Extensive Error Analysis and a Learning-Based Evaluation of Medical Entity Recognition Systems to Approximate User Experience

Isar Nejadgholi, Kathleen C. Fraser and Berry De Bruijn
National Research Council Canada
{isar.nejadgholi, kathleen.fraser,berry.debruijn}@nrc-cnrc.gc.ca

Abstract

When comparing entities extracted by a medical entity recognition system with gold standard annotations over a test set, two types of mismatches might occur, label mismatch or span mismatch. Here we focus on span mismatch and show that its severity can vary from a serious error to a fully acceptable entity extraction due to the subjectivity of span annotations. For a domain-specific BERT-based NER system, we showed that 25% of the errors have the same labels and overlapping span with gold standard entities. We collected expert judgement which shows more than 90% of these mismatches are accepted or partially accepted by the user. Using the training set of the NER system, we built a fast and lightweight entity classifier to approximate the user experience of such mismatches through accepting or rejecting them. The decisions made by this classifier are used to calculate a learning-based F-score which is shown to be a better approximation of a forgiving user's experience than the relaxed F-score. We demonstrated the results of applying the proposed evaluation metric for a variety of deep learning medical entity recognition models trained with two datasets.

1 Introduction

Named entity recognition (NER) in medical texts involves the automated recognition and classification of relevant medical/clinical entities, and has numerous applications including information extraction from clinical narratives (Meystre et al., 2008), identifying potential drug interactions and adverse affects (Harpaz et al., 2014; Liu et al., 2016), and de-identification of personal health data (Dernoncourt et al., 2017).

In recent years, medical NER systems have improved over previous baseline performance by incorporating developments such as deep learning models (Yadav and Bethard, 2018), contextual word embeddings (Zhu et al., 2018; Si et al., 2019), and domain-specific word embeddings (Alsentzer et al., 2019; Lee et al., 2019; Peng et al., 2019). Typically, research groups report their results using common evaluation metrics (most often precision, recall, and F-score) on standardized data sets. While this facilitates exact comparison, it is difficult to know whether modest gains in F-score are associated with significant qualitative differences in the system performance, and how the benefits and drawbacks of different embedding types are reflected in the output of the NER system.

This work aims to investigate the types of errors and their proportion in the output of modern deep learning models for medical NER. We suggest that an evaluation metric should be a close reflection of what users experience when using the model. We investigate different types of errors that are penalized by exact F-score and identify a specific error type where there is high degrees of disagreement between the human user experience and what exact F-score measures: namely, errors where the extracted entity is correctly labeled, but the span only overlaps with the annotated entity rather than matching perfectly. We obtain expert human judgement for 5296 such errors, ranking the severity of the error in terms of end user experience. We then compare the commonly used F-score metrics with human perception, and investigate if there is a way to automatically analyze such errors as part of the system evaluation. The code that calculates the number of different types of errors given the predictions of an NER model and the corresponding annotations is available upon request and will be released at `https://github.com/nrc-cnrc/NRC-MedNER-Eval` after publication. We will also release the collected expert judgements so that other researchers can use it as a benchmark for further investigation about this type of errors.

Proceedings of the BioNLP 2020 workshop, pages 177–186

2 What do NER Evaluation Metrics Measure?

An output entity from an NER system can be incorrect for two reasons: either the span is wrong, or the label is wrong (or both). Although entity-level exact F-score (also called strict F-score) is established as the most common metric for comparing NER models, exact F-score is the least forgiving metric in that it only credits a prediction when both the span and the label exactly match the annotation.

Other evaluation metrics have been proposed. The Message Understanding Conference (MUC) used an evaluation which took into account different types of errors made by the system (Chinchor and Sundheim, 1993). Building on that work, the SemEval 2013 Task 9.1 (recognizing and labelling pharmacological substances in biomedical text) employed four different evaluations: *strict match*, in which label and span match the gold standard exactly, *exact boundary match*, in which the span boundaries match exactly regardless of label, *partial boundary match*, in which the span boundaries partially match regardless of label, and *type match*, in which the label is correct and the span overlaps with the gold standard (Segura Bedmar et al., 2013). The latter metric, also commonly known as *inexact match*, has been used to compute inexact or relaxed F-score in the i2b2 2010 clinical NER challenge (Uzuner et al., 2011). Relaxed F-score and exact F-score are the most frequently used evaluation metrics for measuring the performance of medical NER systems (Yadav and Bethard, 2018). Other biomedical NER evaluations have accepted a span as a match as long as either the right or left boundary is correct (Tsai et al., 2006). In BioNLP shared task 2013, the accuracy of the boundaries is relaxed or measured based on similarity of entities (Bossy et al., 2013). Another strategy is to annotate all possible spans for an entity and accept any matches as correct (Yeh et al., 2005), although this detailed level of annotation is rare.

