U.S. National Library of Medicine

Chemical-protein relation extraction with ensembles of SVM, CNN, and RNN models

Yifan Peng[1], Anthony Rios[1,2], Ramakanth Kavuluru[2,3], Zhiyong Lu[1]

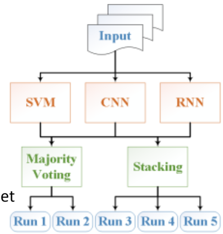
[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health

[2]Department of Computer Science, University of Kentucky

[3]Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky

Outline

- Individual models
  - SVM
  - CNN
  - RNN
- Ensembles of three models
  - Majority voting
  - Stacking
- Experiments
  - 5-fold cross validation on training + dev set
  - Results on test set



Chemical-protein relations

- A multiclass classification problem
- The chemical-protein relations occurring in a single sentence

SVM with rich feature vector

SVM

- Linear kernel
- One-vs-rest scheme

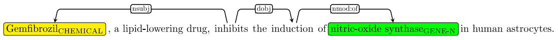
Miwa, M.; Sætre, R.; Miyao, Y. & Tsujii, J. A rich feature vector for protein-protein interaction extraction from multiple corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, **2009**, *1*, 121-130

Rich Feature Vector

- Words/Part-of-speech tags surrounding the chemical and gene mentions
- Bag-of-words between the chemical and gene mentions
- Distance between two entity mentions
- Shortest path in a dependency graph

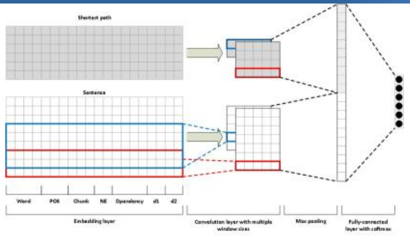
Shortest path in a dependency graph

- Obtained using Bllip parser + Stanford dependencies converter


nsubj
dobj
nmod:of
GemfibrozilCHEMICAL, a lipid-lowering drug, inhibits the induction of nitric-oxide synthaseGENE-N in human astrocytes.

- Vertex walks
  - CHEMICAL – *nsubj* – inhibits
  - inhibits – *dobj* – induction
  - induction – *nmod:of* – GENE
- Edge walks
  - *nsubj* – inhibits – *dobj*
  - *dobj* – induction – *nmod:of*

Convolutional Neural Network

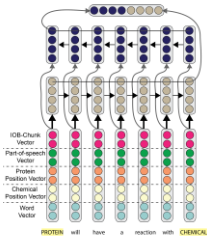

Shortest path
Sentence
Word
POS
Chunk
NE
Dependency
d1
d2
Embedding layer
Convolution layer with multiple window sizes
Max pooling
Fully-connected layer with softmax

Peng, Y. & Lu, Z. Deep learning for extracting protein-protein interactions from biomedical literature. *Proceedings of BioNLP 2017*, **2017**, 29-38

Convolutional Neural Network

- Word embedding: 300
  - trained on PubMed using word2vec
- Part-of-speech, chunk and named entities: one-hot encoding
  - Obtained using Genia Tagger
- Convolutional window size: 3 and 5
- Filters: 300

Recurrent Neural Network



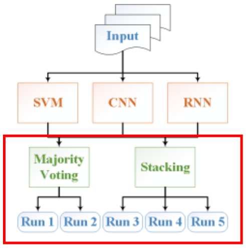
Kavuluru, R.; Rios, A. & Tran, T. Extracting Drug-Drug Interactions with Word and Character-Level Recurrent Neural Networks. *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, **2017**, 5-12

Recurrent Neural Network

- Pairwise ranking loss
  - The output layer has 5 positive classes
  - If all 5 class scores are negative, then we predict the negative class
- Preprocessing
  - Replace word occurs less than 5 times with an UNK token
- Word embedding: 300
  - Obtained from GloVe

Santos, C. N.; Xiang, B. & Zhou, B. Classifying Relations by Ranking with Convolutional Neural Networks. *ACL*, **2015**, 626-634

Ensembles of SVM, CNN, and RNN models



Ensembles of SVM, CNN, and RNN models

Majority voting

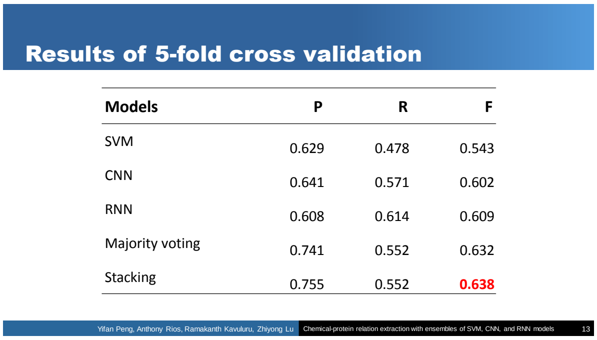
- Select the relations that are predicted by more than 2 models

