TMR: Evaluating NER Recall on Tough Mentions

Jingxuan Tu
Brandeis University
jxtu@brandeis.edu

Constantine Lignos
Brandeis University
lignos@brandeis.edu

Abstract

We propose the *Tough Mentions Recall* (TMR) metrics to supplement traditional named entity recognition (NER) evaluation by examining recall on specific subsets of "tough" mentions: *unseen mentions*, those whose tokens or token/type combination were not observed in training, and *type-confusable mentions*, token sequences with multiple entity types in the test data. We demonstrate the usefulness of these metrics by evaluating corpora of English, Spanish, and Dutch using five recent neural architectures. We identify subtle differences between the performance of BERT and Flair on two English NER corpora and identify a weak spot in the performance of current models in Spanish. We conclude that the TMR metrics enable differentiation between otherwise similar-scoring systems and identification of patterns in performance that would go unnoticed from overall precision, recall, and F1.

1 Introduction

For decades, the standard measures of performance for named entity recognition (NER) systems have been precision, recall, and F1 computed over entity mentions.[1] NER systems are primarily evaluated using exact match[2] F1 score, micro-averaged across mentions of all entity types. While per-entity-type scores available from the `conlleval` scorer (Tjong Kim Sang, 2002) are often reported, there are no widely-used diagnostic metrics that further analyze the performance of NER systems and allow for separation of systems close in F1.

This work proposes *Tough Mentions Recall* (TMR), a set of metrics that provide a fine-grained analysis of the mentions that are likely to be most challenging for a system: *unseen mentions*, ones that are present in the test data but not the training data, and *type-confusable mentions*, ones that appear with multiple types in the test set. We evaluate the performance of five recent popular neural systems on English, Spanish and Dutch data using these fine-grained metrics. We demonstrate that TMR metrics enable differentiation between otherwise similar-scoring systems, and the model that performs best overall might not be the best on the tough mentions. Our NER evaluation tool is publicly available via a GitHub repository.[3]

2 Related Work

Previous work in NER and sequence labeling has examined performance on out-of-vocabulary (OOV) tokens and rare or unseen entities. Ma and Hovy (2016) and Yang et al. (2018) evaluate system performance on mentions containing tokens not present in the pretrained embeddings or training data. Such analysis can be used broadly—Ma and Hovy perform similar analyses for part of speech tagging and NER—and can guide system design around the handling of those tokens.

Augenstein et al. (2017) present a thorough analysis of the generalization abilities of NER systems, quantifying the performance gap between seen and unseen mentions, among many other factors. Their work predates current neural NER models; the newest model they use in their evaluation is SENNA (Collobert et al., 2011). While prior work has considered evaluation on unseen mentions, it has focused on experimenting on English data, and the definition of "unseen" has focused on the tokens themselves being unseen (UNSEEN-TOKENS in our

[1] We use the term *mention* to refer to a specific annotated reference to a named entity—a span of tokens (*token sequence*) and an entity type. We reserve the term *entity* for the referent, e.g. the person being named. The traditional NER F1 measure is computed over mentions ("phrase" F1).

[2] While partial match metrics have been used (e.g. Chinchor and Sundheim, 1993; Chinchor, 1998; Doddington et al., 2004; Segura-Bedmar et al., 2013), exact matching is still most commonly used, and the only approach we explore.

[3] https://github.com/jxtu/EvalNER

Proceedings of the 16th Conference of the European Chapter of the Associationfor Computational Linguistics: Student Research Workshop, pages 155–163
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

TRAINING SET	**Newcastle**$_{\text{[LOC]}}$ is a city in the **UK**$_{\text{[LOC]}}$.
TEST SET	**John Brown**$_{\text{[PER]}}$, the **Newcastle**$_{\text{[ORG]}}$ star from the **UK**$_{\text{[LOC]}}$, has...

