Sebastian, Basti, Wastl?!
Recognizing Named Entities in Bavarian Dialectal Data

Siyao Peng, Zihang Sun, Huangyan Shan, Marie Kolm,
Verena Blaschke, Ekaterina Artemova, Barbara Plank

MaiNLP & MCML, LMU Munich, Germany

LREC-COLING 2024
May 22, 2024

Named Entity Recognition (NER) is a fundamental task.

Problem: Lack of high-quality annotations on non-standard language varieties.

WikiAnn (Pan et al., 2017) silver annotations on 282 languages/dialects, including Bavarian, but on unnatural texts.

bar · 300 rows

Search this dataset

tokens
sequence

```
[ "'", "''", "Mongolei", "''", "'" ]
[ "Dieter", "Hildebrandt", "(", "seit", "2005", ")" ]
[ "'", "''", "Heiliges", "Remisches", "Reich", "''", "'" ]
[ "Weiterleitung", "Josua", "(", "Buach", ")" ]
[ "***", "'", "''", "Ortenberg", "''", "'" ]
[ "'", "''", "Kina", "''", "'" ]
[ "Israel", "Kamakawiwo'ole", "-", "Wejdbekannta", "Sänga" ]
[ "Clark", "County", ",", "Indiana" ]
[ "**", "Rudolph", "Moshammer", ",", "1940-2005" ]
[ "***", "'", "''", "Holland", "''", "'" ]
[ "WEITERLEITUNG", "Gila", "County", ",", "Arizona" ]
[ "X", "Japan", "(", "1997", "aufgelöst", ",", "2007", "/",
"08", "wiedavaeinigt", ")" ]
[ "**", "'", "''", "Toskana", "''", "'" ]
[ "Weiterleitung", "Johannes", "XXIII", "." ]
```

Paper Contributions

Named Entity Recognition Dialects Bavarian German

- The first dialectal NER dataset for German, BarNER;
- 161K tokens annotated on Bavarian Wikipedia articles (*bar-wiki*) and tweets (*bar-tweet*);
- Comparing lexical distribution, syntactic construction, and entity information with 3 German NER datasets on *wiki*, *tweet* and *news*;
- Incorporating German datasets to improve BarNER parsing;
- Multi-task learning with Bavarian-German Dialect Identification.

BarNER

First manually annotated NER dataset
on a German dialect – Bavarian

Annotations and guidelines available:
`https://github.com/mainlp/BarNER`

BarNER – Guidelines

CoNLL06 style (Tjong Kim Sang and De Meulder, 2003)

- PERson, LOCation, ORGanization, MISCellaneous;

GermEval 2014/NoSta-D style (Benikova et al., 2014)

- -deriv/-part for nominal derivation and compounding;
- *Italienroas*$_{\text{LOCpart}}$ 'tour of Italy';
- *eiropäischn*$_{\text{LOCderiv}}$ 'European';

Others

- LANGuage, RELIGION, EVENT, work-of-art (WOA);
- Only flat and named entities, excluding common nouns, pronouns, overlapping, or nested NEs.

Tagset normalized to CoNLL06 for analyses and experiments.

BarNER – Genres

Wikipedia articles

- Carefully written and consistently updated;
- `https://bar.wikipedia.org/wiki/Wikipedia:Hoamseitn`.

Twitter (X) tweets

- noisier, less formal, and more dynamic;
- Snowballed from a list of 17 Bavarian 'seed users' (`http://indigenoustweets.com/bar/`) to their friends;
- Manually classified into *bar*/*de*/*other*/*NA* and only kept *bar*-items;
- Hashtags ($[\#minga]_{\text{LOC}}$) and emojis ($[\text{🇩🇪}]_{\text{LOC}}$) are annotated.

BarNER – Inter-Annotator Agreement (IAA)

- Three graduate students took five months to annotate BarNER;
- 53% of BarNER are **double annotated** for disagreement studies;
- 85+ typed span F1s;
- Entity span detection is harder for tweets but entity typing is easier.

