Automatic Extraction of BEL-Statements based on Neural Networks

Mehdi Ali[1,2], Sumit Madan[2], Asja Fischer[1], Henning Petzka[1] and Juliane Fluck[2]
[1]University of Bonn, 53012 Bonn,
[2]Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, 53754 Sankt Augustin

***Abstract*— The automatic extraction of biomedical relations and entities from text has become extremely important in systems biology. For coding the extracted information, the Biological Expression Language (BEL) can be used. A BEL-statement consists of a subject (entity), a predicate (type of relationship), and an object (entity or a further BEL-statement). This paper describes a system based on neural networks (NNs) to extract BEL-statements in the context of the BioCreAtivE 2017 track 3 (task 1) challenge. In our approach, the overall problem is divided into four subtasks: (i) the detection of named entities (NER), (ii) deciding whether a pair of entities participate in a relation, (iii) determining which of the entities participating in a relation is the subject/object entity, and (iv) extracting the type of the relation. By merging the solutions of the subtasks, the BEL-statements are generated. Except for the named entity recognition, (convolutional) NNs were used to solve the tasks. The results show that a neural net based approach is reasonable to use for the extraction of biomedical relations. The limitations of our system are related to the small size (compared to other NN-based applications) of the data set. We argue that by overcoming this limitation, promising results can be expected from NN-based approaches in future.**

Keywords*— Text Mining; Convolutional Neural Network; Relation Extraction; Biological Expression Language (BEL)

I. INTRODUCTION

To be able to analyze the immense amount of biological data, automated systems are required. In order to perform an automated analysis, the knowledge must be available in a computational representation. Unfortunately, a big amount of the existing knowledge is only available in form of unstructured scientific texts. Therefore, it is of interest to develop systems that extract information from scientific publications and represent the information in a machine interpretable form. The Biological Expression Language (BEL)[1] is a domain-specific language, which makes it possible to capture the extracted knowledge in a structured representation (1). Relationships encoded in BEL are triples consisting of a subject, a predicate, and an object. For example, the BEL statement "p(HGNC:TLR2) increases cat(p(HGNC:CASP1))" can be extracted from the sentence "Interestingly, BLP also activates caspase 1 through TLR2, resulting in proteolysis and secretion of mature IL-1beta" (PMID:10880445) (2).

The BioCreative VI BEL track task 1[2] offers a platform to compare systems for extracting biological knowledge out of free-texts and saving the gathered information as BEL-statements (3). This task is carried out in two stages: i) without information about named entities and ii) with gold standard named entities. The evaluation scheme designed by the authors consider several structural levels of a BEL statement such as term (for e.g. HGNC:CASP1), function (for e.g. catalytic or kinase activity), and relation (for e.g. increases or decreases) level. For the latter two levels, two additional secondary levels are available. Due to time constraints, we did not participate in the function prediction.

In the last years, neural network (NN) based approaches have become popular in Natural Language Processing (NLP) (4–7). Especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are used in text mining. The advantage of neural networks is that they are capable of learning features by themselves (representation learning) by comprising features learned in the different layers of the network in a hierarchical manner. In the low-level layers, basic features are learned and in the higher layers, more complex features are extracted based on the low-level features. This allows to overcome the process of feature engineering or at least to reduce this process (8).

Quan et al. (5) and Hua et al. (9) show that neural networks can successfully be used to extract biomedical relations from scientific publications. Quan et al. use a multichannel CNN approach to extract protein-protein (PPIs) and drug-drug interactions (DDIs). Their system is able to predict associations between pairs of entities without determining which of the entity in an association is the subject or object and without giving details about the type of the relation. For the BioCreative VI BEL task 1, we used this work as a foundation to develop our NN-based system that is capable of extracting BEL-statements out of sentences without involving the complex and time-consuming process of manual feature engineering.

In the following we give a short description of our system in Section II, present the achieved results in Section III, and discuss the results and provide a future outlook in Section IV.