	Newcastle [ORG]	**John Brown**[PER]	**UK**[LOC]
SEEN			✔
UNSEEN-TYPE	✔		
UNSEEN-TOKENS		✔	
UNSEEN-ANY	✔	✔	

Table 1: Example data and how mentions would be classified into unseen and type-confusable mention sets

work). We use the umbrella of "tough mentions" to cover a number of possible distinctions that can be made with regards to how unseen test set data is, and we experiment on multiple languages.

Mesbah et al. (2018) propose an iterative approach for long-tail entity extraction, focusing on entities of two specific types in the scientific domain. Derczynski et al. (2017) propose evaluation on a set of unique mentions, which emphasizes the ability of a system to recognize rarer entities. As entities and their types change quickly (Derczynski et al., 2015), recall on emerging entities is becoming a more critical measure in evaluating progress. Ribeiro et al. (2020) propose CHECKLIST, which can be applied to NER by using invariance tests; for example, replacing a mention with another one of the same entity type should not affect the output of the model. Fu et al. (2020) evaluate the generalization of NER models through breakdown tests, annotation errors and dataset bias. They examine the performance on subsets of entities based on the entity coverage rate between train and test set. They also release *ReCoNLL*, a revised version of CoNLL-2003 English with fewer annotation errors which we use in this work.

3 Unseen and Type-confusable Mentions

3.1 Unseen Mentions

Given annotated NER data divided into a fixed train/development/test split, we are interested in the relationship between the mentions of the training and test sets. We classify mentions into three mutually-exclusive sets described in Table 1: SEEN, UNSEEN-TYPE, and UNSEEN-TOKENS, and a superset UNSEEN-ANY that is the union of UNSEEN-TYPE and UNSEEN-TOKENS. **UK**$_{\text{[LOC]}}$ appears in both the training and test set, so it is a SEEN mention. As there is no mention consisting of the token sequence *John Brown* annotated as any type in the test set, **John Brown**$_{\text{[PER]}}$ is an UNSEEN-TOKENS mention.[4] While there is no mention with the tokens and type **Newcastle**$_{\text{[ORG]}}$ in the training data, the token sequence *Newcastle* appears as a mention, albeit with a different type (LOC). **Newcastle**$_{\text{[ORG]}}$ is an UNSEEN-TYPE mention as the same token sequence has appeared as a mention, but not with the type ORG.

3.2 Type-confusable Mentions

Token sequences that appear as mentions with multiple types in the test set form another natural set of challenging mentions. If **Boston**$_{\text{[LOC]}}$, the city, and **Boston**$_{\text{[ORG]}}$, referring to a sports team[5] are both in the test set, we consider all mentions of exactly the token sequence *Boston* to be *type-confusable mentions* (TCMs), members of TCM-ALL. We can further divide this set based on whether each mention is unseen. TCM-UNSEEN is the intersection of TCM-ALL and UNSEEN-TOKEN; TCM-SEEN is the rest of TCM-ALL.

Unlike Fu et al. (2020), who explore token sequences that occur with different types in the training data, we base our criteria for TCMs around type variation in the test data. Doing so places the focus on whether the model can correctly produce multiple types in the output, as opposed to how it reacted to multiple types in the input. Also, if type confusability were based on the training data, it would be impossible to have TCM-UNSEEN mentions, as the fact that they are type confusable in the training data means they have been seen at least twice in training and thus cannot be considered unseen. As our metrics compute subsets over the gold standard entities, it is natural to only measure recall and not precision on those subsets, as it is not clear exactly which false positives should be considered in computing precision.

3.3 Data Composition

We evaluate using the ReCoNLL English (Fu et al., 2020), OntoNotes 5.0 English (Weischedel et al., 2013, using data splits from Pradhan et al. 2013), CoNLL-2002 Dutch, and CoNLL-2002 Spanish (Tjong Kim Sang, 2002) datasets. We use ReCoNLL (Fu et al., 2020) in our analysis instead

[4] The matching criterion for the token sequence is case sensitive, requires an exact—not partial—match, and only considers mentions. **John Henry Brown**$_{\text{[PER]}}$, **john brown**$_{\text{[PER]}}$, or unannotated *John Brown* appearing in the training set would not make **John Brown**$_{\text{[PER]}}$ a seen mention.