BarNER – Statistics

- Both *bar-wiki* and *bar-tweet* reach 75K+ tokens;
- *bar-tweet* has much fewer entities due to informality and length.

Corpus	#Toks	#Sents	#Ents	Ents/Tok (in %)
bar-wiki	75.7k	3.6k	4.2k	5.5
bar-tweet	86.1k	7.5k	2.5k	2.9

Named Entities Diverge

Bavarian vs. German

$>$

Wikipedia vs. Tweets vs. News

Comparisons – Five German & Bavarian Datasets

	Bavarian	German
wiki	*bar-wiki*	*de-wiki* the wiki portion of NoSta-D (Benikova et al., 2014)
tweet	*bar-tweet*	*de-tweet* MobIE transportation tweets (Hennig et al., 2021)
news	/	*de-news* CoNLL 2006 news (Tjong Kim Sang and De Meulder, 2003)

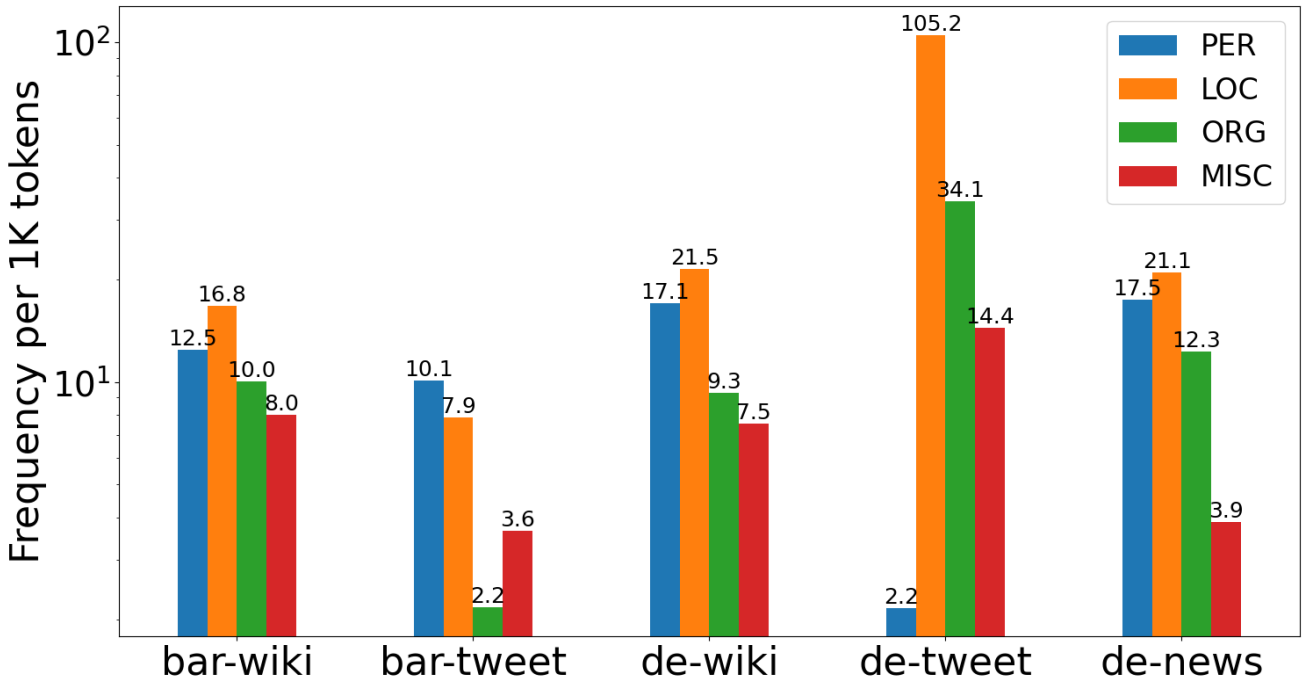
Comparisons – Lexical Similarities

We use **Jaccard Similarity (JS)**:

- $\frac{\text{shared (i.e., intersection) tokens}}{\text{concatenated (i.e., union) tokens}}$ between datasets;
- Compare surface strings to preserve variations;
- German *wiki*×*news* highest 0.417 – formality and well-editedness;
- Similar between tweets and other same-dialect genres:
  - DE *tweet*×*news* 0.229;
  - DE *tweet*×*wiki* 0.195;
  - BAR *tweet*×*wiki* 0.181.

