Predicting Chemical Protein Relations with Biaffine Relation Attention Networks

Patrick Verga and Andrew McCallum
College of Information and Computer Sciences, University of Massachusetts Amherst

Abstract— **Predicting relationships between biological entities is important for drug discovery and precision medicine. The Biocreative VI Task 5 focuses on mining chemical-protein interactions from text. Our submission uses a biaffine relation attention network to encode the full paper abstract and predict relationships between all mention pairs simultaneously. We use no patterns, rules, hard written features, or external resources. Despite this, our best run achieves an test set Micro F1 score of 45.82.**

Keywords—relation extraction, neural networks, chemical protein interactions

I. Introduction

Knowledge bases containing relationships between entieis are powerful tools for downstream tasks such as question answering, query understanding, and exploratory research such as drug discovery and precision medicine. Because these knowledge bases are highly incomplete, methods for knowledge base completion have been developed. These can broadly be broken down into link predition – inferring missing links using existing graph properties, and relation extraction – mining new entities and relationships from text.

Extracting relations between entities is one of the core problems in information extraction. Initial methods relied on hand written patterns and bootstrapping methods. Later rich hand crafted features were fed into machine learning classifiers such as support vector machines. More recently, neural networks have become the state of the art, primarily gated recurrant neural networks and convolutional neural networks.

Previous neural models for relation extraction have formed predictions on a single mention pair at a time constrained to a single sentence. Our model instead produces all predictions for all mention pairs simultaneously by encoding the entire abstract. The biaffine relation attention network (BRAN) encodes the full paper abstract using the Transformer attention-based architecture [2] and then applies a bi-affine operation between all mention pairs with respect to the set of query relations [1].

Our experiments on the Biocreative VI Task 5: Text mining chemical-protein interactions (CHEMPROT) dataset show our models strong performance. Our model encoding the entire abstract outperforms an equivalent sentence level classifier by incorporating a broader context to make its predictions. We improve our performance further by ensembling many versions of our model trained with different random seeds.

II. Model

The BRAN model was first proposed in [1] and was shown to have state of the art performance on the Biocreative V Chemical Disease Relation dataset. The model does not use any handcrafted features or rules. Even tokenization is performed using corpus statistics.

A. Inputs

We tokenize our data into byte-pair encodings using a budget of 7500 [3]. The full abstract is converted to a sequence of token embeddings of dimensions 64. Words are randomly replaced with a special UNK token with probability .15 and we apply dropout to the embeddings with probability .15.

B. Transformer

These token embeddings are then contextually encoded using the Transformer architecture with 2 block repeats and internal dimensions of size 64. The feed-forward component consists of 3 convolutional layers with kernel width 1, 5, and 1 and dimension 256. Dropout is applied to the internal layers with probability .15.

C. Biaffine pairwise scores

For each pair of tokens in our abstract we compute a biaffine operation using a 64 by 64 dimensional matrix per relation type.

D. Training

We train our model using cross entropy over the training set using the Adam optimizer with learning rate .0005, epsilon 1e-4, beta1 .9 and beta2 .9 and a batch size of 8. We additionally apply add gradient noise with standard deviation .1 [4]. Additionally, we use the final output representations of the Transformer to predict named entity labels using BIO encoding. We perform early stopping on the development set optimized for Micro F1.

III. Results

Model	*F1*
Sentence Ensemble	44.6
Abstract Ensemble	46.9

Fig. 1. F1 scores on the development set. Our model trained on full asbtracts outperforms our model trained on just single sentences. Example of a figure caption.

Model	*F1*
Abstract	41.2
Abstract Ensemble	46.9
Abstract+Sentence Ensemble	49.8

Fig. 2. F1 scores on the development set. Ensemble models outperform the single model scores. Combining models trained on the full abstract and those trained on the single sentences improves performance further.

Model	*Precision*	*Recall*	*F1*
Abstract+Sentence Ensemble	47.18	44.53	45.82

Fig. 3. Test set scores.

Figure 1 shows various versions of our models scores on the development set. We tune each of the per-relation decision thresholds separately. We apply a post-processing step to our full abstract models to remove predictions that cross the sentence boundary because the CHEMPROT dataset only considers within sentence relationships. Models denoted with ensemble average the prediction probabilities of 20 models trained with the same hyper parameters but different random seeds.

