Different Tastes of Entities: Investigating Human Label Variation in Named Entity Annotations

Siyao Logan Peng   Zihang Sun   Sebastian Loftus   Barbara Plank

MaiNLP lab, CIS, LMU Munich, Germany

Unimplicit @ EACL 2024
March 21, 2024

Named Entity (NE) annotations differ between versions. Are they:

Annotation Errors?

Guideline Differences?

Text Ambiguity?

Named Entity (NE) annotations differ between versions. Are they:

Annotation Errors?

Guideline Differences?

Text Ambiguity?

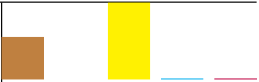
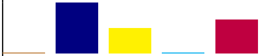
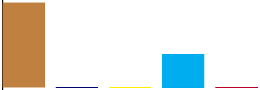
	Sentence	PER LOC ORG MISC O
a.	UK bookmakers [William Hill] said on Friday they ... *original, conllpp, reiss: PER clean: ORG*	
b.	[ALPINE] SKIING – WOMEN 'S WORLD CUP ... *original, conllpp, reiss: O clean: LOC*	
c.	I bear witness that there is no [God] . *original, conllpp, reiss: PER clean: MISC*	

Table 1: CoNLL 2003 Named Entity annotations (`original`), subsequent revisions (`conllpp`, `reiss`, `clean`), and distributions of student annotations.

This work

- Human Label Variation (HLV) are linguistically debatable cases where multiple labels are acceptable (Plank, 2022);
- We study HLV in *expert-annotated NE data* in 3 language(s) variants: *English, Danish, Bavarian*;
- HLV among *iterative published revisions* vs. *independent annotators*.

Datasets & Preprocessing

English:

- `original`: Tjong Kim Sang and De Meulder (2003);
- `conllpp`: Wang et al. (2019)'s revision;
- `reiss`: Reiss et al. (2020)'s revision;
- `clean`: Rücker and Akbik (2023)'s revision.

Danish:

- `plank`: Plank et al. (2020)'s original;
- `hvingelby`: Hvingelby et al. (2020)'s revision.

Bavarian: disagreements between two annotators on the double-annotated subset of BarNER (Peng et al., 2024).

Preprocessing: token alignment and tagset normalization.

Entity-level Disagreements

Four disagreement types – adapted Reiss et al. (2020)'s error types:

- `Tag`: same span but different tags

  [a b]$_{LOC}$ vs. *[a b]*$_{ORG}$

- `Span`: different overlapping spans but same tag,

  [a b]$_{LOC}$ vs. *[a]*$_{LOC}$ *b*

- `Both`: overlapping spans with different tags,

  [a b]$_{LOC}$ vs. *[a]*$_{ORG}$ *b*

- `Missing`: one annotator misses the entity completely

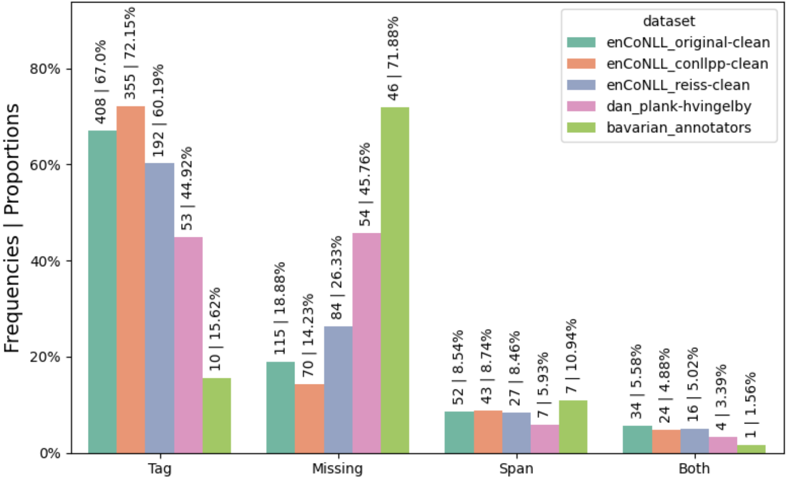
  [a b]$_{LOC}$ vs. *a b*

Five comparison pairs:

- EN `original-clean`, `conllpp-clean`, `reiss-clean`;
- DA `plank-hvingelby`;
- BAR annotators.

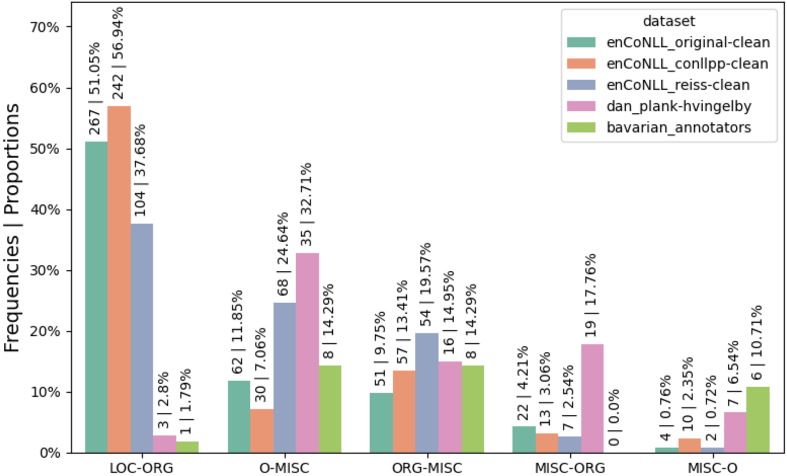
Tag and Missing disagreements are prevalent

- Tag contributes most to English revisions;
- Danish and Bavarian contain more Missing;
- Tag+Missing accounts for 85%+ disagreements in all comparisons.