[1] http://openbel.org

[2] http://www.biocreative.org/tasks/biocreative-vi/track-3/

II. System Description

Our NN-based system has a component based architecture in which each component has a specific task. Since the components can easily be exchanged, the system remains flexible. To extract BEL-statements out of the sentences four subtasks must be solved. The first sub-task (and the only one not solved by an NN in our setting) is entity recognition and normalization. The remaining tasks are formulated as binary classification problems (see Fig. 1) and each one is solved by an NN. For each pair of entities, the system has to predict, whether the sentence describes a relation between them. Then, for the case a relation exists, it must determine which of the entities corresponds to the subject and which to the object, respectively. The last step is to extract the type of the relation ("increases" or "decreases"). Based on the predictions for the four sub-tasks, a BEL-Statement is created for a pair of entities, if the existence of a relation is predicted. We use a multi-channel CNN architecture as proposed by Quan et al. (5) in all NN-based sub-tasks. For each task, we train a separate model. In the following, the four single steps and the applied CNN model are explained in more detail.

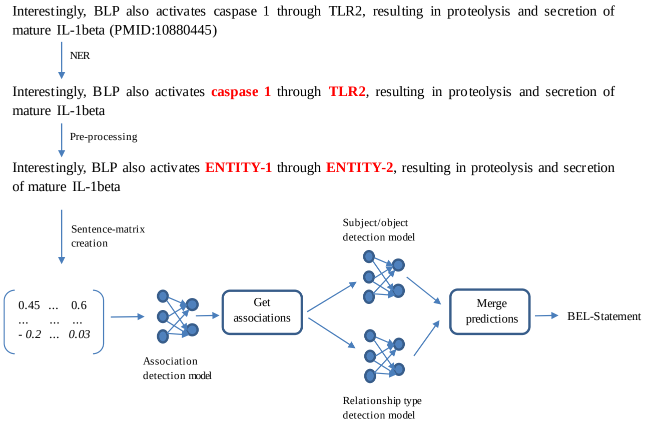
Interestingly, BLP also activates caspase 1 through TLR2, resulting in proteolysis and secretion of mature IL-1beta (PMID:10880445)
NER
Interestingly, BLP also activates caspase 1 through TLR2, resulting in proteolysis and secretion of mature IL-1beta
Pre-processing
Interestingly, BLP also activates ENTITY-1 through ENTITY-2, resulting in proteolysis and secretion of mature IL-1beta
Sentence-matrix creation
0.45 ... 0.6
... ...
- 0.2 ... 0.03
Association detection model
Get associations
Subject/object detection model
Relationship type detection model
Merge predictions
BEL-Statement

Fig. 1. BEL-Statements extraction workflow

A. *Named Entity Recognition (NER)*

The first step is the extraction of named entities from the sentences. For this task, the rule and dictionary based software ProMiner (10) is used. It contains several terminologies to detect named entities. We use ProMiner with the terminologies listed in Table I. On the training set we used ProMiner to find the offsets of the annotated entities. For the test set we consider all detected entities for further predictions.

B. *Association Detection Model*

The second sub-task is to decide, whether a sentence describes an association between two entities or not (independently from any information about subject/object and the relationship type). Since we use a supervised machine learning approach, we need training instances from both classes to train our models. In the training set only positive examples (without any position information of the entities in question within the sentence) are annotated, meaning that only examples of related entities are given. If a sentence does not mention a relation between two occurring entities, this fact is not captured. However, to train a binary classifier, also negative examples are required. Therefore, we create artificial negative examples based on the assumption that, if for a pair of entities no relation is annotated in the training set, the sentence doesn't describe a relation between these entities. This strategy can produce false positives in the sense that a negatively annotated relation might in fact hold because it is not guaranteed that the curator annotated all relations in a sentence (11). The advantage of this approach is that, given a reliable NER tool, the automated creation of negative examples can be realised with very low effort. Our model was trained based on 6,389 instances comprised of 4,633 positive and 1,756 negative examples. To evaluate our model, we applied a 10-fold cross-validation. The results are depicted in Table II. Although the data set is imbalanced the results are quiet promising, especially for the "Association"- class.