[5] For example: ***Boston***$_{\textit{[ORG]}}$ *won the World Series in 2018.*

Set	LOC	ORG	PER	MISC	ALL
UNSEEN-ANY	17.9	45.9	**85.3**	35.5	47.6
UNSEEN-TOK.	17.5	41.6	**85.1**	35.1	46.1
UNSEEN-TYPE	0.4	**4.3**	0.2	0.4	1.5
TCM-ALL	7.1	**13.7**	0.4	1.0	6.3
TCM-SEEN	5.4	**9.5**	0.4	1.0	4.6
TCM-UNSEEN	1.7	**4.2**	0.0	0.0	1.7
All (Count)	1,668	1,661	1,617	702	5,648

Table 2: Percentage of all mentions in each subset, with total mentions in the final row (ReCoNLL English)

Set	LOC	ORG	PER	MISC	ALL
UNSEEN-ANY	24.4	30.8	**68.9**	60.9	39.6
UNSEEN-TOK.	22.4	29.2	**67.1**	58.8	37.8
UNSEEN-TYPE	2.0	1.6	1.8	**2.1**	1.8
TCM-ALL	**23.3**	7.5	1.1	4.7	10.7
TCM-SEEN	**22.6**	6.8	0.8	4.1	10.1
TCM-UNSEEN	**0.7**	**0.7**	0.3	0.6	0.6
All (Count)	1,084	1,400	735	340	3,559

Table 3: Percentage of all mentions in each subset, with total mentions in the final row (CoNLL-2002 Spanish)

of the CoNLL-2003 English data (Tjong Kim Sang and De Meulder, 2003) to improve accuracy as it contains a number of corrections.

Tables 2, 3, and 4 give the total mentions of each entity type and the percentage that fall under the proposed unseen and TCM subsets for the three CoNLL datasets.[6] Across the three languages, 39.6%–54.6% of mentions are unseen, with the highest rate coming from PER mentions. UNSEEN-TYPE contains under 2% of mentions in English and Spanish and almost no mentions in Dutch; it is rare for a token sequence to only appear in training with types that do not appear with it in the test data.

Similarly, TCMs appear in the English (10.7%) and Spanish (6.3%) data, but almost never in Dutch (0.2%). The differences across languages with regards to TCMs may reflect morphology or other patterns that prevent the same token sequence from appearing with multiple types, but they could also be caused by the topics included in the data. In English, the primary source of TCMs is the use of city names as sports organizations, creating LOC-ORG confusion.

[6]Tables for OntoNotes 5.0 English are provided in the appendix (Tables 16-17).

Set	LOC	ORG	PER	MISC	ALL
UNSEEN-ANY	36.8	52.2	**72.6**	51.2	54.6
UNSEEN-TOK.	36.8	52.1	**72.5**	50.9	54.4
UNSEEN-TYPE	0.0	0.1	0.1	**0.3**	0.2
TCM-ALL	0.1	0.0	0.2	**0.3**	0.2
TCM-SEEN	**0.1**	0.0	**0.1**	0.0	0.1
TCM-UNSEEN	0.0	0.0	0.1	**0.3**	0.1
All (Count)	774	882	1,098	1,187	3,941

Table 4: Percentage of all mentions in each subset, with total mentions in the final row (CoNLL-2002 Dutch)

4 Results

4.1 Models and Evaluation

We tested five recent mainstream NER neural architectures that either achieved the state-of-the-art performance previously or are widely used among the research community.[7] The models are CHARCNN+WORDLSTM+CRF[8](CHARCNN), CHARLSTM+WORDLSTM+CRF[8] (CHARLSTM), CASED BERT-BASE[9] (Devlin et al., 2019), BERT-CRF[10] (Souza et al., 2019), and FLAIR (Akbik et al., 2018).[11]

We trained all the models using the training set of each dataset. We fine-tuned English Cased BERT-Base, Dutch (Vries et al., 2019) and Spanish (Cañete et al., 2020) BERT models and used the model from epoch 4 after comparing development set performance for epochs 3, 4, and 5. We also fine-tuned BERT-CRF models using the training data, and used the model from the epoch where development set performance was the best within the maximum of 16 epochs.

