Adversarial Perturbations for Joint Entity and Relation Extraction

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder

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

1 Introduction

The goal of the entity recognition and relation extraction task is to discover relational structures of entity mentions from unstructured texts. It is a central problem in information extraction since it is critical for tasks such as knowledge base population and question answering. In this work, we focus on extending the training procedure of our newly proposed general purpose joint model [4] for entity recognition and relation extraction with adversarial training (AT) [2].

Our model performs the two tasks of entity recognition and relation extraction simultaneously. It achieves state-of-the-art performance in a number of different contexts (i.e., news, biomedical, real estate) and languages (i.e., English, Dutch) without relying on any manually engineered features nor additional NLP tools. In summary, our proposed model: (i) does not rely on external NLP tools nor hand-crafted features, (ii) entities and relations within the same text fragment (typically a sentence) are extracted simultaneously, where (iii) an entity can be involved in multiple relations at once. To evaluate the proposed AT method, we perform the same set of experiments while we apply AT on top of our joint model. Compared to the baseline model, applying AT during training leads to a consistent additional increase in joint extraction effectiveness.

2 Model

Our general purpose joint model for joint entity and relation extraction is described in detail in [4]. In the AT model [2], we compute the worst-case perturbations η of the input embeddings. For the **NER task**, we adopt the BIO encoding scheme using either a (i) a softmax or (ii) a CRF approach. We model the **relation extraction task** as a multi-label head selection problem [3, 4]. In our model, each word w_i can be involved in multiple relations with other words. 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.

We exploit the idea of AT 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. We generate an adversarial example by adding the worst-case perturbation η_{adv} to the original embedding w using the approximation $\eta_{adv} = \epsilon g / \|g\|$, with $g = \nabla_w \mathcal{L}_{\text{JOINT}}(w; \hat{\theta})$, where ϵ is a small bounded norm treated as a hyperparameter. We train on the mixture of original and adversarial examples, so the final loss is computed as: $\mathcal{L}_{\text{JOINT}}(w; \hat{\theta}) + \mathcal{L}_{\text{JOINT}}(w + \eta_{adv}; \hat{\theta})$.

3 Experimental Results

We evaluate our models on four datasets (ACE04, CoNLL04, DREC [1], ADE) using three evaluation schemas [2, 4, 5]. Experimental results show consistent effectiveness of the adversarial training on top of our general purpose joint model. In all of the experiments, AT improves the predictive performance of the baseline model in the joint setting. Moreover, the models reach maximal performance from early training epochs (i.e., faster than without AT). 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 entity classification (assuming entity boundaries are given) and 0.8% for the NER tasks, respectively. For the DREC dataset, in both settings, there is an overall improvement of ∼1% and 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.

References

1. Bekoulis, G., Deleu, J., Demeester, T., Develder, C.: 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). pp. 274–279. Valencia, Spain (3–7 Apr 2017)
2. Bekoulis, G., Deleu, J., Demeester, T., Develder, C.: Adversarial training for multi-context joint entity and relation extraction. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2830–2836 (2018)
3. Bekoulis, G., Deleu, J., Demeester, T., Develder, C.: An attentive neural architecture for joint segmentation and parsing and its application to real estate ads. Expert Systems with Applications **102**, 100–112 (2018)
4. Bekoulis, G., Deleu, J., Demeester, T., Develder, C.: Joint entity recognition and relation extraction as a multi-head selection problem. Expert Systems with Applications **114**, 34–45 (2018)
5. Bekoulis, G., Deleu, J., Demeester, T., Develder, C.: Sub-event detection from twitter streams as a sequence labeling problem. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 745–750 (2019)