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

Siyao Peng[*Λ] Zihang Sun[*] Sebastian Loftus[*] Barbara Plank[*Λ]
[*]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[Λ]Munich Center for Machine Learning (MCML), Munich, Germany
siyaopeng@cis.lmu.de

Introduction

- We study Human Label Variation (HLV) (Plank, 2022) in *expert-annotated NE data* in 3 languages: *English*, *Danish*, *Bavarian*;
- HLV among *iterative published revisions* vs. *independent annotators*;
- *Text ambiguity* and *guideline change* dominate HLV;
- *Student-surveyed annotations* help distributional analyses.

Related Work

Human Label Variation (HLV)

- Linguistically debatable cases where multiple labels are acceptable;
- Gives insights into label distribution and annotators' preferences;
- No HLV analysis on expert-labeled NEs.

Noise in NE datasets

- Noise in `original` English CoNLL 2003 data (Tjong Kim Sang and De Meulder, 2003) $\geq$ error rates of SOTA models;
- Multiple revisions: `conllpp` (Wang et al., 2019), `reiss` (Reiss et al., 2020), and `clean` (Rücker and Akbik, 2023);
- Many are guideline updates, and 2.34% entities remain ambiguous.

Dataset & Preprocessing

- **English**: Manually align test tokens of `original`, `conllpp`, `reiss`, and `clean` -- 46,738 tokens and 5,629, 5,683, 5,636, 5,725 NEs;
- **Danish**: `plank`'s (Plank et al., 2020) test annotation on DDT and `hvingelby`'s (Hvingelby et al., 2020) re-annotation;531 & 564 NEs;
- **Bavarian**: disagreements between two annotators; ~400 NEs.

Entity-level Disagreements

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

- `Tag`: same span but different tags, e.g., *[a b]*$_{LOC}$ vs. *[a b]*$_{ORG}$;
- `Span`: diff. overlapping spans but same tag, *[a b]*$_{LOC}$ vs. *[a]*$_{LOC}$ *b*;
- `Both`: overlapping spans with diff. tags, *[a b]*$_{LOC}$ vs. *[a]*$_{ORG}$ *b*;
- `Missing`: one annotator misses the entity, *[a b]*$_{LOC}$ vs. *a b*.

`Tag` and `Missing` are prevalent -- see Figure 1

- Five paired comparisons: EN `original-clean`, `conllpp-clean`, `reiss-clean`, DA `plank-hvingelby`, and BAR annotators;
- `Tag` contributes most to English revisions;
- Danish and Bavarian contain more `Missing`;
- `Tag+Missing` accounts for 85%+ disagreements in all comparisons.

Top 5 disagreed label pairs in `Tag+Missing` -- see Figure 2

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

Sources of Disagreements

Three sources (adapted Jiang and de Marneffe 2022):

- *Text ambiguity*: uncertainty in sentence meaning with(out) context;
- *Guideline update*: NE type definitions vary across guideline versions;
- *Annotator error*: attention slip or knowledge gap errors.

Case study setup

- 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%.

Observations -- see Table 1

- 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*, e.g., LOC/MISC; and *annotator errors*;
- BAR: 67.2% *annotator error*; but acceptable by some EN guidelines.

