# Cross-Domain Comparison of BERT and BiLSTM for Performing NER

**Caroline Sofie Skovby, Petya Ivanova Petrova, Andreea-Larisa Rosu**
BSc in Data Science
IT University of Copenhagen
[cssk, pety, rosu]@itu.dk
Github: https://github.itu.dk/Pety/NLP_project

## Abstract

This study focuses on three different datasets written in English, used for training, testing, and evaluating the performance of the two deep learning language models BERT and BiLSTM. Our goal is to showcase the strengths and limitations of the two chosen models in the aspect of Named Entity Recognition while considering the diversity of the data used, data that ultimately proved to be crucial in the way we lead our investigation.

## Introduction

The concept of Natural Language Processing (NLP) brings into our view new ways of manipulating data through machine learning specialised in working with lines of text. An ever-growing and vastly used subdomain of NLP is Named Entity Recognition (NER). NER, the act of extracting information and classifying named entities, however, proved to be a challenging task in NLP. From language ambiguity causing inconsistencies in annotations to the presence of unfamiliar words, our aim in this project is to try to look into such issues by finding the best technique for doing Named Entity Recognition.

The techniques spoken of are in the form of language models. We divide our attention between two deep learning models: Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Encoder Representations from Transformers (BERT). With the help of three different datasets, we aim to compare the performance of the two language models in as much detail as possible.

We aim to include cross-domain learning in our research in order to investigate generalizability to new and unseen data.

Therefore, the research question we will be focusing on throughout this investigation is:

**“What are the comparative strengths and weaknesses of BERT and BiLSTM in handling cross-domain data for Named Entity Recognition (NER) tasks, and how do these models differ in their ability to generalize across different linguistic contexts and domains?"**

## Related work

Named entity recognition is a good tool for enhancing information management, decision-making, and communication processes across various sectors of society. It can be adjusted for any task, in any language..

The paper by Hvingelby et al., 2021 offers an overview of what NER can really be used for and focuses on a meticulously annotated dataset tailored for the Danish language, facilitating NER model training and evaluation in a specific linguistic domain. It sheds light on the effectiveness of BERT and BiLSTM in cross-domain NER tasks, and proves to be a kickstart for conducting comparative studies across different linguistic contexts.

The study by Jia, Liang, & Zhang, 2019 addresses the challenge of NER across diverse domains by leveraging cross-domain language modelling techniques. It aims to enhance NER accuracy and generalisation capabilities in heterogeneous data environments. The findings highlight the effectiveness of this approach in improving NER performance across various text sources and domains, offering valuable insights for advancing cross-domain NER techniques.

The paper by Lee, Lu, & Lin, 2022 adds to the growing body of research on NER techniques, specifically in the context of the Chinese language. It displays advanced neural network architectures, including BERT and BiLSTM to improve the accuracy of NER for Chinese text. By combining deep learning models, the goal of the research is to highlight the effectiveness of advanced models in capturing named entities in any language.

Our project focuses on not only the comparison between two deep learning language models (BERT and BiLSTM) and their performance levels in the context of NER, but also the cross-domain status of the investigation, lead by the multitude of data used in order to achieve a more thorough understanding of how well our models would perform with new and unseen data.

## Data

EWT (English Web Treebank): we used a train set of this dataset to train our models and a test set to get predictions that we use as a baseline to compare performance of the other datasets to. The dataset contains sentences from five different web domains: Yahoo! answers, newsgroups, weblogs, local business reviews from Google and Enron mails. (Plank, 2021)

CrossNER: we use the conll2003 train dataset from CrossNER as one of our comparison datasets. The text is a collection of news wire articles from the Reuters Corpus. (Hugging Face CONLL, 2003)

Tweebank: the other comparison dataset we use. It contains anonymised English tweets. (Jiang, Hua, Beeferman, & Roy, 2022)

We chose these datasets because they all have BIO annotation with the same named entity types (‘location’, ‘organisation’, ‘person’) and all sentences are in English, which lets us compare performance regarding a change in text genre. The CrossNER dataset also contains a ‘misc’ entity type, which we changed to ‘O’, to have all datasets contain the same labels.

