

User 8: Summary of task 1A & 1B

Jeremiah Crim, Ryan McDonald and Fernando Pereira

University of Pennsylvania: Biomedical Data Mining Project

Work on statistical IE part of a larger project at University of Pennsylvania

Annotating large set of data on drug development & pediatric oncology

Entities: chemicals, genes, variation events, malignancies

Relations: I.e. gene variation causing malignancy

Treebanking: Adding syntactic structure over entity structure

Propbanking: Adding semantic role information

Have prototype disks available containing:

Data: entity tagged and POS tagged

Tools: annotating tool and POS tagger

Info: Info on the project

Talk with me or Jay if you want a copy

Overview

- Task 1A
 - Conditional Random Fields
 - Feature set comparison
- Task 1B
 - Challenges (use of 1A?)
 - Pattern Matching
 - Maximum Entropy Classification

Task 1A: Gene Identification

Conditional Random Fields

- Conditional probability model:

$$P(\mathbf{t}|\mathbf{o}) = \frac{\exp\{ \sum_i \sum_j \lambda_j f_j(t_{i-1}, t_i, \mathbf{o}) \}}{Z(\mathbf{o})}$$

- \mathbf{o} is the sequence of input tokens
- \mathbf{t} is a sequence of labels (gene, not-gene)
- Train to maximize log-likelihood of training
- Globally normalized (unlike max-ent)
- $f(t_{i-1}, t_i, \mathbf{o})$: maps predicates on input to $\{0, 1\}$
- Want to find features that model problem

Task 1A: Features

- Standard orthographic features
 - [A-Z], [A-Za-z], [A-Z].*, [,\.\.::()?!], etc.
- Word, Prefix, Suffix & CharNGram
- Gene & Gene context lists
- Non-Gene Lists (enzymes, common bio words, organisms, etc.)
- Infrequent trigrams (Tannabe & Wilbur 01)
- Window $\{-1,0,1\}$
- Conjunction features built with feature induction (McCallum `03)

red = part of open system (i.e. uses outside lexicons)

Results

Evaluation data

System	Precision	Recall	F-meas
Closed	0.830	0.773	0.801
Open	0.863	0.787	0.823

Development data

System	Precision	Recall	F-meas
Closed & w/o F.I.	0.793	0.731	0.761
Closed	0.807	0.744	0.774
Trigrams	0.811	0.759	0.784
Non-gene Lex	0.818	0.743	0.778
Gene Lexicons	0.812	0.775	0.793
Open - full	0.817	0.782	0.799

Task 1B:

Initial Investigations

- Idea: tag genes (1A) and match to synonym list
 - Tagger is not that useful on data
 - Mouse OK
 - Fly and Yeast terrible
 - Cannot simply “match” to synonym list
 - Synonym list is not complete
 - Synonyms are often uninformative
 - Sharing of synonyms: *ubx* (3), *amylase* (120), *alcohol dehydrogenase* (111), *amy* (74)
 - 7000 (mostly common) synonyms occur for more than one gene for fly (~10%).

System 1 – Pattern Matching

- Algorithm
- For each gene, g
 - For each synonym, s , in g 's syn list
 - If s is matched in document
 - Add g to documents normalized list

... is required for transvection at the **bithorax** complex
...
zeste may play a role in the normal regulation of **Ubx** and its other target genes.

+

...
FBgn0003944: CG10388 Cbx,
DmUbx, Hm,
Ubx, abx,
bithorax
...

=

FBgn000284
FBgn0003944
FBgn0012923
...

