Gene Name Identification / Mentioning at BioCreative Challenge 2

Presenter: Sai Qian
University of Saarland
saiqian@coli.uni-sb.de

Overview

- Motivation & Introduction
- *Rie Kubota Ando's* system
- *Kou et al.'s* system
- *Huang et al.'s* system
- Inspiration based on *Kou & Huang's* system
- General combination of BioCreative 2
- Conclusion

Motivation & Intruction

- The largest and most reliable source of biomedical knowledge: scientific literature
  - Protein-protein interactions
  - Disease-gene associations
- Initial steps
  - Tagging gene
  - Gene product mentions
- The second BioCreative challenge (BioCreative 2)
  - Gene mention task (GM)
  - Gene normalization task (GN)
  - Protein-protein interaction task (PPI)

Motivation & Intruction

- Gene mention task similar to named entity recognition task for *person names* and *company names*
- Significant difference
  - Quantity: millions of gene names used
  - Creativity: new gene names are created continuously
  - Random: authors do not use standardized names
  - Co-occurrence: co-occur with other types (cell names)
  - Indefinity: expert readers disagree on the result
  - Ambiguity: a sequence of DNA referred by a gene name may vary in nonspecific ways (polymorphism, multiple alleles)

Rie Kubota Ando's System

IBM T.J. Watson Research Center

- A semi-supervised learning method (ASO)
- Automatic induction of high-order features
- Gene name lexicon lookup
- Classifier combination
- Simple post-processing

Rie Kubota Ando's System

IBM T.J. Watson Research Center

- Alternating Structure Optimization (ASO)
  - A multi-task learning algorithm
  - Simultaneously learning multiple tasks that related to each other
- Application of ASO
  - Automatic generation of thousands of prediction problems (auxiliary problems)
  - Their (problems) labeled data info from unlabeled data
- Learning new (and better) feature representation from unlabeled data
- *"A framework for learning predictive structures from multiple tasks and unlabeled data"*

Rie Kubota Ando's System

IBM T.J. Watson Research Center

- Unlabeled data
  - Total number of 500 million words – resource intensive
  - A randomly generated subset – performance marginal
  - Hope: benefit from the unlabeled data with reasonable computational time
- The setting
  - Go through every sentence, count word frequency
  - Choose a sentence if – it contains a word occurring at least *k* times
  - Discard a sentence – otherwise

Rie Kubota Ando's System

IBM T.J. Watson Research Center

- High-order features
  - Combing two or more base features
  - E.g. "current-word='gene' & next-word='*'"
  - Generating all combination: training expensive
- Bi-gram feature
  - Construct bi-gram feature only from misclassified data with pure base feature
  - Retain the positive bi-gram feature, discard the negative ones
  - The best result with the least computational time

Rie Kubota Ando's System

IBM T.J. Watson Research Center

- Performance improvement with combination of several classifiers
  - Classifiers with similar performance but make different mistakes
- Left-to-right & right-to-left chunker
  - Taking a union of the two sets of annotations (BioCreative 1)
  - Remove any annotation that overlaps with another by keeping the longer ones (BioCreative 2)
  - "AAA", "AAABB"

Rie Kubota Ando's System

IBM T.J. Watson Research Center

- Domain lexicon
  - A domain lexicon generated from LocusLink, Swiss-Prot , Mesh
  - A list of names with tags that indicate the information source (e.g. "MESH")
- Simple post-processing
  - Remove annotations that include any unmatched parenthesis
  - e.g. ***)"

Rie Kubota Ando's System

IBM T.J. Watson Research Center

- Result

	Post-processing	Feature induction	Name lexicons	Classifier combination	Unlabeled data	P	R	F	
Baseline	–	–	–	–	–	89.13	79.39	83.98	–
Post-processing	X	–	–	–	–	89.40	79.39	84.10	(+0.12)
Feature induction	–	X	–	–	–	89.11	79.86	84.23	(+0.25)
Name lexicon	–	–	X	–	–	88.89	80.48	84.47	(+0.49)
Classifier combination	–	–	–	X	–	85.14	84.90	85.02	(+1.04)
Unlabeled data	–	–	–	–	X	91.17	81.52	86.07	(+2.09)
Run#3	X	X	X	–	X	**91.54**	81.99	86.50	(+2.52)
Run#1	X	X	–	X	X	88.37	85.94	87.14	(+3.16)
Run#2	X	X	X	X	X	88.48	**85.97**	**87.21**	(+3.23)