Top disagreed label pairs in Tag+Missing

- LOC-ORG, O-MISC, ORG-MISC most frequent (70%+) in English;
- Most (80%+) of Danish concern MISC, e.g., ships & clubs;
- Missing (O) annotations donate the majority (70%+) to Bavarian.



Sources of Disagreements

Three sources – adapted Jiang and de Marneffe (2022) – to understand which factors triggered these disagreements:

- *Text ambiguity*: uncertainty in sentence meaning with or without sufficient context;
- *Guideline update*: NE type definitions vary across guideline versions;
- *Annotator error*: attention slip or knowledge gap errors.

Case study

- Manually annotated a small sample of disagreed NE pairs;
- EN: 200 `original-clean` test; all in DA (118) & BAR (64) test;
- IAA (Kappa) on 50 `original-clean` NEs: 61.73%.

Sources of Disagreements – Observations

- Difficult: lack of information as *annotator error* or *text ambiguity*;
- EN: most (80.0%) are *guideline update*: `clean` is more context-free;
- DA: 52.5% *guideline updates*, LOC/MISC; 41.5% *annotator errors*;
- BAR: 67.2% *annotator error*; but acceptable by some EN guidelines.

Source types	English		Danish		Bavarian	
text ambiguity	19	9.5%	7	6.0%	10	15.6%
guideline update	160	80.0%	62	52.5%	11	17.2%
annotator error	21	10.5%	49	41.5%	43	67.2%
Total	200	100.0%	118	100.0%	64	100.0%

Table 2: Distributions of sources of disagreements.

Classroom Surveyed Annotations

Label variation surfaces in classroom-surveyed annotations.

Sentence	PER	LOC	ORG	MISC	O	*abstained*
LA CLIPPERS AT [NEW YORK]		14 original conllpp reiss	0 clean			
[White House] spokesman Mike McCurry said ...		2 original conllpp reiss	11 clean	1		
The man who kicked [Australia] to defeat with ...		5 original conllpp reiss	9 clean			
The granddaughter of Italy 's [Fascist] dictator			3	3 clean	8 original conllpp reiss	
at about 1:30 A.M. [EST]				10 clean	2 original conllpp reiss	2

Table 3: Classroom surveyed annotations on difficult disagreement cases in CoNLL03 test.

Conclusion & Future Work

Conclusion

- *NE disagreements* across *expert revisions* vs. *individual annotations*;
- *Guideline updates* and *text ambiguities* lead established EN+DA;
- *Annotator errors* dominate the new BAR corpus;
- *Student-surveyed annotations* help distributional analyses.

Ongoing & Future Work

- More statistically meaningful (larger) student-surveyed annotations;
- Separating valid label variations from true errors (Weber-Genzel et al., 2024);
- More dialectal datasets (Blaschke et al., 2024; Peng et al., 2024);
- Inform models regarding conflicts among annotation guidelines.

Questions?
Comments?

Paper

Siyao Logan Peng
MaiNLP lab, CIS, LMU Munich
siyaopeng@cis.lmu.de

This project is supported by ERC Consolidator Grant DIALECT 101043235.

References I

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]

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A Named Entity Resource for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 4597–4604. https://aclanthology.org/2020.lrec-1.565

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating Reasons for Disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics* 10 (2022), 1357–1374. https://doi.org/10.1162/tacl_a_00523 Place: Cambridge, MA Publisher: MIT Press.

Siyao Peng, Zihang Sun, Huangyan Shan, Marie Kolm, Verena Blaschke, Ekaterina Artemova, and Barbara Plank. 2024. Sebastian, Basti, Wastl?! Recognizing Named Entities in Bavarian Dialectal Data. arXiv:2403.12749 [cs.CL]

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 10671–10682. https://doi.org/10.18653/v1/2022.emnlp-main.731

References II

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish Nested Named Entities and Lexical Normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6649–6662. `https://doi.org/10.18653/v1/2020.coling-main.583`

Frederick Reiss, Hong Xu, Bryan Cutler, Karthik Muthuraman, and Zachary Eichenberger. 2020. Identifying Incorrect Labels in the CoNLL-2003 Corpus. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, Raquel Fernández and Tal Linzen (Eds.). Association for Computational Linguistics, Online, 215–226. `https://doi.org/10.18653/v1/2020.conll-1.16`

Susanna Rücker and Alan Akbik. 2023. CleanCoNLL: A Nearly Noise-Free Named Entity Recognition Dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8628–8645. `https://aclanthology.org/2023.emnlp-main.533`

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`

References III

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. CrossWeigh: Training Named Entity Tagger from Imperfect Annotations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 5154–5163. `https://doi.org/10.18653/v1/D19-1519`

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating Annotation Error from Human Label Variation. arXiv:2403.01931 [cs.CL]