In Figure 1 we see that our model that is trained on the full abstract outperforms the model trained on single sentences. Figure 2 shows that ensembling models outperforms single models, and ensembling the sentence level and abstract level models improve performance further. Figure 3 shows the results of our best run (run 4) which ensembles the full abstract and sentence models.

IV. Conclusion

We presented our submission to the Biocreative VI Task 5 Chemical-Protein relation text mining shared task. Our biaffine relation attention model effectively and efficiently extracted chemical protein relationships despite using no hand crafted features or external resources.

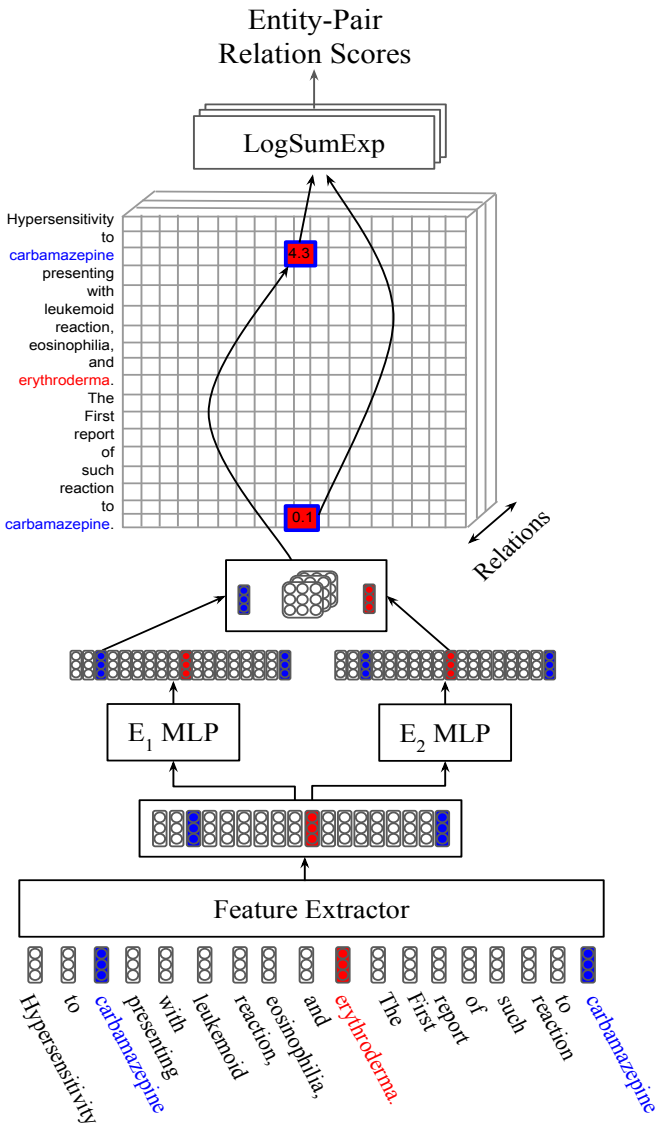

Fig. 4. The relation extraction architecture. Inputs are initially contextually encoded using the transformer. Each transformed token is then passed through an $Entity_1$ and $Entity_2$ MLP to produce two separate versions of each token able to act as either the head or tail of the relationship. A bi-affine operation is then performed between each $Entity_1$ and $Entity_2$ token with respect to each relations embedding matrix producing a token $\times$ token $\times$ relation pairwise affinity tensor. Finally, the scores for cells corresponding to the same entity pair are pooled with a separate LogSumExp operation for each relation to get a final score. The colored tokens are meant only to illustrate calculating the score for a given pair of entities. The model is only given entity information when gathering scores to pool from the affinity matrix.

REFERENCES

1. Verga, P., Strubell, E., Shai, O., McCallum, A.. (2017) Attending to All Mentions for Full Abstract Biological Relation Extraction. *arXiv preprint*.

2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. Attention Is All You Need. arXiv preprint arXiv:1706.03762.

3. Neelakantan, A., Vilnis, L., Le, Q.V., Sutskever, I., Kaiser, L., Kurach, K. and Martens, J., 2015. Adding gradient noise improves learning for very deep networks. arXiv preprint arXiv:1511.06807.

4. Sennrich, R., Haddow, B., and Birch, A.. (2015). Neural machine translation of rare words with subword units. *arXiv preprint*