CoNLL#: Fine-grained Error Analysis and a Corrected Test Set for CoNLL-03 English

Andrew Rueda, Elena Álvarez Mellado, Constantine Lignos

What's left for NER models to get right on CoNLL-03 English?
What errors remain in the test set?

Background

- CoNLL-03 English is used as a benchmark of NER progress (despite many other datasets available)
- What are models getting wrong?
- What's (still) wrong in the test set?

Correcting CoNLL-03 English

- Many prior attempts to correct this dataset
- CoNLL++: corrected 309 labels in test set [1]
- ReCoNLL: corrected 105 labels and 10 sentence boundaries in test set [2]
- CoNLL-CODAIT and CleanCoNLL: changed labels across train/dev/test under simplified annotation guidelines [3, 4]
- Our approach follows original annotation guidelines and does not modify the train and dev sets

Correction Process

- Used CoNLL-CODAIT as baseline for tokenization and sentence boundaries in test set
- Adjudicated CoNLL++ and ReCoNLL label changes, all CoNLL-CODAIT label changes, and errors from each model
- Over 500 corrections in total (Table 1)
- Over 2 points F1 improvement from corrections (Table 2)

Error analysis on CoNLL#

- Economy documents have lowest performance
  - Particularly on data and hybrid format (Table 3)
  - Tough mentions found in economy documents: ambiguous acronyms, unseen mentions, unlikely mentions, capitalized non-entities
- Boundary errors were the most common error type across all models and document types (39%), with adjacent entities especially challenging
- Irregular capitalization was a common source of error

Document Type Annotation

- Annotated all test set documents in two dimensions
- **Domains**: Sports, Economy, World Events
- **Formats**: Text, Data (tables), Hybrid (text and data)
- Lots of sports articles, but amount of pure data articles is unusual for an NER dataset (Figure 1)

Correction	Count	Example
Tokenization	5	*JosepGuardiola* → *Josep Guardiola*
Bad hyphenation	27	*SKIING-WORLD CUP* → *SKIING - WORLD CUP*
Sentence boundaries	63	*Results of National Basketball Association games on Friday*
Labels	457	*Tasmania* LOC → *Tasmania* ORG

Table 1: Test set corrections

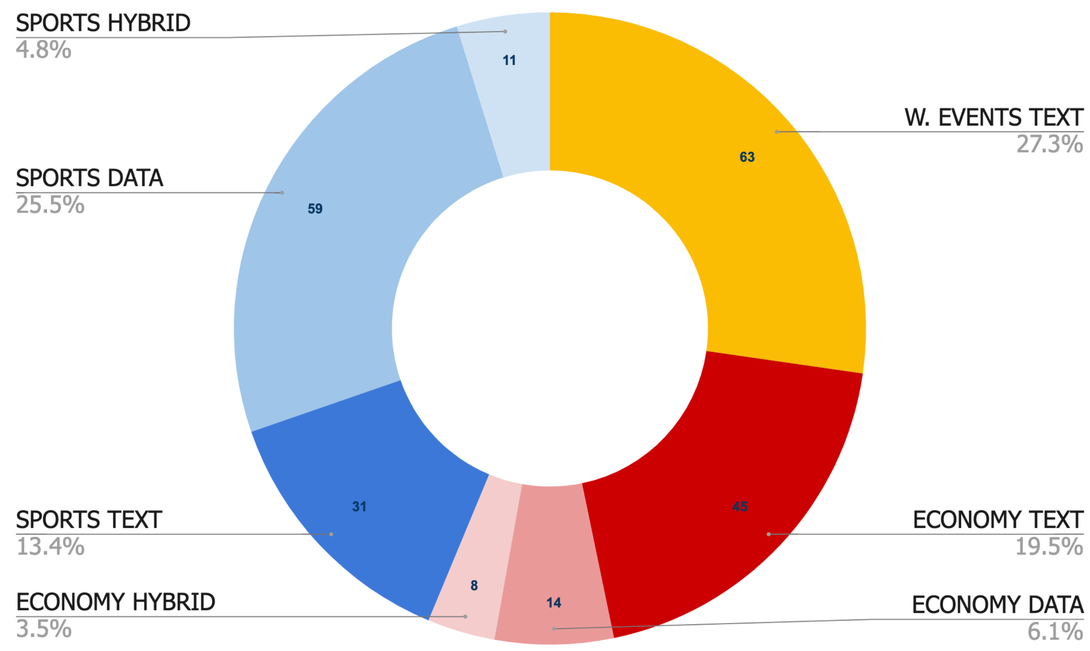

Figure 1: Test set document domains/formats

Model	CoNLL-03	CoNLL#	Change
XLM-R FLERT	93.64	95.98	+2.34
LUKE	**94.44**	**97.10**	+2.66
ASP-T0-3B	93.88	96.50	+2.62

Table 2: Original and corrected test set F1

Model	Sports	World Events	Economy	All
XLM-R FLERT	97.22	97.18	90.43	95.94
LUKE	98.29	97.54	92.67	97.10
ASP-T0-3B	97.05	97.74	93.13	96.50

Table 3: Test set F1 by document domain

Conclusions

- CoNLL#: test set consistent with original annotation guidelines but with significantly less annotation noise
- Please use our corrected test set!
- Models need to improve performance on economy documents and data tables

References

[1] CrossWeigh: Training Named Entity Tagger from Imperfect Annotations (Wang et al., EMNLP-IJCNLP 2019)
[2] Rethinking Generalization of Neural Models: A Named Entity Recognition Case Study (Fu et al., AAAI 2020)
[3] Identifying Incorrect Labels in the CoNLL-2003 Corpus (Reiss et al., CoNLL 2020)
[4] CleanCoNLL: A Nearly Noise-Free Named Entity Recognition Dataset (Rücker & Akbik, EMNLP 2023)