Extraction of BEL-Statements Based on Neural Networks

TEAM: MEHDI ALI[1,2], SUMIT MADAN[2], DR. ASJA FISCHER[1], DR. HENNING PETZKA[1]

[1]UNIVERSITY OF BONN, 53012 BONN
2FRAUNHOFER INSTITUTE FOR ALGORITHMS AND SCIENTIFIC COMPUTING, SCHLOSS BIRLINGHOVEN, 53754 SANKT AUGUSTIN

19.10.17

Contents

Task Description - BioCreative VI Track 3 (Task 1)

PMID: 10880445

Interestingly, BLP also activates caspase 1 through TLR2, resulting in proteolysis and secretion of mature IL-1beta

Entity

Entity

Entity

Entity

p(HGNC:TLR2)   increases   cat(p(HGNC:CASP1))

Subject

Relationship Type

Object

Motivation

Traditional machine learning approach:

- Extract lexical and semantic features: POS, word stems, number of tokens between entities, dependency parsing etc.
- Use a classifier (e.g. SVM) to classify the instances
- Parameter optimization of the classifier

Complex feature engineering necessary!

Neural Networks: Overcome the time-consuming process of complex feature engineering:

- In NLP interesting results based on convolutional neural networks and recurrent neural networks

Fraunhofer SCAI

universitätbonn

Workflow for BEL-Statements Extraction

Interestingly, BLP also activates caspase 1 through TLR2, resulting in proteolysis and secretion of mature IL-1beta

