Partial Evaluation Metrics in Named Entity Recognition

James Kong
Brandeis University
jameskong@brandeis.edu

Charles Sullivan
Brandeis University
charlessullivan@brandeis.edu

Abstract

Exact match mention F1 score is the standard evaluation metric for named entity recognition, and discussion and use of metrics which try to award credit to partially correct predicted mentions is not prevalent. We explore partial evaluation metrics for NER by implementing our own metrics and training and testing an NER system to analyze those metrics, and write a set of guidelines for determining when it might be correct to assign partial credit to an imprecise prediction made by an NER system.

1 Introduction

The typical evaluation metric for named entity recognition is exact match mention F1. Named entity recognition is often a first step in processing data before a downstream task such as relation extraction, and for that task partial matches may still be useful. In addition, when dealing with noisy data where exact matching is difficult, it could be valuable to identify predictions which are "close enough". For example, the test set we use (Derczynski et al., 2016) contains a gold annotation for "the New York Times" [ORG]. Our model predicts "New York Times" [ORG] instead, likely because "the" is not capitalized. When using mention F1, this prediction is counted as both a false positive and a false negative: a mention was predicted which did not appear exactly in the gold data, and the mention in the gold data was not predicted exactly. However, it is evident that this prediction refers to the same entity as the predicted mention. Though it would be better if the model predicted the gold mention exactly, counting this prediction as two mistakes seems harsh. Our goal with using partial credit metrics is to explore how to give some credit for cases like these. Partial credit metrics could also be more sensitive to model improvement, since ideally their scores would be higher for models which are able to identify useful mentions, even if those mentions are not perfect. On the other hand, mention F1 would decrease if a model made more partially correct predictions, which potentially incentivizes predicting fewer mentions. We implement several NER evaluation metrics that give partial credit to predicted mentions which do not exactly match the spans of gold-annotated mentions. We then train a BiLSTM-CRF on a corpus of Twitter data and use those metrics to evaluate the model. Finally, we perform an analysis of partial matches assigned credit by our metrics to explore whether partial credit metrics could be effective for NER. The main contributions of our work are an implementation of three metrics for NER which include awarding partial credit to some predictions and full credit to others, a set of guidelines which outline when we believe it is appropriate to assign partial credit to a predicted mention, and a detailed analysis of the partial credit assigned by our metrics to predictions made on twitter data.

2 Related Work

Our exploration of partial credit metrics for named entity recognition (NER) builds upon previous research that has acknowledged the limitations of strict exact matching, particularly when dealing with noisy data or when approximate matches still hold value. We identified several research studies that informed our approach to implementing and evaluating metrics such as Left boundary match, Right boundary match, and Partial (overlap) match, and our decision to award 0.5 true positives for these partial recognitions.

The concept of assigning partial credit in information extraction tasks was notably present in the Message Understanding Conferences. Specifically, (Chinchor and Sundheim, 1993) detailed the evaluation metrics for MUC-5, where they introduced categories like Correct (COR), Partial (PAR), and Incorrect (INC) for assessing system outputs against an answer key. We adopted a crucial aspect of their

scoring, which was that a Partial (PAR) match was counted as 0.5 true positives in the calculation of recall and precision. While their focus was broader than NER, the MUC-5 methodology for handling non-exact matches provided a precedent for our scoring system.

More directly related to NER, (Tsai et al., 2006) investigated various matching criteria for evaluating biomedical named entity recognition (Bio-NER) systems. They recognized that strict exact matches are not always necessary and systematically surveyed and implemented criteria including boundary relaxation methods like Left match, Right match, and Left/right match, as well as overlap-based methods such as Partial match. Their work evaluated Bio-NER systems using these relaxed criteria, demonstrating that metrics like right match and left match could be suitable alternatives to stricter evaluations, especially when exact boundaries are less critical. The definitions and evaluations of Left match, Right match, and Partial match in their study directly correspond to the metrics we implemented, and their findings supported the idea that such relaxed criteria can offer a more nuanced view of system performance.

Further reinforcing the utility of partial matching in NER evaluation, (Segura-Bedmar et al., 2013) described the SemEval-2013 Task 9, which included a subtask on recognizing pharmacological substances. For this NER subtask, their evaluation methodology explicitly went beyond strict exact boundary and type matching. They reported scores based on exact boundary matching, partial boundary matching (any overlap), and type matching (requiring some overlap and a correct type). They noted that the strict criterion could be overly restrictive and might overlook partial matches that are still valuable for downstream tasks, such as drug-drug interaction extraction. This work provided an example of overlap-based partial matching criteria being used in a community-wide NER challenge and highlighted the practical relevance of considering such matches.

