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

Andrew Rueda*, **Elena Álvarez Mellado**†, **Constantine Lignos***
*Michtom School of Computer Science, Brandeis University
†NLP & IR Group, School of Computer Science, UNED
{andrewrueda,lignos}@brandeis.edu
elena.alvarez@lsi.uned.es

Abstract

Modern named entity recognition systems have steadily improved performance in the age of larger and more powerful neural models. However, over the past several years, the state-of-the-art has seemingly hit another plateau on the benchmark CoNLL-03 English dataset. In this paper, we perform a deep dive into the test outputs of the highest-performing NER models, conducting a fine-grained evaluation of their performance by introducing new document-level annotations on the test set. We go beyond F1 scores by categorizing errors in order to interpret the true state of the art for NER and guide future work. We review previous attempts at correcting the various flaws of the test set and introduce CoNLL#, a new corrected version of the test set that addresses its systematic and most prevalent errors, allowing for low-noise, interpretable error analysis.

Keywords: Named Entity Recognition, Named Entity Annotation, Error Analysis, Error Classification

1. Introduction

In recent years, we have seen substantial improvements in named entity recognition (NER) performance due to new techniques such as pretrained language models, novel approaches that use document-level context and entity embeddings, and a general increase in computational power. These advancements have resulted in a steady rise in state-of-the-art performance on the usual NER benchmarks, with the CoNLL-03 English corpus (Tjong Kim Sang and De Meulder, 2003) being the most popular evaluation set. However, results on these NER benchmarks seem to have plateaued since 2021,[1] and a glass ceiling for the task has been hypothesized (Stanislawek et al., 2019). This raises the question: what is there left to improve for the CoNLL-03 English NER task?

In this paper, we investigate what NER models are still struggling with. In order to do so, we run three of the best performing models for NER on the CoNLL-03 English dataset and conduct a fine-grained error analysis that goes beyond the traditional false positive/false negative distinction that informs span-level F1 score. The goal of this analysis is to hone in on the lingering errors from state-of-the-art NER models.

The first step was to annotate the 231 original documents in the test set by assigning each a document domain and a document format. We train and test three models with recent state-of-the-art F1 scores and report their performances across document domains and formats. Due to the presence of significant annotation errors in the original CoNLL-03 English corpus, we gathered the best known gold label corrections (Section 2), and went through multiple rounds of adjudication to fix data processing errors pervasive in the test set, including faulty sentence boundaries and tokenization errors.

The result is CoNLL#, a new version of the CoNLL-03 English test set which corrects issues more consistently than previous attempts.[2] We rescore the models using this corrected set, showing improved performance across the board. Finally, we carry out a quantitative and qualitative error analysis to find interpretable, meaningful patterns among the models' prediction errors.

The contributions of this paper are as follows. First, we report results on CoNLL-03 English for three state-of-the-art NER models and document their fine-grained performance across different document types (formats and domains). Second, we release a revised version of the CoNLL-03 English test set (CoNLL#) that incorporates the adjudicated corrections made in previous revised versions, along with corrections for the systematic errors we identified that none of the previous versions fixed. Finally, we report results of the three SOTA NER models on the new revised CoNLL# dataset. Our error analysis (both qualitative and quantitative) sheds light on the issues that SOTA models are still struggling with.

[1] https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003

[2] CoNLL# is released at https://github.com/bltlab/conll-sharp.

LREC-COLING 2024, pages 3718–3728
20-25 May, 2024.

2. Previous Work

Recent advances in NER, such as using document-level features (Schweter and Akbik, 2020) and entity embeddings (Yamada et al., 2020) have improved results on popular NER benchmarks, with the best reported F1 scores on CoNLL-03 English surpassing 94. However, after decades of steady progress in NER, recent models seem to have reached a plateau.

Prior work has already pointed out the potential existence of a glass ceiling that looms over NER (Stanislawek et al., 2019). Errors in the gold standard have been pointed out as a possible cause. As a result, there have been significant prior efforts to identify annotation errors and release corrected versions of the data. Stanislawek et al. (2019) investigated errors in state-of-the-art NER models using linguistic knowledge to categorize errors within an original annotation schema for all entity types except MISC. They found that models at the time struggled with predictions that required gathering sentence-level context, as well as document-level co-reference.