Comparisons – Entity Type Distributions

- Frequencies of NE types per 1K tokens (log-scaled);
- *bar-wiki* and *de-wiki*: similar type distributions;
- *de-tweet*: extreme LOC outlier (105.2) – many routes/streets/cities;
- *bar-tweet*: least entities, esp. ORG; PER>LOC – personal chats.



Comparisons – Top Entities

- Shared entities: *Deutschland* or *Deitschland* 'Germany';
- Most common city names differ between dialects: *Minga* 'Munich' in Bavarian vs. *Berlin* and *Frankfurt* in German;
- *bar-wiki*: document titles;
- *de-wiki*: city and country names;
- *bar-tweet*: tweet friends' names;
- *de-tweet*: railway lines, *#S3*, *S3*;
- *de-news*: currency *Mark* and political parties *SPD* or *CDU*.

bar-wiki

de-wiki

bar-tweet

de-tweet

de-news

Comparisons – Annotators' Observations

Person entities:

- Family names come before given names,
  e.g., *Dreßen* is the family name in *Dreßen Thomas*;
- Shortened given names with diminutive suffixes (e.g., *-l*),
  Sebastian becomes *Basti* or *Wastl*;
- Given names are typically preceded by definite articles,
  e.g., *d'Maria* and *da Michel*.

Possessive constructions:

the genitive determiner in German is replaced by combining preposition *vo* 'from' with a dative determiner:

- 'Association of National Olympic Committees' in English
- *Vaeinigung vo de Nationoin Olympischn Komitees* in Bavarian
- *Vereinigung der Nationalen Olympischen Komitees* in German

NER results on Bavarian

Cross-domain, sequential, and joint training with German

Multi-task learning with dialect identification

Experiments – Setup

- MaChAmp (van der Goot et al., 2021) with masked CRF decoder;
- Datasets: five tag-normalized German and Bavarian datasets;
- German GBERT (Chan et al., 2020)
  `https://huggingface.co/deepset/gbert-large`;
- Multilingual XLM (Conneau et al., 2020)
  `https://huggingface.co/xlm-roberta-large`;
- 3-run average on Span F1.

Experiments – In-domain

- More difficult on ***bar-wiki***, ***bar-tweet***, and ***de-tweet*** – smaller datasets and non-mainstream variations (Bavarian and/or tweet);
- XLM-R for later experiments – higher F1s on BAR.

In-domain	***bar-wiki***	***bar-tweet***	*de-wiki*	***de-tweet***	*de-news*
			Corpus Statistics		
#TrainToks	61.4K	71.8K	232.4K	47.0K	207.0K
#TrainEnts	2.7K	1.6K	12.9K	7.3K	10.0K
			In-domain Results		
XLM-R	**72.91**	**77.55**	85.67	77.14	88.35
GBERT	72.17	73.30	**86.68**	**79.75**	**90.23**

Experiments – Out-of-domain (OOD)

- Models trained on the larger *de-wiki* and *de-news* perform badly on *bar-wiki*, *bar-tweet*, and *de-tweet*;
- Models trained on smaller but in-domain Bavarian data are better;
- However, cross-genre degradations between *de-wiki* and *de-news* are relatively small.

Experiments – Sequential

Motivation: *bar-wiki/tweet* suffer from smaller training size.

Sequential: train on another dataset → train+evaluate on target.

- Improved performances on all five datasets;
- *de-wiki* → *bar-wiki* – same genre but more data;
- *bar-wiki* → *bar-tweet* – same dialect but denser entities;
- *de-tweet* → *de-wiki* & *de-news* – topic-heavy entities;
- All other BAR/DE datasets → *de-news* – more diverse data.