Our project integrates and extends these research papers' ideas. We adopt the 0.5 true positive scoring convention established in MUC (Chinchor and Sundheim, 1993) and implement specific boundary-based and overlap-based partial metrics similar to those systematically analyzed by (Tsai et al., 2006). However, our contribution goes beyond applying these concepts to the distinct domain of noisy Twitter data (Derczynski et al., 2016). We researched and developed the explicit guidelines (detailed in Section 5.3) for judging the semantic appropriateness of partial matches, addressing when partial credit should reasonably be given in the context of noisy text. Furthermore, the cited works primarily focused on metric definition and score reporting, while we conducted a detailed manual analysis, guided by these novel guidelines, to quantitatively assess the quality and appropriateness of the partial credit awarded by each metric (Table 5). This analysis provides crucial validation that, for our task and data, these metrics predominantly reward semantically meaningful predictions, directly investigating whether partial credit aligns with human intuition of similarity in a challenging, noisy setting. While prior works established the mechanics and potential utility (Segura-Bedmar et al., 2013) of partial evaluation, our study adds a layer of qualitative validation and provides a concrete framework (our guidelines) for assessing the semantic relevance of partial NER metrics.

3 Data

We train and evaluate our model on the Broad Twitter Corpus, a dataset of tweets collected between 2012 and 2014. We use the recommended train, dev, and test splits. The corpus is annotated with types Person, Organization, and Location, and usernames and hastags can be annotated with any of these types. Mentions include typical organization names like "Facebook" [ORG], as well as mentions with capitalization and spelling inconsistencies ("SUAREZZZZZ" [PER], "the New York Times" [ORG]), and mentions containing twitter usernames and hashtags ("@eBay" [ORG], "@justinbeiber" [PER]). This noise is why we chose to use this dataset for these experiments: it may make it more difficult for our model to precisely identify the spans of mentions when there are capitalization errors, spelling errors, and username mentions - which often do not look like typical person, organization, or location names. As a result, it could be valuable for our model to identify useful portions of mentions, even if it fails to perfectly match the gold data. Counts of the number of tweets in each split and of each mention type are shown in Table 1. SeqScore (Lignos et al., 2023, Palen-Michel et al., 2021) was used to compute some counts, and the rest are from Flair's (Akbik et al., 2019) evaluation of the dataset. The dataset is not divided further into sentences, and predic-

BTC	Train	Dev	Test
Tweets	6338	998	2001
PER	3101	705	1602
LOC	1996	151	602
ORG	2267	270	792

Table 1: Counts of tweets and of mentions of each type in the Broad Twitter Corpus. Counts @ as part of the following mention when both are labeled with a B tag of the same type.

tions are made over whole tweets. By convention, the "@" character is tokenized from usernames, but both the "@" and the following username are annotated with a "B" tag, though they are parts of the same mention (Derczynski et al., 2016). This poses a challenge for evaluation, since when decoding labels this would typically be interpreted as two mentions. We make the following change to the dataset across the train, dev, and test splits: when an "@" is labeled with a "B" tag and the following token is labeled with a "B" tag of the same type, change that "B" tag to an "I". We made this change using a script which can be found on our GitHub repository. We also removed one tweet from the dev set which appeared to be tokenized incorrectly and was causing an error when loading the data using the Flair library.

4 Model

We trained a BiLSTM-CRF model using the `SequenceTagger` implementation from the Flair library[1] (Akbik et al., 2019). We chose to use Flair as it handles the decoding of BIO tags into mentions, which simplified the implementation of our partial evaluation metrics by abstracting away complexities related to different decoding schemes or the handling of invalid label sequences.

For token representations, we utilized a stack of pre-trained embeddings provided by Flair. This stack consisted of 100-dimensional Twitter WordEmbeddings (`WordEmbeddings('twitter')`) combined with contextual string embeddings, specifically the (`FlairEmbeddings('news-forward')` and `FlairEmbeddings('news-backward')`), each contributing 2048 dimensions (**?**). This resulted in a concatenated embedding of 4196 dimensions for each token. To mitigate overfitting, the model incorporated `WordDropout` with a probability of 0.05 and `LockedDropout` with a probability of 0.5 between layers.