Here, we focus on the differences between what F-score measures and the user experience. In the case of a correct label with a span mismatch, it is not always obvious that the user is experiencing an error, due to the subjectivity of span annotations (Tsai et al., 2006; Kipper-Schuler et al., 2008). Existing evaluation metrics treat all such span mismatches equally, either penalizing them all (exact F-score), rewarding them all (relaxed F-score), or based on oversimplified rules that do not generalize across applications and data sets. We use both human judgement and a learning-based approach to evaluate span mismatch errors and the resulting gap between what F-score measures and what a human user experiences. We only consider the information extraction task and not any specific downstream task.

3 Types of Errors in NER systems

While the SemEval 2013 Task 9.1 categorized different types of matches for the purpose of evaluation, we further categorize mismatches for the sake of error analysis. We consider five types of mismatches between annotation and prediction of the NER system. Reporting and comparing the number of these mismatches alongside an averaged score such as F-score can shed light on the differences of NER systems.

- **Mismatch Type-1, Complete False Positive**: An entity is predicted by the NER model, but is not annotated in the hand-labelled text.
- **Mismatch Type-2, Complete False Negative**: A hand labelled entity is not predicted by the model.
- **Mismatch Type-3, Wrong label, Right span**: A hand-labelled entity and a predicted one have the same spans but different tags.
- **Mismatch Type-4, Wrong label, Overlapping span**: A hand-labelled entity and a predicted one have overlapping spans but different tags.
- **Mismatch Type-5, Right label, Overlapping span**: A hand-labelled entity and a predicted one have overlapping spans and the same tags.

We focus on Type-5 errors and show that treating these mismatches is not a trivial task. Previous works have shown that some Type-5 mismatches are completely wrong predictions while others are fully acceptable predictions resulting from the subjectivity and inconsistency of span annotations (Tsai et al., 2006).

Figure 1 shows several examples of error Type-5. In the first example, *an adenosine - thallium stress test* is annotated as a *test*, while the NER system extracts *thallium stress test* as a *test*. Here, what NER extracted is partially correct but misses an

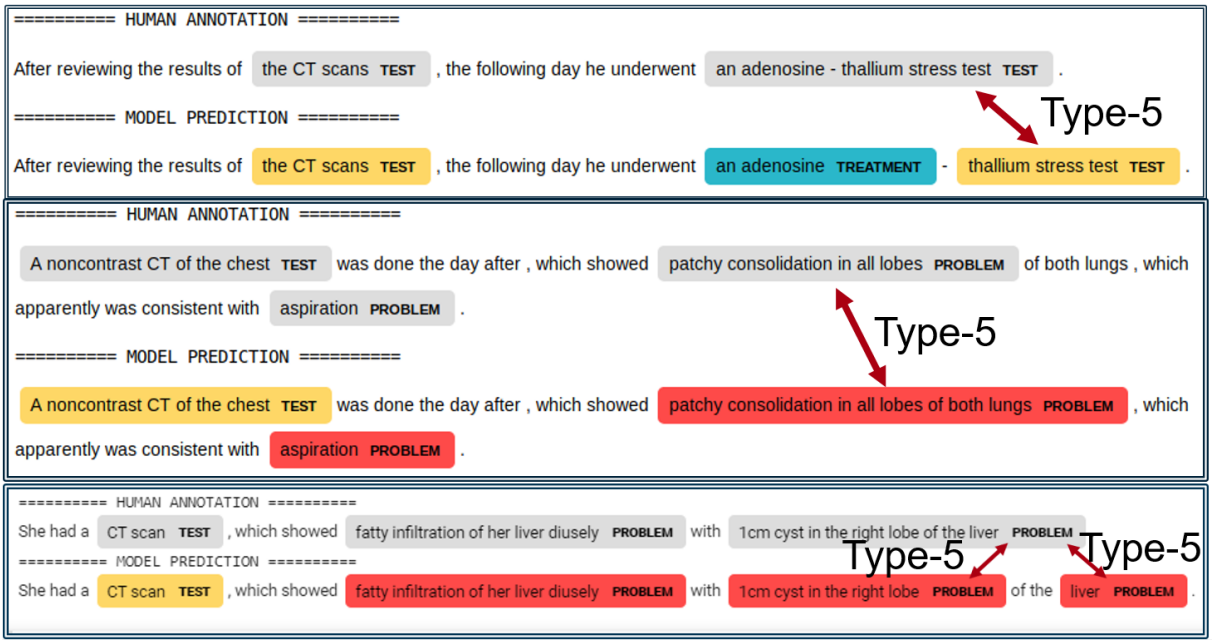


Figure 1: Examples of Type-5 error. We used the visualisation tool developed in (Zhu et al., 2018)

