Fine-Grained Error Analysis and Fair Evaluation of Labeled Spans

Katrin Ortmann

Introduction

Evaluation in NLP serves two main purposes:

(i) Determine how good a system is at a task and compare it with others
(ii) Analyze the system's errors to improve its performance

Problem: The traditional evaluation of labeled spans with precision, recall, and F_1-score leads to **double penalties**.

- Overlapping spans count as two errors, despite being closer to the target annotation than missing/superfluous spans.

Goal: Develop an approach for the fair evaluation of single- and multi-level labeled spans in order to:

(i) Prevent double penalties for a single unit
(ii) Compute meaningful values for precision, recall, and F_1
(iii) Provide detailed insights about a system's weaknesses

Precision, Recall, F_1-score

Fine-grained error types prevent double penalties, but raw errors counts are not comparable across data sets.

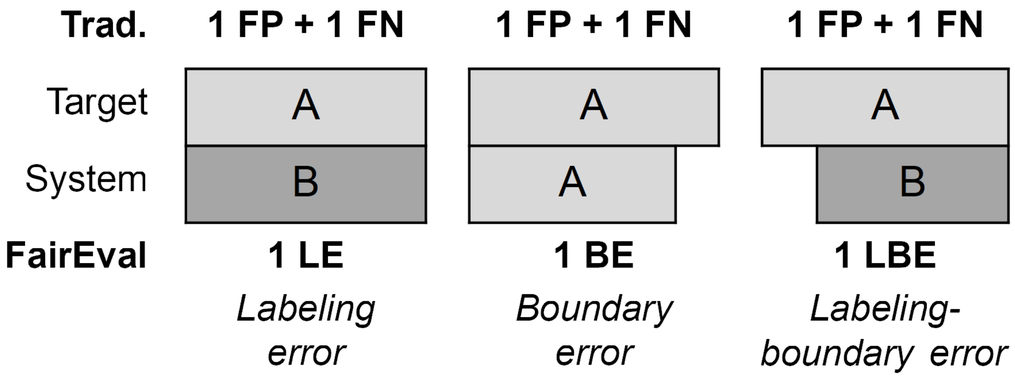
Suggestion: Include them in the calculation of standard evaluation metrics.

- Treating additional error types as 1 FP (Ortmann, 2021) or 1 FN (Read et al., 2012) leads to more realistic F_1-scores, but makes recall and precision hard to interpret.
- Instead, count them as half FP and half FN because they indicate a (partly) missing target annotation and a (partly) incorrect system annotation.

$$1LE = 1BE = 1LBE = 0.5FP + 0.5FN$$

Error Types

False positives (FP) and negatives (FN) should refer only to 1:0 and 0:1 mappings. For overlapping spans, more fine-grained error types are introduced based on Manning (2006):



Additional distinction of boundary errors: The system span is smaller (BE_s), larger (BE_l), or overlaps (BE_o) with the target.

Algorithm for Error Identification

Input: Target and system spans with (begin, end, label, toks)

Step 1: Count 1:1 mappings
- Identical spans → TP
- Identical begin and end, different label → LE

Step 2: Count boundary errors
- Most similar, overlapping spans with identical label → BE
- Remove matches from input to prevent double counting

Step 3: Count labeling-boundary errors
- Repeat step 2 with differing label → LBE

Step 4: Count 1:0 and 0:1 mappings
- Spans only in target annotation → FN
- Spans only in system annotation → FP

Example Evaluation

Exemplary application to 3 different NLP tasks, including flat and multi-level annotations.

- Fair evaluation returns higher overall scores because no error counts twice.
- Annotations that look the same with traditional evaluation can actually result from very different error distributions.

		Prec	Rec	F_1
NER	*Trad.*	86.66	83.51	85.05
	Fair	90.42	87.23	88.80
Chunks	*Trad.*	97.20	96.39	96.79
	Fair	97.86	97.86	97.86
Topol. Fields	*Trad.*	93.41	94.27	93.84
	Fair	94.78	95.92	95.35