CoNLL++ (Wang et al., 2019) is at this point the most widely-adopted corrected test set for CoNLL-03 English. They introduced the CrossWeigh framework to identify potential label mistakes in NER data. CoNLL++ itself corrects 309 labels in the original test set in a way that remains consistent with the original annotation guidelines which date back to the MUC-7 Named Entity task.[3] The authors made the decision not to correct any of the sentence boundaries or tokenization errors present in the test set.

ReCoNLL (Fu et al., 2020) was a similar attempt that used a measure called Entity Coverage Ratio to identify mentions that appeared in the test set with different labels than in the training set, manually correcting those labels when needed. This resulted in 105 corrected labels and 10 corrected sentence boundaries in the test set.

Reiss et al. (2020) embarked on a wider-scale correction effort, using a self-supervised approach to alter 1,320 token labels across the entire corpus. This included effectively making changes to the annotation guidelines to make the annotation more consistent. These changes, which resulted in a corrected set we will refer to as CoNLL-CODAIT, included a concerted effort to correct faulty sentence boundaries and token errors.

After this paper was submitted for review, another corrected version of CoNLL-03 English was released: CleanCoNLL (Rücker and Akbik, 2023). CleanCoNLL is a similar relabeling effort that seeks to fix the existing errors in CoNLL-03 English. CleanCoNLL was derived from CoNLL-CODAIT, and consequently, the training set and development set were also modified. In contrast, our approach tries to conform to the original annotation guidelines and does not modify the training and development sets. As this work was performed concurrently with ours, we do not discuss CleanCoNLL in detail, but Section 5 provides counts for the number of tokens in disagreement between CoNLL#, CleanCoNLL, and CoNLL-CODAIT.

Despite these many efforts, corrected versions do not always agree with one another, and the question of what are NER models struggling with still remains. In this paper, we conduct a fine-grained analysis to identify systematic errors that SOTA NER models make on the CoNLL-03 English benchmark in order to better understand what may be the reason for this recent plateau. We also identify systematic bugs with the original CoNLL-03 English test set and propose a new corrected version that adjudicates and ameliorates previous versions.

[3] https://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html

3. Document Type Annotation

In order to get a general overview of the type of errors models could be struggling with, we first annotated each of the documents in the CoNLL-03 English test set according to its type. Each document was annotated with two labels: one for the domain or genre of the news piece, and another for the document format (explained below). The motivation behind this decision was that CoNLL-03 English suffers from data selection biases, most notably the overrepresentation of sports articles compared to other domains (Chiticariu et al., 2010; Nagesh et al., 2012). By annotating the domain and format on the document level, we wanted to see if models were making systematic errors on certain document types. Table 1 summarizes the number of documents in the CoNLL-03 English test set per domain and format.

3.1. Document Domains

Documents in CoNLL-03 English were extracted from online newswire articles in December 1996. Beyond the well-known skew toward sports articles, Agerri and Rigau (2016) also note the inclusion of "lots of financial and sports data in tables," but to our knowledge no one has quantified these findings on the document level. We annotated the 231 test documents and found that the documents fell neatly within 3 major categories:

Sports 101 out of 231 documents were indeed news items covering international sporting events.

Format	Domain			
	World Events	Economy	Sports	Total
Text Article	63	45	31	139
Data Report	0	14	59	73
Hybrid	0	8	11	19
Total	63	67	101	231

Table 1: Test set documents per domain and format

These documents were a mixture of text news stories, sports schedules, and scores extracted from tabular data (see Section 3.2).

Economy 67 of the documents involved international trade, business, and market updates. Similarly to sports, some of the articles in this domain were primarily tabular data (stock indexes, market prices, etc.).

World Events 63 documents involved general news and events from around the world. This includes politics, crime, natural disasters, etc. Though slightly varied in nature, they share the quality that none of them are in the aforementioned data formats, and are all text-based articles written in complete sentences.

3.2. Document Formats

The documents in CoNLL-03 English test set were annotated into the following three categories:

Text article News articles written in complete English sentences.

Data report News reports that, beyond having an informative title, consisted solely of strings extracted from tabular data.

Hybrid Articles that had a section written in text, as well as a section in data format.