The datasets are in various written styles which is useful as a domain shift as text appears in a variety of different ways depending on the medium it was made for. We train our models using the EWT as a baseline as it contains a genre of text that is in between the two comparison datasets. The CrossNER dataset is written in a formal text genre that follows standardised writing often seen in publishing, and the Tweebank dataset contains more informal writing that is often seen on social media which includes spelling errors, slang and emojis. We consider the EWT dataset to be a middle ground between the two other datasets.

Our assumption is that of the two comparison datasets, the models will perform better at NER on the CrossNER dataset, as it is a little closer in text genre to the baseline dataset than the Tweebank dataset.

## Method

We got a CRF-BiLSTM model from the group 10abc. Our BERT model was made following the tutorial at Hugging Face.

We measured the performance of the BiLSTM and BERT baselines by calculating their span-f1 scores. Our passing criteria for the baseline models was to have a span-f1 score of at least 30% on the EWT dev data. We have used EWT dev data as a test data because we did not have the gold labels of the test data. We then preprocessed the two comparison datasets to be of the same shape as the EWT dataset to get predictions from our models.

We will use quantitative and qualitative methods to analyse and compare the models performance.

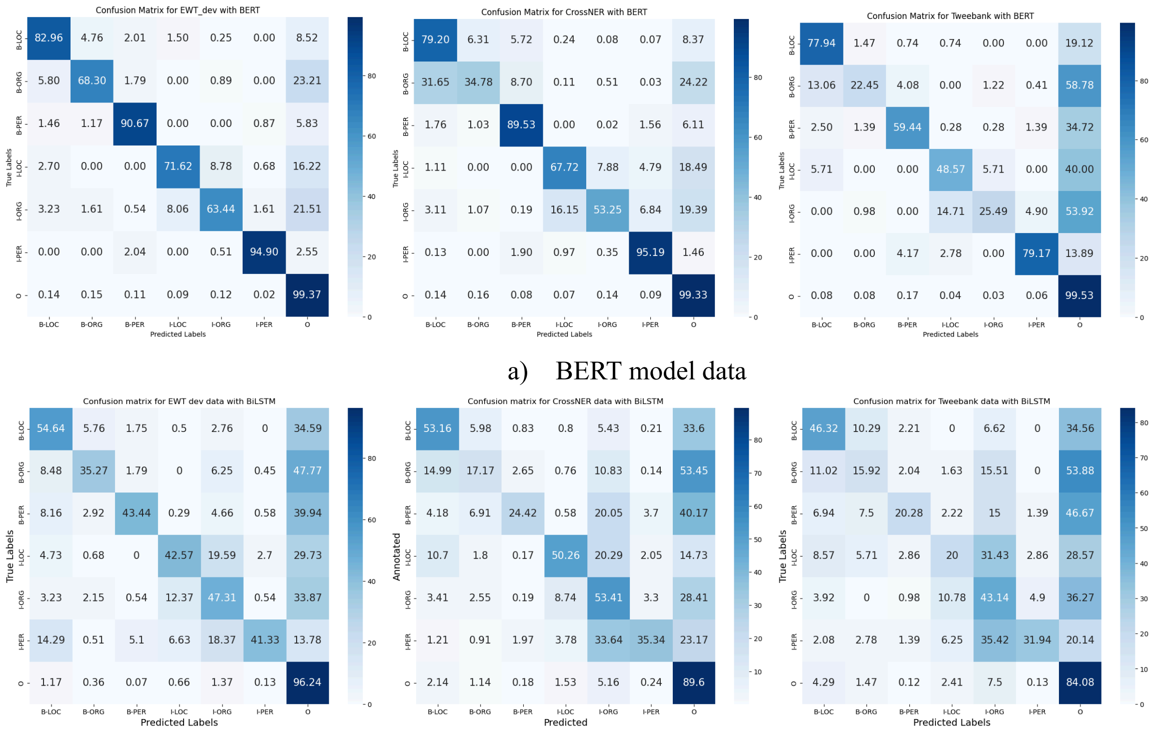
Quantitative: span-F1 score to see how many of the named entities were correctly found. Confusion matrices to look at distributions of predictions of labels for the models and see if there are types that are more often misclassified in general or with only some classes.