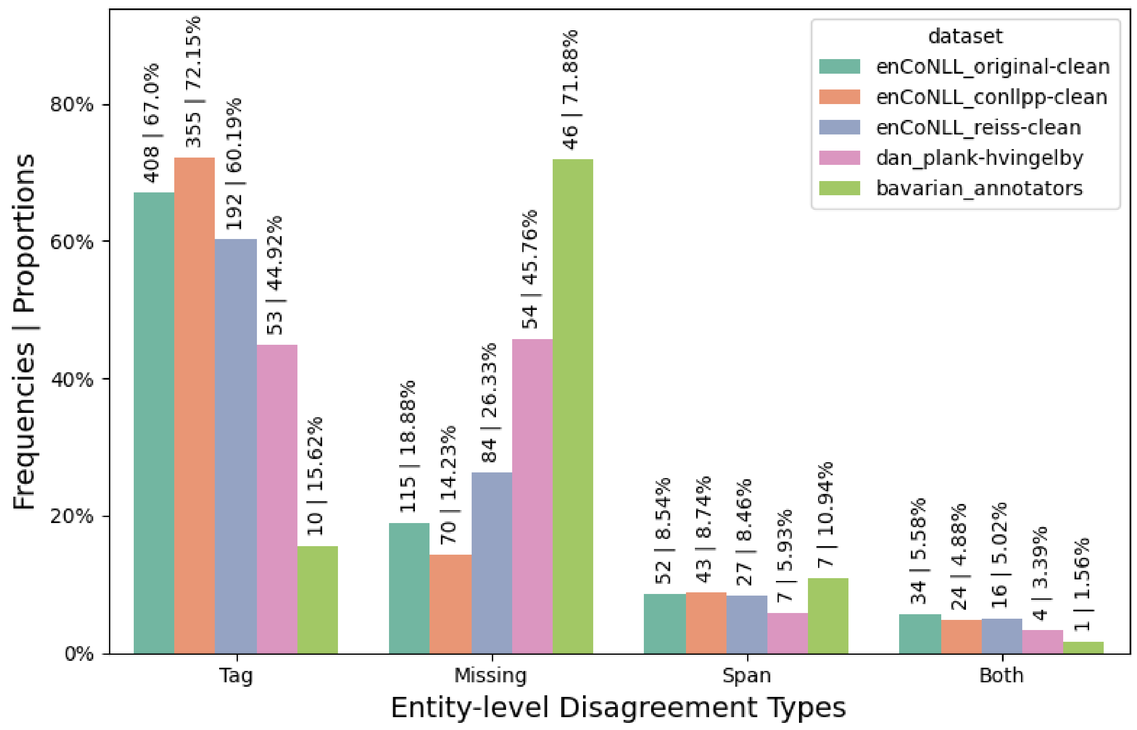

Figure 1: Proportion of entity-level disagreements.

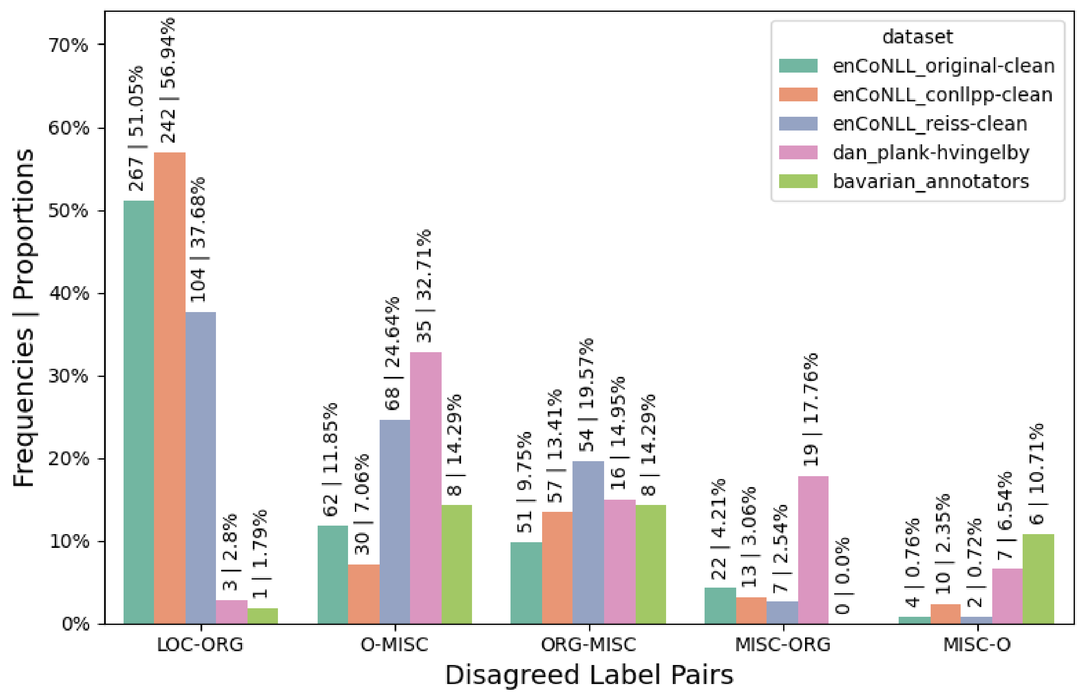

Figure 2: Proportions of top 5 label pairs in `Tag` and `Missing`.

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 1: Distributions of sources of disagreements.

Student Surveyed Annotations

- 27 student-surveyed annotations on EN and BAR;
- Label variation is prevalent in student-surveyed annotations.

Sentence	PER	LOC	ORG	MISC	O
a. UK bookmakers [William Hill] ...	■		■	—	—
b. [ALPINE] SKIING ...	—	■	—	■	■
c. ... that there is no [God] .	■	—	—	■	—

Table 2: Distribution of student-surveyed annotations.

Future Work

- Conducting much larger scale student-surveyed annotations to get statistically meaningful NE distributions for NER models;
- Separating valid label variations from true annotation mistakes;
- Remedying conflicts among versions of annotation guidelines.

Paper Data MaiNLP

Presented at UnImplicit @ EACL 2024. March 21, 2024. Malta.
This project is supported by ERC Consolidator Grant DIALECT 101043235.