# Adversarial training for multi-context joint entity and relation extraction

**Giannis Bekoulis**  **Johannes Deleu**  **Thomas Demeester**  **Chris Develder**

Ghent University - imec, IDLab
Department of Information Technology
`{firstname.lastname}@ugent.be`

## Abstract

Adversarial training (AT) is a regularization method that can be used to improve the robustness of neural network methods by adding small perturbations in the training data. We show how to use AT for the tasks of entity recognition and relation extraction. In particular, we demonstrate that applying AT to a general purpose baseline model for jointly extracting entities and relations, allows improving the state-of-the-art effectiveness on several datasets in different contexts (i.e., news, biomedical, and real estate data) and for different languages (English and Dutch).

## 1 Introduction

Many neural network methods have recently been exploited in various natural language processing (NLP) tasks, such as parsing (Zhang et al., 2017), POS tagging (Lample et al., 2016), relation extraction (dos Santos et al., 2015), translation (Bahdanau et al., 2015), and joint tasks (Miwa and Bansal, 2016). However, Szegedy et al. (2014) observed that intentional small scale perturbations (i.e., adversarial examples) to the input of such models may lead to incorrect decisions (with high confidence). Goodfellow et al. (2015) proposed adversarial training (AT) (for image recognition) as a regularization method which uses a mixture of clean and adversarial examples to enhance the robustness of the model. Although AT has recently been applied in NLP tasks (e.g., text classification (Miyato et al., 2017)), this paper — to the best of our knowledge — is the first attempt investigating regularization effects of AT in a joint setting for two related tasks.

We start from a baseline joint model that performs the tasks of named entity recognition (NER) and relation extraction at once. Previously proposed models (summarized in Section 2) exhibit

several issues that the neural network-based baseline approach (detailed in Section 3.1) overcomes: (i) our model uses automatically extracted features without the need of external parsers nor manually extracted features (see Gupta et al. (2016); Miwa and Bansal (2016); Li et al. (2017)), (ii) all entities and the corresponding relations within the sentence are extracted at once, instead of examining one pair of entities at a time (see Adel and Schütze (2017)), and (iii) we model relation extraction in a multi-label setting, allowing multiple relations per entity (see Katiyar and Cardie (2017); Bekoulis et al. (2018a)). The core contribution of the paper is the use of AT as an extension in the training procedure for the joint extraction task (Section 3.2).

To evaluate the proposed AT method, we perform a large scale experimental study in this joint task (see Section 4), using datasets from different contexts (i.e., news, biomedical, real estate) and languages (i.e., English, Dutch). We use a strong baseline that outperforms all previous models that rely on automatically extracted features, achieving state-of-the-art performance (Section 5). Compared to the baseline model, applying AT during training leads to a consistent additional increase in joint extraction effectiveness.

## 2 Related work

**Joint entity and relation extraction:** Joint models (Li and Ji, 2014; Miwa and Sasaki, 2014) that are based on manually extracted features have been proposed for performing both the named entity recognition (NER) and relation extraction subtasks at once. These methods rely on the availability of NLP tools (e.g., POS taggers) or manually designed features leading to additional complexity. Neural network methods have been exploited to overcome this feature design issue and usually involve RNNs and CNNs (Miwa and Bansal,

2016; Zheng et al., 2017). Specifically, Miwa and Bansal (2016) as well as Li et al. (2017) apply bidirectional tree-structured RNNs for different contexts (i.e., news, biomedical) to capture syntactic information (using external dependency parsers). Gupta et al. (2016) propose the use of various manually extracted features along with RNNs. Adel and Schütze (2017) solve the simpler problem of entity classification (EC, assuming entity boundaries are given), instead of NER, and they replicate the context around the entities, feeding entity pairs to the relation extraction layer. Katiyar and Cardie (2017) investigate RNNs with attention without taking into account that relation labels are not mutually exclusive. Finally, Bekoulis et al. (2018a) use LSTMs in a joint model for extracting just one relation at a time, but increase the complexity of the NER part. Our baseline model enables simultaneous extraction of multiple relations from the same input. Then, we further extend this strong baseline using adversarial training.