TABLE I. Terminologies used for NER

Entity Class	*Resources*	*OpenBEL Namespace*
Human genes/proteins	EntrezGene/Uniprot	HGNC
Mouse genes/proteins	EntrezGene/Uniprot	MGI
Chemicals	ChEBI	CHEBI
Diseases	MeSH disease subtree	MESHD
Biological processes	Gene Ontology biological processes	GOBP

TABLE II. 10-fold cross-validation results for association detection model

Class	*Recall*	*Precision*	*F1-Score*
Association	96.5%	90.7%	93.4 %
No-Association	73.9%	88.9%	80.7%

C. *Subject/Object Detection Model*

The third sub-task is to decide which of the entities in an entity pair is the subject and which the object. To train this model, only positive examples are needed. Since the subject and object information is contained in the training set, a training instance can be directly created as follows. For each relation, the subject and the object are extracted. If the subject occurs before the object in the sentence, the instance is assigned to the class "Subject First" otherwise to the class "Object First". This training set consists of 4,633 examples from which 3,156 instances belong to the class "Subject First" and 1,477 instances to the class "Object First". This data set is imbalanced, too, and the results for the 10-fold cross-validation (see Table III) show that there is still room for improvement.

TABLE III. 10-fold cross-validation results for subject/object detection-model

Class	*Recall*	*Precision*	*F1-Score*
Subject First	88.5%	81.5%	84.9%
Object First	56.9%	70.3%	62.3%

D. *Relationship Type Detection Model*

The last subtask is to determine the type of the relationship of an entity pair participating in an association. As for the subject/object detection subtask no artifical negative instances have to be created. The training set contains the relationship types “increases”, “directly increases”, “decreases” and “directly decreases”, but for the BioCreative VI BEL task 1 “directly increases” is mapped to “increases” and “directly decreases” is mapped to “decreases”, so that there are only two types of relationships to predict. The neural network for this task is trained based on 4,325 instances consisting of 3,103 “increases” and 1,222 “decreases” examples. The results are shown in Table IV.

TABLE IV. 10-fold cross-validation results for relationship type detection model

Class	*Recall*	*Precision*	*F1-Score*
Increases	91.8%	80.8%	85.7%
Decreases	39.5%	66.4%	48.2%

The model lacks in the correct identification of the class “decreases”, especially the recall is low. We argue that this is due to the unbalanced data set. The class “increases” contains 71.7% of the instances while only 28.3% of the instances belong to the class “decreases”. To solve this issue, we experimented with an oversampling strategy by simply copying the “decreases”-examples. However, no significant improvement could be achieved, but the time for training the model rose, so that we decided to train the final model with the initial data set.

E. *Architecture of the Multichannel CNN*

For all three models described in Section II. B-D multichannel CNNs are trained (see Fig. 2). The embedding layer contains the representations of the input sentence. The idea of the multichannel CNN is to use different input channels for different representations of the sentence. Different Word2Vec models (4) are used to transform each word of the sentence into a vector representation. Based on these word-vectors, a sentence-matrix is generated and passed to the network as input. In the sentence-matrix, each column represents a word and the number of rows indicate the dimension of each word-vector. For our networks, we use four Word2Vec models[3] trained by (12), which are based on PubMed, PubMed Central (PMC), and Wikipedia texts, respectively.

In the convolutional layer, local features are computed by applying a convolutional operation on each sentence-representation (see Fig. 2). For each sentence-representation we retrieve a scalar value every time the sliding window of the convolution operation is shifted. The scalar values of each shift are summed up and passed to the activation function. The rows of a new matrix represent the results of the convolutions created by different kernels/filters (in this case 4 filters). At the end of the convolutional layer a max-pooling operation creates a feature vector by extracting the most significant features, taking for each column the biggest value. The feature vector is passed to a fully-connected layer and the result of the fully-connected layer is given as input to a softmax classifier producing the predictions. In the convolutional layer and in the fully-connected layer the Exponential Linear Unit (ELU) is used as the non-linear activation function.