important part of the entity. Whether the extracted entity is acceptable may depend on the downstream task. In the next sentence, *patchy consolidation in all lobes* is annotated as a *problem*, but the NER system extracted *patchy consolidation in all lobes of both longs* as the *problem*. Here, the system's prediction is more complete than the annotated entity, and so it appears to be a fully acceptable prediction. In the last example, according to human annotation, *1cm cyst in the right lobe of the liver* is a *problem*, but the NER system extracts two entities from the same phrase, 1) *1cm cyst in the right lobe* as a *problem* and 2) *liver* as another *problem*. While the first extracted entity is correct and may be acceptable the second one is completely wrong.

4 Datasets and Models

We consider two medical text datasets, one clinical and the other biomedical. We analyse the errors of three models for each dataset to cover a variety of deep learning models.

i2b2 dataset: The i2b2 dataset of annotated clinical notes was introduced by (Uzuner et al., 2011) in a shared task on entity recognition and relation extraction. The texts, consisting of de-identified discharge summaries, have been annotated for three entity types: problems, tests, and treatments. There are two versions of this dataset, as the version that was released to the wider NLP community contains fewer texts than in the original shared task. We use the second version, which has become an important benchmark in the literature on clinical NER (Bhatia et al., 2019; Zhu et al., 2018). There are 170 documents (16520 entities) in the i2b2 train set and 256 documents (31161 entities) in its test set.

The i2b2 dataset was annotated by community annotators with carefully crafted guidelines. The ground truth generated by the community obtained F-measures above 0.90 against the ground truth of the experts (Uzuner et al., 2011).

MedMentions dataset: The MedMentions dataset was released in 2019 and contains 4,392 abstracts from biomedical articles on PubMed (Mohan and Li, 2019). The abstracts are annotated for UMLS concepts and semantic types. The fully annotated dataset contains 127 semantic types and these classes are highly-imbalanced. The creators of the dataset also provide a version which has been annotated with only a subset of the most relevant concepts, called 'st21pv' (*21 semantic types from preferred vocabularies*); we consider this version in the current work. While fewer papers have been published on MedMentions to date, it represents an interesting challenge to NLP systems due to its imbalanced and high number of classes, and some observed inconsistencies in the annotations (Fraser et al., 2019). There are 3513 documents (162,908 entities) in the st21pv train set and 879 documents (40,101 entities) in the test set.

MedMentions was annotated by a team of profes-

sional annotators with rich experience in biomedical content curation. The precision of the annotation in MedMention is estimated as 97.3% (Mohan and Li, 2019).

Model Structures: We explore a variety of NER deep learning models. For all the models we follow the commonly used deep learning structure consisting of a pretrained embedding model and supervised prediction layers. For embedding, we explore three different models: a non-contextualized embedding model (Glove), general domain contextualized embedding model (BERT pretrained on general domain text) and a domain-specific contextualized embedding model (BERT pretrained on domain-specific text corpora). For the i2b2 dataset, we consider *Glove+bi-LSTM+CRF* (Pennington et al., 2014), *BERT+linear* (Devlin et al., 2018) and *ClinicalBERT+linear* (Alsentzer et al., 2019) models. For the st21pv MedMentions dataset, we consider *Glove+bi-LSTM+CRF*, *BERT+linear* and *BioBERT+linear* models (Lee et al., 2019). Clinical BERT is pretrained on clinical notes (similar to i2b2) and BioBERT is pretrained on biomedical articles from PubMed (similar to st21pv).

5 Analysis of Error Types Across Models and Datasets

Further investigation of Type-5 errors is only worthwhile if a significant proportion of the errors belong to this group. We looked at the distribution of error types across datasets and NER models, described in Section 4, and visualized the results in Figure 2. By calculating the distribution of error types, we observed that for all assessed models at least 20% of the errors are recognized as Type-5 mismatches.

Moreover, for both datasets, we observed that better NER models generate more Type-5 errors. Models based on general BERT outperform glove-based models in terms of both exact and relaxed f-score and they also generate relatively more Type-5 errors. Same pattern is observed when comparing domain-specific BERT models with general BERT models. This observation may be explained with the fact that contextualized embeddings combine the meaning of words through attention mechanism and the span information might be more vague in the resulting representation. Figure 3 shows exact F-score, relaxed F-score and the proportion of Type-5 mismatches to the total number of errors, for all the models and datasets. This analysis implies that proper handling of Type-5 errors becomes more important for comparison of modern strong NER systems.

6 Expert Judgement on Type-5 Errors

We considered an information extraction task and asked a medical doctor to assess the Type-5 errors made by the BioBERT NER model on the st21pv dataset and either confirm or reject the extracted entity with granular scores. Our goal is to: 1) investigate the proportion of Type-5 extracted entities that are acceptable, 2) set a benchmark of human experience from Type-5 errors.