The core architecture was a BiLSTM-CRF. The BiLSTM layer had a hidden state size of 256 units (for each direction, resulting in a 512-dimensional output from the BiLSTM). A Conditional Random Field (CRF) layer was used on top of the BiLSTM to jointly decode the optimal sequence of tags.

Our final model configuration was determined after a search of embedding and learning rate combinations. The selected model was trained with an initial learning rate of 0.05 and a mini-batch size of 32. We trained for a maximum of 25 epochs. The learning rate was managed using Flair's `AnnealOnPlateau` scheduler, which reduced the learning rate by a factor of 0.5 if no improvement on the development set's F1-score was observed for 3 consecutive epochs, with a minimum learning rate of 0.0001. The model was trained using Google Colab, and the training notebook detailing the hyperparameter search and final model training is available on our GitHub repository.

5 Results

5.1 Partial Evaluation Metrics

We implemented three metrics which award partial credit to predictions made which have different spans than the gold-annotated data.

- Left boundary match: left boundary of predicted span matches gold, type matches
- Right boundary match: right boundary of predicted span matches hold, type matches
- Partial match: Any overlap between predicted and gold spans, type matches

For all of these metrics, we awarded 1 true positive to predictions which exactly matched gold annotations in boundary and type, and 0.5 true positives to predictions which in type and the specified part of the span. In this way, we hope to avoid penalizing close predictions while still providing incentive for improvement: close predictions could still be valuable, but an exact match will always be worth more. Letting *EXACT* be the number of predictions with correct type and span, *PARTIAL* be the number of predictions given partial credit, *PREDICTIONS* be the number of predictions made by the model, and *REFERENCES* be the number of gold-annotated

[1] https://flairnlp.github.io/

mentions, equations 1 and 2 describe how we calculated precision and recall for each metric. These follow the same idea as metrics implemented for SemEval 2013 (Segura-Bedmar et al., 2013) in that they award the equivalent of 1 true positive to exactly correct mentions and give partial matches half the weight. Since these calculations change how true positives are counted, we cannot use the same denominator for precision and recall as is typically used, since the reduced weight to partial matches would be canceled out. Instead, we use the number of predictions made to calculate precision, and the number of mentions in the gold data to calculate recall. This reflects the fact that when our model predicts a partially correct mention, it misses half a point of available credit. F1-score is calculated in the standard way.

$$\text{Precision} = \frac{EXACT + 0.5 * PARTIAL}{PREDICTIONS} \quad (1)$$

$$\text{Recall} = \frac{EXACT + 0.5 * PARTIAL}{REFERENCES} \quad (2)$$

When assigning partial credit, up to one prediction can match each gold annotation. This is a given for the left and right boundary matching schemes since predicted spans do not overlap, but it is possible for multiple partially correct spans to overlap the same gold annotation: in that case it could be possible to award the same credit to two partial matches as to a single exact match. We prevent this by only giving credit to the first partial match found for a gold-annotated mention. It is also possible for one prediction to overlap two gold mentions: in that case, the prediction is only allowed to match one of the mentions.

5.2 Scores on Dev and Test Sets

We evaluate our model using standard mention level matching and each of our implemented metrics. Table 2 shows the F1 scores for each metric on the dev and test sets of the Broad Twitter Corpus. To further understand the composition of these scores, Figure 1 visualizes the number of matches receiving full (1.0) versus partial (0.5) credit on the test set. The number of exact matches found (2109) is consistent across the different strategies, as expected since they all count exact matches identically. The variation comes from the additional partial matches captured: 107 for the Overlap strategy, 46 for Left Boundary, and 72 for Right Boundary.

	Exact	Left	Right	Partial
Dev	74.59	75.32	75.22	75.85
Test	74.56	75.38	75.84	76.45

Table 2: F1 scores for each metric on the dev and test sets of the Broad Twitter Corpus.

Though we cannot directly compare our F1 score to others since we modified the dataset, our results are close enough to other reported results (Agarwal and Nenkova, 2021) that we can conclude the model is working. A comparison of precision, recall, and F1 score for each metric on the test set is shown in Figure 2. F1 scores for left, right, and partial match metrics are within 1-2 points higher than exact match on the dev and test sets, indicating that our model is finding some partial matches but not many. Scores are very similar between the dev and test sets, showing that our model appears to generalize well to unseen data. As expected, the score for partial matches is higher than that for left and right matches, since by definition partial matches can include left or right matches.