**Adversarial training (AT)** (Goodfellow et al., 2015) has been proposed to make classifiers more robust to input perturbations in the context of image recognition. In the context of NLP, several variants have been proposed for different tasks such as text classification (Miyato et al., 2017), relation extraction (Wu et al., 2017) and POS tagging (Yasunaga et al., 2018). AT is considered as a regularization method. Unlike other regularization methods (i.e., dropout (Srivastava et al., 2014), word dropout (Iyyer et al., 2015)) that introduce random noise, AT generates perturbations that are variations of examples easily misclassified by the model.

## 3 Model

### 3.1 Joint learning as head selection

The baseline model, described in detail in Bekoulis et al. (2018b), is illustrated in Fig. 1. It aims to detect (i) the type and the boundaries of the entities and (ii) the relations between them. The input is a sequence of tokens (i.e., sentence) $w = w_1, ..., w_n$. We use character level embeddings to implicitly capture morphological features (e.g., prefixes and suffixes), representing each character by a vector (embedding). The character embeddings are fed to a bidirectional LSTM (BiLSTM) to obtain the character-based representation of the word. We also use pre-trained word embeddings.
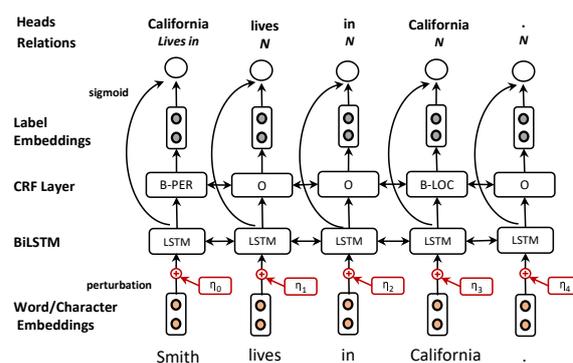


Figure 1: Our model for joint entity and relation extraction with adversarial training (AT) comprises (i) a word and character embedding layer, (ii) a BiLSTM layer, (iii) a CRF layer and (iv) a relation extraction layer. In AT, we compute the worst-case perturbations $\eta$ of the input embeddings.

Word and character embeddings are concatenated to form the final token representation, which is then fed to a BiLSTM layer to extract sequential information.

For the **NER task**, we adopt the BIO (Beginning, Inside, Outside) encoding scheme. In Fig. 1, the B-*PER* tag is assigned to the beginning token of a 'person' (*PER*) entity. For the prediction of the entity tags, we use: (i) a softmax approach for the entity classification (EC) task (assuming entity boundaries given) or (ii) a CRF approach where we identify both the *type* and the boundaries for each entity. During decoding, in the softmax setting, we greedily detect the entity *types* of the tokens (i.e., independent prediction). Although independent distribution of *types* is reasonable for EC tasks, this is not the case when there are strong correlations between neighboring tags. For instance, the BIO encoding scheme imposes several constraints in the NER task (e.g., the B-*PER* and I-*LOC* tags cannot be sequential). Motivated by this intuition, we use a linear-chain CRF for the NER task (Lample et al., 2016). For decoding, in the CRF setting, we use the Viterbi algorithm. During training, for both EC (softmax) and NER tasks (CRF), we minimize the cross-entropy loss $\mathcal{L}_{NER}$. The entity tags are later fed into the relation extraction layer as label embeddings (see Fig. 1), assuming that knowledge of the entity *types* is beneficial in predicting the relations between the involved entities.