Human Judgement Scheme: The following scoring scheme is used by the expert for scoring the acceptability of Type-5 mismatches for the BioBERT-based model trained with the st21vp dataset. The Type-5 mismatches are identified and the expert is given the original sentence in the test set, the annotated (gold-standard) entity, and the entity predicted by the NER model for all 5296 Type-5 mismatches.

SCORE = 1: The predicted entity is wrong and gets rejected. For example, while *gene transfer* is annotated as a *research_activity* in the test set, the NER extracted *gene* as *research_activity*.
SCORE = 2: The predicted entity is correct but an important piece of information is missing when seen in the full sentence. The prediction is partially accepted by the expert. For example, *injury of lung* is labeled as *injury_or_poisoning* in the test set, but the NER extracts only the word *injury* as *injury_or_poisoning*.
SCORE = 3: The predicted entity is correct but could be more complete. The prediction is accepted by the expert. The entity *normal HaCaT lines* is annotated as *anatomical_structure* in the test set but the NER extracts only *HaCaT lines* with the same label.
SCORE = 4: The predicted entity is equally correct and is accepted by the expert. As an example the annotated entity in test set is *196b-5p,* as an *anatomical_structure* but the NER extracts *-196b-5p,* as an entity with the same tag.
SCORE = 5: The predicted entity is more complete than the annotated entity and is accepted by the expert. The annotated entity in the test set is *drugs* with the tag *chemical* and the NER extracts *Alzheimer's drugs* with the same tag.

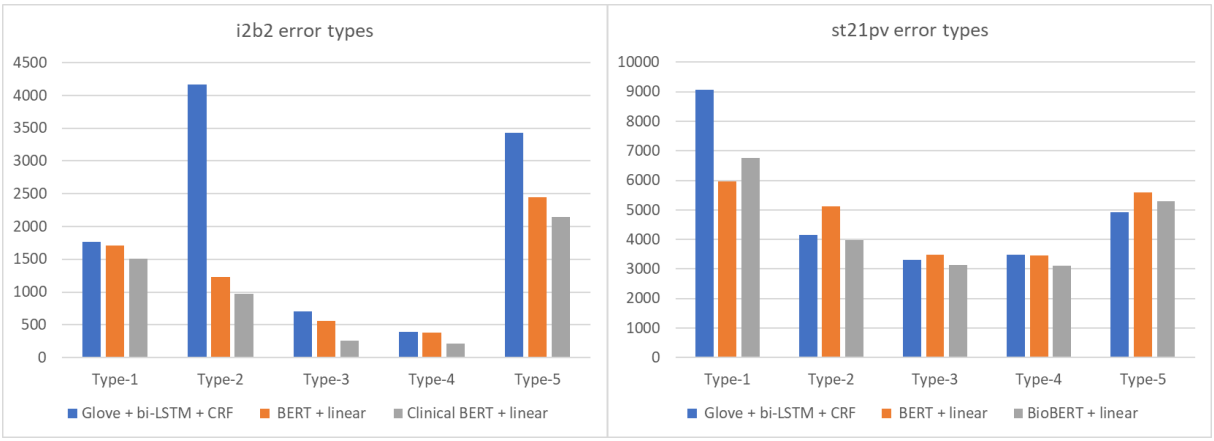


Figure 2: Types of errors made on the i2b2 and MedMentions-st21pv datasets

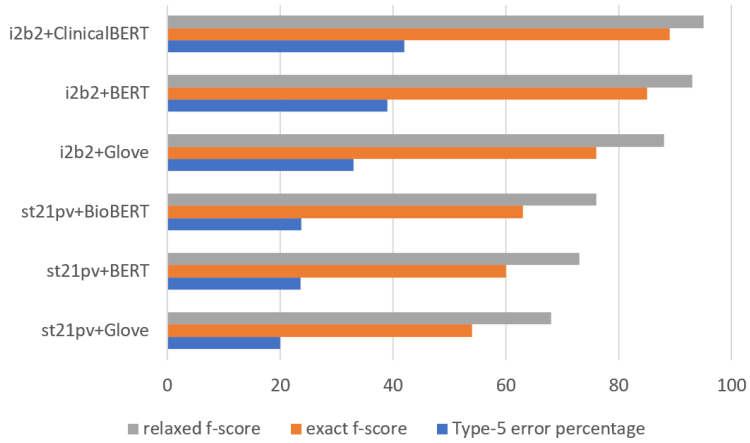


Figure 3: The change of relative proportion of Type-5 errors across dataset and models as the f-scores change

Results of Human Judgement Analysis: The results of the expert judgement are summarized in Figure 4.

