Mining for Medical Relations in Research Articles:

Training Models

Hannes Berntsson

Purpose

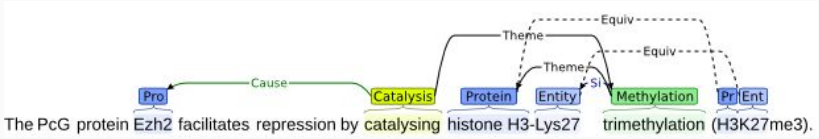
- Process and tag millions of medical abstracts and texts quickly.
- Save biomedical scientists decades of work.

Goals

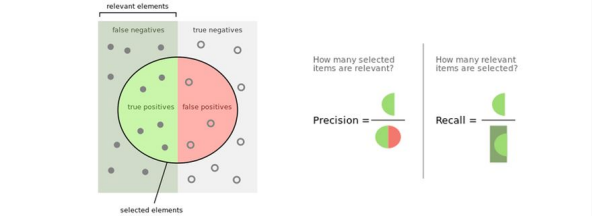
- Create a baseline model for relations extraction.
- Proof of concept with issues and future solutions.

Overview

1. Training Data
2. Similar Projects
3. Models and Results
4. Future Iterations



UNIVERSITY OF TURKU



Training Data

Different Approaches

- Gold Standard
  - Excellent
  - Very costly
- Silver Standard
  - Might work great
  - Complicated
- No Labeled Data
  - Distant Supervision [1]

[1] Mintz, et al. (2009). Distant supervision for relation extraction without labeled data. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pp.1003-1011.

Training Data

Data Used

- BioInfer [1]
  - Gold standard
  - Binarized version
  - What I used for 95% of the project
  - ~2500 examples
- Data From Project
  - Silver standard
  - ~5500 examples
- TAC 2018, Drug-Drug Interaction [2]
  - Gold standard
  - Initially used
  - Ultimately not relevant

[1] Pyysalo, S. et al. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1).
[2] https://bionlp.nlm.nih.gov/tac2018druginteractions/

Training Data

Example

BioInfer:

alpha-catenin inhibits beta-catenin signaling by preventing formation of a beta-catenin*T-cell factor*DNA complex -> NEG
[no_interaction, POS, NEG]

Project:

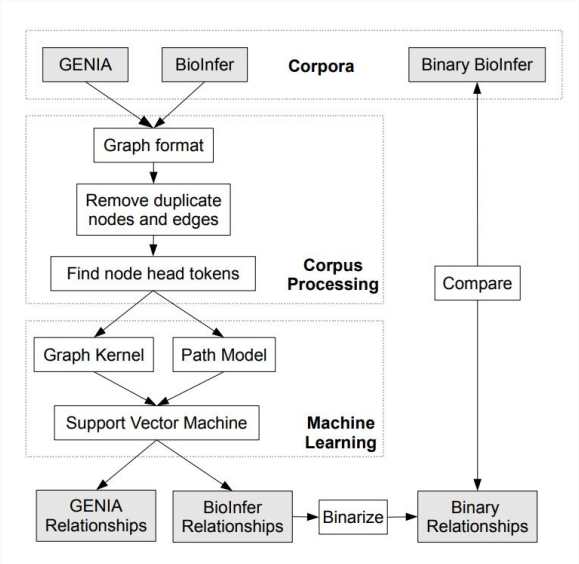
Phentolamine, an alpha blocker, completely blocked the NE-stimulated VO2 ... -> N
[no_interaction, P, N]

Similar Projects

- Multiple projects on NLP relation extraction
- Several for medical/biomedical texts. [1, 2]

Here's a similar project using the BioInfer Corpus:

Learning to Extract Biological Event and Relation Graphs [1]



			untyped undirected				typed directed		
corpus	parse	features	P	R	F	AUC	P	R	F
BioInfer	GS	PM	84.4	82.1	83.1±2.3	89.4±1.8	78.7	76.7	77.7±2.6
		GK	74.9	70.6	72.6±2.6	82.6±2.2	72.6	56.8	63.6±2.5
	CL	PM	76.6	67.3	71.5±4.6	81.4±2.6	73.5	61.9	67.0±3.7
		GK	66.8	61.4	63.8±2.4	77.3±1.5	64.2	47.1	54.1±4.1
GENIA	GS	PM	75.5	63.1	68.7±1.5	80.5±1.2	70.2	60.9	65.2±2.4
	CL	PM	72.3	57.4	63.8±2.8	77.6±2.1	65.6	55.5	60.1±3.0

[1] Björne, J. and Ginter, F. (2019). Learning to Extract Biological Event and Relation Graphs. *NODALIDA 2009 Conference Proceedings*, pp.18 - 25.