Qualitative: we will look at some of the sentences with errors for each model and dataset and discuss potential possibilities as to why the models made an error.

## Results and Analysis

| | BERT | | | BiLSTM | | |
|---|---|---|---|---|---|---|
| | Sp. F1 | Prec. | Rec. | Sp. F1 | Prec. | Rec. |
| EWT (dev) | 77% | 75% | 79% | 40% | 40% | 40% |
| CrossNER | 68% | 69% | 66% | 31% | 36% | 28% |
| Tweebank | 55% | 64% | 48% | 10% | 7% | 17% |

Table 1: Detailed results of the BERT and BILSTM models on the test sets



a) BERT model data

b) BiLSTM model data

Figure 1: Normalized confusion matrices, left to right: EWT, CrossNER, Tweebank

Both models perform the best on the EWT dev dataset, which was expected as this is the same type of text as what the models were trained on. The BERT model got the highest span-F1 score for all three datasets, which was also expected. Both models had a small difference between their precision and recall for the EWT and CrossNER datasets at <10% while the Tweebank datasets had a greater difference, which supports our assumption of the Tweebank dataset being the most difficult one to predict on as the performance is more irregular. The span-F1 scores for the Tweebank dataset also has a far greater difference between the EWT dev dataset than the CrossNER. Both models have a 9% difference between the span-F1 score of the EWT dataset and the CrossNER dataset, which is a more significant loss for the BiLSTM, as it predicted fewer named entities correctly to begin with on the baseline domain.

To further inspect the predictions of the models' we plotted confusion matrices for each set of predictions, Figure 1. The matrices have been normalised to show the distribution of the true labels in percentages, as there are far more non-named entities, 'O', in each dataset than the other types. The correct predictions are shown on the diagonal of a confusion matrix and the percentage of a class predicted as 'O' is on the rightmost column. The BERT model's predictions have a much darker diagonal than the BiLSTM and the rightmost column is darker than the rest of the confusion matrix, but not by much. In comparison, the BiLSTM has the diagonal and rightmost column visible, with the colours being closer to one another, with the rightmost column being slightly darker. This means that both models will generally classify to 'O' when they miss a named entity, which is expected as it is the most common type in all datasets. Both models also have an issue with predicting the beginning of an organisation as the beginning of a location on the two comparison datasets. This could be because locations and organisations can be very similar or even have a different label depending on the context. There appears to be a highlighted column on the BiLSTM matrices for 'I-ORG' predictions, which is fainter on the baseline dataset and more clear on the two comparison datasets. Since the same error doesn't appear for the 'B-ORG' it could be due to the model making very long spans when predicting organisations, so they overlap with a lot of other named entities. This would mean that the BiLSTM has less of an idea of where a named entity starts and begins besides not just predicting to the right kind of named entity. The BERT model has a tendency to confuse 'B-LOC' with 'B-ORG' and 'I-LOC' with 'I-ORG', which again probably comes from the two types being similar in a lot of ways, especially compared to a person which is often more distinct. The confusion between 'I-LOC' and'I-ORG' looks more prominent, but that is because there are fewer of them than 'B-LOC' and 'B-ORG' since they depend on them to show up, so getting an 'I' wrong will have a proportionally greater impact than getting a 'B' wrong. This could mean that the BERT model is good at finding a named entity in its entire span, but it will sometimes get the type of named entity wrong. It also makes sense that 'person' is the named entity that gets mistaken the least for other named entities, as names often follow a similar structure of first name, middle name(s), last name, often with the first letter of each name capitalised and words used for names, at least in English, are less often used as anything but a name for a person than many words that are used for names and locations, which BERT would know from its pretrained tokenizer.

Qualitative analysis: The errors the BERT model makes are more easy to reason with in terms of misunderstanding than the errors the BiLSTM model makes, which are without any immediate logic or patterns. The BERT model will often correctly identify the presence of a named entity but when it makes mistakes it is easier to come up with linguistic reasons for why these mistakes happen. It could be single words in the middle of a sentence that are capitalised are likely a named entity but if it is uncommon it will guess what type it is.

