CNNBiF: CNN-based Bigram Features for Named Entity Recognition

Chul Sung, Vaibhava Goel, Etienne Marcheret, Steven J. Rennie, and David Nahamoo
Pryon.com, Brooklyn, New York.
{csung, vgoel, emarcheret, srennie, dnahamoo}@pryon.com

Abstract

Transformer models fine-tuned with a sequence labeling objective have become the dominant choice for named entity recognition tasks. However, a self-attention mechanism with unconstrained length can fail to fully capture local dependencies, particularly when training data is limited. In this paper, we propose a novel joint training objective which better captures the semantics of words corresponding to the same entity. By augmenting the training objective with a group-consistency loss component we enhance our ability to capture local dependencies while still enjoying the advantages of the unconstrained self-attention mechanism. On the CoNLL2003 dataset, our method achieves a test F1 of 93.98 with a single transformer model. More importantly our fine-tuned CoNLL2003 model displays significant gains in generalization to out of domain datasets: on the OntoNotes subset we achieve an F1 of 72.67 which is 0.49 points absolute better than the baseline, and on the WNUT16 set an F1 of 68.22 which is a gain of 0.48 points. Furthermore, on the WNUT17 dataset we achieve an F1 of 55.85, yielding a 2.92 point absolute improvement.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in knowledge extraction that detects named entities in text and assigns them to pre-defined categories such as persons, organizations, and locations. It plays a critical role in various applications including question answering, information retrieval, co-reference resolution, and topic modeling (Yadav and Bethard, 2019). Pre-trained transformers fine-tuned with a sequence labeling objective have become the de facto standard for the NER task because these models have shown state-of-the-art performance without the human effort of feature engineering.

Despite these achievements, fine-tuning of pre-trained transformer models has two potential weaknesses: first, unconstrained self-attention implements a global receptive field for all interactions, with no inductive bias toward focusing on and composing local dependencies hierarchically (Dehghani et al., 2019; Wang et al., 2019), and second, with small amounts of labeled data, training such models end-to-end is susceptible to overfitting.

To address these limitations we propose a novel joint sequence labeling objective, inspired by BERT's next sentence prediction (NSP) objective (Devlin et al., 2019). In contrast with the NSP objective, which evaluates sentence pairs, we design a word level objective specifically for the NER task. On top of the conventional sequence labeling objective, our novel objective enables modeling of the relationship of adjacent words based on a new tagging scheme, which helps the model to better capture local dependencies in a sequence.

For the additional objective, we employ a simple convolutional architecture based on CNN bigram features (in short, CNNBiF) to better capture the relationships between adjacent words. Under the single loss objective of the conventional sequence labeling approach we have observed that the predictions output by pre-trained transformers quickly converge to the training target labels. Our joint learning approach regularizes these models to encourage them to better capture the semantic and syntactic dependencies between nearby words.

Our key contributions in this paper are:

- We propose a novel joint training objective to better capture the semantic and syntactic patterns of text through a single model architecture. The novel objective employs a new tagging scheme and a convolutional neural network architecture.
- We present results illustrating the efficacy of our model, showing (1) a performance increase over strong baseline models on two standard benchmark datasets and (2) further

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1016–1021
November 7–11, 2021.