We model the **relation extraction task** as a multi-label head selection problem (Bekoulis

et al., 2018b; Zhang et al., 2017). In our model, each word $w_i$ can be involved in multiple relations with other words. For instance, in the example illustrated in Fig. 1, "Smith" could be involved not only in a *Lives in* relation with the token "California" (head) but also in other relations simultaneously (e.g., *Works for, Born In* with some corresponding tokens). The goal of the task is to predict for each word $w_i$, a vector of heads $\hat{y}_i$ and the vector of corresponding relations $\hat{r}_i$. We compute the score $s(w_j, w_i, r_k)$ of word $w_j$ to be the head of $w_i$ given a relation label $r_k$ using a single layer neural network. The corresponding probability is defined as: $\mathbb{P}(w_j, r_k \mid w_i; \theta) = \sigma(s(w_j, w_i, r_k))$, where $\sigma(.)$ is the sigmoid function. During training, we minimize the cross-entropy loss $\mathcal{L}_{\text{rel}}$ as:

$$\sum_{i=0}^{n} \sum_{j=0}^{m} - \log \mathbb{P}(y_{i,j}, r_{i,j} \mid w_i; \theta) \qquad (1)$$

where $m$ is the number of associated heads (and thus relations) per word $w_i$. During decoding, the most probable heads and relations are selected using threshold-based prediction. The final objective for the joint task is computed as $\mathcal{L}_{\text{JOINT}}(w; \theta) = \mathcal{L}_{\text{NER}} + \mathcal{L}_{\text{rel}}$ where $\theta$ is a set of parameters. In the case of multi-token entities, only the last token of the entity can serve as head of another token, to eliminate redundant relations. If an entity is not involved in any relation, we predict the auxiliary "N" relation label and the token itself as head.

## 3.2 Adversarial training (AT)

We exploit the idea of AT (Goodfellow et al., 2015) as a regularization method to make our model robust to input perturbations. Specifically, we generate examples which are variations of the original ones by adding some noise at the level of the concatenated word representation (Miyato et al., 2017). This is similar to the concept introduced by Goodfellow et al. (2015) to improve the robustness of image recognition classifiers. We generate an adversarial example by adding the worst-case perturbation $\eta_{adv}$ to the original embedding $w$ that maximizes the loss function:

$$\eta_{adv} = \operatorname*{argmax}_{\|\eta\| \leq \epsilon} \mathcal{L}_{\text{JOINT}}(w + \eta; \hat{\theta}) \qquad (2)$$

where $\hat{\theta}$ is a copy of the current model parameters. Since Eq. (2) is intractable in neural networks, we use the approximation proposed in Goodfellow et al. (2015) defined as: $\eta_{adv} = \epsilon g / \|g\|$, with $g =$

$\nabla_w \mathcal{L}_{\text{JOINT}}(w; \hat{\theta})$, where $\epsilon$ is a small bounded norm treated as a hyperparameter. Similar to Yasunaga et al. (2018), we set $\epsilon$ to be $\alpha \sqrt{D}$ (where $D$ is the dimension of the embeddings). We train on the mixture of original and adversarial examples, so the final loss is computed as: $\mathcal{L}_{\text{JOINT}}(w; \hat{\theta}) + \mathcal{L}_{\text{JOINT}}(w + \eta_{adv}; \hat{\theta})$.

## 4 Experimental setup

We evaluate our models on four datasets, using the code as available from our github codebase.[1] Specifically, we follow the 5-fold cross-validation defined by Miwa and Bansal (2016) for the ACE04 (Doddington et al., 2004) dataset. For the CoNLL04 (Roth and Yih, 2004) EC task (assuming boundaries are given), we use the same splits as in Gupta et al. (2016); Adel and Schütze (2017). We also evaluate our models on the NER task similar to Miwa and Sasaki (2014) in the same dataset using 10-fold cross validation. For the Dutch Real Estate Classifieds, DREC (Bekoulis et al., 2017) dataset, we use train-test splits as in Bekoulis et al. (2018a). For the Adverse Drug Events, ADE (Gurulingappa et al., 2012), we perform 10-fold cross-validation similar to Li et al. (2017). To obtain comparable results that are not affected by the input embeddings, we use the embeddings of the previous works. We employ early stopping in all of the experiments. We use the Adam optimizer (Kingma and Ba, 2015) and we fix the hyperparameters (i.e., $\alpha$, dropout values, best epoch, learning rate) on the validation sets. The scaling parameter $\alpha$ is selected from $\{5e-2, 1e-2, 1e-3, 1e-4\}$. Larger values of $\alpha$ (i.e., larger perturbations) lead to consistent performance decrease in our early experiments. This can be explained from the fact that adding more noise can change the content of the sentence as also reported by Wu et al. (2017).

