END-TO-END NAMED ENTITY RECOGNITION AND RELATION EXTRACTION USING PRE-TRAINED LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Named entity recognition (NER) and relation extraction (RE) are two important tasks in information extraction and retrieval (IE & IR). Recent work has demonstrated that it is beneficial to learn these tasks jointly, which avoids the propagation of error inherent in pipeline-based systems and improves performance. However, state-of-the-art joint models typically rely on external natural language processing (NLP) tools, such as dependency parsers, limiting their usefulness to domains (e.g. news) where those tools perform well. The few neural, end-to-end models that have been proposed are trained almost completely from scratch. In this paper, we propose a neural, end-to-end model for jointly extracting entities and their relations which does not rely on external NLP tools and which integrates a large, pre-trained language model. Because the bulk of our model’s parameters are pre-trained and we eschew recurrence for self-attention, our model is fast to train. On 5 datasets across 3 domains, our model matches or exceeds state-of-the-art performance, sometimes by a large margin.

1 INTRODUCTION

The extraction of named entities (named entity recognition, NER) and their semantic relations (relation extraction, RE) are key tasks in information extraction and retrieval (IE & IR). Given a sequence of text (usually a sentence), the objective is to identify both the named entities and the relations between them. This information is useful in a variety of NLP tasks such as document summarization, question answering, knowledge base population, and semantic search (Jiang [2012]). In the biomedical domain, NER and RE facilitate large-scale biomedical data analysis, such as network biology (Zhou et al. [2014]), gene prioritization (Aerts et al. [2006]), drug repositioning (Wang & Zhang [2013]) and the creation of curated databases (Li et al. [2015]). In the clinical domain, NER and RE can aid in disease and treatment prediction, readmission prediction, de-identification, and patient cohort identification (Miotto et al. [2017]).

Most commonly, the tasks of NER and RE are approached as a pipeline, with NER preceding RE. There are two main drawbacks to this approach: (1) Pipeline systems are prone to error propagation between the NER and RE systems. (2) One task is not able to exploit useful information from the other (e.g. the type of relation identified by the RE system may be useful to the NER system for determining the type of entities involved in the relation, and vice versa). More recently, joint models which simultaneously learn to extract entities and relations have been proposed, alleviating the aforementioned issues and achieving state-of-the-art performance (Miwa & Sasaki [2014], Miwa & Bansal [2016, 2017], Zhang et al. [2017], Bekoulis et al. [2018a,b], Nguyen & Verspoor [2019], Li et al. [2019]).

Many of the proposed joint models for entity and relation extraction rely heavily on external NLP tools such as dependency parsers. For instance, Miwa & Bansal [2016] propose a recurrent neural network (RNN)-based joint model that uses a bidirectional long-short term memory network (BiLSTM) to model the entities and a tree-LSTM to model the relations between entities; Li et al. [2017] propose a similar model for biomedical text. The tree-LSTM uses dependency tree information extracted using an external dependency parser to model relations between entities. The use of these external NLP tools limits the effectiveness of a model to domains (e.g. news) where those
NLP tools perform well. As a remedy to this problem, Bekoulis et al. (2018a) propose a neural, end-to-end system which jointly learns to extract entities and relations without relying on external NLP tools. In Bekoulis et al. (2018b), they augment this model with adversarial training. Nguyen & Verspoor (2019) propose a different, albeit similar end-to-end neural model which makes use of deep biaffine attention (Dozat & Manning, 2016). Li et al. (2019) approach the problem with multi-turn question answering, posing templated queries to a BERT-based QA model (Devlin et al., 2018) whose answers constitute extracted entities and their relations and achieve state-of-the-art results on three popular benchmark datasets.