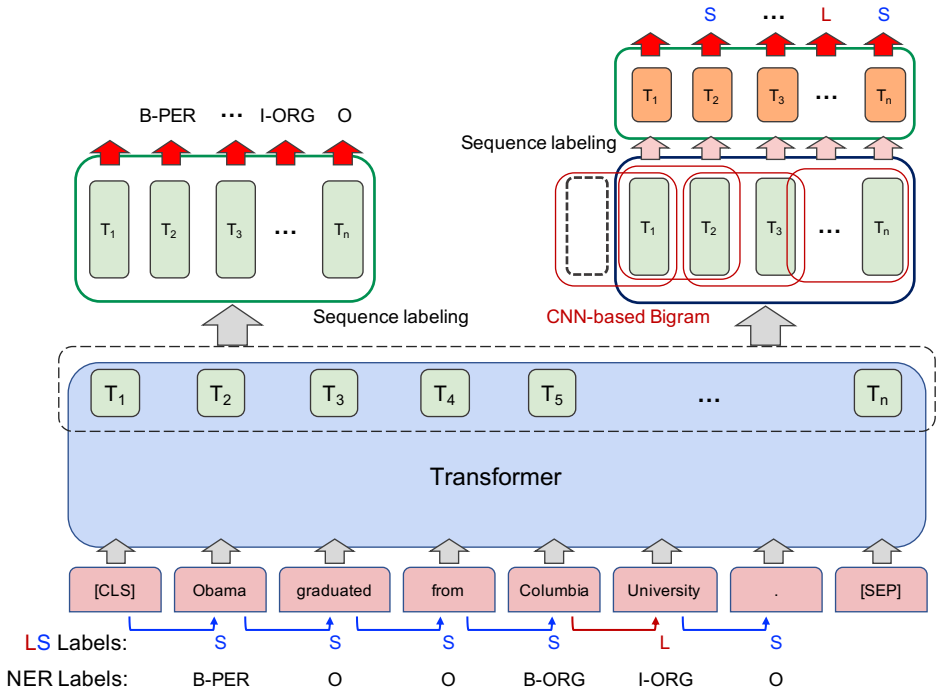

Figure 1: Proposed model architecture including CNN-based bigram features (CNNBiF), with joint training objectives under `IOB2`, and our proposed group consistency loss based on linkage-separation (LS) labeling. The LS objective and CNN-based bigram features regularize model training, and lead to sequence representations that better capture the relationships between adjacent words.

performance gains on out-of-domain datasets, which shows that our approach is effective at reducing overfitting.

2 Related Work

Recent applications of multi-task objectives with one of the objectives being named entity recognition has demonstrated improved performance on the NER task. Zheng et al., 2017 applied a multi-task objective learning to named entity recognition and relation extraction to show improvements over individual tasks. Martins et al., 2019 performed joint learning of NER and entity linking tasks in order to leverage the information in two related tasks, using an LSTM model architecture. Similarly, Eberts and Ulges, 2019 presented a joint learning model based on a single transformer network to leverage interrelated signals between the NER and entity relationship tasks.

Prior to the advent of transformer-based networks, CNN networks were applied successfully to various NLP classification tasks. Kim, 2014 reports on the effectiveness of these networks where a one-layer CNN is applied to pre-trained word vectors (Mikolov et al., 2013).

3 Proposed Approach

As illustrated in Figure 1 our model leverages a pre-trained transformer network. This network is fine tuned with two sequence labeling objectives applied to the single NER task. The first sequence labeling objective is a standard NER objective with the `IOB2` tagging scheme as described in the following. Given an input sequence of n words $X = [x_1, x_2, ..., x_n]$, we perform a prediction on every word x_i to obtain a corresponding NER-tag sequence $Y_e = [y_1, y_2, ..., y_n]$, where $y_n \in D_e = \{\texttt{O}, \texttt{B-PER}, \texttt{I-PER}, \texttt{B-ORG}, \texttt{I-ORG}, \ldots\}$ such that every new entity instance starts with a `B` tag and all subsequent words belonging to that entity instance are marked with an `I` tag. Given the example sentence "Obama graduated from Columbia University .", the expected NER-tag sequence is "`B-PER O O B-ORG I-ORG O`" as shown in Figure 1. The NER objective aims to learn the function $F_e(\Theta) : X \rightarrow Y_e$.