All models were trained five times each on a single NVIDIA TITAN RTX GPU. The mean and standard deviation of scores over five training runs are reported for each model. It took approximately 2 hours to train each of FLAIR and NCRF++ on each of the CoNLL-2002/3 datasets, 12 hours to train FLAIR, and 4 hours to train NCRF++ on OntoNotes 5.0 English. It took less than an hour to fine-tune BERT or BERT-CRF models on each dataset. Hyperparameters for Spanish and Dutch models implemented using NCRF++ were taken from Lample et al. (2016). FLAIR does not provide hyperparameters for training CoNLL-02 Spanish, so we used

[7]We could not include a recent system by Baevski et al. (2019) because it was not made publicly available.

[8]Using the NCRF++ (Yang and Zhang, 2018) implementations: https://github.com/jiesutd/NCRFpp.

[9]NER implementation from https://github.com/kamalkraj/BERT-NER.

[10]A Cased BERT-Base Model with an additional CRF layer.

[11]https://github.com/flairNLP/flair

Model	Precision	Recall	F1
CHARLSTM	91.92 (±0.29)	91.90 (±0.31)	91.91 (±0.28)
CHARCNN	92.13 (±0.18)	91.93 (±0.18)	92.03 (±0.17)
FLAIR	**93.00 (±0.15)**	**93.66 (±0.08)**	**93.33 (±0.12)**
BERT	91.04 (±0.11)	92.36 (±0.13)	91.70 (±0.14)
BERT-CRF	91.13 (±0.15)	92.29 (±0.04)	91.70 (±0.08)

Table 5: Standard P/R/F1 (ReCoNLL-2003 English)

Model	Precision	Recall	F1
CHARLSTM	87.12 (±0.42)	86.38 (±0.36)	86.90 (±0.40)
CHARCNN	86.94 (±0.27)	86.28 (±0.33)	86.61 (±0.25)
FLAIR	**88.56 (±0.12)**	89.42 (±0.09)	**88.99 (±0.10)**
BERT	87.52 (±0.09)	**89.84 (±0.12)**	88.67 (±0.10)
BERT-CRF	87.29 (±0.33)	89.32 (±0.19)	88.29 (±0.26)

Table 6: Standard P/R/F1 (OntoNotes 5.0 English)

those for CoNLL-02 Dutch. We did not perform any other hyperparameter tuning.

4.2 Baseline Results

We first examine the performance of these systems under standard evaluation measures. Tables 5 and 6 give performance on ReCoNLL and OntoNotes 5.0 English datasets using standard P/R/F1. In English, Flair attains the best F1 in both datasets, although BERT attains higher recall for OntoNotes.[12]

BERT attains the highest F1 in Dutch (91.26) and Spanish (87.36); due to space limitations, tables are provided in the appendix (Tables 14-15). BERT-CRF performs similar or slightly worse than BERT in all languages, but generally attains lower standard deviation in multiple training runs, which suggests greater stability from using a CRF for structured predictions. The same observation also holds for Flair which also uses a CRF layer. We are not aware of prior work showing results from using BERT-CRF on English, Spanish, and Dutch. Souza et al. (2019) shows that the combination of Portuguese BERT Base and CRF does not show better performance than bare BERT Base, which agrees with our observations. F1 rankings are otherwise similar across languages. The performance of CharLSTM and CharCNN cannot be differentiated in English, but CharLSTM substantially outperforms CharCNN in Spanish (+2.53) and Dutch (+2.15).

[12]We are not aware of any open-source implementation capable of matching the F1 of 92.4 reported by Devlin et al. (2019). The gap between published and reproduced performance likely stems from the usage of the "maximal document context," while reimplementations process sentences independently, as is typical in NER. Performance of Flair is slightly worse than that reported in the original paper because we did not use the development set as additional training data.

