ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Entity-Pair Embeddings for Improving Relation Extraction in the Biomedical Domain

Farrokh Mehryary[1,2], Hans Moen[1], Tapio Salakoski[1], Filip Ginter[1]

1- Turku NLP Group, Department of Future Technologies,
University of Turku, Turku, Finland

2- University of Turku Graduate School (UTUGS), Turku, Finland

Abstract. We introduce a new approach for training named-entity pair embeddings to improve relation extraction performance in the biomedical domain. These embeddings are trained in an unsupervised manner, based on the principles of distributional semantics. By adding them to neural network architectures, we show that improved F-Scores are achieved. Our best performing neural model which utilizes entity-pair embeddings along with a pre-trained BERT encoder, achieves an F-score of 77.19 on CHEMPROT (Chemical-Protein) relation extraction corpus, setting a new state-of-the-art result for the task.

1 Introduction

The significant amount and the increasing publication rate in the biomedical domain make it difficult for biomedical researchers to acquire and maintain all information that is necessary for their research. Biomedical relation extraction systems aim to address this problem. These systems can periodically scan the whole publicly available literature (PubMed article abstracts and PubMed Central Open Access (PMCOA) full article texts) and extract relations and interactions of biomedical named entities from the texts and build up-to-date relation databases or molecular interaction networks to facilitate biomedical research. A number of shared task challenges have been organized to promote the development and evaluation of such systems. For example, the BioCreative VI shared task [1] was recently organized, which provided the CHEMPROT corpus for chemical-protein relations extraction. Since the CHEMPROT corpus is fairly new, relatively large and carefully annotated, it has become an important benchmark for evaluating modern relation extraction systems.

So far the best results for relation extraction have been achieved using system ensemble approaches. On the CHEMPROT corpus, the best result during the BioCreative VI challenge was obtained by Peng et al. [2] (an F-score of 64.10) by using an ensemble system consisting of a recurrent neural network system, a convolutional neural network system and a support vector machine system. Recently, the introduction of transformer-based language representation models such as BERT [3], impacted the field and resulted in unprecedented jumps in F-score on many data sets. The state-of-the-art result on CHEMPROT is recently achieved by Lee et al. [4] (an F-score of 76.46), by pre-training the BERT encoder on PubMed sentences and fine-tuning it with a decision layer on the CHEMPROT data for relation extraction. This 12 percentage points

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

increase in F-score is substantial and raises the question to what extent further improvements on top of BERT can be achieved.

Biomedical literature includes a lot of information about the relations and interactions of biomedical named entities (e.g. genes, proteins, chemicals, and drugs). We aim to leverage this literature-wide information using unsupervised methods and for every unique named-entity pair (E_i, E_j), capture all stated information about E_i and E_j and their relations and build embeddings (vector representations) of entities and entity pairs. Similarly to word2vec [5] for ordinary words, our objective is for similar proteins, chemicals and protein-chemical pairs to obtain similar embeddings. We are especially interested in investigating the possible effects of incorporating these entity and entity-pair embeddings into neural models, in order to improve the performance in relation extraction tasks in the biomedical domain. Given that manually annotated training data for such tasks is usually limited, the domain is an obvious target for transfer learning through pre-trained embeddings. We hypothesize that by pre-training and incorporating the entity-pair embeddings into neural networks, we can improve the performance in relation extraction tasks at hand, potentially better than using individual entity embeddings alone.

In this paper, we explore different approaches for pre-training vector representations for biomedical entities and entity pairs. We concentrate on the chemical and protein named entities in the CHEMPROT corpus and train different types of entity and pair embeddings. We show that when these embeddings are added into BERT-based neural architectures, they can boost the performance of relation extraction. Our approaches are inspired by the work of Levy and Goldberg [6], using richer contexts to extend the skip-gram architecture of word2vec model introduced by Mikolov et al. [5].

2 Method

We propose to pre-train embeddings for named entities and named-entity pairs using a word2vec skip-gram style training, whereby the named entities, or entity pairs are given as the focus terms, and elements from their contexts are predicted. We will investigate two ways to define the context: a simple linear context of words as in the base word2vec, and as an alternative a rich set of features extracted from the context. These features have previously been shown to be useful in supervised relation extraction and one might therefore expect they result in embeddings informative for relation extraction. More specifically, we will rely on the Turku Event Extraction System (TEES) [7] to generate these features, a system that has achieved numerous top ranks in biomedical relation extraction tasks.