5.3 Analysis of Partial Matches

When evaluating our model on the dev set, we saved predicted mentions which matched gold mentions for each metric with how much credit each prediction was assigned. A sample is shown in Table 3, and the entire list can be found on GitHub.

Gold	Prediction	Credit
@ firefox	firefox	0.5
@ Forbes	@ Forbes	1.0
Microsoft	Microsoft	1.0

Table 3: Sample of partial matches output by our scorer.

We then filtered our results for matches worth 0.5 credit, and manually determined whether it was appropriate or not to assign partial credit to each match. We created guidelines before analyzing the matches, but refined them while performing our review. Our goal with these guidelines is to define matches which are meaningful and which humans can readily identify as the same or closely related to the gold mention as appropriate, and all others as inappropriate. The rules we used are as follows:

1. Partial match clearly refers to the same entity.

   For example, predictions can omit an @ token at the start of a username mention,

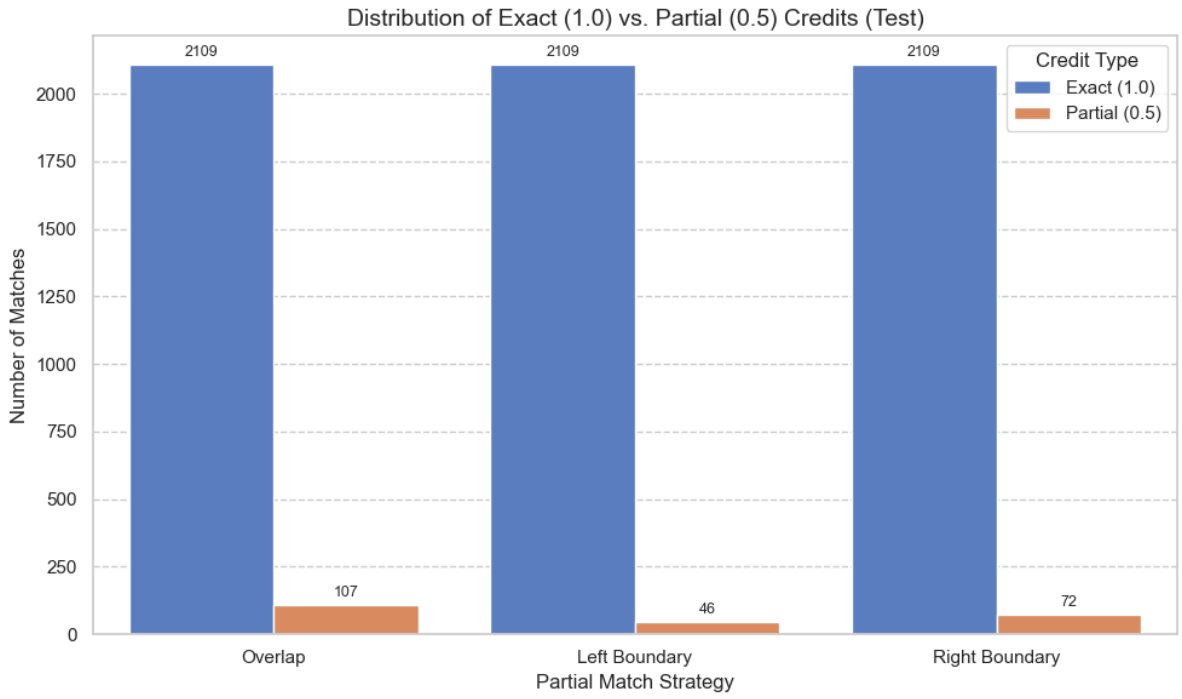

Figure 1: Distribution of exact (1.0 credit) and partial (0.5 credit) matches on the BTC test set for different partial matching strategies. Counts are shown above each bar.

but the mention is otherwise exactly identified. In some cases, a prediction captures a different valid name for the same entity: "VERIZON" [ORG] clearly refers to the same entity as "VERIZON WIRELESS" [ORG].

2. Partial match refers to a more general entity which contains the entity mentioned and is clearly related.

   We also think it is appropriate to award partial credit to predictions like "Philips" [ORG] for "Philips AVENT" [ORG]. There is a clear relation between these entities, since one belongs to the other.