Model	ALL	TCM-ALL	TCM-SEEN	TCM-UNSEEN
CHARLSTM	91.90	85.52 (±1.09)	87.36 (±0.70)	80.61 (±3.00)
CHARCNN	91.93	85.58 (±1.08)	87.55 (±1.11)	80.36 (±3.37)
FLAIR	**93.66**	**88.47 (±0.51)**	**89.75 (±0.73)**	**87.76 (±1.86)**
BERT	92.36	88.28 (±0.74)	89.69 (±0.89)	85.46 (±1.74)
BERT-CRF	92.29	87.02 (±0.71)	89.43 (±0.76)	79.59 (±1.25)

Table 7: Recall over all mentions and each type-confusable mention subset (ReCoNLL-2003 English)

Model	ALL	U-ANY	U-TOK.	U-TYPE
CHARLSTM	91.90	86.94 (±0.58)	87.32 (±0.63)	75.29 (±2.54)
CHARCNN	91.93	87.06 (±0.21)	87.48 (±0.18)	74.41 (±1.48)
FLAIR	**93.66**	**89.93 (±0.25)**	**90.31 (±0.19)**	78.53 (±2.94)
BERT	92.36	87.94 (±0.29)	88.02 (±0.31)	**85.29 (±2.04)**
BERT-CRF	92.29	87.55 (±0.14)	87.73 (±0.12)	82.12 (±1.53)

Table 8: Recall over all mentions and each unseen (U-) mention subset (ReCoNLL-2003 English)

4.3 TMR for English

We explore English first and in greatest depth because its test sets are much larger than those of the other languages we evaluate, and we have multiple well-studied test sets for it. Additionally, the CoNLL-2003 English test data is from a later time than the training set, reducing train/test similarity.

Revised CoNLL English. One of the advantages of evaluating using TMR metrics is that systems can be differentiated more easily. Table 7 gives recall for type-confusable mentions (TCMs) on ReCoNLL English. As expected, recall for TCMs is lower than overall recall, but more importantly, recall is less tightly-grouped over the TCM subsets (range of 8.17) than all mentions (1.76). This spread allows for better differentiation, even though there is a higher standard deviation for each score. For example, BERT-CRF generally performs very similarly to BERT, but scores 5.87 points lower for TCM-UNSEEN, possibly due to how the CRF handles lower-confidence predictions differently (Lignos and Kamyab, 2020). Flair has the highest all-mentions recall and the highest recall for TCMs, suggesting that when type-confusable mentions have been seen in the training data, it is able to effectively disambiguate types based on context.

Table 8 gives recall for unseen mentions. Although Flair attains higher overall recall, BERT attains higher recall on UNSEEN-TYPE, the set on which all models perform their worst. While there are few (85) mentions in this set, making assessment of statistical reliability challenging, this set allows us to identify an advantage for BERT in this specific subset: a BERT-based NER model is better able to produce a novel type for a token sequence

Model	ALL	TCM-ALL	TCM-SEEN
CHARLSTM	86.38	80.65 (±0.46)	82.24 (±0.46)
CHARCNN	86.28	79.80 (±0.41)	81.49 (±0.40)
FLAIR	89.42	**86.00 (±0.44)**	**87.39 (±0.51)**
BERT	**89.84**	84.72 (±0.18)	85.61 (±0.00)
BERT-CRF	89.32	85.46 (±0.40)	86.83 (±0.46)

Table 9: Recall over all mentions and each type-confusable mention subset (OntoNotes 5.0 English)

Model	ALL	U-ANY	U-TOKENS
CHARLSTM	86.38	72.71 (±0.80)	74.34 (±0.80)
CHARCNN	86.28	72.50 (±0.76)	74.10 (±0.75)
FLAIR	89.42	77.56 (±0.21)	79.05 (±0.16)
BERT	**89.84**	**79.97 (±0.11)**	**81.14 (±0.14)**
BERT-CRF	89.32	78.46 (±0.56)	79.63 (±0.61)

Table 10: Recall over all mentions and each unseen mention subset (OntoNotes 5.0 English)

only seen with other types in the training data.