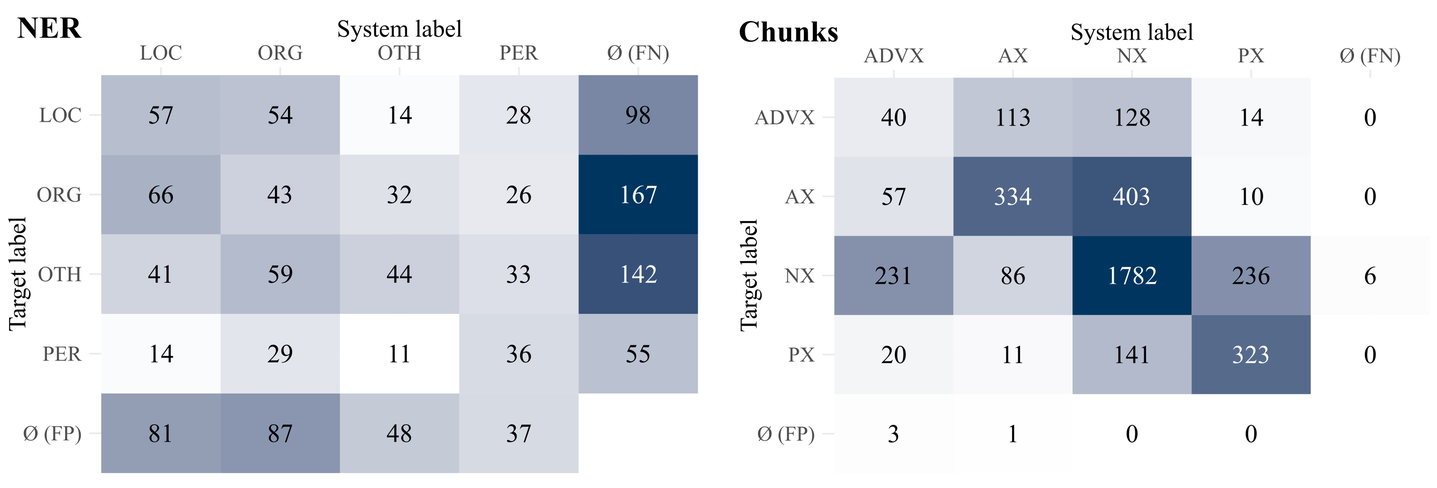


Figure: Confusion matrices for the (main) labels of the first two annotation tasks. Only errors are included, i.e., the diagonal displays boundary errors. False positives and negatives are shown in the bottom row and the right-most column, respectively. The remaining cells represent labeling and labeling-boundary errors.

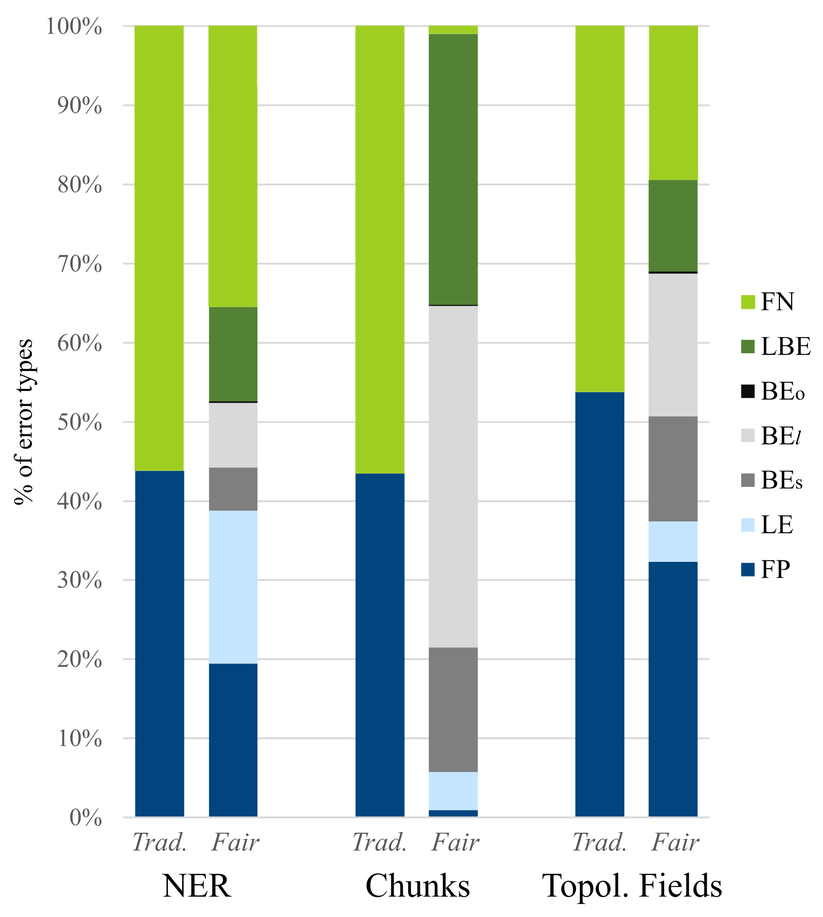


Figure: Distribution of error types for the three annotation tasks according to traditional and fair evaluation.

References

Implementation and data sets: https://github.com/rubcompling/FairEval

Manning, C. (2006). *Doing Named Entity Recognition? Don't optimize for F1.* Retreived from https://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html.

Ortmann, K. (2021). Chunking historical German. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Reykjavik, Iceland (Online), pp. 190–199.

Read, J., E. Velldal, L. Øvrelid, and S. Oepen (2012). Uio1: Constituent-based discriminative ranking for negation resolution. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 310–318.

Project-ID 232722074 – SFB 1102

LREC 2022

Gefördert durch DFG Deutsche Forschungsgemeinschaft