We use three types of evaluation, namely: (i) *S(trict)*: we score an entity as correct if both the entity boundaries and the entity *type* are correct (ACE04, ADE, CoNLL04, DREC), (ii) *B(oundaries)*: we score an entity as correct if only the entity boundaries are correct while the entity *type* is not taken into account (DREC) and (iii) *R(elaxed)*: a multi-token entity is considered correct if at least one correct *type* is assigned to the tokens comprising the entity, assuming that the

---

[1] `https://github.com/bekou/multihead_joint_entity_relation_extraction`

2832

| | Settings | Features | Eval. | Entity | Relation | Overall |
|---|---|---|---|---|---|---|
| ACE 04 | Miwa and Bansal (2016) | ✓ | S | 81.80 | 48.40 | 65.10 |
| | Katiyar and Cardie (2017) | ✗ | S | 79.60 | 45.70 | 62.65 |
| | baseline | ✗ | S | 81.16 | 47.14 | 64.15 |
| | baseline + AT | ✗ | S | **81.64** | **47.45** | **64.54** |
| CoNLL 04 | Gupta et al. (2016) | ✓ | R | 92.40 | 69.90 | 81.15 |
| | Gupta et al. (2016) | ✗ | R | 88.80 | 58.30 | 73.60 |
| | Adel and Schütze (2017) | ✗ | R | 82.10 | 62.50 | 72.30 |
| | baseline EC | ✗ | R | **93.26** | 67.01 | 80.14 |
| | baseline EC + AT | ✗ | R | 93.04 | **67.99** | **80.51** |
| | Miwa and Sasaki (2014) | ✓ | S | 80.70 | 61.00 | 70.85 |
| | baseline | ✗ | S | 83.04 | 61.04 | 72.04 |
| | baseline + AT | ✗ | S | **83.61** | **61.95** | **72.78** |
| DREC | Bekoulis et al. (2018a) | ✗ | B | 79.11 | 49.70 | 64.41 |
| | baseline | ✗ | B | 82.30 | 52.81 | 67.56 |
| | baseline + AT | ✗ | B | **82.96** | **53.87** | **68.42** |
| | baseline | ✗ | S | 81.39 | 52.26 | 66.83 |
| | baseline + AT | ✗ | S | **82.04** | **53.12** | **67.58** |
| ADE | Li et al. (2016) | ✓ | S | 79.50 | 63.40 | 71.45 |
| | Li et al. (2017) | ✓ | S | 84.60 | 71.40 | 78.00 |
| | baseline | ✗ | S | 86.40 | 74.58 | 80.49 |
| | baseline + AT | ✗ | S | **86.73** | **75.52** | **81.13** |

Table 1: Comparison of our method with the state-of-the-art in terms of $F_1$ score. The proposed models are: (i) baseline, (ii) baseline EC (predicts only entity classes) and (iii) baseline (EC) + AT (regularized by AT). The ✓ and ✗ symbols indicate whether the models rely on external NLP tools. We include different evaluation types (*S*, *R* and *B*).

boundaries are known (CoNLL04), to compare to previous works. In all cases, a relation is considered as correct when both the relation *type* and the argument entities are correct.

## 5 Results

Table 1 shows our experimental results. The name of the dataset is presented in the first column while the models are listed in the second column. The proposed models are the following: (i) *baseline*: the baseline model shown in Fig. 1 with the CRF layer and the sigmoid loss, (ii) *baseline EC*: the proposed model with the softmax layer for EC, (iii) *baseline (EC) + AT*: the baseline regularized using AT. The final three columns present the $F_1$ results for the two subtasks and their average performance. Bold values indicate the best results among models that use only automatically extracted features.

