatically tagged documents in the annotated document collection, we do have gold data for general LOC entities. To establish which approach performs best on this dataset for the coarse LOC entities, we compare three NER systems: SpaCy and Stanford (Section 2), and an approach based on BERT¹³. The SpaCy models are trained on Ontonotes 5 and Common Crawl (English; `en_core_web_md`) and WikiNER and TIGER (German; `de_core_news_md`). The Stanford models are trained on the CoNLL 2003 data [18]. BERT-NER is trained on WikiNER [11]. The results are shown in Table 3.

■ **Table 3** Results for LOC entity recognition on the annotated document collection.

		Precision	Recall	F1
SpaCy	German	54.56	80.05	65.05
	English	77.94	54.57	64.19
Stanford	German	91.51	58.74	71.55
	English	84.75	50.06	62.95
BERT-NER	German	55.56	81.09	65.97
	English	70.71	59.97	64.90

Given that the Stanford output is the basis for our manual annotation, we expect a bias toward this system, and indeed we see that for German, this system outperforms the other two by approx. 6 points in F-score. For English however, the BERT-NER system performs best, though the difference with the other two is much smaller. Furthermore, having the initial character in uppercase is generally a distinguishing feature of named entities (and consequently a strong feature for many NER systems), but this indicator is not as strong in German, since nouns are by default in upper case. Still, the Stanford system performs considerably better for German than for English. The other systems do not exhibit the same disparity, and in fact perform better on English than on German. We consider looking into this an important venue for future work. Once we have annotated more data for the particular sub-classes, based on these intermediate results, we plan to train fine-grained modules, and perform a new comparison.

6 Conclusions and Future Work

We develop a dataset annotated with fine-grained geographical named entities (toponyms) that can be used to train named entity recognition modules that identify location-type entities in text documents. Regular NER modules typically distinguish four classes of entities (persons, organisations, locations, other). In our guided tour use case, knowledge about more detailed sub-classes allows for more relevant content and recommendations. In this first stage, we automatically tag a corpus crawled specifically for our use case and manually correct the output to obtain gold annotations for fine-grained entity types (i. e., bootstrap a small, gold

¹³<https://github.com/kamalkraj/BERT-NER>

corpus). Currently the dataset encompasses 2,682 (for German) and 777 (for English) LOC type entities. We report on performance for three NER systems on this dataset, although only using the coarse LOC type due to the size of the corpus.

In terms of future work, we plan to increase the volume of annotated data, both by annotating the remaining section of our crawled corpus and by double-annotating at least part of it to obtain inter-annotator figures. Once the corpus is in a more definitive state, we will examine how to make it available through the European Language Grid [12]. We will evaluate our model on the more detailed sub-classes using this final version of the corpus.

References

- 1 Beatrice Alex, Claire Grover, Richard Tobin, and Jon Oberlander. Geoparsing historical and contemporary literary text set in the City of Edinburgh. *Language Resources and Evaluation*, 53(4):651–675, 2019.
- 2 Irma Arts, Anke Fischer, Dominic Duckett, and René van der Wal. Information technology and the optimisation of experience – The role of mobile devices and social media in human-nature interactions. *Geoforum*, 122:55–62, 2021. doi:10.1016/j.geoforum.2021.03.009.
- 3 Grant DeLozier, Jason Baldrige, and Loretta London. Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI’15, page 2382–2388. AAAI Press, 2015.
- 4 Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. A Pragmatic Guide to Geoparsing Evaluation. *CoRR*, abs/1810.12368, 2018. arXiv:1810.12368.
- 5 Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. What’s missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623, 2018.
- 6 Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 368:3875–89, August 2010. doi:10.1098/rsta.2010.0149.
- 7 Morteza Karimzadeh, Wenyi Huang, Siddhartha Banerjee, Jan Oliver Wallgrün, Frank Hardisty, Scott Pezanowski, Prasenjit Mitra, and Alan M. MacEachren. GeoTxt: A Web API to Leverage Place References in Text. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, GIR ’13, page 72–73, New York, NY, USA, 2013. Association for Computing Machinery. doi:10.1145/2533888.2533942.
- 8 Elena Leitner, Georg Rehm, and Julián Moreno-Schneider. Fine-grained Named Entity Recognition in Legal Documents. In Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, editors, *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019)*, number 11702 in Lecture Notes in Computer Science, pages 272–287, Karlsruhe, Germany, 2019. Springer. 10/11 September 2019.
- 9 Elena Leitner, Georg Rehm, and Julián Moreno-Schneider. A Dataset of German Legal Documents for Named Entity Recognition. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odiijk, and Stelios Piperidis, editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France, May 2020. European Language Resources Association (ELRA). Accepted for publication. Submitted version available as preprint.
- 10 Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia spotlight: shedding light on the web of documents. In Chiara Ghidini, Axel-Cyrille Ngonga Ngomo, Stefanie N. Lindstaedt, and Tassilo Pellegrini, editors, *I-SEMANTICS*, ACM International Conference Proceeding Series, pages 1–8. ACM, 2011. URL: <http://dblp.uni-trier.de/db/conf/i-semantics/i-semantics2011.html#MendesJGB11>.