While demonstrating strong performance, end-to-end systems like Bekoulis et al. (2018a;b) and Nguyen & Verspoor (2019) suffer from two main drawbacks. The first is that most of the models parameters are trained from scratch. For large datasets, this can lead to long training times. For small datasets, which are common in the biomedical and clinical domains where it is particularly challenging to acquire labelled data, this can lead to poor performance and/or overfitting. The second is that these systems typically contain RNNs, which are sequential in nature and cannot be parallelized within training examples. The multi-pass QA model proposed in Li et al. (2019) alleviates these issues by incorporating a pre-trained language model, BERT (Devlin et al., 2018), which eschews recurrence for self-attention. The main limitation of their approach is that it relies on hand-crafted question templates to achieve maximum performance. This may become a limiting factor where domain expertise is required to craft such questions (e.g., for biomedical or clinical corpora). Additionally, one has to create a question template for each entity and relation type of interest.

In this study, we propose an end-to-end model for joint NER and RE which addresses all of these issues. Similar to past work, our model can be viewed as a mixture of a NER module and a RE module (Figure 1). Unlike most previous works, we include a pre-trained, transformer-based language model, specifically BERT (Devlin et al., 2018), which achieved state-of-the-art performance across many NLP tasks. The weights of the BERT model are fine-tuned during training, and the entire model is trained in an end-to-end fashion.

Our main contributions are as follows: (1) Our solution is truly end-to-end, relying on no hand-crafted features (e.g. templated questions) or external NLP tools (e.g. dependency parsers). (2) Our model is fast to train (e.g. 12 minutes on a single GPU for the CoNLL04 corpus), as most of its parameters are pre-trained and we avoid recurrence. (3) We match or exceed state-of-the-art performance for joint NER and RE on 5 datasets across 3 domains.

2 THE MODEL

Figure 1 illustrates the architecture of our approach. Our model is composed of an NER module and an RE module. The NER module is identical to the one proposed by Devlin et al. (2018). In short, for a given input sequence \( s \) of \( N \) word tokens \( w_1, w_2, \ldots, w_N \), the pre-trained BERT\(_{BASE}\) model first produces a sequence of vectors \( x_1^{(NER)}, x_2^{(NER)}, \ldots, x_N^{(NER)} \) which are then fed to a feed-forward neural network (FFNN) for classification.

\[
s_i^{(NER)} = \text{FFNN}_{\text{NER}}(x_i^{(NER)})
\]

The output size of this layer is the number of BIOES-based NER labels in the training data, \( |C^{(NER)}| \).

In the BIOES tag scheme, each token is assigned a label, where the B- tag indicates the beginning of an entity span, I- the inside, E- the end and S- is used for any single-token entity. All other tokens are assigned the label O.

During training, a cross-entropy loss is computed for the NER objective,

\[
\mathcal{L}_{\text{NER}} = - \sum_{n=1}^{N} \log \left( \frac{e^{s_n^{(NER)}}}{\sum_{c \in C^{(NER)}} e^{s_{n,c}^{(NER)}}} \right)
\]

where \( s_n^{(NER)} \) is the predicted score that token \( n \in N \) belongs to the ground-truth entity class and \( s_{n,c}^{(NER)} \) is the predicted score for token \( n \) belonging to the entity class \( c \in C^{(NER)} \).
In the RE module, the predicted entity labels are obtained by taking the argmax of each score vector $s_1^{(NER)}, s_2^{(NER)}, \ldots, s_N^{(NER)}$. The predicted entity labels are then embedded to produce a sequence of vectors $e_1^{(NER)}, e_2^{(NER)}, \ldots, e_N^{(NER)}$ and concatenated with the hidden states from the final layer in the BERT model.

$$x_i^{(RE)} = x_i^{(NER)} \parallel e_i^{(NER)}$$

Following Miwa & Bansal (2016) and Nguyen & Verspoor (2019), we incrementally construct the set of relation candidates, $R$, using all possible combinations of the last word tokens of predicted entities, i.e. words with E- or S- labels. An entity pair is assigned to a negative relation class (NEG) when the pair has no relation or when the predicted entities are not correct. Once relation candidates are constructed, classification is performed with a deep bilinear attention mechanism (Dozat & Manning, 2016), as proposed by Nguyen & Verspoor (2019).