[2] Rinaldi, F. and Andronis, C. et al., (2004). Mining relations in the GENIA corpus. *In Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics*, held in conjunction with ECML/PKDD in Pisa, Italy. 24 September 2004.

SVM with NLP Tags using sciSpacy[1]

alpha-catenin inhibits beta-catenin signaling by preventing formation of a beta-catenin*T-cell factor*DNA complex.

Tokens, PoS and dependency tags surrounding the two entities:

Tokens:
{None, None, inhibits, beta-catenin, signaling}
{signaling, preventing, formation, None, None}

POS:
{None, None, VBZ, NP... }

Same for dependency tags.

Results on BioInfer:

F-Score: 57.3

[1] https://allenai.github.io/scispacy/

Entity Replacement Bigram/Trigrams in Dense Keras-net

ENTITY1 inhibits beta-catenin signaling by preventing formation of a ENTITY2.

5000 most common bigrams/trigrams (Bag of Words):

“ENTITY1 inhibits”

“to reduce ENTITY2”

“blocks ENTITY2”

“prevents ENTITY2 production”

“ENTITY2 was inhibited”

“inhibited by ENTITY1”

… etc.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 100)               500100
_________________________________________________________________
dense_2 (Dense)              (None, 100)               10100
_________________________________________________________________
dense_3 (Dense)              (None, 3)                 303
=================================================================
Total params: 510,503
Trainable params: 510,503
Non-trainable params: 0
_________________________________________________________________