The second sequence labeling objective applies a group-consistency loss component with a new `Linkage or Separation` (shortly LS) tagging scheme. Given the NER-tag sequence Y_e, we generate a corresponding LS-tag sequence Y_{LS} la-

Dataset	OntoNotes (PLONER ver.)	WNUT16 (PLONER ver.)	WNUT17 (test set)
Domain	Telephone Conversations (TC), Newswire (NW), Broadcast News (BN), Broadcast Conversation (BC), Weblogs (WB), Pivot Text (PT), and Magazine Genre (MZ)	Twitter	StackExchange and Reddit
Entity Types	`Person, Location, Organization`	`Person, Location, Organization`	`Person, Location, Corporation, Group, Product, Creative work`
Raw Sent. #	2,501	750	1,287

Table 1: Open-domain evaluation datasets overview.

beling a word as `L` when it is internal to a mention (i.e., its NER tag has prefix `I-`), otherwise the word is labeled as `S`. Given the example in Figure 1, we label 'University' as `L` because the word is in the same entity with the previous word, labeling other words as `S` as shown in Figure 1. Furthermore, the feature vector for this word is computed by applying a convolutional network with a 2×1 kernel to the transformer output features for the current and preceding words. The group-consistency objective aims to learn the function $F_{LS}(\Theta) : X \rightarrow Y_{LS}$.

For training, two loss functions are computed: $L_e = -\sum \log p(y_i^e)$ for the NER labeling objective and $L_{LS} = -\sum \log p(y_i^{LS})$ for the LS labeling objective. The total loss is given by an unweighted sum: $L = L_e + L_{LS}$. The input sentence is tokenized by byte-pair encoded (BPE) tokens (Sennrich et al., 2016), and some individual words can be represented by multiple tokens. When a word consists of multiple BPE tokens, we select the first token as its feature vector.

4 Experiments

We fine-tune the pre-trained transformer model on two popular annotated English NER datasets (CoNLL2003 (Tjong Kim Sang and De Meulder, 2003) and OntoNotes 5.0[1]) along with inclusion of the CNN-based bigram features. The resulting models are tested with their repective test datasets as an in-domain evaluation.

Next, to assess generalization to out-of-domain data, we use the fine-tuned CoNLL2003 model and evaluate its performance on out-of-domain benchmark datasets: PLONER (Fu et al., 2020), which is a cross-domain generalization evaluation set with three entity types (`Person, Location, Organization`), and WNUT17[2].

Benchmark datasets. We benchmark the two popular NER datasets:

- **CoNLL2003**: The CoNLL2003 dataset[3] contains sentences with part-of-speech (POS), syntactic chunk, and named entity annotations from newswire articles. The named entity tags consist of four categories (`Person, Location, Organization` and `Miscellaneous` for non-inclusive entities of the previous three groups). We directly employ the training and test set without any change.
- **OntoNotes 5.0**: The OntoNotes 5.0 dataset[4] is comprised of 1,745k English text data from various text genres (such as telephone conversations, newswire, newsgroups, broadcast news, broadcast conversation, weblogs, and religious texts), providing deeper 18 named entity categories. The dataset is converted into the `IOB2` tagging scheme with open source code[5].

The benchmark datasets are partitioned into a training, development and test set, with development set used for hyperparameter tuning and test set for evaluation.

Out-of-domain datasets. We employ the OntoNotes and WNUT16 datasets of PLONER (Fu et al., 2020) and WNUT17 test data[6] to evaluate the proposed approach on unseen domains with a fine-tuned model on CoNLL2003 training data. The out-of-domain evaluation datasets are summarized in Table 1.

[1] *https://catalog.ldc.upenn.edu/LDC2013T19*

[2] *https://noisy-text.github.io/2017/emerging-rare-entities.html*

[3] *https://www.clips.uantwerpen.be/conll2003/ner/*

[4] *https://catalog.ldc.upenn.edu/LDC2013T19*

[5] *https://github.com/yuchenlin/OntoNotes-5.0-NER-BIO*

[6] *https://noisy-text.github.io/2017/emerging-rare-entities.html*

Fine-tuning Approach	CoNLL2003			OntoNotes 5.0		
	Pre.	Rec.	F1	Pre.	Rec.	F1
RoBERTa-L	92.49	93.57	93.02	91.01	90.92	90.96
+ CNNBiF with LS	92.99	93.73	93.36	**91.07**	91.35	**91.21**
FLERT (XLM-R-L)	93.06	94.44	93.74	90.40	**91.39**	90.90
+ CNNBiF with LS	**93.33**	**94.64**	**93.98**	90.60	91.23	90.91

Table 2: Results of different fine-tuning approaches on two benchmark test sets.