For ACE04, the baseline outperforms Katiyar and Cardie (2017) by ∼2% in both tasks. This improvement can be explained by the use of: (i) multi-label head selection, (ii) CRF-layer and (iii) character level embeddings. Compared to Miwa and Bansal (2016), who rely on NLP tools, the baseline performs within a reasonable margin (less than 1%) on the joint task. On the other hand, Li et al. (2017) use the same model for

the ADE biomedical dataset, where we report a 2.5% overall improvement. This indicates that NLP tools are not always accurate for various contexts. For the CoNLL04 dataset, we use two evaluation settings. We use the *relaxed* evaluation similar to Gupta et al. (2016); Adel and Schütze (2017) on the EC task. The baseline model outperforms the state-of-the-art models that do not rely on manually extracted features (>4% improvement for both tasks), since we directly model the whole sentence, instead of just considering pairs of entities. Moreover, compared to the model of Gupta et al. (2016) that relies on complex features, the baseline model performs within a margin of 1% in terms of overall $F_1$ score. We also report NER results on the same dataset and improve overall $F_1$ score with ∼1% compared to Miwa and Sasaki (2014), indicating that our automatically extracted features are more informative than the hand-crafted ones. These automatically extracted features exhibit their performance improvement mainly due to the shared LSTM layer that learns to automatically generate feature representations of entities and their corresponding relations within a single model. For the DREC dataset, we use two evaluation methods. In the *boundaries* evaluation, the baseline has an improvement of ∼3% on both tasks compared to Bekoulis et al. (2018a), whose quadratic scoring layer complicates NER.

Table 1 and Fig. 2 show the effectiveness of the adversarial training on top of the baseline model. In all of the experiments, AT improves the predictive performance of the baseline model in the joint setting. Moreover, as seen in Fig. 2, the performance of the models using AT is closer to maximum even from the early training epochs. Specifically, for ACE04, there is an improvement in both tasks as well as in the overall $F_1$ performance (0.4%). For CoNLL04, we note an improvement in the overall $F_1$ of 0.4% for the EC and 0.8% for the NER tasks, respectively. For the DREC dataset, in both settings, there is an overall improvement of ∼1%. Figure 2 shows that from the first epochs, the model obtains its maximum performance on the DREC validation set. Finally, for ADE, our AT model beats the baseline $F_1$ by 0.7%.

Our results demonstrate that AT outperforms the neural baseline model consistently, considering our experiments across multiple and more diverse datasets than typical related works. The im-
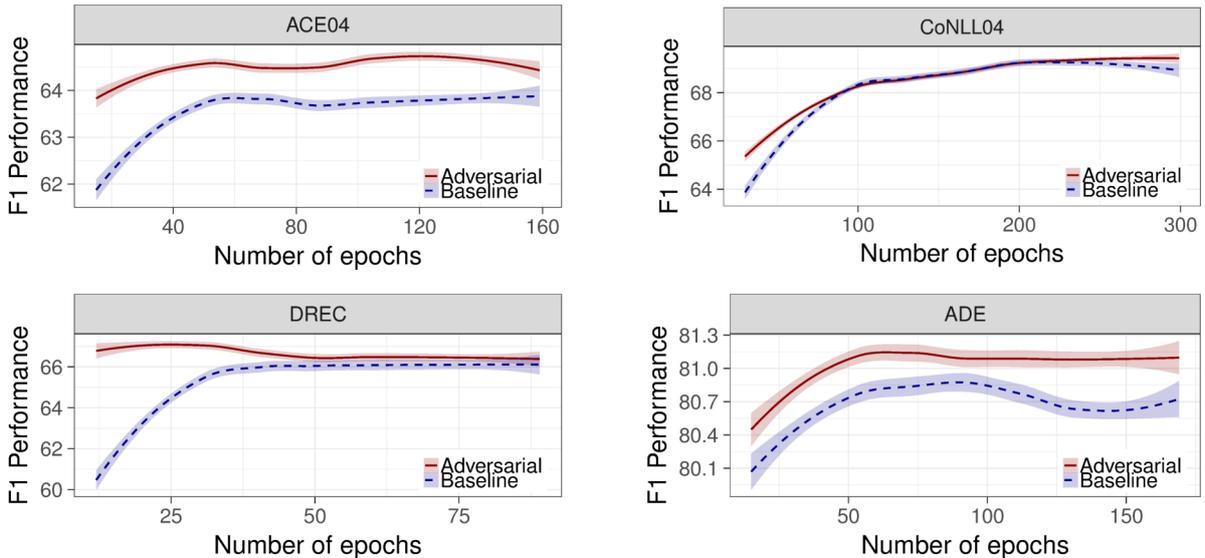
Figure 2: $F_1$ performance of the baseline and the AT models on the validation sets from 10-30 epochs onwards depending on the dataset. The smoothed lines (obtained by LOWESS smoothing) model the trends and the 95% confidence intervals.