Stacking

- Random Forest classifier
- 17 features:
  - 6 from SVM
  - 6 from CNN
  - 5 from RNN (pariwise ranking loss)

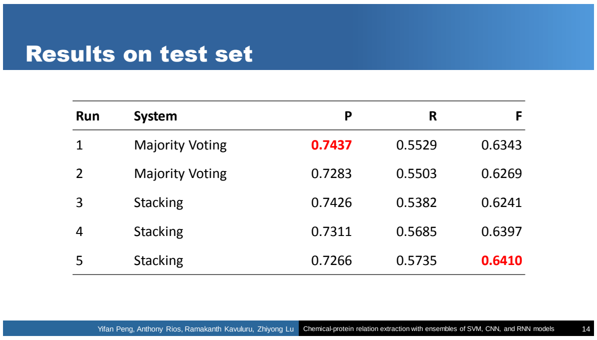
Results for 5-fold cross validation

- Combine training and development sets
- 5-fold cross validation
  - 60% for training
  - 20% for development (also used to train the stacking systems)
  - 20% for test



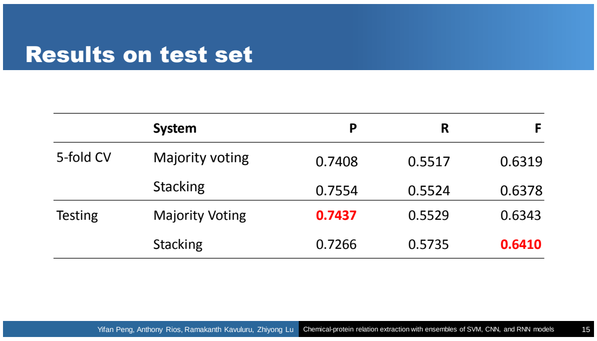
Results of 5-fold cross validation

Models	P	R	F
SVM	0.629	0.478	0.543
CNN	0.641	0.571	0.602
RNN	0.608	0.614	0.609
Majority voting	0.741	0.552	0.632
Stacking	0.755	0.552	0.638



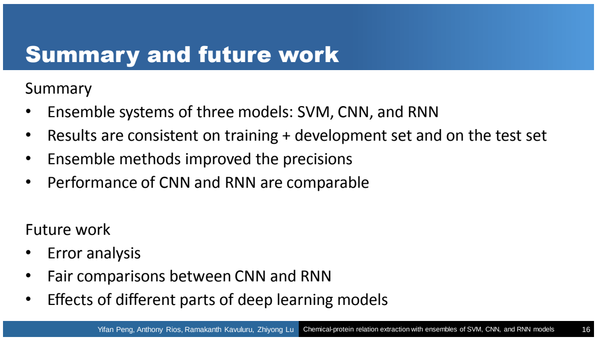
Results on test set

Run	System	P	R	F
1	Majority Voting	0.7437	0.5529	0.6343
2	Majority Voting	0.7283	0.5503	0.6269
3	Stacking	0.7426	0.5382	0.6241
4	Stacking	0.7311	0.5685	0.6397
5	Stacking	0.7266	0.5735	0.6410



Results on test set

	System	P	R	F
5-fold CV	Majority voting	0.7408	0.5517	0.6319
	Stacking	0.7554	0.5524	0.6378
Testing	Majority Voting	0.7437	0.5529	0.6343
	Stacking	0.7266	0.5735	0.6410



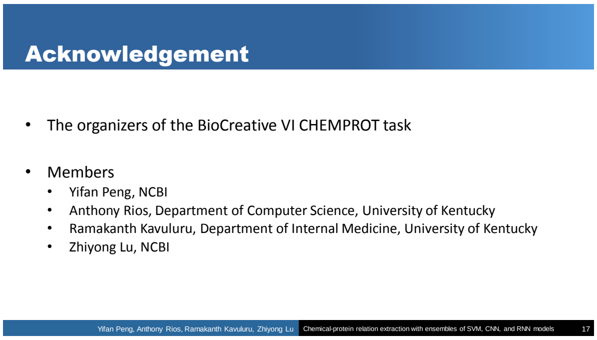
Summary and future work

Summary

- Ensemble systems of three models: SVM, CNN, and RNN
- Results are consistent on training + development set and on the test set
- Ensemble methods improved the precisions
- Performance of CNN and RNN are comparable

Future work

- Error analysis
- Fair comparisons between CNN and RNN
- Effects of different parts of deep learning models



Acknowledgement

- The organizers of the BioCreative VI CHEMPROT task
- Members
  - Yifan Peng, NCBI
  - Anthony Rios, Department of Computer Science, University of Kentucky
  - Ramakanth Kavuluru, Department of Internal Medicine, University of Kentucky
  - Zhiyong Lu, NCBI

Thank You!

yifan.peng@nih.gov