First, to encode directionality, the mechanism uses two FFNNs to project each $x_i^{(RE)}$ into head and tail vector representations, which correspond to whether the $i^{th}$ word serves as the head or tail argument of the relation.

$$h_i^{(head)} = \text{FFNN}_{\text{head}}(x_i^{(RE)})$$
$$h_i^{(tail)} = \text{FFNN}_{\text{tail}}(x_i^{(RE)})$$

These projections are then fed to a biaffine classifier,

$$s_{j,k}^{(RE)} = \text{Biaffine}(h_j^{(head)}, h_k^{(tail)})$$

$$\text{Biaffine}(x_1, x_2) = x_1^T U x_2 + W(x_1 \parallel x_2) + b$$

where $U$ is an $m \times |C^{(RE)}| \times m$ tensor, $W$ is a $|C^{(RE)}| \times (2 \times m)$ matrix, and $b$ is a bias vector. Here, $m$ is the size of the output layers of FFNN$_{\text{head}}$ and FFNN$_{\text{tail}}$ and $C^{(RE)}$ is the set of all relation classes (including NEG). During training, a second cross-entropy loss is computed for the RE objective

$$L_{RE} = - \sum_{r=1}^{R} \log \left( \frac{e^{s_r^{(RE)}}}{\sum_c e^{s_{r,c}^{(RE)}}} \right)$$
where $s_r^{(RE)}$ is the predicted score that relation candidate $r \in R$ belongs to the ground-truth relation class and $s_{r,c}^{(RE)}$ is the predicted score for relation $r$ belonging to the relation class $c \in C^{(RE)}$.

The model is trained in an end-to-end fashion to minimize the sum of the NER and RE losses.

$$
\mathcal{L} = \mathcal{L}_{\text{NER}} + \mathcal{L}_{\text{RE}}
$$

(9)

2.1 ENTITY PRETRAINING

In Miwa & Bansal (2016), entity pre-training is proposed as a solution to the problem of low-performance entity detection in the early stages of training. It is implemented by delaying the training of the RE module by some number of epochs, before training the entire model jointly.

Our implementation of entity pretraining is slightly different. Instead of delaying training of the RE module by some number of epochs, we weight the contribution of $\mathcal{L}_{\text{RE}}$ to the total loss during the first epoch of training.

$$
\mathcal{L} = \mathcal{L}_{\text{NER}} + \lambda \mathcal{L}_{\text{RE}}
$$

(10)

where $\lambda$ is increased linearly from 0 to 1 during the first epoch and set to 1 for the remaining epochs. We chose this scheme because the NER module quickly achieves good performance for all datasets (i.e., within one epoch). In early experiments, we found this scheme to outperform a delay of a full epoch.

2.2 IMPLEMENTATION

We implemented our model in PyTorch (Paszke et al., 2017) using the BERT$_{\text{BASE}}$ model from the PyTorch Transformers library. Our model is available at our GitHub repository. Furthermore, we use NVIDIA’s automatic mixed precision (AMP) library Apex to speed up training and reduce memory usage without affecting task-specific performance.

3 EXPERIMENTAL SETUP

3.1 DATASETS AND EVALUATION

To demonstrate the generalizability of our model, we evaluate it on 5 commonly used benchmark corpora across 3 domains. All corpora are in English. Detailed corpus statistics are presented in Table A.1 of the appendix.

3.1.1 ACE04/05

The Automatic Content Extraction (ACE04) corpus was introduced by Doddington et al. (2004), and is commonly used to benchmark NER and RE methods. There are 7 entity types and 7 relation types. ACE05 builds on ACE04, splitting the Physical relation into two classes (Physical and Part-Whole), removing the Discourse relation class and merging Employment-Membership-Subsidiary and Person-Organization-Affiliation into one class (Employment-Membership-Subsidiary).

For ACE04, we follow Miwa & Bansal (2016) by removing the Discourse relation and evaluating our model using 5-fold cross-validation on the bnews and nwire subsets, where 10% of the data was held out within each fold as a validation set. For ACE05, we use the same test split as Miwa & Bansal (2016). We use 5-fold cross-validation on the remaining data to choose the hyperparameters. Once hyperparameters are chosen, we train on the combined data from all the folds and evaluate on the test set. For both corpora, we report the micro-averaged F$_1$ score. We obtained the pre-processing scripts from Miwa & Bansal (2016).

https://github.com/huggingface/pytorch-transformers
https://github.com/NEURAL/APEX
https://github.com/tticoin/LSTM-ER/tree/master/data
3.1.2 CoNLL04

The CoNLL04 corpus was introduced in Roth & Yih [2004] for learning the problem of simultaneously recognizing entities and relations, and consists of articles from the Wall Street Journal (WSJ) and Associated Press (AP). There are 4 entity types and 5 relation types.

