Transfer Learning for Biomedical Relation Extraction

Block seminar

Mario Sänger (WBI, HU Berlin)

saengema@informatik.hu-berlin.de

Outline

- Introduction
  - Evaluation setting
  - Results (Preview)
- Solution presentations
  - Group 3: CNNs with pre-trained biomedical word embeddings (Hanjo, Danielle, Dennis)
  - Group 1: (Bio-) BERT for Relation Extraction (Phuc, Duy)
- Feedback

Introduction

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" />
```

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	191	425
Negative	904	1198
	17% Positive	26% Positive

Results (preview)

- AiMed

Team	Model	P	R	F1
Group 3 (CNN)	Without-Embs	0.250	0.330	0.284
Group 1 (BERT)	Alibaba-SciBert	**0.711**	0.759	0.734
	Lee-BioBert	0.672	**0.859**	**0.754**
	Lin-Bert	0.674	0.780	0.723

- BioInfer

Team	Model	P	R	F1
Group 3 (CNN)	Without-Embs	0.365	0.191	0.250
Group 1 (BERT)	Alibaba-SciBert	**0.831**	0.671	**0.742**
	Lee-SciBert	**0.831**	0.649	0.729
	Lin-BioBert	0.815	**0.675**	0.739

Comparison with competitors

- AiMed

Author	Model	P	R	F1
	Lee-BioBert	0.672	0.859	0.754
Quan et al. [1]	Multi-Channel CNN	0.764	0.690	0.725
Hsieh et al. [2]	Bi-LSTM	0.788	0.752	0.769
Yadav et al. [3]	Bi-LSTM on DT + Words	0.911	0.822	**0.865**

- BioInfer

Author	Model	P	R	F1
	Alibaba-SciBert	0.831	0.671	0.742
Quan et al. [1]	Multi-Channel CNN	0.813	0.781	0.796
Hsieh et al. [2]	Bi-LSTM	0.870	0.874	**0.872**
Yadav et al. [3]	Bi-LSTM on DT + Words	0.724	0.831	0.774

* Figures taken from original published results

Feedback

Thank you for your attention!

References

[1] Quan, Chanqin, et al. "*Multichannel convolutional neural network for biological relation extraction." BioMed research international* 2016 (2016).

[2] Hsieh, Yu-Lun, et al. "Identifying protein-protein interactions in biomedical literature using recurrent neural networks with long short-term memory." *Proceedings of the eighth international joint conference on natural language processing (volume 2: short papers)*. 2017.

[3] Yadav, Shweta, et al. "Feature Assisted bi-directional LSTM Model for Protein-Protein Interaction Identification from Biomedical Texts." *arXiv preprint arXiv:1807.02162* (2018).