Train on 4712 samples, validate on 832 samples
Epoch 1/100, Batch size 10
```

Entity Replacement Bigram/Trigrams in Dense Keras-net

Results on BioInfer:

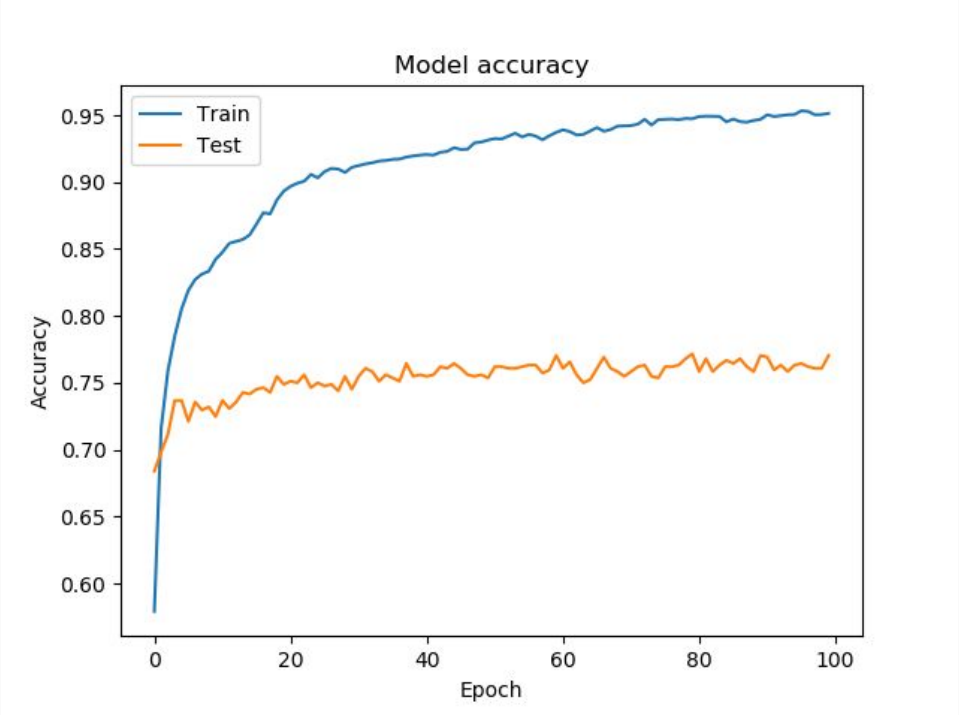
Accuracy: 77.0%
Loss: 85.3 (categorical cross-entropy)

Recall: 69.3
Precision: 72.7
F-Score: 70.8

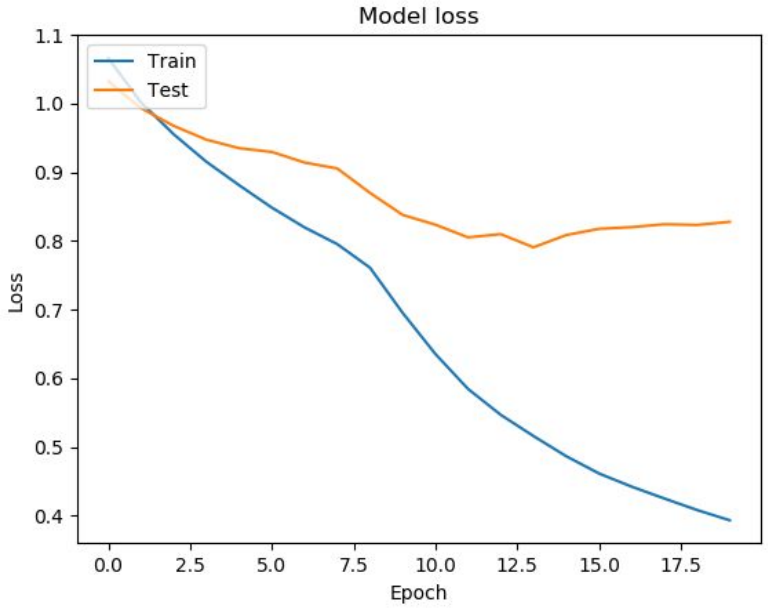
Results on Project Data:

Accuracy: 67.7%
Loss: 82.8 (categorical cross-entropy)

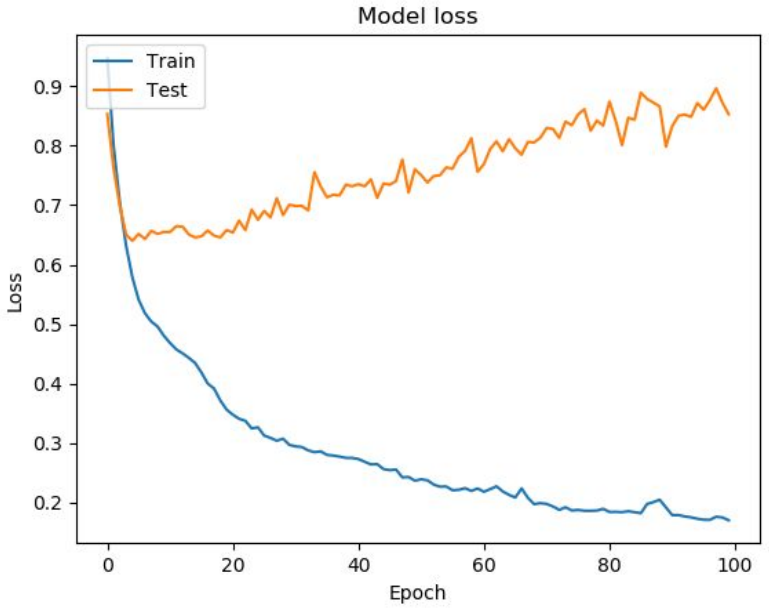
Recall: 63.8
Precision: 64.7
F-Score: 64.1



Model accuracy on the BioInfer corpus



Model loss on the project data



Model loss on the BioInfer corpus (overtrained)

Model Loss on the BioInfer and Project Data

Future Iterations

Improvements and Plans

- Dependency Path, LSTM, Embeddings (very nearly done)
- Run predictions on PubMed corpus
- Pair with an entity tagger model
- Tag the whole relation (more like a NER task)

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, None)	0	
embedding_1 (Embedding)	(None, None, 200)	853800	input_1[0][0]
input_2 (InputLayer)	(None, None, 2)	0	
concatenate_1 (Concatenate)	(None, None, 202)	0	embedding_1[0][0] input_2[0][0]
bidirectional_1 (Bidirectional)	(None, 400)	644800	concatenate_1[0][0]
dense_1 (Dense)	(None, 64)	25664	bidirectional_1[0][0]
batch_normalization_1 (BatchNor	(None, 64)	256	dense_1[0][0]
dropout_1 (Dropout)	(None, 64)	0	batch_normalization_1[0][0]
dense_2 (Dense)	(None, 3)	195	dropout_1[0][0]

Thanks!

Hannes Berntsson

dat15hbe@student.lu.se

LUND UNIVERSITY