As there are no official splits, we use the same test set split as Miwa & Sasaki [2014]. We use 5-fold cross-validation on the remaining data to choose hyperparameters. Once hyperparameters are chosen, we train on the combined data from all folds and evaluate on the test set, reporting the micro-averaged $F_1$ score.

3.1.3 ADE

The adverse drug event corpus was introduced by Gurulingappa et al. [2012] to serve as a benchmark for systems which aim to identify adverse drug events from free-text. It consists of the abstracts of medical case reports retrieved from PubMed. There are two entity types, Drug and Adverse effect and one relation type, Adverse drug event.

Similar to previous work (Li et al., 2016; 2017; Bekoulis et al., 2018b), we remove $\sim$130 relations with overlapping entities and evaluate our model using 10-fold cross-validation, where 10% of the data within each fold was used as a validation set, 10% as a test set and the remaining data is used as a train set. We report the macro $F_1$ score averaged across all folds.

3.1.4 i2b2

The 2010 i2b2/VA dataset was introduced by Uzuner et al. [2011] for the 2010 i2b2/Va Workshop on Natural Processing Challenges for Clinical Records. The workshop contained an NER task focused on the extraction of 3 medical entity types (Problem, Treatment, Test) and an RE task for 8 relation types.

In the official splits, the test set contains roughly twice as many examples as the train set. To increase the number of training examples while maintaining a rigorous evaluation, we elected to perform 5-fold cross-validation on the combined data from both partitions. We used 10% of the data within each fold as a validation set, 20% as a test set and the remaining data was used as a train set. We report the micro $F_1$ score averaged across all folds.

To the best of our knowledge, we are the first to evaluate a joint NER and RE model on the 2010 i2b2/VA dataset. Therefore, we decided to compare to scores obtained by independent NER and RE systems. We note, however, that the scores of independent RE systems are not directly comparable to the scores we report in this paper. This is because RE is traditionally framed as a sentence-level classification problem. During pre-processing, each example is permutated into processed examples containing two “blinded” entities and labelled for one relation class. E.g. the example: “His PCP had recently started ciprofloxacin TREATMENT for a UTI PROBLEM” becomes “His PCP had recently started @TREATMENT$ for a @PROBLEM$”, where the model is trained to predict the target relation type, “Treatment is administered for medical problem” (TrAP).

We argue that this task is inherently easier than the joint setup, for two reasons: relation predictions are made on ground-truth entities, as opposed to predicted entities (which are noisy) and the model is only required to make one classification decision per pre-processed sentence. In the joint setup, a model must identify any number of relations (or the lack thereof) between all unique pairs of predicted entities in a given input sentence. To control for the first of these differences, we report scores from our model in two settings, once when predicted entities are used as input to the RE module, and once when ground-truth entities are used.

5 http://cistern.cis.lmu.de/globalNormalization/globalNormalization_all.zip
3.2 HYPERPARAMETERS

Besides batch size, learning rate and number of training epochs, we used the same hyperparameters across all experiments (see Table A.2). Similar to Devlin et al. (2018), learning rate and batch size were selected for each dataset using a minimal grid search (see See Table A.3).

One hyperparameter selected by hand was the choice of the pre-trained weights used to initialize the BERT BASE model. For general domain corpora, we found the cased BERT BASE weights from Devlin et al. (2018) to work well. For biomedical corpora, we used the weights from BioBERT (Lee et al., 2019), which recently demonstrated state-of-the-art performance for biomedical NER, RE and QA. Similarly, for clinical corpora we use the weights provided by Peng et al. (2019), who pre-trained BERT BASE on PubMed abstracts and clinical notes from MIMIC-III.