↓ 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

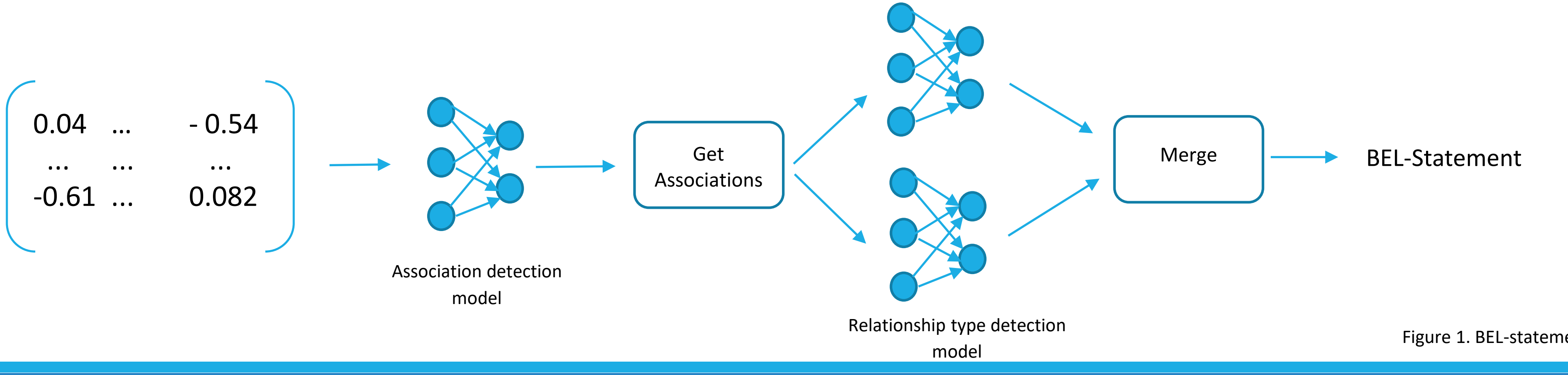

Figure 1. BEL-statements extraction workflow

Multichannel CNN Architecture

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

Fully-connected and classification

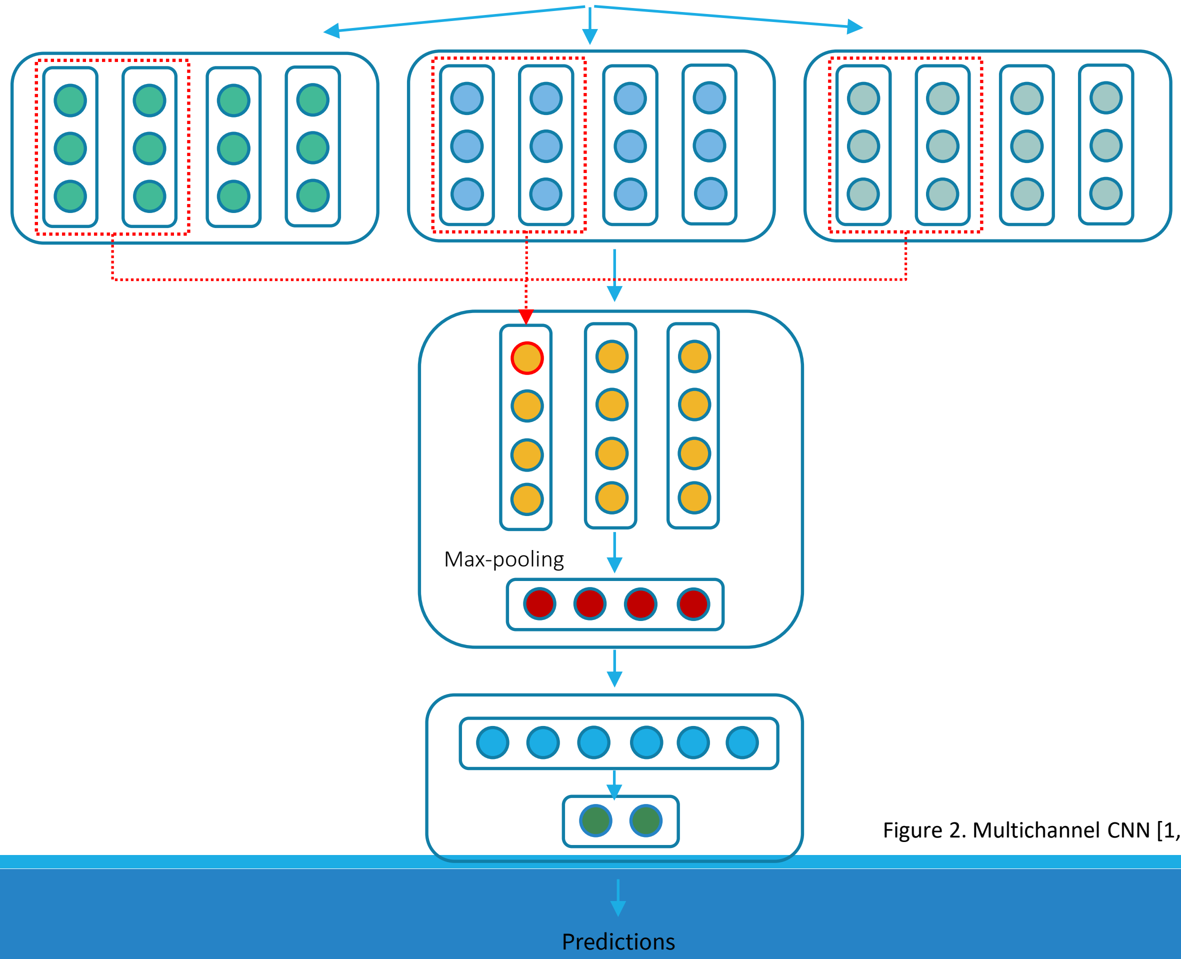

Figure 2. Multichannel CNN [1,2]

Datasets – Creation of Instances

Association detection model:

- 4633 “association” and 1756 “no association” instances
- No negative examples annotated → Create artificial negative instances

Subject/object detection model:

- 3156 “subject first“ and 1477 “object first” instances

Relationship type detection model:

- 3103 “increases” and 1222 “decreases” instances

Fraunhofer SCAI

universitätbonn

Results for BioCreative 2017 – Track 3 (Task 1)

Table 1: Results on test set 2017 without gold standard entities (stage 1)

Class	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 %
Statement	20.61 %	16.1 %	18.08 %

Table 2: Results on test set 2017 with gold standard entities (stage 2)

Class	Recall	Precision	F1-Score
Term	84.6 %	99.23 %	91.33 %
Relation-Secondary	83 %	90.05 %	86.36 %
Relation	45,61 %	41.6 %	43.51 %
Statement	22.37 %	25 %	23.61 %

Prediction of Functions

Table 3: Predictions of functions without gold standard entities

Class	Recall	Precision	F1-Score
Function-Secondary	37.33 %	62.22 %	46.67 %
Function	28.42 %	40.91 %	33.54 %

Summary and Outlook

- Information extraction system to extract BEL-statements
- Usage of multichannel CNN-based architecture [1]
- For NER a dictionary and rule-based system called ProMiner [3] is used
- Relation extraction task is divided into three subtasks
- Results indicate that a NN-based approach is reasonable

- Create further models to predict BEL functions
- Evaluate new, updated and fine-tuned Word2Vec models
- Use more data from other tasks (such as BioNLP, and also BioCreative)
- Investigate different neural network architectures (e.g. recurrent neural networks)

Fraunhofer SCAI

universitätbonn

Literature

1. Quan, C., Hua, L., Sun, X., et al. (2016) Multichannel Convolutional Neural Network for Biological Relation Extraction, Biomed Res. Int., 2016, 1–10

2. Hua, L. and Quan, C. (2016) A Shortest Dependency Path Based Convolutional Neural Network for Protein-Protein Relation Extraction, Biomed Res. Int., 2016

3. Fluck, J., Mevissen, H.-T., Dach, H., et al. (2007) ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries, *Proc. Second BioCreative Chall. Eval. Work.*, 149–151.