- **PLONER**: The PLONER dataset is reproduced for cross-domain generalization evaluation from different domain NER datasets including WNUT16, OntoNotes-bn, OntoNotes-wb, OntoNotes-mz, OntoNotes-nw, and OntoNotes-bc. These datasets contain three types of entities (`Person`, `Location`, `Organization`) while the other categories are dropped. We combine all OntoNotes-xx datasets into a single test set.

- **WNUT17**: The WNUT17 dataset provides emerging and rare entities from newly-emerging texts such as newswire or social media. The named entity classes consist of six categories (`Person`, `Location`, `Corporation`, `Group`, `Creative work`, `Product`). We merge `Corporation` and `Group` into `Organization`, and `Creative work` and `Product` into `Miscellaneous` to align with the four CoNLL2003 categories.

Following previous work, we measure the precision, recall, and F1 score for each entity category and report the micro-averaged values for each dataset. We use the `RoBERTa-Large` (RoBERTa-L) transformer model (Liu et al., 2019) with a simple linear classifier for sequence labeling as a baseline model. We include the CNNBiF component on the baseline architecture and train the model with two sequence labeling objectives. We also employ the FLERT model proposed by Schweter and Akbik, 2020 to evaluate our approaches. The FLERT model leverages document-level features for state-of-the-art NER task results. To reproduce FLERT results, we stay with their proposed `XLM-RoBERTa-Large` (XLM-R-L) transformer model (Conneau et al., 2020) and fine-tuning configurations. We add the CNNBiF component on top of that and train the model with two sequence labeling objectives.

As the representation of each word given input sequence we use the last layer of the transformer and a common subword pooling strategy `first` (Devlin et al., 2019). To fine-tune the transformers we use the AdamW (Loshchilov and Hutter, 2019) optimizer with the fixed same number of 20 epochs. For the RoBERTa-L transformer we use a linear warmup and linear decay learning rate schedule with a learning rate of 1e-5 and for the FLERT (XLM-R-L) model we use a one-cycle training strategy with a learning rate of 5e-6 as suggested in their paper. We use the (RoBERTa-L) transformer model from HuggingFace[7] and FLERT model from flairNLP[8].

CNN-based Bigram Feature Component. On top of the two baseline models (RoBERTa-L and FLERT) we add our proposed CNNBiF along with the NER sequence labeling classifier. The input is the representations of individual words adding padding vectors to both sides. For each pair's representation we employ a simple CNN layer with a 2×1 kernel filter considering the previous word as the pair. After we truncate the last representation paired with the last padding vector, we produce the same length and same dimension of input representations. On top of the CNNBiF layer we add a linear classifier to predict `Linkage or Separation` tags of individual pair representations.

Results & Analysis. First, to gain understanding of the impact of CNN-based bigram features, we conduct a comparative evaluation on fine-tuning of RoBERTa-L and FLERT models with and without the CNNBiF module. As Table 2 shows, we find that addition of the CNNBiF approach in the RoBERTa-L model with LS objective outperforms the conventional sequence labeling approach across the CoNLL2003 and OntoNotes 5.0 benchmark data. Similarly, we observe even stronger performance increases in the FLERT model when we include the CNNBiF approach with LS objective, achieving a test F1 of 93.98 on the CoNLL2003 test data.