4 RESULTS

4.1 JOINTLY LEARNING NER AND RE

Table 1 shows our results in comparison to previously published results, grouped by the domain of the evaluated corpus. We find that on every dataset besides i2b2, our model improves NER performance, for an average improvement of \(\sim 2\%\). This improvement is particularly large on the ACE04 and ACE05 corpora (3.98% and 2.41% respectively). On i2b2, our joint model performs within 0.29% of the best independent NER solution. For relation extraction, we outperform previous methods on 2 datasets and come within \(\sim 2\%\) on both ACE05 and CoNLL04. In two cases, our performance improvement is substantial, with improvements of 4.59% and 10.25% on the ACE04 and ADE corpora respectively. For i2b2, our score is not directly comparable to previous systems (as discussed in section 3.1.4) but will facilitate future comparisons of joint NER and RE methods on this dataset. By comparing overall performance, we find that our approach achieves new state-of-the-art performance for 3 popular benchmark datasets (ACE04, ACE05, ADE) and comes within 0.2% for CoNLL04.

4.2 ABLATION ANALYSIS

To determine which training strategies and components are responsible for our models performance, we conduct an ablation analysis on the CoNLL04 corpus (Table 2). We perform five different ablations: (a) Without entity pre-training (see section 2.1), i.e. the loss function is given by equation 9. (b) Without entity embeddings, i.e. equation 3 becomes \(x_i^{(\text{RE})} = x_i^{(\text{NER})}\). (c) Replacing the two feed-forward neural networks, FFNN\(_{\text{head}}\) and FFNN\(_{\text{tail}}\) with a single FFNN (see equation 4 and 5). (d) Removing FFNN\(_{\text{head}}\) and FFNN\(_{\text{tail}}\) entirely. (e) Without the bilinear operation, i.e. equation 7 becomes a simple linear transformation.

Removing FFNN\(_{\text{head}}\) and FFNN\(_{\text{tail}}\) (see equation 4 and 5) has, by far, the largest negative impact on performance. Interestingly, however, replacing FFNN\(_{\text{head}}\) and FFNN\(_{\text{tail}}\) with a single FFNN has only a small negative impact. This suggests that while these layers are very important for model performance, using distinct FFNNs for the projection of head and tail entities (as opposed to the same FFNN) is relatively much less important. The next most impactful ablation was entity pre-training, suggesting that low-performance entity detection during the early stages of training is detrimental to learning (see section 2.1). Finally, we note that the importance of entity embeddings is surprising, as a previous study has found that entity embeddings did not help performance on the CoNLL04 corpus specifically (Bekoulis et al., 2018a), although their architecture was markedly different. We conclude that each of our ablated components is necessary to achieve maximum performance.

4.3 ANALYSIS OF THE WORD-LEVEL ATTENTION WEIGHTS

One advantage of including a transformer-based language model is that we can easily visualize the attention weights with respect to some input. This visualization is useful, for example, in detecting model bias and locating relevant attention heads (Vig, 2019). Previous works have used such visualizations to demonstrate that specific attention heads mark syntactic dependency relations and

