Transfer Learning for Biomedical Relation Extraction

Flash presentations

Mario Sänger (WBI, HU Berlin)

saengema@informatik.hu-berlin.de

Outline

- Flash presentations
  - Group 3: CNNs with pre-trained biomedical word embeddings (Hanjo, Danielle, Dennis)
  - Group 1: (Bio-) BERT for Relation Extraction (Phuc, Duy)

- Evaluation
  - Training data
  - Evaluation data
  - Result submission

Evaluation

Training data

- You have two distinct training data sets for the competition
  - Both corpora contain protein-protein interactions
  - Usage of a uniform, consistent XML-format

	AIMed	BioInfer
#Documents	180	669
#Sentences	1554	894
#Entities	3407	3611
Distinct	933	952
#Pairs	4680	8043
Positive	800	2109
Negative	3880	5934

Evaluation setting

- Your approaches will be evaluated on hold-out sets (~ test data) of the two corpora
  - The (XML-) format is equivalent to the training data set
  - Of course, the test sets do not contain the gold standard labels
    - => No *interaction* attribute!

```
<pair e1="AIMed.d0.s7.e1" e2="AIMed.d0.s7.e3" id="AIMed.d0.s7.p4" />
<pair e1="AIMed.d0.s7.e2" e2="AIMed.d0.s7.e3" id="AIMed.d0.s7.p5" />
```

- Test data will be released at start of January

Evaluation setting

- Your approaches will be evaluated on hold-out sets (~ test data) of the two corpora

	AIMed	BioInfer
#Documents	44	167
#Sentences	389	286
#Entities	795	810
Distinct	288	298
#Pairs	1095	1623
Positive	?	?
Negative	?	?

Result submission

- You have to submit the results (~ prediction of your models) via an CSV-file
  - One file per corpus and one prediction per line
  - 1. Column: pair identifier (e.g. *AIMed.d0.s7.p1*)
  - 2. Column: prediction of your model (True / False)
  - Use tab stop ("\t") as separator

```
AIMed.d0.s7.p0      True
AIMed.d0.s7.p1      False
```

- You are allowed (and encouraged) to submit multiple results originating from different variants of your model

Thank you for your attention!