OntoNotes 5.0 English. Examination of the OntoNotes English data shows that Flair outperforms BERT for type-confusable mentions, but BERT maintains its lead in overall recall when examining unseen mentions. Tables 9 and 10 give recall for type-confusable and unseen mentions.[13]

Summary. Table 11 gives a high-level comparison between BERT and Flair on English data. Using the TMR metrics, we find that the models that attain the highest overall recall may not perform the best on tough mentions. However, the results vary based on the entity ontology in use. In a head-to-head comparison between Flair and BERT on ReCoNLL English, despite Flair having the highest overall and TCM recall, BERT performs better than Flair on UNSEEN-TYPE, suggesting that BERT is better at predicting the type for a mention seen only with other types in the training data. In contrast, on OntoNotes 5.0 English, BERT attains the highest recall on UNSEEN mentions, but performs worse than Flair on TCMs. The larger and more precise OntoNotes ontology results in the unseen and type-confusable mentions being different than in the smaller CoNLL ontology. In general, Flair performs consistently better on TCMs while BERT performs better on UNSEEN mentions.

[13]We do not display results for TCM-UNSEEN and UNSEEN-TYPE as they each represent less than 1% of the test mentions. BERT's recall for TCM-UNSEEN mentions is 19.51 points higher than any other system. However, as there are 41 mentions in that set, the difference is only 8 mentions.

4.4 TMR for CoNLL-02 Spanish/Dutch

Tables 12 and 13 give recall for type-confusable and unseen mentions for CoNLL-2002 Spanish and Dutch.[14] The range of the overall recall for Spanish (11.80) and Dutch (17.13) among the five systems we evaluate is much larger than in English (1.76), likely due to systems being less optimized for those languages. In both Spanish and Dutch, BERT has the highest recall overall and in every subset.

While our proposed TMR metrics do not help differentiate models in Spanish and Dutch, they can provide estimates of performance on subsets of tough mentions from different languages and identify areas for improvement. For example, while the percentage of UNSEEN-TYPE mentions in Spanish (1.8) and ReCoNLL English (1.5) is similar, the performance for BERT for those mentions in Spanish is 34.04 points below that for ReCoNLL English. By using the TMR metrics, we have identified a gap that is not visible by just examining overall recall.

Compared with ReCoNLL English (6.3%) and Spanish (10.7%), there are far fewer type-confusable mentions in Dutch (0.2%). Given the sports-centric nature of the English and Spanish datasets, which creates many LOC/ORG confusable mentions, it is likely that their TCM rate is artificially high. However the near-zero rate in Dutch is a reminder that either linguistic or data collection properties may result in a high or negligible number of TCMs. OntoNotes English shows a similar rate (7.7%) to ReCoNLL English, but due to its richer ontology and larger set of types, these numbers are not directly comparable.

5 Conclusion

We have proposed Tough Mentions Recall (TMR), a set of evaluation metrics that provide a fine-grained analysis of different sets of formalized mentions that are most challenging for a NER system. By looking at recall on specific kinds of "tough" mentions—unseen and type-confusable ones—we are able to better differentiate between otherwise similar-performing systems, compare systems using dimensions beyond the overall score, and evaluate how systems are doing on the most difficult subparts of the NER task.

We summarize our findings as follows. For

[14]In Table 12, TCM-UNSEEN is not shown because it includes less than 1% of the test mentions (0.6%); in Table 13 UNSEEN-TYPE (0.2%) and TCM (0.2%) are not shown.