Below is an example of part of a document we labeled with the sports domain and data report format:

```
Chelsea B-ORG
2 O
Everton B-ORG
2 O
Conventry B-ORG
1 O
Tottenham B-ORG
2 O
```

Model	Precision	Recall	F1
XLM-R FLERT	92.87	**94.53**	93.64
LUKE	**95.64**	94.51	**94.44**
ASP-T0-3B	93.65	94.15	93.88

Table 2: Replicated results obtained on the original CoNLL-03 English test set

4. Modeling

We then trained and tested three NER SOTA models on the CoNLL-03 English dataset. The document-level annotation described on Section 3 was only devised for the error analysis step and performed on the test set, so it was not used in any way during the training of the models.

4.1. Models

We chose three of the best NER models that have recently pushed the state-of-the-art results, each with a reported F1 score greater than 93.0.

XLM-R FLERT Cross-lingual RoBERTa embeddings fine-tuned on CoNLL-03 English using document context (Schweter and Akbik, 2020). We trained this model using the published best configuration and the provided random seed.

LUKE Embeddings that represent both words and "entities" (contextualized mentions-strings) using the BERT Masked Language Model objective (Yamada et al., 2020). We carried out the steps given by the authors to load their best model.[4]

ASP-T0-3B A 3-billion parameter T5 model fine-tuned using a novel structure-building approach to capture dependencies (Liu et al., 2022). We trained this model using 10 random seeds, and extracted the best test labels.[5]

All three models were trained on the *original* (non-corrected) CoNLL-03 English training set. For each model, we followed the configurations made available by the authors. We trained XLM-R FLERT and ASP-T0-3B and when necessary, augmented the original code in order to extract the per-token predicted labels.

	Sports	World Events	Economy	All domains
XLM-R FLERT				
Text Article	94.09	94.62	90.75	93.37
Data Report	94.94	-	74.48	93.64
Hybrid	97.51	-	77.31	95.45
All Formats	95.17	94.62	87.68	93.69
LUKE				
Text Article	94.61	94.72	91.02	93.62
Data Report	95.52	-	75.44	94.27
Hybrid	99.33	-	97.35	99.14
All Formats	95.93	94.72	89.22	94.44
ASP-T0-3B				
Text Article	91.8	95.55	90.69	93.23
Data Report	94.56	-	84.98	93.98
Hybrid	97.68	-	90.63	96.83
All Formats	94.48	95.55	89.93	93.88

Table 3: CoNLL-03 English test set F1 across document formats and domains

4.2. Results

Table 2 summarizes the performance of each model on the CoNLL-03 English test set. Table 3 displays results split across domains and formats, using the annotation from Section 3. These results show that, despite attention paid to sports articles in past work, the economy domain is in fact the lowest-performing domain. This will be explored in greater detail in Section 6.

5. Towards CoNLL#: Adjudicating among Previous Corrections and Making Additional Corrections

A cursory analysis of the output produced by the models from Section 4 revealed that some of the errors were not in fact the models' fault, but instead annotation mistakes in the gold standard. Errors in the annotation prevent us from doing a reliable diagnosis on what NER SOTA models are really struggling with. Therefore we decided to manually correct these annotation errors on the CoNLL-03 English test set and rerun the models on a cleaner version of the test set, so that our results would truly illuminate the CoNLL-03 English instances that NER models find most challenging.

We partly based our corrections on previously-published corrected versions of the CoNLL-03 English test set: CoNLL++ (Wang et al., 2019), ReCoNLL (Fu et al., 2020), and CoNLL-CODAIT (Reiss et al., 2020). We decided to perform an adjudication process among all three corrected test sets, comparing labels across versions, and making final decisions when they disagreed.

These three corrected versions of CoNLL-03 English took very different approaches to the correction process. Consequently, their results vary greatly. For instance, CoNLL++ and ReCoNLL did not attempt any token corrections, such as repairing token typos, or splitting faulty tokens. As a result, they have a perfect one-for-one token overlap, in contrast with CoNLL-CODAIT, which sought to fix all token errors, as well as faulty sentence boundaries (see Table 4).

With this in mind, we decided to proceed with our adjudication process as follows. First, we compared the labels of CoNLL++ and ReCoNLL, making adjudication decisions for each of their disagreements. Second, we used CoNLL-CODAIT as a starting point for correcting *all* of the sentence and token boundary errors in the test set. Finally, we compare token-level label disagreements with our new test set and CoNLL-CODAIT.

The result of this process is CoNLL#, a corrected version of the CoNLL-03 English test set that adjudicates disagreements among previous corrected versions and includes fixes for errors none of the previous versions considered.