[7] *https://huggingface.co/transformers/*
[8] *https://github.com/flairNLP/flair*

Fine-tuning Approach	OntoNotes (PLONER ver.)			WNUT16 (PLONER ver.)			WNUT17 (test set)		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
RoBERTa-L	67.45	77.61	72.18	63.56	72.50	67.74	47.56	59.68	52.93
+ LS	67.43	77.96	72.32	64.15	72.61	68.12	48.08	**62.92**	54.51
+ CNNBiF	67.49	77.91	72.33	**64.52**	71.86	67.99	50.00	57.18	53.35
+ CNNBiF with LS	**67.84**	**78.24**	**72.67**	64.22	**72.76**	**68.22**	**51.31**	61.26	**55.85**
FLERT (XLM-R-L)	64.81	75.70	69.83	59.53	68.67	63.78	50.40	57.83	53.86
+ CNNBiF with LS	65.68	76.18	70.54	59.09	69.82	64.01	49.57	59.41	54.05

Table 3: Results of different fine-tuning approaches with CoNLL2003 training data on out-of-domain test data.

To investigate the impact of CNN-based bigram features on out-of-domain data, we fine-tune the RoBERTa-L and FLERT models on the CoNLL2003 training set and then evaluate these models on the out-of-domain datasets including OntoNotes (PLONER version), WNUT16 (PLONER version), and WNUT17. The results are shown in Table 3. We provide additional experiments for the ablation study of the RoBERTa-L model. When we use the CNNBiF layer for the LS task training jointly, we observe a much larger performance gain over the RoBERTa-L sequence labeling model. We see the only IOB2 sequence labeling task shows more mismatching predictions in the multi-word entity mentions compared to the unigram entity mentions and the LS joint task alleviates the weakness of the IOB2 sequence labeling task. To better handle the LS task we see that a single representation of adjacent tokens via a convolutional layer better captures their relationship and brings much higher performance in the LS task. Moreover, the FLERT model clearly show that the addition of the CNNBiF layer with the LS joint task significantly improves performance on the unseen-domains. Interestingly, we observe that the FLERT model is slightly worse than the RoBERTa-L model. We conjecture this is because this model brings more contextual information and therefore it is more susceptible to overfitting and less generalizable to out-of-domain sets.

Fine-tuning Approach	WNUT17	
	Single-word Entity (total ent. # 718)	Multiple-word Entity (total ent. # 361)
RoBERTa-L	475	169
+ CNNBiF with LS	**477**	**184**

Table 4: Effect of CNNBiF fine-tuning approach on different entity spans (single- and multiple-word entities) of WNUT17 test set.

Table 4 shows how the CNNBiF layer leverages the fine-tuning procedure and the impact on the prediction of singleton and multiple-word entities of WNUT17 test set. Very interestingly, we observe that there is a slight performance improvement in singleton entity examples, and a much larger performance gain in multiple-word entities, demonstrating the importance of capturing local dependency patterns for entity recognition task.

5 Conclusion

We propose a novel joint training objective for the NER task that, together with CNN-based bigram features (CNNBiF), aims to better capture local dependencies in the transformer architecture. Our results show that CNNBiF achieves near state-of-the-art F1 score with a single transformer model on the CoNLL2003 and OntoNotes 5.0 benchmark datasets. More importantly, we demonstrate that the proposed model achieves significant gains over the baseline in generalization to out-of-domain datasets. In the future we plan to investigate how the CNNBiF component impacts smaller labeled data training sets and other sequence labeling problems such as part-of-speech tagging, word segmentation, and layout extraction from documents by joint modeling of language and document image.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2019. Universal transformers.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training.

Jinlan Fu, Pengfei Liu, Qi Zhang, and Xuanjing Huang. 2020. Rethinking generalization of neural models: A named entity recognition case study.

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2019. Joint learning of named entity recognition and entity linking.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Stefan Schweter and Alan Akbik. 2020. Flert: Document-level features for named entity recognition.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Zhiwei Wang, Yao Ma, Zitao Liu, and Jiliang Tang. 2019. R-transformer: Recurrent neural network enhanced transformer.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236, Vancouver, Canada. Association for Computational Linguistics.