- Almost 40% of the Type-5 errors are scored as 5. This means that in 40% of the cases the prediction of the NER is more complete than the entity labeled in its test set.
- 70% of the extracted entities scored 3 or above and are fully accepted by the expert.
- 21% of the Type-5 mismatches are scored as 2. These are accepted as a correct entity extraction when seen out of the context, but in the context of a given sentence they lack an important piece of information. Depending on the downstream tasks, they might be an acceptable prediction or not.
- Only 9% of the extracted entities are totally rejected by the expert.

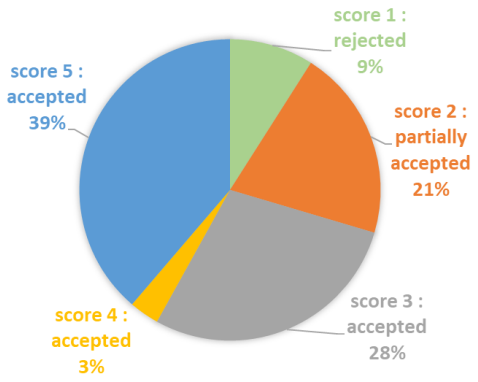
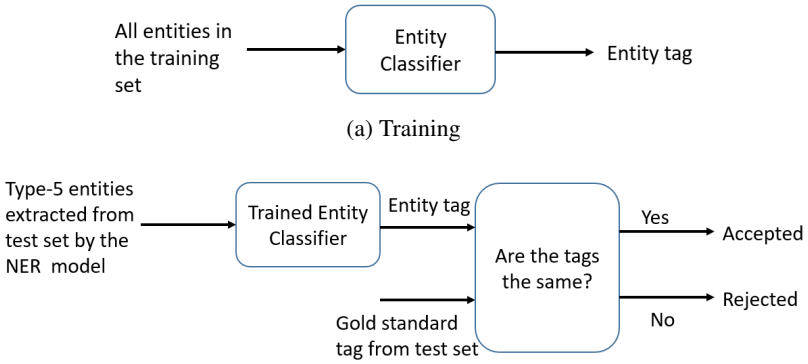


Figure 4: Results of expert judgement for Type-5 mismatches of the BioBERT-based NER model trained with MedMentions-st21pv dataset.

7 Entity Classifier for Automatic Refining of Type-5 Mismatches

We propose that an entity classifier can be trained to predict the tag of entities extracted by the NER model and the predicted tag can be used to distin-



(a) Training

(b) Using for evaluation of the NER model

Figure 5: Workflow of the proposed entity classifier

guish between acceptable and unacceptable Type-5 errors. Figure 5 shows the workflow of the proposed method. Using the training dataset of the NER model, we train an entity classifier with gold standard entities as inputs and their assigned tags as outputs. For this classifier, the span is given and the tag is the only information that has to be learned. Although the full context of the sentence helps the NER model to learn a better representation of the entity, many entities can be classified without seeing the full sentence and this is what the entity classifier learns.

For Type-5 entities, the human annotators and the NER already agree on the tag and it is only the span that is in disagreement. So, the intuition here is that the entity classifier can confirm or reject the tag predicted by NER, given the identified span. This classifier is meant to play a third party role that has seen the variety of span annotations in the training dataset and performs the task that the human expert did in Section 6. This classifier is trained once for each dataset and is not dependent on the type of the NER model.

7.1 Building the Training Data for the Entity Classifier

In order to build a training dataset for the entity classifier, we extracted pairs of *(entity, tag)* from the IOB annotated dataset. The entity classifier should also be able to identify cases where the extracted entity does not belong to any of the pre-defined tags. For this reason we add the label *other* to the list of tags of the classifier. To find examples of the *other* class, we used the spaCy library (Honnibal and Montani, 2017) to extract all the noun chunks that are out of the boundaries of tagged entities and randomly chose a number of them. We limited the size of the *other* class to the average size of classes related to the existing tags.

7.2 Classifier Structure

For the classifier structure, we chose to use a DistilBERT model (Sanh et al., 2019) with a linear prediction layer. DistilBERT is a distilled version of BERT that is an optimum choice when fast inference is required. Since this classifier is going to be used for evaluation and error analysis and is not the main focus of building an NER model, the lightweight and fast inference is an important practical criterion. We train the classifier only one epoch for both datasets. When trained on the train set and tested on the test set, we achieved 89% F-score for i2b2 and 77% F-score for st21pv dataset.