3. Partial match does not add or omit context which significantly changes its meaning.

   In a couple cases we looked at, partial matches missed important information for identifying an entity, or included extra information which changed the meaning of the entity. For example, the prediction for the gold mention "New Hampshire" [LOC] is "Hampshire" [LOC], which could refer to a completely different place. In another case, a prediction contains two separate mentions. We decide that this is too significant an error to find partial credit appropriate, since it is not clear which entity the mention is referring to.

4. Partial match captures one of multiple valid entities in a mention.

   This case was added because some annotations in the BTC group mulitple username mentions into one. In those cases, partial credit is acceptable for identifying one of the entities in the group. For future work, we would make changes to the dataset so that no single annotation contains distinct username mentions.

5. Partial matches must be assigned to one of two categories: "Good" or "Bad".

   This is done so we can measure the proportion of appropriate to inappropriate partial credit.

We experimented with metrics which are type agnostic, but found that requiring type matches allows us to be more permissible with the boundaries of a prediction and makes it easier to identify that

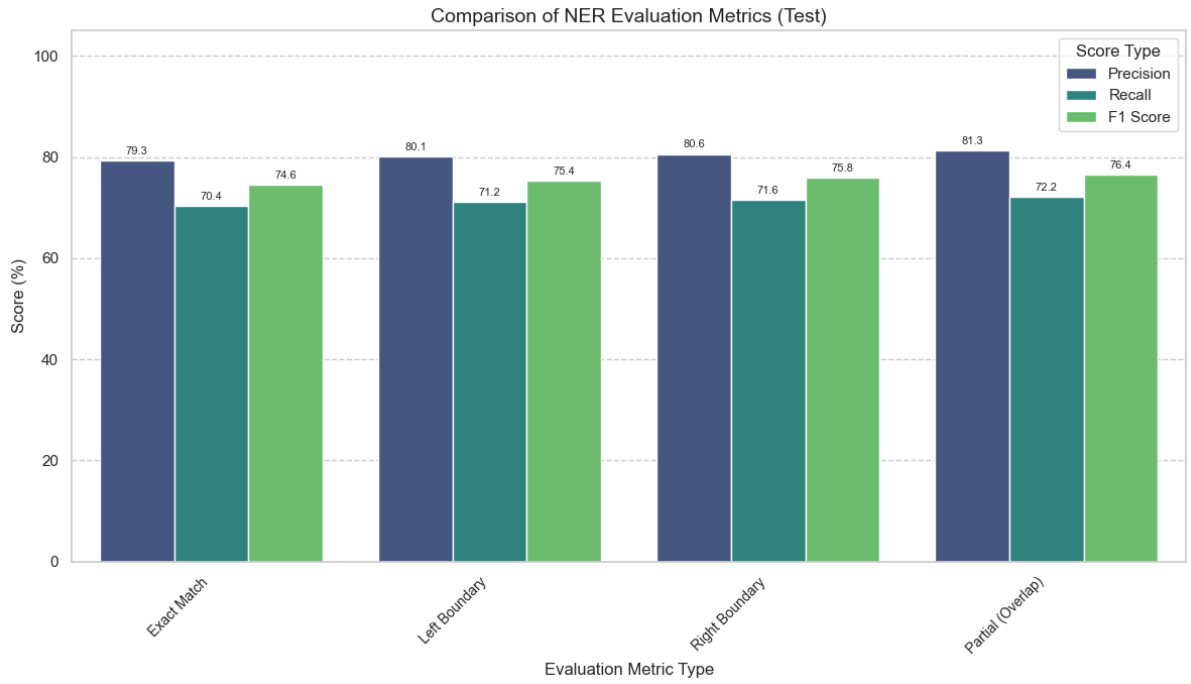

Figure 2: Comparison of evaluation metrics on the BTC test set.

a prediction and annotation refer to the same entity. For example, take the LOC in the BTC dev set "Cassese's". If the model predicts "Cassese", it could be argued that whether that is an acceptable partial match depends on the predicted type: if the prediction is of type PERSON, it no longer refers to the same entity, if it is of type LOC, it seems appropriate to give partial credit: a LOC with a very similar name has been identified. Also, it is typical for NER evaluation to consider prediction types, so by requiring type matching we make our metrics more comparable to the standard. We collectively reviewed the partial matches for each metric. Table 5 shows the results of our analysis, and Table 4 shows some example predictions and how we evaluated them. These samples demonstrate some robustness to noise in our metrics: the @ token is missed in the "@firefox" mention, and the hashtag is omitted in the prediction "Austin Mahone", but the left and right match metrics still count these predictions.