Another\Target	*bar-wiki*	*bar-tweet*	*de-wiki*	*de-tweet*	*de-news*
bar-wiki	—	**79.27**	-	**77.26**	+
bar-tweet	-	—	-	-	+
de-wiki	**73.67**	+	—	-	+
de-tweet	-	-	**86.08**	—	**88.89**
de-news	-	+	-	-	—
In-domain	72.91	77.55	85.67	77.14	88.35

Experiments – Joint

Joint: train on all five → dev/test on target.
Joint+seq: train on five → train on target → dev/test on target.

- Joint improves vastly on *bar-wiki* 8.82 ↑ & mildly on *bar-tweet* 0.62 ↑;
- Joint+seq improves *bar-wiki* by another 2.36 ↑.

	bar-wiki	*bar-tweet*	*de-wiki*	*de-tweet*	*de-news*
joint	**81.73**	**78.17**	**85.89**	-	-
joint+seq	**84.09**	+	-	-	**88.67**
In-domain	72.91	77.55	85.67	77.14	88.35

Experiments – Multi-Task Learning (MTL)

Dialect Identification (DID): Classifying *tweet* and *wiki* as BAR or DE.

MTL with 5 NER and 2 DID tasks: Vastly improve *bar-wiki* 11.26 ↑.

Overall best results:

- *bar-wiki* by multi-task and other four by sequential;
- Still, Bavarian and tweets are more difficult.

	bar-wiki	*bar-tweet*	*de-wiki*	*de-tweet*	*de-news*
In-domain	72.91	77.55	85.67	77.14	88.35
Best model	multi-task	seq-bar-wiki	seq-de-tweet	seq-bar-wiki	seq-de-tweet
Improvement	11.26↑	1.72↑	0.41↑	0.12↑	0.54↑
Final result	84.17	79.27	86.08	77.26	88.89

Conclusion & Future Work

Conclusion

- BarNER – manually annotated named entity corpus for Bavarian;
- Lexical and entity-level distinctions between DE and BAR;
- SOTA results from sequential training and multi-task learning;
- Diversity – genres, topics, and dialects – helps BAR and DE.

Future Work

- Alignments between mainstream languages and dialects;
- More fine-grained sub-dialectal (sub-regional) variations;
- Translation- vs. transfer-based approaches in dialectal NLP;
- Call for more dialectal datasets.

More Bavarian at LREC-COLING 2024:
Now @ Poster I: Slot Intent Detection (Winkler et al., 2024)
Fri 11am @ Poster I: Universal Dependencies (Blaschke et al., 2024)

Questions?
Comments?

Paper

Siyao Logan Peng
MaiNLP & MCML, LMU Munich
siyaopeng@cis.lmu.de

This project is supported by ERC Consolidator Grant DIALECT 101043235.

References I

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. *LREC* (2014).

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank. arXiv:2403.10293 [cs.CL]

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 6788–6796. `https://doi.org/10.18653/v1/2020.coling-main.598`

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8440–8451. `https://doi.org/10.18653/v1/2020.acl-main.747`

Leonhard Hennig, Phuc Tran Truong, and Aleksandra Gabryszak. 2021. MobIE: A German Dataset for Named Entity Recognition, Entity Linking and Relation Extraction in the Mobility Domain. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*. KONVENS 2021 Organizers, Düsseldorf, Germany, 223–227. `https://aclanthology.org/2021.konvens-1.22`

References II

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1946–1958. https://doi.org/10.18653/v1/P17-1178

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*. 142–147. https://aclanthology.org/W03-0419

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive Choice, Ample Tasks (MaChAmp): A Toolkit for Multi-task Learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 176–197. https://doi.org/10.18653/v1/2021.eacl-demos.22

Miriam Winkler, Virginija Juozapaityte, Rob van der Goot, and Barbara Plank. 2024. Slot and Intent Detection Resources for Bavarian and Lithuanian: Assessing Translations vs Natural Queries to Digital Assistants. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.