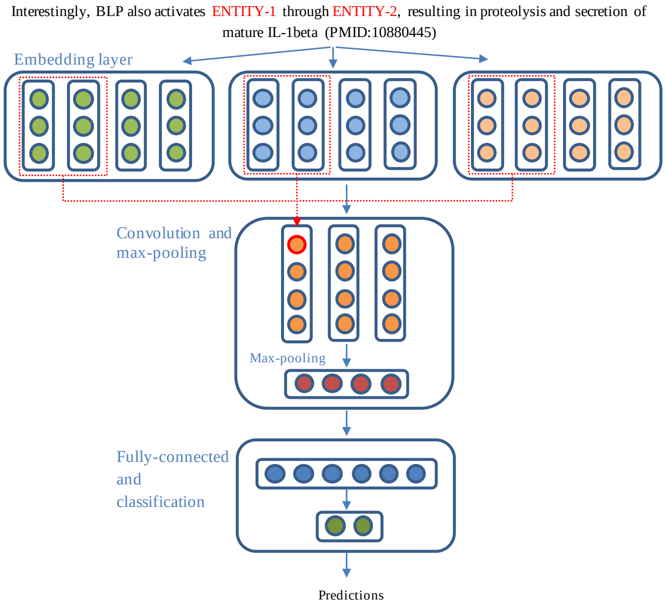
Interestingly, BLP also activates ENTITY-1 through ENTITY-2, resulting in proteolysis and secretion of mature IL-1beta (PMID:10880445)
Embedding layer
Convolution and max-pooling
Max-pooling
Fully-connected and classification
Predictions

Fig. 2. Multichannel convolutional neural network architecture (5,9)

F. *Hyper-Parameters of the Multichannel Convolutional Neural Network*

The selection of appropriate hyper-parameters has a big influence on the performance of the model. We tested different values, but we did not apply any automated hyper-parameter optimization techniques. The selected set of parameters that is used for all three detection models (B-D) is depicted in Table V.

TABLE V. Hyper-Parameters of the trained models

Parameter	*Value*
Learning rate	0.01
Number of filters	200
Filter size	7
Dimension of word-vectors	200
Hidden neurons in the fully connected layer	500
Activation function in convolution layer	Exponential Linear Unit
Activation function in fully conn. layer	Exponential Linear Unit
Loss function	Cross-Entropy
Optimizer	Adagrad
Batch size	20
Training steps	100.000

[3] http://bio.nlplab.org/

III. Results on the Test Set 2017

For each stage of the BioCreative VI BEL task 1 challenge, we handed in three submissions. The difference between Submission 1 and 3 is that the models used in Submission 3 are trained on a reduced training set. For the models of Submission 1, all sentences of the training set with a length up to 75 tokens were considered whereas for the models in Submission 3 the maximum sentence length was chosen to be 65 tokens. We followed this strategy to handle two different distributions of sentence lengths. For Submission 2, we use the models of Submission 1 with the additional rule that whenever two entities predicted to be in a relation have a distance greater than 10 tokens, the prediction is omitted. We argue that an entity which occurs in one part of the sentence is likely not related to an entity which occurs in a distant position of the sentence, especially considering the fact that sentences can contain the same entities multiple times at different positions. The best results are reached with the first submission and its results are presented in Table VI for stage (i) and in Table VII for stage (ii). The most significant observation is the decrease of performance from the Relation-Secondary level to the Relation level, which again directly causes a decrease in the full statement level. Also not predicting any function of entities has a direct impact on the full statement level.