5.1. Comparing CoNLL++ and ReCoNLL

CoNLL++ and ReCoNLL had 276 token-level disagreements. We compared them side by side and manually adjudicated them. CoNLL++ was generally more aggressive in its relabeling efforts, but did so with high precision, as 68.95% of the disagreements were judged to be in favor of CoNLL++, with ReCoNLL having the correct la-

[4] https://colab.research.google.com/github/studio-ousia/luke/blob/master/notebooks/huggingface_conll_2003.ipynb

[5] The original paper provides contradictory information on the best configuration. Our experiments yielded higher performance on ASP-T0-3B than on ASP-T5-3B.

	CoNLL++	ReCoNLL	CoNLL-CODAIT	CoNLL#
CoNLL-03	309	105	565	457
CoNLL++		276	544	261
ReCoNLL			599	360
CoNLL-CODAIT				494

Table 4: Differences in token labels across test sets

Error fix	Count	Example
Token splits	5	*JosepGuardiola* → *Josep Guardiola*
Bad hyphen fixes	27	*SKIING-WORLD CUP* → *SKIING - WORLD CUP*
Sentence boundary fixes	63	[*Results of National Basketball*] [*Association games on Friday*]
Label fixes	457	*Tasmania* `LOC` → *Tasmania* `ORG`

Table 5: Total number of fixes per type in CoNLL# compared to CoNLL-03

bel in 27.44% of disagreements, and 10 cases (3.61%) where neither were correct.

Highlighting some examples, CoNLL++ did a better job of correcting the labels of domestic sports organizations to ORG, such as the *Tasmania* and *Victoria* Australian Rugby clubs, as well as an Egyptian soccer team with the nickname *ARAB CONTRACTORS*. CoNLL++ also had superior labels for MISC mentions, such as properly labeling *Czech* as a MISC in the sentence *Czech ambassador to the United Nations, Karel Kovanda, told the daily media...*

We also identified an invalid label transition in the ReCoNLL dataset (from `O` to `I-PER`), by using SeqScore's (Palen-Michel et al., 2021) validation.

5.2. Repairing Token and Sentence Boundary Errors

Many of the incorrect labels in the CoNLL-03 English test set stem from sentence boundary errors in the original test set. For example, the following sentence boundary (shown with brackets) in the CoNLL-03 English test set interrupts the mention *National Basketball Association*, making it impossible to have a single mention that spans both sentences: [*Results of National Basketball*] [*Association games on Friday*].

CoNLL++ did not attempt to correct any of these sentence boundary errors. In the test set, ReCoNLL corrected 10 sentence boundaries, and CoNLL-CODAIT corrected 26.

For CoNLL#, we attempted to fix all of sentence boundary errors in the 231 test documents—whether or not they happened to interrupt a mention—to allow NER models to be able to predict on correct sentence boundaries. We used CoNLL-CODAIT's sentence correction as a starting point, but through manual effort, found many more errors. We found a systematic sentence boundary error among documents that were labeled in our document-level annotation as sports data reports. Among the 59 test documents that were sports data reports, 43 of them had a faulty sentence boundary in their initial headline between the 16th and 20th characters. This processing error is not observed within any other document type. For CoNLL#, we repaired all 70 sentence boundary errors we identified in the test set.

A similar, systematic error was also found within sports data reports. The sports headlines from these documents feature a hyphen between the name of the sport and the headline of the article, such as the token *SKIING-GOETSCHL* in the test sentence *ALPINE SKIING-GOETCHL WINS WORLD CUP DOWNHILL*. If this hyphen were treated as intended, effectively as a colon, the token *GOETCHL* should have been labeled as `B-PER`.

This type of error occurred 27 times in the original test set, almost all within sports data reports. Only CoNLL-CODAIT attempted to fix these tokenization errors. Our manual inspection found that CoNLL-CODAIT corrected 14 out of 27 of these tokens; for CoNLL#, we fixed all 27 errors.

CoNLL-CODAIT also departed from the CoNLL-03 English tokenization and annotation guidelines by splitting some hyphen-joined entities into two. For example, in the original dataset *UK-US* in a context like *UK-US open skies talks end* should be a single token annotated as `B-MISC`. CoNLL-CODAIT changed 8 instances of tokens like this to be three tokens (*UK - US*), annotated as `B-LOC O B-LOC`. In CoNLL#, we maintained the original tokenization and labels for these tokens.