[https://mimic.physionet.org/](https://mimic.physionet.org/)
Table 1: Comparison to previously published F1 scores for joint named entity recognition (NER) and relation extraction (RE). Ours: our full, end-to-end model (see section 2). Ours (gold): our model, when gold entity labels are used as input to the RE module. Bold: best scores. Subscripts denote standard deviation across three runs. ∆: difference to our overall score.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Entity</th>
<th>Relation</th>
<th>Overall</th>
<th>∆</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General</strong></td>
<td>ACE04</td>
<td>Miwa &amp; Bansal (2016)</td>
<td>81.80</td>
<td>48.40</td>
<td>65.10</td>
</tr>
<tr>
<td></td>
<td>Bekoulis et al. (2018b)</td>
<td>81.64</td>
<td>47.45</td>
<td>64.54</td>
<td>-6.25</td>
</tr>
<tr>
<td></td>
<td>Li et al. (2019)</td>
<td>83.60</td>
<td>49.40</td>
<td>66.50</td>
<td>-4.29</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>87.58</td>
<td>53.99</td>
<td>70.79</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>(gold)</td>
<td>87.21</td>
<td>58.63</td>
<td>72.92</td>
<td>–</td>
</tr>
<tr>
<td><strong>ACE05</strong></td>
<td>Miwa &amp; Bansal (2016)</td>
<td>83.40</td>
<td>55.60</td>
<td>69.50</td>
<td>-3.42</td>
</tr>
<tr>
<td></td>
<td>Bekoulis et al. (2017)</td>
<td>83.61</td>
<td>61.95</td>
<td>72.78</td>
<td>-5.37</td>
</tr>
<tr>
<td></td>
<td>Li et al. (2019)</td>
<td>87.80</td>
<td>68.90</td>
<td>78.35</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>89.46</td>
<td>66.83</td>
<td>78.15</td>
<td>–</td>
</tr>
<tr>
<td>CoNLL04</td>
<td>Miwa &amp; Sasaki (2014)</td>
<td>80.70</td>
<td>61.00</td>
<td>70.85</td>
<td>-7.30</td>
</tr>
<tr>
<td></td>
<td>Bekoulis et al. (2018b)</td>
<td>83.61</td>
<td>61.95</td>
<td>72.78</td>
<td>-5.37</td>
</tr>
<tr>
<td></td>
<td>Li et al. (2019)</td>
<td>87.80</td>
<td>68.90</td>
<td>78.35</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>89.56</td>
<td>85.77</td>
<td>87.66</td>
<td>–</td>
</tr>
<tr>
<td>Biomedical</td>
<td>ADE</td>
<td>Li et al. (2016)</td>
<td>79.50</td>
<td>63.40</td>
<td>71.45</td>
</tr>
<tr>
<td></td>
<td>Li et al. (2017)</td>
<td>84.60</td>
<td>71.40</td>
<td>78.00</td>
<td>-9.66</td>
</tr>
<tr>
<td></td>
<td>Bekoulis et al. (2018b)</td>
<td>86.73</td>
<td>75.52</td>
<td>81.13</td>
<td>-6.53</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>89.55</td>
<td>85.77</td>
<td>87.66</td>
<td>–</td>
</tr>
<tr>
<td>Clinical</td>
<td>i2b2*</td>
<td>Si et al. (2019)**</td>
<td>89.55</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Peng et al. (2019)**</td>
<td>–</td>
<td>76.40</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Ours (gold)</td>
<td>–</td>
<td>72.03</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>89.26</td>
<td>63.02</td>
<td>76.14</td>
<td>–</td>
</tr>
</tbody>
</table>

*To the best of our knowledge, there are no published joint NER and RE models that evaluate on the i2b2 2010 corpus. We compare our model to the state-of-the-art for each individual task (see section 3.1.4).

** We compare to the scores achieved by their BERT BASE model.

Table 2: Ablation experiment results on the CoNLL04 corpus. Scores are reported as a micro-averaged F1 score on the validation set, averaged across three runs of 5-fold cross-validation. (a) Without entity pre-training (section 2.1). (b) Without entity embeddings (Eq. 3). (c) Using a single FFNN in place of FFNN head and FFNN tail (Eq. 4 and 5) (d) Without FFNN head and FFNN tail (e) Without the bilinear operation (Eq. 7). Bold: best scores. Subscripts denote standard deviation across three runs. ∆: difference to the full models score.

<table>
<thead>
<tr>
<th>Model</th>
<th>Entity</th>
<th>Relation</th>
<th>Overall</th>
<th>∆</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model</td>
<td>86.32</td>
<td>60.29</td>
<td>73.30</td>
<td>–</td>
</tr>
<tr>
<td>(a) w/o Entity pre-training</td>
<td>85.91</td>
<td>59.44</td>
<td>72.67</td>
<td>-0.63</td>
</tr>
<tr>
<td>(b) w/o Entity embeddings</td>
<td>86.07</td>
<td>59.52</td>
<td>72.80</td>
<td>-0.51</td>
</tr>
<tr>
<td>(c) Single FFNN</td>
<td>86.02</td>
<td>60.13</td>
<td>73.07</td>
<td>-0.23</td>
</tr>
<tr>
<td>(d) w/o Head/Tail</td>
<td>83.93</td>
<td>54.67</td>
<td>69.30</td>
<td>-4.00</td>
</tr>
<tr>
<td>(e) w/o Bilinear</td>
<td>86.35</td>
<td>59.60</td>
<td>72.98</td>
<td>-0.33</td>
</tr>
</tbody>
</table>

that lower layers tend to learn more about syntax while higher layers tend to encode more semantics (Raganato & Tiedemann, 2018).