Dataset	Model	ALL	U-ANY	U-TOK.	U-TYPE	TCM-ALL	TCM-SEEN	TCM-UNSEEN
ReCoNLL-English	BERT				✔			
	FLAIR	✔	✔	✔		✔	✔	✔
Ontonotes 5.0	BERT	✔	✔	✔	N/A			N/A
	FLAIR				N/A	✔	✔	N/A

Table 11: Performance comparison between BERT and Flair on English data. A ✔ indicates higher recall under a metric. No comparisons are made for UNSEEN-TYPE and TCM-UNSEEN using OntoNotes due to data sparsity.

Model	ALL	U-ANY	U-TOK.	U-TYPE	TCM-ALL
CHARLSTM	79.76	70.56 (±0.93)	71.72 (±0.94)	46.25 (±3.86)	70.31 (±0.84)
CHARCNN	77.05	67.28 (±0.69)	68.13 (±0.51)	49.38 (±4.76)	68.48 (±0.68)
FLAIR	87.47	79.89 (±0.59)	81.65 (±0.50)	42.81 (±3.05)	77.02 (±1.23)
BERT	**88.85**	**83.04 (±0.58)**	**84.55 (±0.58)**	**51.25 (±3.39)**	**80.00 (±0.78)**
BERT-CRF	88.70	82.36 (±0.42)	83.93 (±0.40)	49.38 (±1.78)	79.74 (±0.63)

Table 12: Recall over all mentions and unseen and type-confusable mention subsets (CoNLL-2002 Spanish)

Model	ALL	U-ANY	U-TOKENS
CHARLSTM	77.35	66.32 (±0.23)	66.46 (±0.23)
CHARCNN	74.55	64.50 (±0.37)	64.61 (±0.32)
FLAIR	89.43	82.86 (±0.26)	83.00 (±0.26)
BERT	**91.68**	**86.65 (±0.17)**	**86.74 (±0.20)**
BERT-CRF	91.26	85.88 (±0.58)	85.94 (±0.58)

Table 13: Recall over all mentions and unseen mention subsets (CoNLL-2002 Dutch)

English, the TMR metrics provide greater differentiation across systems than overall recall and are able to identify differences in performance between BERT and Flair, the best-performing systems in our evaluation. Flair performs better on type-confusable mentions regardless of ontology, while performance on unseen mentions largely follows the overall recall, which is higher for Flair on ReCoNLL and for BERT on OntoNotes.

In Spanish and Dutch, the TMR metrics are not needed to differentiate systems overall, but they provide some insight into performance gaps between Spanish and English related to UNSEEN-TYPE mentions.

One challenge in applying these metrics is simply that there may be relatively few unseen mentions or TCMs, especially in the case of lower-resourced languages. While we are interested in finer-grained metrics for lower-resourced settings, data sparsity issues pose great challenges. As shown in Section 3.3, even in a higher-resourced setting, some subsets of tough mentions include less than 1% of the total mentions in the test set. We believe that lower-resourced NER settings can still benefit from our work by gaining information on pretraining or tuning models towards better performance on unseen and type-confusable mentions.

For new corpora, these metrics can be used to guide construction and corpus splitting to make test sets as difficult as possible, making them better benchmarks for progress. We hope that this form of scoring will see wide adoption and help provide a more nuanced view of NER performance.

Acknowledgments

Thanks to two anonymous EACL SRW mentors, three anonymous reviewers, and Chester Palen-Michel for providing feedback on this paper.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition. *Comput. Speech Lang.*, 44(C):61–83.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. 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 5360–5369, Hong Kong, China. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained BERT model

and evaluation data. In *Proceedings of the Practical ML for Developing Countries Workshop at ICLR 2020*.

Nancy Chinchor. 1998. Appendix b: MUC-7 test scores introduction. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

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*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Constantine Lignos and Marjan Kamyab. 2020. If you build your own NER scorer, non-replicable results will come. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 94–99, Online. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Sepideh Mesbah, Christoph Lofi, Manuel Valle Torre, Alessandro Bozzon, and Geert-Jan Houben. 2018. TSE-NER: An iterative approach for long-tail entity extraction in scientific publications. In *International Semantic Web Conference*, pages 127–143. Springer.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

F. Souza, Rodrigo Nogueira, and R. Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. *ArXiv*, abs/1909.10649.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

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.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *arXiv:1912.09582 [cs]*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. OntoNotes release 5.0 LDC2013T19.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jie Yang and Yue Zhang. 2018. NCRF++: An open-source neural sequence labeling toolkit. In *Proceedings of ACL 2018, System Demonstrations*, pages 74–79, Melbourne, Australia. Association for Computational Linguistics.