Overall, CoNLL-CODAIT made 68 corrections on the test set with regards to sentence boundaries and tokenization, of which we accepted 60 in our adjudication process. CoNLL# contains an additional 42 similar corrections.

5.3. Comparing Labels to CoNLL-CODAIT

CoNLL-CODAIT included a far-reaching reannotation project for each of the CoNLL-03 English training, development, and test sets. For example, they changed national sports teams from LOC to ORG in sentences like *Japan began the defense of their Asian Cup Title with a lucky 2-1 win against Syria in a Group C Championship match on Friday.* Following the original annotation guidelines, local sports teams that are referred to using a location should be annotated as ORG, while national sports teams referred to using a country name should be annotated as LOC.[6] The change to annotate both as ORG reduces the number of arbitrary distinctions models must make, but is not in keeping with the original annotation guidelines.

Table 4 shows that at a token level, CoNLL-CODAIT labels have many more disagreements relative to the other test sets in question. This is due to the fact that the CoNLL-CODAIT approach included not just correcting annotation errors, but also changing the annotation guidelines.

For instance, as noted by both the authors of CoNLL++ and CoNLL-CODAIT, there are 41 examples in which upcoming sports games displayed in the common format of *ANAHEIM AT BUFFALO* labeled the latter team as LOC. This labeling decision is present in all of the CoNLL-03 English training, dev, and test data, and so a proper correction should either overturn all of these labels in the corpus (as CoNLL-CODAIT did), or leave all of them the same (as CoNLL++ did).

Although there are differences between what the annotation guidelines required and what the CoNLL-03 English annotators did, our analysis suggests that no single interpretation of schema for NER is infallible, and are in many cases subjective for tough or ambiguous mentions. Given that we did not aim to modify any annotations in the training data, we decided to not change the annotation guidelines governing the test data. This approach ensured that the labeling decisions implemented in the test data would be in line with the annotation decisions in the training data, contributing to low-noise error analysis. As a result, our round of label adjudication with CoNLL-CODAIT maintained most of the labels decided upon in our first round of adjudication between CoNLL++ and ReCoNLL: we adopted the CoNLL-CODAIT label in only 4.86% of the disagreements.

The result of these two rounds of adjudication, repairs of tokens and sentence boundaries, and arbitration of label disagreements among the corrected test sets, is CoNLL#, our corrected version of the CoNLL-03 English dataset. Table 5 summarizes the number of fixes per type in CoNLL# compared to CoNLL-03. Token-level comparison between CoNLL# and the other test sets can be seen in Table 4.

Table 6 summarizes the types and number of label disagreements between CLEANCONLL, CoNLL-CODAIT and CoNLL#.

[6] See section A.2.2 from the original MUC-7 guidelines: https://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html.

Agreed	Disagreed	Count
CoNLL-CODAIT CleanCoNLL CoNLL#		49,593
	CoNLL-CODAIT CleanCoNLL CoNLL#	15
CoNLL-CODAIT CleanCoNLL	CoNLL#	291
CleanCoNLL CoNLL#	CoNLL-CODAIT	180
CoNLL-CODAIT CoNLL#	CleanCoNLL	316

Table 6: Differences between CLEANCONLL, CoNLL-CODAIT, and CoNLL#

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

Table 7: F1 scores for 3 SOTA models on CoNLL-03 and CoNLL#

6. Results and Error Analysis on CoNLL#

We reran the three SOTA NER models from Section 4 on our new corrected CoNLL-03 English test set, CoNLL#, and evaluated the results. Table 7 shows the results obtained on the new CoNLL# test set compared to previous results obtained on the original CoNLL-03 test set. Table 8 shows recall results over previously seen (during NER training) and unseen entities. With the new CoNLL#, the overall F1 for each of the models increased by more than 2 points.

Table 9 summarizes the results obtained across document domains and document formats. The performance gains made by testing on CoNLL# seem to have a roughly uniform effect across all of the different document types. The predictions on economy documents have improved, but these documents are still the predominant source

	CoNLL-03		CoNLL#	
	Seen	Unseen	Seen	Unseen
XLMR	96.49	**92.36**	98.47	93.98
LUKE	**96.90**	91.88	**99.22**	**94.64**
ASP-T0-3B	96.42	91.64	98.57	94.63

Table 8: Recall results for seen and unseen mentions en CoNLL-03 and CoNLL#

of lingering errors. In fact, while performance on economy text articles increased by multiple points for each model, their respective performances on data reports and hybrid articles went *down*, which is unique to those two document types.