In Figure 2 we visualize the attention weights of select layers and attention heads from an instance of BERT fine-tuned within our model on the CoNLL04 corpus. We display four patterns that are easily interpreted: paying attention to the next and previous word, paying attention to the word itself, and paying attention to the end of the sentence. These same patterns have been found in pre-trained
Figure 2: Visualization of the attention weights from select layers and heads of BERT after it was fine-tuned within our model on the CoNLL04 corpus. Darker squares indicate larger attention weights. Attention weights are shown for the input sentence: "Ruby fatally shot Oswald two days after Kennedy was assassinated.". The CLS and SEP tokens have been removed. Four major patterns are displayed: paying attention to the next word (first image from the left) and previous word (second from the left), paying attention to the word itself (third from the left) and to the end of the sentence (fourth from the left).

BERT models that have not been fine-tuned on a specific, supervised task (Vig, 2019; Raganato & Tiedemann, 2018), and therefore, are retained after our fine-tuning procedure.

To facilitate further analysis of our learned model, we make available Jupyter and Google Colaboratory notebooks on our GitHub repository, where users can use multiple views to explore the learned attention weights of our models. We use the BertViz library (Vig, 2019) to render the interactive, HTML-based views and to access the attention weights used to plot the heat maps.

5 DISCUSSION AND CONCLUSION

In this paper, we introduced an end-to-end model for entity and relation extraction. Our key contributions are: (1) No reliance on any hand-crafted features (e.g. templated questions) or external NLP tools (e.g. dependency parsers). (2) Integration of a pre-trained, transformer-based language model. (3) State-of-the-art performance on 5 datasets across 3 domains. Furthermore, our model is inherently modular. One can easily initialize the language model with pre-trained weights better suited for a domain of interest (e.g. BioBERT for biomedical corpora) or swap BERT for a comparable language model (e.g. XLNet (Yang et al., 2019)). Finally, because of (2), our model is fast to train, converging in approximately 1 hour or less on a single GPU for all datasets used in this study.

Our model out-performed previous state-of-the-art performance on ADE by the largest margin (6.53%). While exciting, we believe this corpus was particularly easy to learn. The majority of sentences (~68%) are annotated for two entities (drug and adverse effect, and one relation (adverse drug event). Ostensibly, a model should be able to exploit this pattern to get near-perfect performance on the majority of sentences in the corpus. As a test, we ran our model again, this time using ground-truth entities in the RE module (as opposed to predicted entities) and found that the model very quickly reached almost perfect performance for RE on the test set (~98%). As such, high performance on the ADE corpus is not likely to transfer to real-world scenarios involving the large-scale annotation of diverse biomedical articles.

In our experiments, we consider only intra-sentence relations. However, the multiple entities within a document generally exhibit complex, inter-sentence relations. Our model is not currently capable of extracting such inter-sentence relations and therefore our restriction to intra-sentence relations will limit its usefulness for certain downstream tasks, such as knowledge base creation. We also ignore the problem of nested entities, which are common in biomedical corpora. In the future, we would like to extend our model to handle both nested entities and inter-sentence relations. Finally, given that multilingual, pre-trained weights for BERT exist, we would also expect our model’s performance to hold across multiple languages. We leave this question to future work.
REFERENCES