- 11 Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. Learning Multilingual Named Entity Recognition from Wikipedia. *Artif. Intell.*, 194:151–175, 2013. doi:10.1016/j.artint.2012.03.006.
- 12 Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajic, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiljevs, Orians Anvari, Andis Lagzdīņš, Jūlija Meļņika, Gerhard Backfried, Erinc Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampfer, Dorothea Thomas-Aniola, José Manuel Gómez Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. European Language Grid: An Overview. In Nicoletta Calzolari et al., editor, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3359–3373, Marseille, France, 2020. European Language Resources Association (ELRA).
- 13 Georg Rehm, Peter Bourgonje, Stefanie Hegele, Florian Kintzel, Julián Moreno Schneider, Malte Ostendorff, Karolina Zaczynska, Armin Berger, Stefan Grill, Sören Räuchle, Jens Rauenbusch, Lisa Rutenburg, André Schmidt, Mikka Wild, Henry Hoffmann, Julian Fink, Sarah Schulz, Jurica Seva, Joachim Quantz, Joachim Böttger, Josefine Matthey, Rolf Fricke, Jan Thomsen, Adrian Paschke, Jamal Al Qundus, Thomas Hoppe, Naouel Karam, Frauke Weichhardt, Christian Fillies, Clemens Neudecker, Mike Gerber, Kai Labusch, Vahid Rezaezhad, Robin Schaefer, David Zellhöfer, Daniel Siewert, Patrick Bunk, Lydia Pintscher, Elena Aleynikova, and Franziska Heine. QURATOR: Innovative Technologies for Content and Data Curation. In Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus, and Lydia Pintscher, editors, *Proceedings of QURATOR 2020 – The conference for intelligent content solutions*, Berlin, Germany, 2020. CEUR Workshop Proceedings, Volume 2535. 20/21 January 2020.
- 14 Georg Rehm, Karolina Zaczynska, Peter Bourgonje, Malte Ostendorff, Julián Moreno-Schneider, Maria Berger, Jens Rauenbusch, André Schmidt, Mikka Wild, Joachim Böttger, Joachim Quantz, Jan Thomsen, and Rolf Fricke. Semantic Storytelling: From Experiments and Prototypes to a Technical Solution. In Tommaso Caselli, Eduard Hovy, Martha Palmer, and Piek Vossen, editors, *Computational Analysis of Storylines: Making Sense of Events*. Cambridge University Press, 2021. In print.
- 15 Georg Rehm, Karolina Zaczynska, Julián Moreno Schneider, Malte Ostendorff, Peter Bourgonje, Maria Berger, Jens Rauenbusch, André Schmidt, and Mikka Wild. Towards Discourse Parsing-inspired Semantic Storytelling. In Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus, and Lydia Pintscher, editors, *Proceedings of QURATOR 2020 – The conference for intelligent content solutions*, Berlin, Germany, 2020. CEUR Proc., Vol. 2535. 20/21 Jan. 2020.
- 16 M. Reznicek. Linguistische Annotation von Nichtstandardvarietäten: Guidelines und Best Practices: Guidelines NER. Technical report, Humboldt-Universität zu Berlin, September 2013.
- 17 Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 Shared Task: Language-Independent NER. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan, 2002.
- 18 Erik F. Tjong Kim Sang and Fien De Meulder. 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 - Volume 4*, CONLL '03, page 142–147, USA, 2003. Association for Computational Linguistics. doi:10.3115/1119176.1119195.
- 19 Won, Miguel and Murrieta-Flores, Patricia and Martins, Bruno. Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, 5:2, 2018. doi:10.3389/fdigh.2018.00002.