We conducted an annotated error analysis on outputs of the three state-of-the-art English NER models. This was done in a side-by-side analysis of the gold token-level BIO labels and the predicted labels in spreadsheet format. For each token-level label mismatch, context of the nearby tags was used to classify each error within the following schema from Chinchor and Sundheim (1993): **Missed** for a full false negative, **Spurious** for a full false positive, **Boundary Error** when a mention was detected but with imperfect overlap, and **Type Error** for when the boundaries were perfect but the tag type was incorrect. For each Type Error, the sub-type was also recorded (i.e. (LOC, ORG) when a LOC mention was wrongly predicted as an ORG mention).

We also recorded invalid label transitions for the token-level predictions. This was only applicable for XLM-R FLERT, as the other two made span-label predictions instead of token-level. Table 10 summarizes the results obtained by the three models across document domains.

6.1. Recurrent Errors in economy documents

As introduced in Section 4.2, all state-of-the-art NER models that we tested performed significantly worse on economy test documents than the others. We can also see that these mostly come from data reports and hybrid articles in the economy domain.

Using SeqScore's error counts feature (Palen-Michel et al., 2021), we counted the mention-level errors of all three models on economy documents (Table 11).

Our manual error analysis revealed that the economy domain documents had a high density of tough mentions, which we classified as follows.

Ambiguous acronyms Economy articles were much more packed with acronyms and initialisms that the models struggled with. Acronyms are predominantly a mixture of ORG and MISC mentions. Examples: *NYMEX*, *ADRs*, *BTPs*, *CEFTA*, *CST*, *CBOT*, *ORE*.

Obscure / unseen mentions Much more than in the other domains, economy documents contained mentions which were not only unseen in the training data, but perhaps too rare even for the pretrained embeddings to help. This is crucially made worse by the fact that they often lacked sufficient context within the sentence or even document. This is especially true amongst economy data reports, which was the document type with worst overall F1 scores. This ranges from commercial product names [*Arabian Light*, MISC], to obscure international companies [*Thai Resource*, ORG], to purely esoteric mentions [*Algoa Day*, MISC (name of a ship)].

Unlikely mentions In many cases, tokens that likely have high correlation with one type are present in economy documents with an idiosyncratic type. One clear example of this is the mention *Manitoba*, which co-refers to an organization named *Manitoba Pork* named earlier in the document. All of the models mislabeled *Manitoba* as a LOC in this string: *Manitoba 's Hog Price Range : 84.00-86.00 per cwt*

Capitalized non-entities Table 10 shows that models committed the most SPURIOUS errors in the economy domain. An example of this type of error occurs in the stock price data reports, in which assets that are not named entities are capitalized, such as in the following string: *Wheat 121 130 121.3 121 Maize (Flint) 113 114 113.7 112 Maize (Dent) 113 114 113.7 112*

According to the section A.4.5.2 of the original MUC guidelines for Named Entities, "sub-national regions when referenced only by compass-point modifier" should not be tagged as locations. The CoNLL-03 English corpus stayed true to this directive, and has labeled unnamed regions such as *East Coast* and *West Coast* with the O tag. However, most likely due to their capitalization, all of the models mislabeled them as LOC mentions.

6.2. Other Recurrent Errors

Compound mentions The most common error type across all models and document types were boundary errors (39.0%). In many cases, the models would get confused by adjacent mentions, and could not parse if they were two distinct mentions, or part of one mention. Strings such as *Nazi German*, *Algerian Moslem*, and *UK Department of Transport* are two separate mentions each, but all three models incorrectly treated them as single mentions of two-token length. Conversely, the

	Sports	World Events	Economy	All Domains
XLM-R FLERT				
Text Article	95.00	97.18	93.59	95.61
Data Report	97.72	-	78.38	96.46
Hybrid	98.09	-	76.27	95.88
All Formats	97.22	97.18	90.43	95.94
LUKE				
Text Article	96.70	97.54	95.26	96.68
Data Report	98.42	-	76.66	97.05
Hybrid	99.81	-	95.58	99.40
All Formats	98.29	97.54	92.67	97.10
ASP-T0-3B				
Text Article	93.34	97.74	95.32	95.97
Data Report	97.96	-	81.91	96.92
Hybrid	98.22	-	90.27	97.46
All Formats	97.05	97.74	93.13	96.50