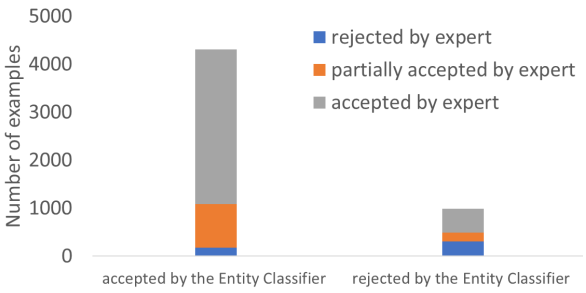


Figure 6: Comparison of decisions made by the human expert and the entity classifier for the Type-5 mismatches of BioBERT NER and st21pv dataset.

7.3 Using the Entity Classifier for Refining Type-5 Mismatches

By building the entity classifier, our goal is to refine the Type-5 errors and separate the acceptable predictions of the NER from the unacceptable. For instance, in the last example shown in Figure 1

Annotated in test set	Tag in test set	Extracted by NER	Tag from Entity classifier	Decision
Central pathology	biomedical discipline	Central	Spatial concept	Reject
Therapies	healthcare activity	Agonist Therapies	healthcare activity	Accept

Table 1: Examples of accepted and rejected Type-5 mismatches using the entity classifier (st21pv dataset).

there are two Type-5 errors. We feed the two extracted entities *'1 cm cyst in the right lobe'* and *'liver'* to the entity classifier trained for i2b2 dataset. The classifier predicts the tag *'problem'* for the extracted entity *'1 cm cyst in the right lobe'* and *'Other'* for the extracted entity *'liver'*. Using these predictions we decide that the first entity is acceptable, since although the span of the extracted entity does not match the annotation, the classifier still recognizes it as a member of the correct class. We reject the extracted entity *'liver'* as a *'problem'* since the classifier recognizes it as not being a *'problem'*. Table 1 shows examples of rejected and accepted Type-5 mismatches from the st21pv dataset.

7.4 Comparing the Classifier and the Expert

Figure 6 shows the comparison between the expert's judgment and the classifier's judgement about Type-5 mismatches for the BioBERT NER model on st21 pv dataset.

Our analysis shows that 96% of the entities accepted by the classifier are also accepted or partially accepted by the expert, and 86% of the entities accepted or partially accepted by the expert are accepted by the classifier as well. The classifier and the expert disagree about 17% of the entities. In 24% of the disagreements, the probabilities assigned to the tags generated by the entity classifier are low (less than 0.5) and our manual investigation shows that the classifier's prediction is mostly wrong in these cases. These mistakes mostly occurs in 5 classes namely *anatomical_structure*, *biologic_function*, *chemical*, *finding* and *health_care_activity*.

We also observed that this classifier is not able to distinguish between accepted and partially accepted entities extracted by the NER model, which is one of the limitations of this method. The probabilities assigned to the tags is 0.89 ± 0.17 for accepted entities, 0.88 ± 0.17 for partially accepted entities, and 0.78 ± 0.23 for rejected entities.

8 Refining Type-5 Mismatches Across Datasets and Models

Figure 7 shows how the entity classifier refines Type-5 errors across models and datasets. Consistently, a significant proportion of Type-5 errors are accepted by the entity classifier. For example, for the Glove-based model trained on i2b2 dataset, the entity classifier accepts 90% of Type-5 errors which is 26.6% of the total number of the errors penalized by the exact f-score. The proportion of accepted Type-5 mismatches to the total number of errors is 31.11% for *i2b2+BERT*, 33.23% for *i2b2+ClinicalBERT*, 17.95% for *st21pv+Glove*, 19.55% for *st21pv+BERT* and 19.36% for *st21pv+BioBERT*. To sum up, about 20% to 30% of the mismatches penalized by exact f-score are accepted by the entity classifier.

9 Learning-Based F-score

The trained entity classifier can be leveraged for F-score calculation. Here, instead of penalizing all the type-5 mismatches as in exact F-score or rewarding all of them in relaxed F-score, we penalize the type-5 mismatches that are rejected by the classifier and reward the rest of them. In other words, this F-score penalizes errors of Type-1, Type-2, Type-3, Type-4 and the Rejected Type-5 mismatches. Accepted Type-5 mismatches and exact matches are rewarded.

