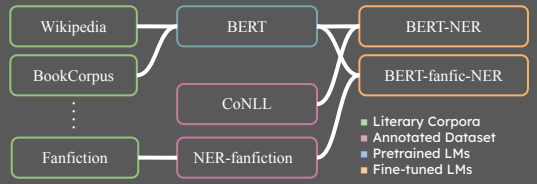
Fine Tuning a Pretrained BERT-based Model for Named Entity Recognition to the Domain of Fanfiction

N.P.A.Kindt@student.tudelft.nl
Author: Nathan Kindt
Supervisors: Chenxu Hao
Ivan kondyurin
Responsible Prof.: Hayley Hung

1. Introduction

Pretrained Language Models (PLMs) have revolutionised the field of Natural Language Processing (NLP) and paved the way for many new, exciting large-scale studies for various areas of research. One such field presents itself in the emerging digital literary corpus that is fanfiction, providing research opportunities within the fields of (NLP), Computational (Socio-) Linguistics, the Social Sciences and Digital Humanities. However, because of the unique linguistic characteristics of this domain many modern NLP solutions utilizing PLMs encounter difficulties when applied on fanfiction texts. This study aims to 1) measure the performance of mainstream PLMs on fanfiction texts and 2) increase the performance of one such model by fine-tuning on in-domain data. The NLP task of Named Entity Recognition (NER) is chosen as the focus of this study because of the foundational nature of this task for various other downstream NLP methods.

Overview of relations BERT-fanfic-NER and BERT-NER



*NER Annotated Fanfiction Dataset Split Sizes**

Split	Training	Evaluation	Test	Overfit Test
Size	120	14	15	11

* 1 dataset entry contains 256 tokens padded up to the context window size of 512.

2. Background

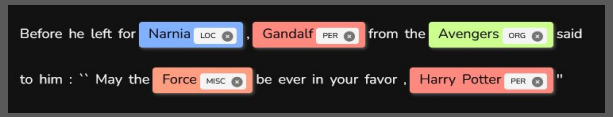
A. Language Models

Pretrained Language Models (PLMs) are LMs that have been pre-trained on large amounts of text, learning to model the context-dependent semantic meaning of words through the Attention mechanism. Numerous different PLMs have been developed with varying architectures, all with their own applications, challenges and limitations which are an active area of research. For NLP, the BERT model especially led to breakthroughs for several NLP tasks through its bidirectional attention mechanism.

A. Named Entity Extraction

The NER task involves extracting named entities out of unstructured text by assigning class labels: Person (PER), Organization (ORG), Location (LOC) and Miscellaneous (MISC). Tokens not labeled as any entity are given the 'Outside entity class' (O) label. Many downstream IE tasks depend on NER as a first step such as Entity Linking, Relation Extraction and Coreference Resolution.

Example of the NER problem within fanfiction



A. Adapting PLMs

A major advantage of PLMs is that they can be further be trained for specific downstream tasks or into specialized domains through fine-tuning or additional pre-training. BioBERT is one such example of an model adapted for the biomedical domain through continued pre-training, after which it could be fine-tuned to perform several downstream tasks with increased performance on in-domain data. As of yet there exist no PLM adapted to the domain of fanfiction texts. Current studies on fanfiction rely on PLMs trained on general data such as BERT-NER, a BERT-based model fine-tuned on the CoNLL dataset for NER. While achieving impressive results for numerous NLp tasks, these models might perform worse then stated when applied to fanfiction texts since they lack knowledge of this domain.

3. Experiments

A. Dataset Collection. In order to evaluate the performance of PLMs and train one for NER, a annotated dataset was needed. 10 fanfictions were collected and manually annotated for NER. These text where then chunked and divided in a training, evaluation and test split. Since the test set was randomly selected out of all chunks of each story, entities present in the test set had a high chance of also being present in the training set. Thus one story was separated in order to more accurately be able to test the models ability to generalize to unseen data.

B. Training. The training of BERT-fanfic-NER took a total of 32 minutes and all 16 epochs (1 epoch = 7.5 training steps) on a single Nvidia Quadro P1000 GPU.

C. Evaluation. The model was evaluated using the F1 score, which is calculated as the harmonic mean of precision and recall thus ignoring True Negatives. This is a more meaningful metric than regular accuracy since for NER missed entities are as important as incorrectly labeled ones, and Precision is skewed towards the accuracy of the far more numerous 'O' label making this metric arbitrary.

Evaluation Scores of BERT-fanfic-NER (B-FF) and BERT-NER (B-NER)

Dataset	CoNLL		TEST		OVERFIT		ALL FF	
Model	B-NER	B-FF	B-NER	B-FF	B-NER	B-FF	B-NER	B-FF
PER F1	0.96	0.46	0.92	0.96	0.90	0.95	0.91	-
LOC F1	0.93	0.03	0.40	0.56	0.73	0.60	0.68	-
ORG F1	0.90	0.07	0.17	0.44	0.0	0.0	0.37	-
MISC F1	0.80	0.0	0.51	0.20	0.8	0.0	0.35	-
overall Precision	0.90	0.39	0.84	0.89	0.87	0.92	0.84	-
overall Accuracy	0.92	0.17	0.82	0.90	0.86	0.89	0.84	-
overall F1 Score	0.91	**0.24**	0.83	**0.89**	0.86	**0.90**	0.84	

4. Results & Discussion

A. BERT-NER on Fanfiction. The baseline provided by BERT-NER suggests that fanfiction texts indeed differ from other types of texts, since the performance of BERT-NER dropped by 7% (from 91% F1 score to 84%) when evaluated on the fanfiction dataset. This suggests our hypothesis that fanfiction writing exhibits significant distinct linguistic characteristics in regards to regular fictional writing, and therefore NLP techniques utilizing PLMs developed for regular fictional literature under-perform when applied to the domain of fanfiction is correct.

B. BERT-fanfic-NER. An increase in performance of 6% F1 score was gained in comparison to BERT-NER, indicating that in-domain fine-tuning can lead to increased performance of PLMs on NLP tasks. However, we are sceptical of these results.

C. Data Diversity and Contamination. The poor performance of BERT-fanfic-NER on CoNLL implies that the model hasn't generalised well to other data outside of fanfiction, yet the high performance on the overfit test set indicates otherwise. However the overfit test is relatively small and has little diversity, potentially leading to erroneous results. Similarly, due to the limited amount of fanfiction stories the model could have overfitted to the training data. In combination with data contamination, caused by the test set containing only pieces of stories from the training set, this could have led to higher results than what the model would achieve on a truly unseen and independent test set.

5. Recommendations

The evaluation of BERT-NER on fanfiction text implies the need for further investigation of the performance of NLP tasks by PLMs on fanfiction texts, since they are used in numerous interdisciplinary studies on fanfiction for NLP tasks and the results of those studies might be affected by a worse-than-reported accuracy of these models. We also indicate that while this study was limited by the lack of sufficient annotated data for fine-tuning, the performance of NLP by PLMs could also be improved by further pre-training an existing PLM on fanfiction, which can be fine-tuned with existing task-specific non-fanfiction data for downstream NLP tasks.