2.1 Entity and entity-pair embeddings pre-training

We obtain the list of all chemical-protein pairs in the CHEMPROT corpus and find all sentences in PubMed and PMCOA texts [8] that contain at least one pair. For simplicity, we use exact matching approach when searching for the entities

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

in the texts. We then extract a set of features for each pair using the TEES system, including (1) word/lemma/POS-tag and dependency-type N-grams along the shortest path connecting the two entities in the sentence dependency parse graph, (2) word/lemma/part-of-speech N-grams of the words that are located within [-3,+3] words of the two entities, and (3) type and location of all biomedical entities occurring in the sentence. We use the word2vecf toolkit [6] for training the embeddings and use either surrounding words as the context or TEES features. Since simultaneous training of embeddings for entities and entity pairs can impact the final model (due to the shared output layer in the word2vec model), we train separate embedding models that include only entity pairs, only entities, or both pairs and entities, resulting in six models (see Table 1).

Model	**Content**	**Context for training**
`P_TEES`	only pair embeddings	TEES features
`P_Words`	only pair embeddings	union of the words surrounding the two entities
`E_TEES`	only entity embeddings	union of TEES features for every pair that includes the entity
`E_Words`	only entity embeddings	words surrounding the entity
`PE_TEES`	pair and entity embeddings	TEES features
`PE_Words`	pair and entity embeddings	surrounding words

Table 1: Description of the different embedding models.

2.2 Relation extraction with entity and entity-pair embeddings

We incorporate the pre-trained embeddings into the following neural network architectures: (1) `BERT_MASK`: this architecture is developed by Lee et al. [4] and has achieved the state-of-the-art on CHEMPROT corpus. A BERT encoder pre-trained on PubMed sentences is fine-tuned on the CHEMPROT training set with a decision layer for relation extraction task. This layer predicts one of the five possible relation types between the two entities, or a negative label for no relation. We replicate the method as well as use the pre-trained BERT model of Lee et al. [4]. In the `BERT_MASK` method, the entities are replaced with predefined tags (e.g. @PROTEIN$) to inform the classifier where the two entities are located in a sentence; (2) `BERT_MARK`: the `BERT_MASK` model hides all information about the two entities in the sentence as a consequence of its masking strategy. Since the entity and pair embedding vectors we pre-train provide information about the entities, an improvement on top of the `BERT_MASK` model might be due to this fact. Therefore, as a fairer baseline, we introduce the `BERT_MARK` model (identical to the `BERT_MASK`) except we mark the two entities using the special "unused" symbols in BERT vocabulary[1] (e.g. [unused1]17β-estradiol benzoate[unused2]); (3) `BERT+Pairs`: this model is similar to the `BERT_MARK`

[1]This provided better results compared to using normal characters to mark entity spans (which was used by Lee et al. [4]) since the pre-trained BERT has no notion of the unused symbols.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

model, except the pair embedding vector is concatenated to the BERT sentence representation vector (*[CLS]* token), transformed through a 1024-dimensional dense layer with *tanh* activation, and then presented to the decision layer. The dense layer with the non-linear activation function learns to combine BERT features with the pair embedding features; (4) `BERT+Entities`: this model is similar to the `BERT+Pairs` model, except we concatenate the chemical and the protein embedding vectors (not the pair vector); (5) `BERT+Pairs+Entities`: This model is similar to previous models, except we concatenate chemical, protein, and pair vectors. In all models, we use the exact hyper-parameters used by Lee et al. [4] and optimize the learning-rate by grid search on the development set.

3 Evaluation and results

We evaluate all approaches on the CHEMPROT corpus which contains 4,157 training examples, 2,416 examples in the development set and 3,458 examples in the test set. Chemical-protein pairs can have one of 5 positive relations (e.g up_regulation) or no interaction at all (negative). We use the official evaluation script provided by the task organizers which calculates the micro-averaged F-score of the positive classes as the task metric. Since initial random weights of a neural model can slightly impact the final F-score, we repeat each experiment (training on the training set and predicting development or test set) for 10 times which results in obtaining 10 F-scores for each approach. We report the average and standard deviation of the F-scores. We use the two-tailed two-sample independent t-test (Welch's t-test) to establish statistical significance.

3.1 Model selection

Table 2 summarizes the results on the development set, reporting statistical significance at $p = 0.1$. `BERT_MARK` outperforms `BERT_MASK`, suggesting that marking should be preferred over the masking approach. In fact, based on column G1, all models that utilize marking (rows 2-9) outperform the approach of Lee et al. [4]. However, as discussed previously, for us `BERT_MARK` is considered the baseline and as column G2 shows, the models that used only entity embeddings (rows 3,4), do not achieve statistically better results than the baseline. Similarly, the model that used pairs (trained with words contexts, row 5) does not achieve a better result, in contrast to the model that used pair embeddings (trained with TEES context, row 6). All models that utilized pre-trained entity *and* pair embeddings (rows 7,8) achieve statistically better results than the baseline. We further conduct another experiment and test the effect of randomly initializing entity and pair embeddings instead of using pre-trained embeddings, to check if the neural model can efficiently learn these embeddings from scratch. However, this model is not able to outperform the baseline (row 9). Thus we conclude pre-training embeddings on the literature is indeed useful. Based on these development set results, only 3 approaches outperform the baseline (rows 6-8).

