A Comparison of BERT and CRF Models on Named Entity Recognition Tasks

Maria Karamihaylova and William Oliver

May 26 2022

Abstract

In this paper, we compare the performance of BERT, a transformer-based model, and a conditional random field model (CRF), a non-transformer based model, on named entity recognition (NER) tasks. Specifically, we apply the two models to two different data sets, the Groningen Meaning Bank (GMB) Corpus and CoNNL-2003, to determine performance. Results indicate that while the non-transformer model, CRF, works reasonably well at predicting named entities, the transformer model, BERT, outperforms across all evaluation metrics, including recall, precision, and F1.

Introduction

Named entity recognition is a subfield of natural language processing which consists of identifying and extracting important components in a given text and classifying them according to categories (i.e., person, location, currency, time, etc.). NER has a wide-range of real-world applications, including automation of customer support services, recommendation engines, and even resume processing. As such, there is a demand to increase computational effectiveness when performing such tasks. Therefore, for this project, we will compare how different models perform NER tasks. Specifically, we will investigate how the transformer model BERT performs on NER tasks compared to the performance of a conditional random field model, which is a non-transformer model.

In this paper, we first overview the relevant literature on transformers, BERT, and BERT's performance on NER tasks. Next, we discuss the experiments we conducted by applying the CRF and BERT models to the GMB Corpus and the CoNNL-2003 datasets (Eric et al., 2003; Bos et al., 2017). Then, we review the results of our experiments. Finally, we conclude with a discussion of our experiment and results.

Related Work

In our experiment, we compare BERT with a non-transformer CRF model on an NER task. The CRF model was first presented as a foundation for constructing probabilistic models for the purposes of analyzing and predicting sequential data in Lafferty et al. (2001). Linear-chain CRF models have long been utilized for a range of natural language processing tasks such as speech tagging and text chunking as well as NER (Sutton & McCallum, 2006). With the advent of more sophisticated machine learning technologies, it is worth exploring how transformer models can improve upon the predictive performance of simpler CRF models. Vaswani et al. (2017) introduced the transformer neural network architecture, which BERT is based on, and demonstrated its effectiveness with success on English-to-German and English-to-French machine translation tasks. We seek to validate existing research that suggests transformer models outperform non-transformer models on NER tasks.

Transformers are deep learning models which have advanced the machine learning field due to their use of a self-attention mechanism, which extracts a sequence of word embeddings in parallel by determining the relevance and relationship between every word with every other word in a given sequence. Unlike previous models that would determine the word embeddings one word at a time in a linear fashion, the transformer model performs this process by viewing all the words of a sequence and calculating their embeddings simultaneously. This model contains two component parts: the encoder, which creates the word embeddings, and the decoder that uses the word embeddings to make predictions.

The following year, Devlin et al. (2018) introduced BERT, which stands for the Bidirectional Encoder Representations from Transformers. What makes BERT different from preexisting models is the added element of bidirectionality. Earlier deep learning pre-transformer models read text sequentially from left-to-right or right-to-left and transformer models read the entire text at once and so were, in effect, non-directional. In contrast, the BERT model is trained bidirectionally, as the model processes text in both directions simultaneously. The model is then trained with masked language modeling and next sentence prediction. Masked language modeling training involves hiding around 15% of the words in the text and having the program predict what the hidden words are from context, and next sentence prediction involves having the program predict whether two randomly selected sentences are sequentially adjacent. When Devlin et al. (2018) tested BERT on NLP tasks such as question/answering and cloze activities, the model performed significantly better than earlier models.

Baumann (2019) investigates fine-tuning methods applied to multilingual BERT architecture for language modeling as it pertains to NER in English and German. The author applied a universal language model fine-tuning approach for text classification first proposed by Howard & Ruder (2018) to BERT, but this revealed mixed results. The CoNNL-2003 NER data set (Sang & De Meulder, 2003) was utilized as it contains datasets in both target languages. When

the discriminative fine-tuning approach was implemented, results did not improve, indicating that the pre-trained BERT model was already effective.

Lothritz et al. (2020) compared the performance of pre-trained transformer-based models with that of non-transformer models on NER tasks. The task focused specifically on fine-grained NER, which contains a much larger number of entity labels than traditional NER models. The researchers utilized the English Wikipedia Named Entity Recognition and Text Categorization dataset, which contains 49 tags that were automatically annotated. These tags include fashion, finance, food, government, music, and people. Results showed that the transformer-based models (BERT, RoBERTa, XLNet) outperformed the non-transformer-based models (CRF, BiLSTM-CNN-CRF) in most of the entity domains with regard to F1 and recall scores. However, the CRF model outperformed the other models with regard to precision scores. Also, the BERT model generally demonstrated stronger performance compared to the other two transformer-based models. In sum, the authors concluded that selecting a given model for an NER task comes with trade-offs regarding performance because one model did not outperform the others on all tasks.

Experiments

For our experiments, we ran BERT and CRF models on the GMB and CoNNL-2003 datasets. The GMB corpus was developed at the University of Groningen in order to be used for NLP tasks and, specifically, for NER tasks. This dataset is an annotated corpus that contains over one million named-entity tags (as well as POS and lexical category tags among others that were not relevant for the task in this paper). Named-entity tags include geographical entity, organization, person, geopolitical entity, time indicator, artifact, event, and natural phenomenon, and words were labeled with a computer annotator and then checked by humans. There are 8 different tags for entities, and words that do not fall into one of the entity categories are left unmarked. Due to the dataset's named entity tags, it is a useful tool for training and evaluating NER tasks.