Looking at the BiLSTM's predictions of a few tweets we see that few named entities are predicted as just one word, with most being at least two words and sometimes up to 15 words long. In contrast the gold labels show that most named entities span only one word. It appears to often classify words at the end of a sentence that begin with 'URL' as a location when they should not be classified as a named entity. There doesn't seem to be a pattern of tokens beginning with '#' or '@' to be classified as just one specific class or if they indicate if it is the beginning of a named entity or inside. Hashtags and usernames should according to the annotation guideline not be classified as a named entity. BiLSTM also failed to correctly classify name entities in tweets that are written as normal text. The writing style of the tweets varied with some tweets capitalising the first letter of each word, which we thought might make the model classify more often to named entities, but it did not appear like there were any significant influence on the model's arbitrary looking predictions. If several but not all words in a tweet have the first letter capitalised it could appear that it would lead the model to classify those spans as a named entity. BiLSTM would long spans of named entities for a lot of tweets that were written in a headline-style where the first letter of each word is capitalised, which could explain why that model had such a low performance on that dataset.

From inspecting the predictions the BiLSTM made on the EWT data, it appears that it also has an issue of labelling the end of a sentence as a location. 14 of the first 34 sentences are predicted to end with a location (although 2 were one word sentences) the others were predicted to end with 'O' (the correct label for all except two that indeed ended with a location), not any of the other labels. This could come from the training data of the model, which might contain a large number of sentences ending with a location. Many of these predicted locations also include the punctuation at the end when there is one, which there aren't for all sentences. The BERT model does not have this issue.

We find a sentence where four fast food locations are named, each separated with a comma, and the BiLSTM model only correctly separates after the first comma, and then gets the remaining locations wrong as organisations, which you could also argue for them being. The BERT model managed to separate these named entities correctly but classified all of them as organisations.

The BiLSTM's performance on the CrossNER dataset appears to do well at finding countries, which come up often in this dataset, but it struggles to find persons. It sometimes finds an organisation and labels it as a location. It appears to be better on this dataset than the others at finding a named entity span (of one or more words) but will often label it as the wrong entity type. It correctly labelled the person 'John Lloyd Jones' (the preceding word was 'chairman'), but Spanish Farm Minister 'Loyola de Palacio' was labelled as an organisation, same for spokesman 'Nikolaus van der Pas'. All three persons have their full name, are correctly identified as one entity, and have the preceding word in the sentence be their job title, which could be a hint that the named entity is a person, but only one was correctly labelled as a person. This could be because the name 'John Lloyd Jones' is more English than the other two names, and that the model might have seen similar English sounding names before in the train data and the model either doesn’t know the meaning of chairman, minister and spokesmen or it cannot take it into account. The model also correctly labels the person ‘Peter Blackburn’ as a person, so it would appear that the model can classify to English sounding names that it is familiar with from the train data but cannot find non-English sounding names as well, even with context. The BERT model manages to find these mentioned persons as well and labels them correctly, possibly not restricted to only English names by having context on how different names can look. The BiLSTM has difficulties finding persons that only have one name, such as being referred to by their last name, but the BERT model can still find these persons, again, probably because it has a better understanding of context, with people in a professional setting often being referred to by only last name.

The BERT model does not have the same issue as the BiLSTM of labelling many of the unusual words, such as those beginning with ‘#’, ‘@’ or ‘URL’, as a named entity. It appears to be much better at finding the entire span of a named entity, and not making unknown words into one big named entity and it also does not have the issue of ending a lot of sentences with a location. This is likely because BERT has a better understanding of what the words mean, how they relate to each other, and how sentence structure comes from that.

## Discussion

We were not able to train a BiLSTM model to a suitable performance without a CRF layer. Having a CRF layer gives a build in advantage to the BiLSTM over our BERT model regarding correct NER notation. We thought the BERT might learn by itself that these rules exist, but it clearly does not as it in some few cases breaks these rules. Despite not knowing these rules the BERT model still clearly outperformed the BiLSTM, even in domain shifts.

Maybe the BiLSTM’s poor performance on the Tweebank dataset could be because when it was trained, got a vocabulary from the train data and dropped the most uncommon words so it could generalize better to words not in its vocabulary, but if many of the named entities in the train data appeared few times, they were often the words dropped and that could lead the model to associate unknown words with named entities, which is why it makes so many named entity predictions on the Tweebank dataset, as it contains the most “foreign” vocabulary.