9.1 Evaluation of the Learning-Based F-score

We use the expert judgement collected in Section 6 to quantify human experience for the BioBERT-based NER model on st21pv dataset and then use that as a benchmark to evaluate the proposed learning-based F-score. We consider two scenarios based on the scores described in Section 6, 1) a strict user that only accepts scores equal to or above 3, 2) a forgiving user that accepts scores equal to or above 2. We calculated the F-score for each scenario and investigated the error of exact F-score, relaxed F-score and the proposed F-score to each of these scenarios. Table 2 shows that in applications where strict evaluation of the NER

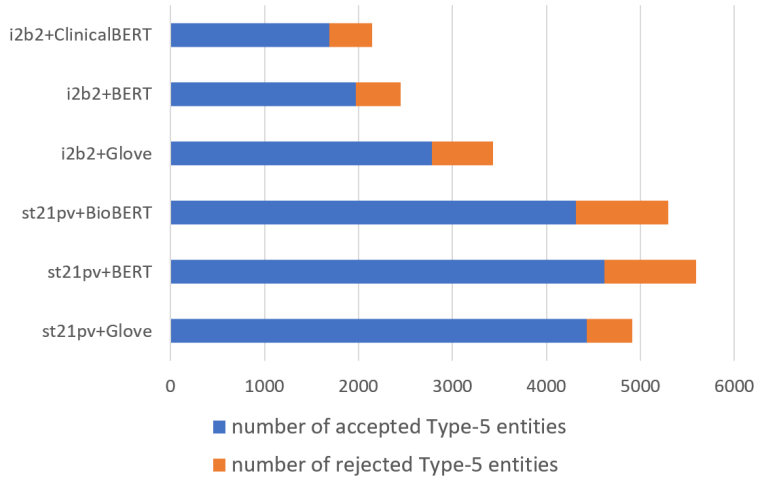


Figure 7: Number of accepted/rejected Type-5 mismatches by the entity classifier

F-score	err. wtr strict user	err. wtr forgiving user
Exact	**-4.3%**	-5.5%
Proposed	5.9%	**4.7%**
Relaxed	8.2%	7.1%

Table 2: Comparing F-scores with human experience.

is needed, exact F-score is better than both proposed and relaxed f-score and results in the least error with respect to the human experience. However, in cases that partially accepted entities can be considered as useful predictions, the proposed method results in the least disagreement with human experience. A better classifier would be able to model human preferences better, and thus make the learning-based F-score a stronger alternative to exact or relaxed F-scores. Another important finding from Table 2 that when choosing between exact and relaxed F-score, exact is the better metric to choose.

Figure 8 shows how the proposed F-score can be compared with exact and relaxed F-score. We only have annotations for the BioBERT+stpv dataset and for the rest of the models we cannot evaluate the F-score with respect to human experience. As expected, from this figure we observe that for all the models, the proposed F-score is a forgiving one and is much closer to the relaxed F-score than the exact F-score.

10 Discussion

We highlighted the fact that when we evaluate NER systems by comparing extracted and annotated entities across a test set, for a significant part of the errors that are penalized by the exact F-score, the label is recognized correctly and the span has overlap with the annotated entity. We referred to this type of error as Type-5 mismatch and for six NER models (3 model structures and 2 datasets) showed that at least 20% of the errors belong to this category. The previous literature has raised the issue that in the case of medical NER, many such predictions are valid and useful entity extractions and penalizing them is a flaw of evaluation metrics. However, distinguishing between acceptable and unacceptable predictions when the label is correct and the span overlaps is not trivial.

We argue that the best evaluation metric is the one that reflects the human experience of the system best. We collected human judgement about a all Type-5 errors made by a NER model based on BioBERT embeddings, trained with st21pv dataset and showed that almost 70% of such errors are completely acceptable and only 10% of them are rejected by the user. The rest of the predictions are acceptable entities for the associated tags but lack important information when seen in the context.

Setting human experience as a benchmark, we suggested that expert judgement can be approximated by a decision made by an entity classifier. The entity classifier can be trained using the training set of an NER. While the NER model looks at the context and identifies the type of the entity and a partially correct span, this classifier looks at the extracted entities out of context and decides whether with the partially correct span, the extracted entity can still belong to the predicted class or not. The entity classifier trained on st21pv dataset accepts more than 80% of Type-5 errors made by BioBERT-based NER model trained with the same

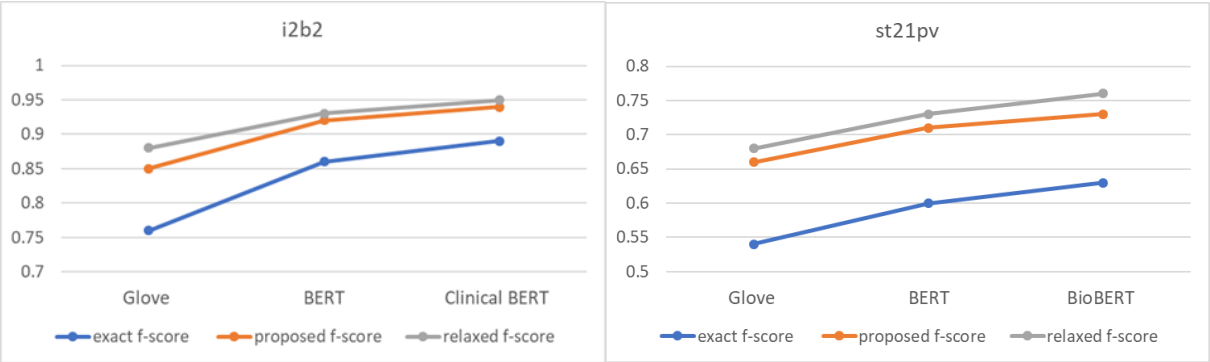


Figure 8: Comparison of f-scores.