The CoNNL-2003 dataset consists of annotated text from the Reuters corpus, which is a large collection of Reuters News stories (Sang & De Moulder, 2003). This dataset is already annotated and split into training, development, and test files. Moreover, this dataset was designed for NER research; accuracy scores for NER tasks with the CoNNL-2003 dataset have become benchmarks in the field, so it is appropriate for our purposes. Named entity tags in this dataset are limited to four entities: Organization, Person, Location, and Miscellaneous.

A CRF model is a type of statistical model that takes neighboring items into account to make predictions about patterns. A CRF model takes the input vectors, the position of the predicted data point, the label of the data point, and the label of the previous data point as input values. The Maximum Likelihood Estimation is then applied to predict the parameters. CRFs are models commonly utilized in NLP tasks for modeling data that are sequential. As such, they are often used in named entity recognition tasks to predict the

most likely labels corresponding to input sequences.

We utilized the crfsuite model via Scikit-learn to implement the conditional random field. We preprocessed the data and extracted the features. We selected the limited-memory BFGS optimization algorithm with Elastic Net regularization (L1 and L2), which is the default parameter for the CRF model. As with the BERT model above, we leveraged the train_test_split function from Scikit-learn to train 80% of the day and to evaluate the other 20%.

BERT is a deep-learning transformer model for natural language processing that, unlike previous language models that read input text from left to right or right to left, reads input text all at once. It was pretrained on language modeling (where 15% of the tokens were masked and BERT had to predict them from context) and next sentence prediction (where BERT predicts whether two sentences would be sequential in a text). This pretraining allows BERT to learn contextual embeddings for words and, thus, complete natural language processing tasks such as named entity recognition.

We ran a pre-trained BERT model on the GMB and CoNNL-2003 datasets. We used the version of BERT available in the Simple Transformers library. The creators of the Simple Transformers library took transformer models available in the Hugging Face transformer library and made them more user-friendly. Using this library, we first preprocessed the datasets with Scikit-learn's train_test_split function with 80% of the data for training and 20% for evaluation. The model we ran consisted of two epochs and training and evaluation batch sizes of 16.

Results

Our results demonstrated that BERT outperformed CRF on recall, F1, and precision scores. When applied to the GMB dataset, the CRF model achieved an F1 score of 0.674, a recall score of 0.643, and a precision score of 0.733. Applying the CRF model to the CoNNL–2003 dataset yielded an F1 score of 0.755, a recall score of 0.736, and a precision score of 0.784. When applied on the GMB dataset, the BERT transformer model achieved an F1 score of 0.797, a recall score of 0.768, and a precision score of 0.827. Applying the BERT model to the CONNL-2003 dataset yielded an F1 score of 0.848, a recall score of 0.836, and a precision score of 0.862. The BERT model performed better than the CRF model with regard to all the evaluation metrics.

Table 1

CRF and BERT Results on the CoNNL-2003

	Precision	Recall	F1
CRF	0.784	0.736	0.755
BERT	0.862	0.836	0.849

Table 2

CRF and BERT Results on the GMB

	Precision	Recall	F1
CRF	0.733	0.643	0.674
BERT	0.827	0.768	0.797

Discussion

Our results found that the BERT transformer model outperformed the CRF model across all evaluation metrics on both of the the CoNNL-2003 and GMB datasets. These results differed from Lothritz et. al's (2020) findings that had the transformer models yielding better recall and F1 scores but the CRF model yielding better precision scores. We question whether these differences have to do with the number of tagged domains. The Lothritz et al. (2020) study investigated a fine-grained annotated corpus consisting of 49 different tagged domains, whereas the GMB corpus contained 8 different tags and the CoNNL-2003 dataset contained only 4. Additionally, the BERT model applied to the CoNNL-2003 dataset yielded a higher F1 score than when applied to the GMB corpus, further indicating that the number of tagged domains impacts evaluation metric scores. Future research can investigate how tagging the dataset impacts NER task results. That is, what are the tradeoffs with less tags and higher accuracy compared to more tags and lower accuracy? If it is the case that fewer tags produce higher results, is it possible that fewer but more relevant tags would produce more meaningful results. For example, if the NER task has to do with automobiles, maybe it makes more sense to only tag car-related entities and ignore the others. Or, maybe more tags are still optimal because they lead to a more comprehensive understanding of the input. Future research can also investigate the reliability of datasets that are automatically tagged and how computer automation of tagging datasets can impact the performance of transformer models of NER tasks.

References

Baumann, A. (2019). Multilingual language models for named entity recognition in german and english. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 21–27.

Bos, J., Basile, V., Evang, K., Venhuizen, N. J., and Bjerva, J. (2017). The groningen meaning bank. In *Handbook of linguistic annotation*, pages 463–496. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lothritz, C., Allix, K., Veiber, L., Bissyandé, T. F., and Klein, J. (2020). Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3750–3760.

Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Sutton, C., McCallum, A., et al. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.