Limitations: size of the comparison datasets, amount of datasets to compare. The confusion matrices can only tell us on an individual word level if we predicted correctly, it doesn’t indicate if the entire named entity is correctly identified or only part of it.

## Conclusion

In conclusion, this study explored Name Entity Recognition (NER) using two well-known models, BERT and BiLSTM, with the goal of identifying their advantages and disadvatages. We aimed to understanding how well they handle cross domain data through evaluation of their performance on three different datasets: EWT, CrossNER and Tweebank.

Our investigation highlighted several key finidings. Firstly, both models performed much better on the EWT- dev dataset, when tested, which was expected because both models were trained using the EWT test data, so the data was familiar. Across all datasets, BERT significantly outperformed BiLST, achieving higher span-F1 score.This demonstrates a more robust ability to generalize different contexts. The performance gap between the two models is more pronounsed on the Tweebank dataset, making it more difficult to predict.

Confusion matrices showed that both models had a tendency to misclassify some of the lables with the most common lable "O". Both models struggled to distinguish between beggining of organization and locations on the tweebank datasets and the crossNER due to thier similarity in different contexts. Often "B-LOC" labels were misclassified as "B-ORG" and "I-LOC" with an "I-ORG" due to their similarity.

Thanks to the qualitative analysis, we found that the mistakes that BERT does when classifying are easier to explain, and usually they are due to linguistic explanations. The model frequently accurately detect the presence of a named entity. On the other hand, those mistkase made by BiLSTM lack any abvious patterns or logic.

Overall, BiLSTM struggled the most with the Tweebank dataset, often getting entity classifications wrong and having trouble with how sentences were structured. Even though it had an advantage with its CRF layer for following NER rules, BERT performed better overall, especially when dealing with different types of text.

## Bibliography

Hvingelby, R., Pauli, A. B., Barrett, M., Rosted, C., Lidegaard, L. M., & Søgaard, A. (2021). DaNE: A Named Entity Resource for Danish. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://aclanthology.org/2020.lrec-1.565.pdf

Jia, C., Liang, X., & Zhang, Y. (2019). Cross-Domain NER using Cross-Domain Language Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2464–2474). Florence, Italy: Association for Computational Linguistics. https://aclanthology.org/P19-1236.pdf

Jiang, H., Hua, Y., Beeferman, D., & Roy, D. (2022). Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)* (pp. 7199–7208). European Language Resources Association (ELRA). https://aclanthology.org/2022.lrec-1.780.pdf

Lee, L.-H., Lu, C.-H., & Lin, T.-M. (2022). NCUEE-NLP at SemEval-2022 Task 11: Chinese Named Entity Recognition Using the BERT-BiLSTM-CRF Model. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 1597–1602). Association for Computational Linguistics. https://aclanthology.org/2022.semeval-1.220.pdf

Plank, B. (2021). Cross-Lingual Cross-Domain Nested Named Entity Evaluation on English Web Texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 1808–1815). Association for Computational Linguistics.
https://aclanthology.org/2021.findings-acl.158.pdf

Hugging Face. (n.d.). CoNLL 2003 dataset. Retrieved from https://huggingface.co/datasets/conll2003

Hugging Face. (n.d.). Token Classification. Retrieved from https://huggingface.co/docs/transformers/tasks/token_classification#inference

## Appendix

All group members contributed equally to this project.

| | EWT (dev) | CrossNER | Tweebank |
|---|---|---|---|
| Number of sentences | 2,002 | 14,041 | 1,639 |
| Number of tokens | 35,444 | 300,677 | 44,946 |
| O | 23,678 | 174,171 | 23,731 |
| B-LOC | 399 | 7,129 | 136 |
| I-LOC | 148 | 1,168 | 35 |
| B-PER | 343 | 6,600 | 360 |
| I-PER | 196 | 4,528 | 144 |
| B-ORG | 224 | 6,297 | 245 |
| I-ORG | 186 | 3,728 | 102 |

Tabel 2: Detailed token overview for each dataset used.