Rie Kubota Ando's System

IBM T.J. Watson Research Center

- Period conclusion
  - Semi-supervised learning based on ASO algorithm
  - Equipped with classifier combination, automatic generation of high-order features, domain lexicon, and simple post-processing
  - Useful for a huge amount of unlabeled data

Kou et al.'s System

Taipei, Taiwan

- Conditional Random Fields (CRFs)
  - *A type of discriminative probabilistic model most often used for the labeling or parsing of sequential data. – Wikipedia*
  - Dominant performance of tagging gene and mentioning protein in BioCreative 1
- Rich feature set
  - 5,059,368 predicates as the features
  - Feature defined based on hundreds of trails
  - E.g. exclude prefix and suffix predicates in previous tagger

Kou et al.'s System

Taipei, Taiwan

- Example of features

Feature	*Example*
Word	proteins
StemmedWord	protein
PartOfSpeech	NN
InitCap	Kinase

- Combination of several taggers
  - Forward tagger – right to left
  - Backward tagger – left to right
  - Backward better than Forward?
  - Union (recall ↑) vs. Intersection (precision ↑)

Kou et al.'s System

Taipei, Taiwan

- Adjacent Ten Union
  - A nearly perfect recall (0.9810) with union of the adjacent 10 tagging solutions
- Procedure
  - Parse sentence in both directions, select the adjacent 10 solutions for each direction
  - Compute the intersection of bidirectional parsing, discard the one which minimizes the sum of the output scores
  - For the rest 18, select the labeled terms appearing in a dictionary with its length > 3

Kou et al.'s System

Taipei, Taiwan

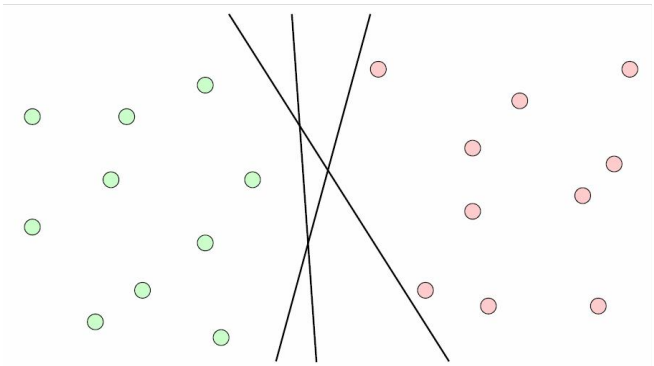
- Result

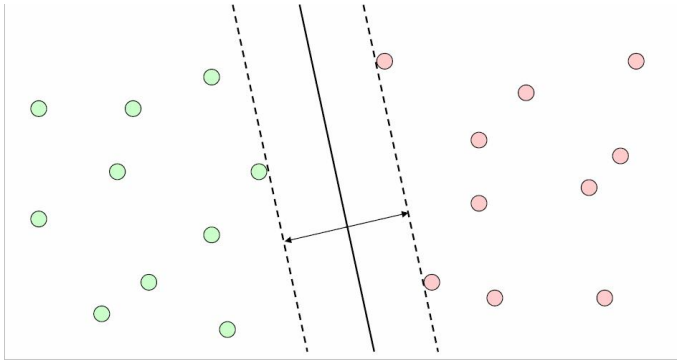
System	Precision	Recall	F-Measure
Forward	0.8660	0.8077	0.8359
Backward	0.8733	0.8118	0.8414
Union	0.8349	0.8578	0.8462
Intersection	0.9076	0.7186	0.8021
Adjacent Ten Union + Dictionary	0.8773	0.8263	0.8510

Huang et al.'s System

Taipei, Taiwan

- Support Vector Machines (SVM)
  - Find a decision surface (hyperplane) in the vector space that separates the document vectors of two categories
- The "best": maximum-margin hyperplane
  - Equal distance to both document sets
  - Margin between hyperplane and document sets is maximal





Huang et al.'s System

Taipei, Taiwan

- SVM – binary classifier
  - One vs. all: Train a binary classifier for each class against all other classes
  - One vs. one: Train a binary classifier for each pair of classes and select the class appearing in the most output
- CRF also trained
- Backward better than Forward
  - More important "signal" at the end of the entities

Huang et al.'s System

Taipei, Taiwan

- Result so far

Table 3: Performance comparison for different models and parsing directions

Model	Forward	Backward
SVM+One vs.All	P:82.81% R:78.27% F:80.48%	P:86.99% R:75.79 %F:81.01%
SVM+One vs.One	P:82.41% R:78.11% F:80.20%	P:85.49% R:79.25 %F:82.25%
CRF	P:86.52% R:79.44% F:82.83%	P:86.77% R:80.39% F:83.46%

P,R and F denote precision, recall, and f-score, respectively.