A Additional Tables

Please see the following pages for additional tables.

	Precision	Recall	F1
CHARCNN	76.74 (±0.36)	74.55 (±0.27)	75.63 (±0.26)
CHARLSTM	78.21 (±0.34)	77.35 (±0.21)	77.78 (±0.27)
FLAIR	90.11 (±0.15)	89.43 (±0.13)	89.77 (±0.14)
BERT	**91.26 (±0.23)**	**91.68 (±0.18)**	**91.47 (±0.18)**
BERT-CRF	90.75 (±0.47)	91.26 (±0.18)	91.00 (±0.32)

Table 14: Standard precision/recall/F1 for all types for each model trained on the CoNLL-2002 Dutch dataset

	Precision	Recall	F1
CHARCNN	77.75 (±0.22)	77.05 (±0.21)	77.40 (±0.20)
CHARLSTM	80.09 (±0.59)	79.76 (±0.63)	79.93 (±0.61)
FLAIR	86.96 (±0.23)	87.47 (±0.19)	87.21 (±0.20)
BERT	**87.36 (±0.52)**	**88.85 (±0.39)**	**88.10 (±0.45)**
BERT-CRF	87.25 (±0.38)	88.70 (±0.20)	87.97 (±0.29)

Table 15: Standard precision/recall/F1 for all types for each model trained on the CoNLL-2002 Spanish dataset

Mentions	ALL	GPE	PER	ORG	DATE	CARD	NORP	PERC	MONEY
UNSEEN-ANY	30.3	10.5	48.9	41.4	20.3	15.3	12.4	29.5	61.8
UNSEEN-TOKENS	29.4	9.9	48.0	40.8	19.7	14.9	12.0	29.5	60.2
UNSEEN-TYPE	0.9	0.6	0.9	0.6	0.6	0.4	0.4	0.0	1.6
TCM-ALL	7.7	11.5	1.7	4.9	3.2	15.4	18.5	0.0	5.1
TCM-SEEN	7.3	11.1	1.6	3.8	3.2	15.2	18.4	0.0	5.1
TCM-UNSEEN	0.4	0.4	0.1	1.1	0.1	0.2	0.1	0.0	0.0
Total (Count)	11,265	2,241	1,991	1,795	1,604	936	842	349	314

Table 16: Percentage of all mentions in each subset, with total mentions in the final row (OntoNotes 5.0 English). Due to space constraints, types are split across this table and the following one.

Mentions	TIME	ORD	LOC	WA	FAC	QUAN	PROD	EVENT	LAW	LANG
UNSEEN-ANY	41.5	3.6	39.1	**83.1**	80	73.3	52.6	47.6	75.0	22.7
UNSEEN-TOKENS	39.6	3.1	34.1	**78.9**	74.8	73.3	48.7	47.6	57.5	4.5
UNSEEN-TYPE	1.9	0.5	5.0	4.2	5.2	0.0	3.9	0.0	**17.5**	18.2
TCM-ALL	7.5	12.8	14	5.4	15.5	0.0	0.0	7.9	0.0	**54.5**
TCM-SEEN	7.5	12.8	14	2.4	14.8	0.0	0.0	7.9	0.0	**54.5**
TCM-UNSEEN	0.0	0.0	0.0	**3.0**	0.7	0.0	0.0	0.0	0.0	0.0
Total (Count)	212	195	179	166	135	105	76	63	40	22

Table 17: Percentage of all mentions in each subset, with total mentions in the final row (OntoNotes 5.0 English). Due to space constraints, types are split across this table and the preceding one.