TABLE VI. Results of stage (i) on test set 2017

Evaluation-Level	*Recall*	*Precision*	*F1-Score*
Term	72.13%	81.18%	76.39%
Relation-Secondary	70.74%	60.45%	65.19%
Relation	35.96%	25.55%	29.87%
Full statement	20.61%	16.10%	18.08%

TABLE VII. Results of stage (ii) on test set 2017

Evaluation-Level	*Recall*	*Precision*	*F1-Score*
Term	84.60%	99.23%	91.33%
Relation-Secondary	83.00%	90.05%	86.36%
Relation	45.61%	41.60%	43.51%
Full statement	22.37%	25.00%	23.61%

IV. Conclusion and Future Work

We have presented the results of our participation in the BioCreative VI BEL track task 1. Our BEL extraction workflow is based on a multichannel CNN architecture motivated by Quan et al. (5). As features for our models, we used several pre-trained Word2Vec embeddings that were created on PubMed, PubMed Central, and Wikipedia texts. Our architecture reached for the relation level an F-score of 29.9% in stage (i) and 43.5% in stage (ii). The results indicate that for the extraction of BEL statements from natural language sentences a NN-based approach is reasonable.

The main difficulty when training NNs in this setting is the limited amount of training data. Especially for the relationship type detection model, we think that an increasing amount of training data can lead to a steep improvement of performance. Also the detection of functions of entities, which wasn't tackled in this work, should increase the performance on the full statement level.

Furthermore, it would be interesting to examine how the system performs with new, updated, and fine-tuned Word2Vec models. The models that were used during the training are already four years old. In particular the work of Bojanowski et al. (13) introducing an extended Word2Vec model seems promising for our use case. Besides, a hyper-parameter optimization approach will be integrated in our workflow, so that a better set of hyper-parameters can be detected. A further aspect to consider is the type of the neural network. Different architectures (e.g. recurrent neural nets) should be investigated.

We also plan to perform a detailed analysis of the predictions to detect the problems causing the performance decline. Finally, for use cases where limited training data is available a hybrid system containing machine learning components and rule-based modules should be taken into account.

Acknowledgement

We thank Linara Adilova and Sven Giesselbach for the fruitful discussions and their beneficial inputs, especially in generating and evaluating different Word2Vec models.

References

1. Slater, T. and Song, D. Saved by the BEL: ringing in a common language for the life sciences. 2012.
2. Aliprantis, A. O., Yang, R. B., Weiss, D. S., et al. (2000) The apoptotic signaling pathway activated by Toll-like receptor-2., *EMBO J.*, 19, 3325–36.
3. Rinaldi, F., Ellendorff, T. R., Madan, S., et al. (2016) BioCreative V track 4: a shared task for the extraction of causal network information using the Biological Expression Language, *Database*, 2016.
4. Mikolov, T., Chen, K., Corrado, G., et al. (2013) Distributed representations of words and hrases and their compositionality, *NIPS*, 1–9.
5. Quan, C., Hua, L., Sun, X., et al. (2016) Multichannel Convolutional Neural Network for Biological Relation Extraction, *Biomed Res. Int.*, 2016, 1–10.
6. Sutskever, I., Vinyals, O. and Le, Q. V (2014) Sequence to sequence learning with neural networks, *Adv. Neural Inf. Process. Syst.*, 3104–3112.
7. Zeng, D., Liu, K., Lai, S., et al. (2014) Relation Classification via Convolutional Deep Neural Network, *Coling*, 2335–2344.
8. Goodfellow, Ian, Bengio, Yoshua, Courville, A. Deep Learning http://www.deeplearningbook.org/.
9. Hua, L. and Quan, C. (2016) A Shortest Dependency Path Based Convolutional Neural Network for Protein-Protein Relation Extraction, *Biomed Res. Int.*, 2016, 1–9.
10. Hanisch, D., Fundel, K., Mevissen, H.-T., et al. (2005) ProMiner: rule-based protein and gene entity recognition., *BMC Bioinformatics*, 6 Suppl 1, S14.
11. Fluck, J., Madan, S., Ansari, S., et al. (2016) Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (BEL)., *Database (Oxford).*, 2016.
12. Pyysalo, S., Ginter, F., Moen, H., et al. (2013) Distributional semantics resources for biomedical text processing, *Proc. Lang. Biol. Med.*
13. Bojanowski, P., Grave, E., Joulin, A., et al. (2016) Enriching word vectors with subword information, *arXiv Prepr. arXiv1607.04606*.