The vast majority of partial credit assigned was judged to be appropriate using our guidelines. Left matches often captured most of entity names, or variations on them, which clearly referenced the same entity. Right matches were very effective for finding predictions which missed the @ token at the start of a username. All overlap matches found were either in the set of left matches or the set of right matches. Note that left and right matches do not sum to partial matches because in two cases, our model made separate predictions which matched the left and right boundary respectively of the same gold mention. In that case, the overlap metric can only count one for partial credit. We find our results promising because they illustrate some robustness to noisy data, and a very high percentage of partial matches were readily identified as the same or a closely related entity, indicating that they could be useful for downstream tasks.

6 Conclusion

In this work, we explored the utility of partial evaluation metrics for Named Entity Recognition, moving beyond the often too-strict exact match F1 score, particularly for noisy datasets like the Broad Twitter Corpus. Our primary contributions include: the implementation of three distinct partial credit metrics (Left boundary, Right boundary, and general Partial (overlap) match) which award 0.5 true positives for inexact but relevant predictions, a set of guidelines developed to assess the appropriateness of such partial credit, and a detailed analysis of these metrics on a BiLSTM-CRF model trained on Twitter data.

Our results indicate that these partial evaluation metrics offer a more nuanced view of model perfor-

Gold	Prediction	Match Type	Analysis
Philips AVENT	Philips	Left	Good
@firefox	firefox	Right	Good
Austin Mahone #AustinMahone	Austin Mahone	Left	Good
Grimsby	Yume & Co Grimsby	Right	Bad
Hampshire	New Hampshire	Right	Bad

Table 4: Caption

Metric	% Partial	Good	% Good
Left	1.92	15/15	100.00
Right	1.67	11/13	84.62
Overlap	3.28	24/26	92.31

Table 5: Analysis of partial matches on the dev set.

mance. The F1 scores for our implemented metrics were consistently 1-2 points higher than the standard exact match F1 score on both development and test sets. While this increment is modest, it signifies that the model does make a number of predictions that, while not perfectly aligned with gold annotations, are semantically valuable. Crucially, our manual analysis, guided by our developed guidelines, revealed that the vast majority (over 84% for all metrics, and 100% for Left matches observed within Table 5) of the partial credit assigned by our metrics was indeed appropriate. This suggests that these metrics successfully identify and reward predictions that capture the core entity, such as variations in entity names or omissions of prefixes like "@" in usernames, which would be entirely penalized by exact matching.

For future work, the guidelines for awarding partial credit could be further refined and validated across diverse datasets, domains, and potentially through broader annotator agreement studies. Experimentation with more sophisticated weighting schemes for partial credit varying by the type or degree of overlap could offer better evaluation.

Limitations

Our findings are primarily based on the English Broad Twitter Corpus (BTC) from 2012-2014, and generalizability to other languages, domains, or newer data requires further study. Necessary modifications to the BTC also limit direct comparability with prior work using the original dataset. Furthermore, our analysis employed a specific BiLSTM-CRF model so results might differ with other models.

Methodologically, we focused on three specific partial match definitions (Left, Right, Overlap) with a fixed 0.5 credit, leaving exploration of alternative criteria or weighting schemes for future work. Manual validation of partial credit semantic valuability involves subjectivity as well and is difficult to scale.

References

Oshin Agarwal and Ani Nenkova. 2021. The Utility and Interplay of Gazetteers and Entity Segmentation for Named Entity Recognition in English. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3990–4002, Online. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Nancy Chinchor and Beth Sundheim. 1993. Muc-5 evaluation metrics. In *Proceedings of the 5th Conference on Message Understanding*, MUC5 '93, page 69–78, USA. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.

Constantine Lignos, Maya Kruse, and Andrew Rueda. 2023. Improving NER research workflows with SeqScore. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 147–152, Singapore. 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.

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.

Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung, and Wen-Lian Hsu. 2006. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7(1):92.

A Example Appendix

We tuned our model over the following learning rates: [0.05, 0.01, 0.005, 0.0025, 0.001]. We also used different combinations of embeddings to tune the model. For word embeddings, we found that the twitter word embeddings provided by Flair outperformed GloVe embeddings, and that the model improved when using both forward and backward contextual string embeddings.