provement of AT over our baseline (depending on the dataset) ranges from ~0.4% to ~0.9% in terms of overall $F_1$ score. This seemingly small performance increase is mainly due to the limited performance benefit for the NER component, which is in accordance with the recent advances in NER using neural networks that report similarly small gains (e.g., the performance improvement in Ma and Hovy (2016) and Lample et al. (2016) on the CoNLL-2003 test set is 0.01% and 0.17% $F_1$ percentage points, while in the work of Yasunaga et al. (2018), a 0.07% $F_1$ improvement on CoNLL-2000 using AT for NER is reported). However, the relation extraction performance increases by ~1% $F_1$ scoring points, except for the ACE04 dataset. Further, as seen in Fig. 2, the improvement for CoNLL04 is particularly small on the evaluation set. This may indicate a correlation between the dataset size and the benefit of adversarial training in the context of joint models, but this needs further investigation in future work.

## 6   Conclusion

We proposed to use adversarial training (AT) for the joint task of entity recognition and relation extraction. The contribution of this study is twofold: (i) investigation of the consistent effectiveness of AT as a regularization method over a multi-context baseline joint model, with (ii) a large scale experimental evaluation. Experiments show that AT improves the results for each task separately, as well as the overall performance of the baseline joint model, while reaching high performance already during the first epochs of the training procedure.

## Acknowledgments

## References

Heike Adel and Hinrich Schütze. 2017. Global normalization of convolutional neural networks for joint entity and relation classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference for Learning Representations*, San Diego, USA.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2017. Reconstructing the house from the ad: Structured prediction on real estate classifieds. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: (Volume 2, Short Papers)*, pages 274–279, Valencia, Spain.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018a. An attentive neural architecture for joint segmentation and parsing and its application to real estate ads. *Expert Systems with Applications*, 102:100–112.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018b. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, volume 2, page 1, Lisbon, Portugal.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*, San Diego, USA.

Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2537–2547.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Arzoo Katiyar and Claire Cardie. 2017. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, San Diego, USA.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California.

Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, 18(1):1–11.

Fei Li, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Joint models for extracting adverse drug events from biomedical text. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2838–2844, New York, USA. IJCAI/AAAI Press.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412, Baltimore, USA.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1858–1869, Doha, Qatar. Association for Computational Linguistics.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *Proceedings of the International Conference on Learning Representations*, Toulon, France.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 1–8, Boston, USA. Association for Computational Linguistics.

Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 626–634, Beijing, China.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, Banff, Canada.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783, Copenhagen, Denmark. Association for Computational Linguistics.

Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, USA.

Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. 2017. Dependency parsing as head selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: (Volume 1, Long Papers)*, pages 665–676, Valencia, Spain.

Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. 2017. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257:59–66.