Table 9: Model performance across domain and format on CoNLL#

	Sports	Economy	World Events
XLM-R FLERT			
Missed	1	14	10
Spurious	16	49	14
Boundary Error	67	48	20
Type Error	77	61	16
LUKE			
Missed	11	39	16
Spurious	4	28	7
Boundary Error	54	42	22
Type Error	54	39	22
ASP-T0-3B			
Missed	11	16	8
Spurious	46	23	7
Boundary Error	124	43	25
Type Error	61	48	13

Table 10: Error types per document domain on CoNLL#

mention *1993 World Cup*, all three models incorrectly left off 1993 and tagged the mention as *1993* [*World Cup*]$_{\text{MISC}}$.

Irregular capitalization Modern state-of-the-art NER systems are still confounded by irregular capitalization cues, at least when trained on the CoNLL-03 English corpus. This is recurrent across all models when dealing with all-caps headlines (all three models labeled *CITY OF HARTFORD* as a LOC, instead of simply HARTFORD), spuriously tagged capitalized non-entities (*Business Policy*), and missed mentions that were in lower case (*world wide web*).

7. Limitations

While still a fixture of NER evaluation 20 years later, the CoNLL-03 English corpus remains imperfect as a benchmark. Many of the errors that we identified persist within the training and development sets. Though past correction efforts worked to ameliorate these errors (Reiss et al., 2020; Fu et al., 2020), the overwhelming majority of novel NER models are trained on the original data.

Due to limited resources for analysis and paper length limits, we have only focused on CoNLL-03 English data for this paper. The CoNLL 2002–3 NER shared tasks included languages beyond English (Dutch, German, and Spanish). Our early investigations showed that the other languages also show systematic, impactful annotation errors. In particular, as previously noted by Agerri and Rigau (2016), the CoNLL-02 Spanish test set has many

XLM-R FLERT			
Count	Error	Type	Text
3	FP	ORG	Chicago
3	FN	LOC	Chicago
3	FP	MISC	Select
2	FN	MISC	ACCESS
2	FP	ORG	Busang
2	FN	LOC	Busang
2	FN	MISC	Canadian
2	FP	MISC	Choice
2	FP	ORG	Busang
2	FN	LOC	Busang
2	FP	ORG	Ministry
LUKE			
Count	Error	Type	Text
3	FN	ORG	NYMEX
2	FN	MISC	ACCESS
2	FN	MISC	Canadian
2	FP	LOC	Canadian West Coast
2	FP	ORG	Durum
2	FP	MISC	GDR
2	FN	ORG	Manitoba
2	FN	ORG	Manitoba Pork
ASP-T0-3B			
Count	Error	Type	Text
4	FN	MISC	trans-Atlantic
3	FN	ORG	NYMEX
2	FP	ORG	ACCESS
2	FN	MISC	ACCESS
2	FN	MISC	Canadian
2	FP	LOC	Canadian West Coast
2	FP	LOC	Iowa-S Minn
2	FN	MISC	Iowa-S Minn
2	FN	ORG	Manitoba

Table 11: Most frequent false positive and false negative errors for each of the three NER models' predictions on economy documents

mentions that include bordering quotation marks, which goes against common NER conventions.

It is just as important for NER models in other languages to be tested on data that is as consistent and clean as possible, so that we can learn more about the lingering NER errors in languages beyond English. For example, our analysis of the CoNLL-02 Spanish data found that state-of-the-art models still struggle with parsing the preposition "de" within mentions.

There are of course many NER datasets beyond the CoNLL 2002–3 shared tasks, and those recently developed for less-resourced languages are of particular interest to us. We hope to collaborate with speakers of those languages to extend our study far beyond English.

Finally, a lingering issue in NLP evaluation is to what extend the high results obtained by models may be caused by data contamination. It is likely that the models saw the CoNLL-03 English data during pretraining, which may explain the the high results obtained on this task. Exactly what this means for our evaluation is a matter of debate, and the impact that data contamination may be having on evaluation is currently an ongoing line of research (Sainz et al., 2023).

8. Conclusion

In this paper, we conducted a full-scale error analysis of state-of-the-art named entity recognition taggers on the CoNLL-03 English dataset. Our document type annotations revealed clear trends, including the relatively poor performance that state-of-the-art models achieve on the 67 economy documents in the test set. We evaluated and adjudicated the various corrected English CoNLL test sets to create CoNLL#, a version of the CoNLL-03 English test set with significantly less annotation-error noise. Finally, we tested three state-of-the-art NER models on this corrected test set, and combed through their errors to get a full sense of what they continue to struggle with, and detail where future models can gain those last F1 points.