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

#	Neural model	Embeddings model	F-score (mean)	F-score (std)	G1	G2
1	BERT_MASK	-	78.41	0.53	-	Yes
2	BERT_MARK	-	78.96	0.41	Yes	-
3	BERT+Entities	E_Words	79.23	0.43	Yes	No
4	BERT+Entities	E_TEES	79.15	0.42	Yes	No
5	BERT+Pairs	P_Words	79.27	0.43	Yes	No
6	BERT+Pairs	P_TEES	79.36	0.32	Yes	Yes
7	BERT+Pairs+Entities	PE_Words	79.47	0.37	Yes	Yes
8	BERT+Pairs+Entities	PE_TEES	**79.55**	0.40	Yes	Yes
9	BERT+Pairs+Entities	Randomly initialized	79.05	0.46	Yes	No

Table 2: Results on CHEMPROT development set. Columns G1 and G2 show if based on the statistical test, the F-score mean is significantly different from the F-score mean of BERT_MASK and BERT_MARK models respectively.

3.2 Final evaluation

We compare our best models selected on the development set (rows 6-8 in Table 2) with the best previous result of Lee et al. [4] (an F-score of 76.46) on the test set. To assess the statistical significance, we use the one-sample t-test ($p = 0.05$). We also evaluate the BERT_MASK model to check how well we have been able to replicate the method of Lee et al. [4].

#	Neural model	Embeddings model	F-score (mean)	F-score (std)	G1
	Lee et al. [4]	-	76.46	-	-
1	BERT_MASK	-	76.41	0.72	No
2	BERT+Pairs	P_TEES	77.13	0.53	Yes
3	BERT+Pairs+Entities	PE_Words	76.71	0.77	No
4	BERT+Pairs+Entities	PE_TEES	**77.19**	0.49	Yes

Table 3: Results on CHEMPROT test set. Column G1 shows if based on the statistical test, F-score mean is significantly different from the F-score of Lee et al. [4].

The test set results (Table 3) validate our replication of the Lee et al. method (row 1). The model that uses surrounding words as the context (row 3) does not outperform the baseline (at $p = 0.05$), however the models that use TEES features (rows 2,4), outperform the best previous result, suggesting that pair-embeddings with rich feature-based context can improve upon a strong BERT-based baseline. Our best model (row 4) sets a new state-of-the-art for the task, improving the best previous score by 0.73 percentage points.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

4 Conclusion and future work

We compared different approaches for pre-training entity and entity-pair embeddings to improve relation extraction performance in the biomedical domain. We have shown that (1) incorporation of these embeddings into neural models helps in achieving better performance, (2) using rich features as context (instead of using the surrounding words, i.e. the normal word2vec approach) leads to better results; (3) using pair embeddings with/without entity embeddings leads to better results compared to using entity embeddings alone. Our best model achieves an F-score of 77.19, improving the best previous result by +0.73pp over a strong baseline, and setting a new state-of-the-art for the task. As future work, we aim to investigate the effect of entity and entity-pair embeddings on other biomedical relation extraction data sets.

5 Acknowledgements

We would like to thank Dr. Sampo Pyysalo for his invaluable recommendations and Dr. Jari Björne for his help in running TEES software. The research was partly funded by the Finnish Cultural Foundation and Academy of Finland (315376).

References

[1] Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martín Pérez-Pérez, Jesus Santamaría, et al. Overview of the BioCreative VI chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146, Bethesda, MD, USA, 2017.

[2] Yifan Peng, Anthony Rios, Ramakanth Kavuluru, and Zhiyong Lu. Extracting chemical-protein relations with ensembles of SVM and deep learning models. *Database*, 2018, 07 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 09 2019.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates Inc., 2013.

[6] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 302–308, 2014.

[7] Jari Björne. *Biomedical Event Extraction with Machine Learning*. PhD thesis, TUCS Dissertations, 2014.

[8] Kai Hakala, Suwisa Kaewphan, Tapio Salakoski, and Filip Ginter. Syntactic analyses and named entity recognition for PubMed and PubMed Central –up-to-the-minute. In *Proceedings of the 2016 Workshop on Biomedical Natural Language Processing*, pages 102–107. Association for Computational Linguistics, 2016.