dataset, 96% of which is also accepted by the expert user. The proposed entity classifier is trained for each NER training set once and can be used to evaluate any NER model trained on that dataset, regardless of the structure of the NER model being evaluated. We used a computationally inexpensive model structure and encourage researchers to use this model in order to automatically evaluate Type-5 mismatches. Reporting the distribution of errors across all error types and also accepted and rejected Type-5 errors, will allow us to compare our models in a variety of dimensions and sheds light on how these models behave differently for detecting labels and spans.

Accepting some Type-5 errors as useful predictions can be translated to F-score calculation by not penalizing the accepted entity extractions. We did this calculation separately for the cases that were accepted by human expert or the classifier, and showed that the F-score resulting from the classifier is closer to the judgement of a forgiving user than both the exact and the relaxed F-score. In cases where a strict evaluation of the system is desired, exact F-score is a better approximation of human experience, due to the fact that the entity classifier is a forgiving one and accepts most of the cases that are partially accepted by the expert.

We only collected human judgement on the decisions made by NER model for one model and one dataset. Further investigation is needed to confirm or reject our observations and to investigate the limitations and potential capabilities of training an entity classifier alongside a NER model and using that for error analysis. Also, further research is needed to find a way of distinguishing between partially accepted and accepted entity extractions, which is a necessary tool for measuring the experience of a strict user. Using extra sources of training data other than the NER training dataset may be a way to improve the judgements of the entity classifier. We used this classifier for error analysis and refining of Type-5 errors. In future work, we can look at the possibility of using this classifier as a refining tool for all types of mismatches or a post-processing tool without the need for annotation to identify the types of mismatches.

11 Conclusion

Medical NER systems that are based on most recent deep learning structures generate a high amount of outputs that match with the hand-labelled entities in terms of tag but only overlap in the span. While the exact f-score penalizes all of these predictions and relaxed f-score credits all of them, a human user accepts a significant proportion of them as valid entities and rejects the rest.
A reformatted version of the NER training dataset can be used to train an entity classifier for evaluation of extracted entities with right label and overlapping span. We showed that there is a high degree of agreement between human expert and this entity classifier in accepting or rejecting span mismatches. This classifier is used to calculate a learning-based evaluation metric that outperforms relaxed F-score in approximating the experience of a forgiving user.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–

78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Parminder Bhatia, Busra Celikkaya, and Mohammed Khalilia. 2019. Joint entity extraction and assertion detection for clinical text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 954–959.

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Philippe Bessières, and Claire Nédellec. 2013. BioNLP shared task 2013–an overview of the bacteria biotope task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 161–169.

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*.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kathleen C Fraser, Isar Nejadgholi, Berry De Bruijn, Muqun Li, Astha LaPlante, and Khaldoun Zine El Abidine. 2019. Extracting UMLS concepts from medical text using general and domain-specific deep learning models. *EMNLP-IJCNLP 2019*, page 157.

Rave Harpaz, Alison Callahan, Suzanne Tamang, Yen Low, David Odgers, Sam Finlayson, Kenneth Jung, Paea LePendu, and Nigam H Shah. 2014. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug Safety*, 37(10):777–790.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Karin Kipper-Schuler, Vinod Kaggal, James Masanz, Philip Ogren, and Guergana Savova. 2008. System evaluation on a named entity corpus from clinical notes. In *Language resources and evaluation conference, LREC*, pages 3001–3007.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. 2016. Drug-drug interaction extraction via convolutional neural networks. *Computational and Mathematical Methods in Medicine*.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, 17(01):128–144.

Sunil Mohan and Donghui Li. 2019. MedMentions: A large biomedical corpus annotated with UMLS concepts. *arXiv preprint arXiv:1902.09476*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Isabel Segura Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). volume 2, pages 341–350. Association for Computational Linguistics.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

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.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.

Alexander Yeh, Alexander Morgan, Marc Colosimo, and Lynette Hirschman. 2005. BioCreAtIvE task 1a: gene mention finding evaluation. *BMC Bioinformatics*, 6(S1):S2.

Henghui Zhu, Ioannis Ch Paschalidis, and Amir Tahmasebi. 2018. Clinical concept extraction with contextual word embedding. *arXiv preprint arXiv:1810.10566*.