A APPENDIX
Table A.1: Detailed entity and relation counts for each corpus used in this study.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Entity classes (count)</th>
<th>Relation classes (count)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>ACE04 Person (12508), Organization (4405), Geographical Entities (4425), Location (614), Facility (688), Weapon (119), and Vehicle (209)</td>
<td>Physical (1202), Person-Social (362), Employment-Membership-Subsidiary (1591), Agent-Artifact (212), Person-Organization-Affiliation (141), Geopolitical Entity-Affiliation (517)</td>
</tr>
<tr>
<td>ACE05</td>
<td>Person (20891), Organization (5627), Geographical Entities (7455), Location (1119), Facility (1461), Weapon (911), and Vehicle (919)</td>
<td>Physical (1612), Part-Whole (1060), Person-Social (615), Agent-Artifact (703), Employment-Membership-Subsidiary (1922), Geopolitical Entity-Affiliation (730)</td>
</tr>
<tr>
<td>CoNLL04</td>
<td>Location (4765), Organization (2499), People (3918), Other (3011)</td>
<td>Kill (268), Live in (521), Located in (406), OrgBased in (452), Work for (401)</td>
</tr>
<tr>
<td>Biomedical</td>
<td>ADE Drug (4979), Adverse effect (5669)</td>
<td>Adverse drug event (6682)</td>
</tr>
<tr>
<td>Clinical</td>
<td>i2b2 Problem (19664), Test (13831), Treatment (14186)</td>
<td>PIP (2203), TeCP (504), TeRP (3053), TrAP (2617), TrCP (526), TrIP (203), TrNAP (174), TrWP (133)</td>
</tr>
</tbody>
</table>

Table A.2: Hyperparameter values and model details used across all experiments.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagging scheme</td>
<td>BIOES</td>
<td>Single token entities are tagged with an S- tag, the beginning of an entity span with a B- tag, the last token of an entity span with an E- tag, and tokens inside an entity span with an I- tag.</td>
</tr>
<tr>
<td>Dropout rate</td>
<td>0.1</td>
<td>Dropout rate applied to the output of all FFNNs and the attention heads of the BERT model.</td>
</tr>
<tr>
<td>Entity embeddings</td>
<td>128</td>
<td>Output dimension of the entity embedding layer.</td>
</tr>
<tr>
<td>FFNN head / FFNN tail</td>
<td>512</td>
<td>Output dimension of the FFNN head and FFNN tail layers.</td>
</tr>
<tr>
<td>No. layers (NER module)</td>
<td>1</td>
<td>Number of layers used in the FFNN of the NER module.</td>
</tr>
<tr>
<td>No. layers (RE module)</td>
<td>2</td>
<td>Number of layers used in the FFNNs of the RE module.</td>
</tr>
<tr>
<td>Optimizer</td>
<td>AdamW</td>
<td>Adam with fixed weight decay regularization (Loshchilov &amp; Hutter, 2017).</td>
</tr>
<tr>
<td>Gradient normalization</td>
<td>$\Gamma = 1$</td>
<td>Rescales the gradient whenever the norm goes over some threshold $\Gamma$ (Pascaru et al. 2013).</td>
</tr>
<tr>
<td>Weight decay</td>
<td>0.1</td>
<td>L2 weight decay.</td>
</tr>
</tbody>
</table>

Table A.3: Hyperparameter values specific to individual datasets. Similar to Devlin et al. (2018), a minimal grid search was performed over the values 16, 32 for batch size and 2e-5, 3e-5, and 5e-5 for learning rate.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. Epochs</th>
<th>Batch size</th>
<th>Learning rate</th>
<th>Initial BERT weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE04</td>
<td>15</td>
<td>16</td>
<td>2e-5</td>
<td>BERT-Base (cased) (Devlin et al., 2018)</td>
</tr>
<tr>
<td>ACE05</td>
<td>15</td>
<td>32</td>
<td>3e-5</td>
<td>BERT-Base (cased)</td>
</tr>
<tr>
<td>CoNLL04</td>
<td>10</td>
<td>16</td>
<td>3e-5</td>
<td>BERT-Base (cased)</td>
</tr>
<tr>
<td>ADE</td>
<td>7</td>
<td>16</td>
<td>2e-5</td>
<td>BioBERT (cased) (Lee et al., 2019)</td>
</tr>
<tr>
<td>i2b2</td>
<td>12</td>
<td>16</td>
<td>2e-5</td>
<td>NCBI-BERT (uncased) (Peng et al., 2019)</td>
</tr>
</tbody>
</table>