9. Ethics and Broader Impact

Benchmark datasets play a key role in NLP research. Improvements in benchmark results are generally accepted as overall progress on a given task. However, this can also lead to benchmark chasing, which reduces the difficulty of a given task to a matter of gaining tenths of a point on a leaderboard, without truly gaining any insight or making true advancements on the task (Raji et al., 2021).

In addition, no benchmark can ever fully capture the complexities of the linguistic phenomenon in question (Paullada et al., 2021). Equating NER progress simply to gains in F1 score on a given benchmark is a reductionist approach. As popular as CoNLL-03 English may be, it suffers from obvious limitations: the genre of the documents are exclusively newswire texts covering a limited set of topics (sports, economy, world events) from a short span of time (1996) and only certain language varieties are represented.

We hope the broader impact of this work will be that progress on the CoNLL-03 English dataset can be better measured due to a lower-noise version of the test set and that others will be able to adapt our methodology to other datasets and languages.

10. Bibliographical References

Rodrigo Agerri and German Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artif. Intell.*, 238:63–82.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1002–1012, Cambridge, MA. Association for Computational Linguistics.

Ajay Nagesh, Ganesh Ramakrishnan, Laura Chiticariu, Rajasekar Krishnamurthy, Ankush Dharkar, and Pushpak Bhattacharyya. 2012. Towards efficient named-entity rule induction for customizability. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 128–138, Jeju Island, Korea. Association for Computational Linguistics.

Chester Palen-Michel, Nolan Holley, and Constantine Lignos. 2021. SeqScore: Addressing barriers to reproducible named entity recognition evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 40–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

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*, pages 8628–8645, Singapore. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Stefan Schweter and Alan Akbik. 2020. FLERT: Document-level features for named entity recognition. *arXiv preprint arXiv:2011.06993*.

Tomasz Stanislawek, Anna Wróblewska, Alicja Wójcicka, Daniel Ziembicki, and Przemyslaw Biecek. 2019. Named entity recognition - is there a glass ceiling? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 624–633, Hong Kong, China. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

11. Language Resource References

Jinlan Fu, Pengfei Liu, and Qi Zhang. 2020. Rethinking generalization of neural models: A named entity recognition case study. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7732–7739.

Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, pages 215–226, Online. Association for Computational Linguistics.

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*, pages 142–147.

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)*, pages 5154–5163, Hong Kong, China. Association for Computational Linguistics.

A. Data statement

We document the information concerning CoNLL# following the data statement format proposed by Bender and Friedman (2018).
Data set name: CoNLL#
Data set developers: Andrew Rueda, Elena Álvarez Mellado, and Constantine Lignos
Dataset license: Our modifications to the original data are distributed under CC BY 4.0. The original license terms apply to the original data.
Link to dataset: `https://github.com/bltlab/conll-sharp`

A.1. Curation rationale

CoNLL# is a reannotation and adjudication of the English section of the CoNLL-03 test set.

A.2. Language variety

The language of this corpus is English (ISO 639-3 `eng`), of the variety used in international newswire.

A.3. Speaker demographic

No detailed information was collected regarding the demographics of the authors of the original texts from CoNLL-03. However, we can infer that the authors of the text were English-speaking journalists aged between 20-65.

A.4. Adjudicator demographic

The annotator and adjudicator of CoNLL# was a 20-30 year-old male graduate student from the USA, trained in linguistics and computational linguistics, whose native language is English.

A.5. Speech situation

The English section of the CoNLL-03 dataset is taken from the Reuters Corpus, which consists of a collection of English journalistic texts written between 1996 and 1997. For a full description of the CoNLL-03 dataset see Tjong Kim Sang and De Meulder (2003).

A.6. Text characteristics

The texts from CoNLL-03 English come from the Reuters Corpus. This means that the texts in CoNLL# are from the newswire domain. Consequently, we can assume that all the texts in CoNLL# are carefully, well-edited texts that follow the rules of "standard" English.

The articles in the test set belong to the sports domain, economy or world events (see section 3.1). In terms of format, the documents in CoNLL# are text articles, data reports (tabulated data) or a mix between the two (see section 3.2).

A.7. Recording quality

N/A

A.8. Other

N/A