- Integration
  - Union – including more tagging results from different models
  - Intersection – filtering out false positives

Huang et al.'s System

Taipei, Taiwan

- Final scheme and result
  - A mixture of intersection & union

Run	Ensemble	Performance		
1	M1∪M3	P:83.27(3)	R:89.34(1)	F:86.20(1)
2	M2∪M3	P:82.98(3)	R:89.58(1)	F:86.15(1)
3	(M1∩M2)∪M3	P:84.93(3)	R:88.28(1)	F:86.57(1)

Inspiration based on *Kou & Huang's* system

Taipei, Taiwan

- Feature Selection
  - Difference when implementing in MALLET & CRF++
  - Removing a subset of features, observing the result (Prefix & Suffix features; orthographic features)
  - Selection of best features depends on the CRF package
- Testing Backward and Forward parsing in CRF++
  - No distinct difference in F-Score like in MALLET
  - Backward parsing is not always superior
  - Bidirectional parsing – wider variety of complementary models

Inspiration based on *Kou & Huang's* system

Taipei, Taiwan

- Post processing
  - Problems caused by unpaired parenthesis
- Example
  - *… implicated the NIMA (never in mitosis, gene A)-related kinase-6 (NEK6)……*
  - "*gene A)-related kinase-6*"
- Procedure
  - Find the left parenthesis
  - stop word (*the*) or parenthesis at the left side of the left parenthesis
  - Extend the original tagging
  - "*the NIMA (never in mitosis, gene A)-related kinase-6*"

Inspiration based on *Kou & Huang's* system

Taipei, Taiwan

- Prominent feature of *Kou et al.'s* & *Huang et al.'s* system
  - Combining divergent but high performance models always improve the performance
- Model Integration
  - Intersection of forward & backward parsing by MALLET with L-BFGS algorithm
  - Forward parsing by CRF++ with L-BFGS algorithm
  - Forward parsing by CRF++ with CTJPGIS algorithm
  - Forward parsing by SVM model

Inspiration based on *Kou & Huang's* system

Taipei, Taiwan

- Result

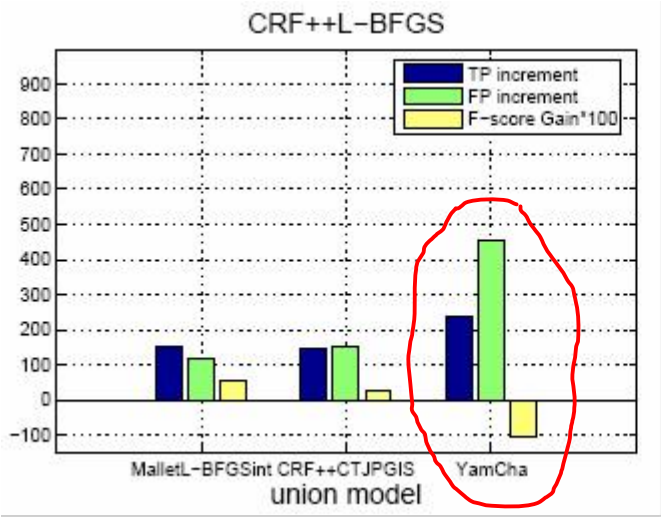
Model		MalletL-BFGSint	CRF++L-BFGS	CRF++CTJPGIS	YamCha
MalletL-BFGSint	Precision	92.11	**88.67**	88.65	84.98
	Recall	75.69	**86.68**	86.40	87.03
	F-score	83.10	**87.67**	87.51	85.99
CRF++L-BFGS	Precision		90.15	88.21	84.16
	Recall		84.28	86.59	88.01
	F-score		87.12	87.39	86.05
CRF++CTJPGIS	Precision			90.60	84.39
	Recall			82.96	87.73
	F-score			86.61	86.03
YamCha	Precision				86.96
	Recall				80.70
	F-score				83.71

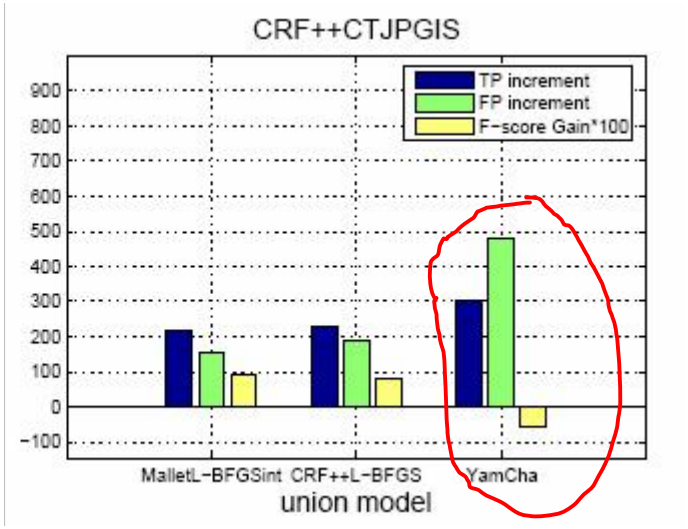
Inspiration based on *Kou & Huang's* system

Taipei, Taiwan

	truly YES	truly NO
system YES	true positives	false positives
system NO	false negatives	true negatives

- What affects F-Score?
  - True Positive ↑ & False Positive ↓
  - F-Score ↑





General Combination of BioCreative 2

National Center for Biotechnology Information, Maryland

- Review of BioCreative 2
  - Total 21 participants
  - F-Score (87.2 – 48.2)
- Materials in BioCreative 2
  - MEDLINE: both training and testing
  - Sentences likely to contain gene name = sentences not likely to contain gene name
- How would the systems work in other situation?
- Artificial sets of sentences from 2 databases
  - Random MEDLINE: F-Score lower
  - Random Trans. Factors: F-Score higher

General Combination of BioCreative 2

National Center for Biotechnology Information, Maryland

- Improvement on the best score?
- With the help of all submitted systems
  - Machine learning to predict gene mentions
  - Holding out 25 sentences, training on 4975 sentences, fusion of the 25 results
  - Boosted Decision Tree & Conditional Random Field
- Highest F-Score – 90.66 (87.2)
  - Future systems should be able to achieve improved performance
  - Refining the corpus & Improving systems design through collaboration

Conclusion

- Semi-supervised system – Rie's
  - 5 components
  - Useful for unlabeled data
- *Kou et al.'s* & *Huang et al.'s*
  - CRF & SVM with bidirectional parsing
- Inspired by *Kou* & *Huang*
  - Feature selection; backward vs. forward; post-processing
- Combination of all 21 systems
  - Machine learning method, highest F-Score

References

- Rie Kubota Ando *BioCreative Ⅱ Gene Mention Tagging System at IBM Watson*
- Roman K. Christoph M.F. Juliane F. Martin H.A *Named Entity Recognition with Combinations of Conditional Random Fields*
- John Wilbur, Larry Smith, Lorrie Tanabe *BioCreative 2. Gene Mention task*
- Yu-Ming Chang, Cheng-Ju Kou, Han-shen Huang, Yu-Shi Lin, Chun-Nan Hsu *Analysis and Enhancement of Conditional Random Fields Gene Mention Taggers in BioCreative Ⅱ Challenge Evaluation*
- Cheng-Ju Kou, Yu-Ming Chang, Han-Shen Huang, Kuan-Ting Lin, Bo-Hou Yang, Yu-Shi Lin, Chun-Nan Hsu, I-Fang Chung *Rich Feature Set, Unification of Bidirectional Parsing and Dictionary Filtering for High F-Score Gene Mention Tagging*
- Han-Shen Huang, Yu-Shi Lin, Kuan-Ting Lin, Cheng-Ju Kou, Yu-Ming Chang, Bo-Hou Yang, I-Fang Chung, Chun-Nan Hsu *High-Recall Gene Mention Recognition by Unification of Multiple Backward Parsing Models*

The End

Thanks for your attention!