System 1 – Pattern Matching

Fly

System	Precision	Recall	F-meas
Simple	0.033	0.861	0.063

- Problem – matches everything!
 - Synonyms that are common words
 - Synonyms that occur for more than one gene
- Use training data to increase accuracy

System 1 – Useful Synonyms

- A synonym, s , for a gene, g , is useful iff for training set D :

$$\frac{\sum_{d \in D} \text{match}(s,d) \times \text{labels}(g,d)}{\sum_{d \in D} \text{match}(s,d)} > \delta$$

$$\text{match}(s,d) = \begin{cases} 1 & \text{if } s \text{ is matched in } d \\ 0 & \text{otherwise} \end{cases}$$

$$\text{labels}(g,d) = \begin{cases} 1 & \text{if } g \text{ is in } d\text{'s gene list} \\ 0 & \text{otherwise} \end{cases}$$

- Meant to represent the conditional probability of g labeling a document given a match on synonym s
- Tuning on development returned $\delta = 0.4$

System 1 – Candidate List

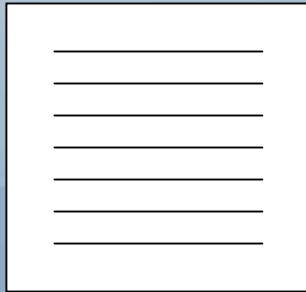
- For the fly organism we must consider 36,000 different genes (each with 3+ synonyms)
- Reduce the number of genes under consideration
- Only consider synonyms for genes in a documents “candidate list”

System 1 – Candidate List

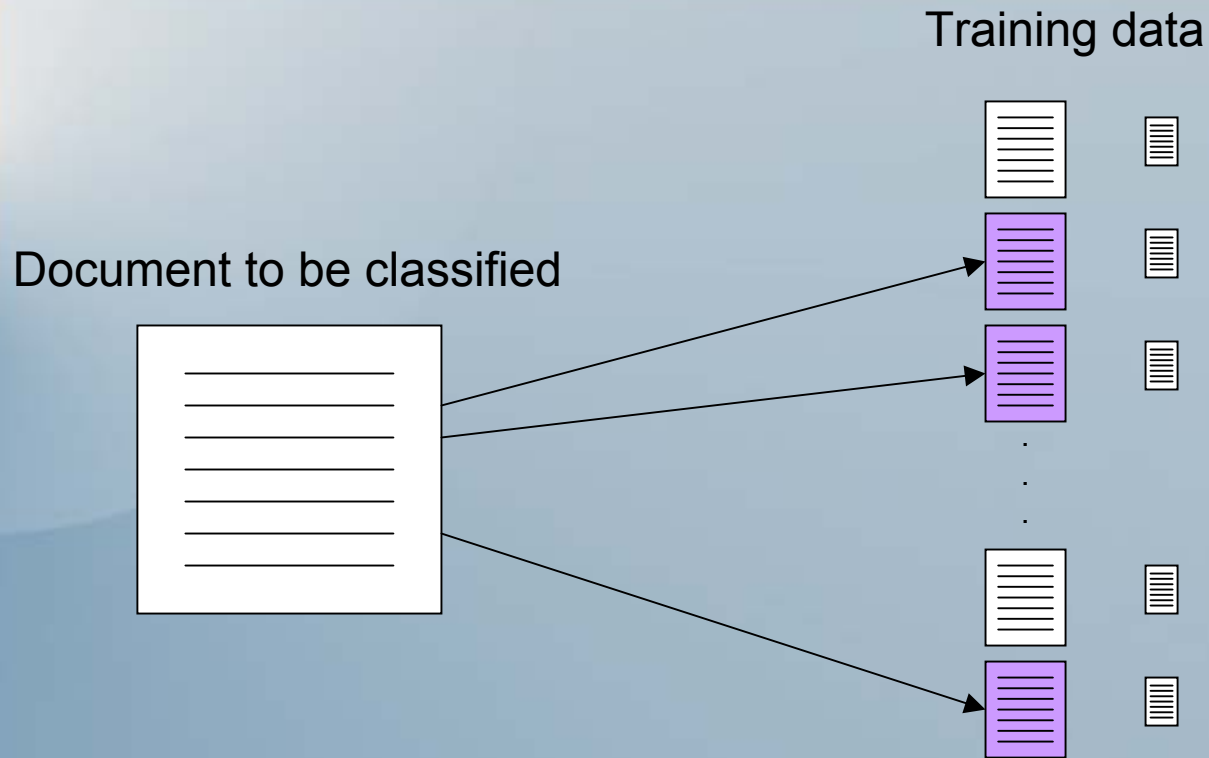
Training data



Document to be classified

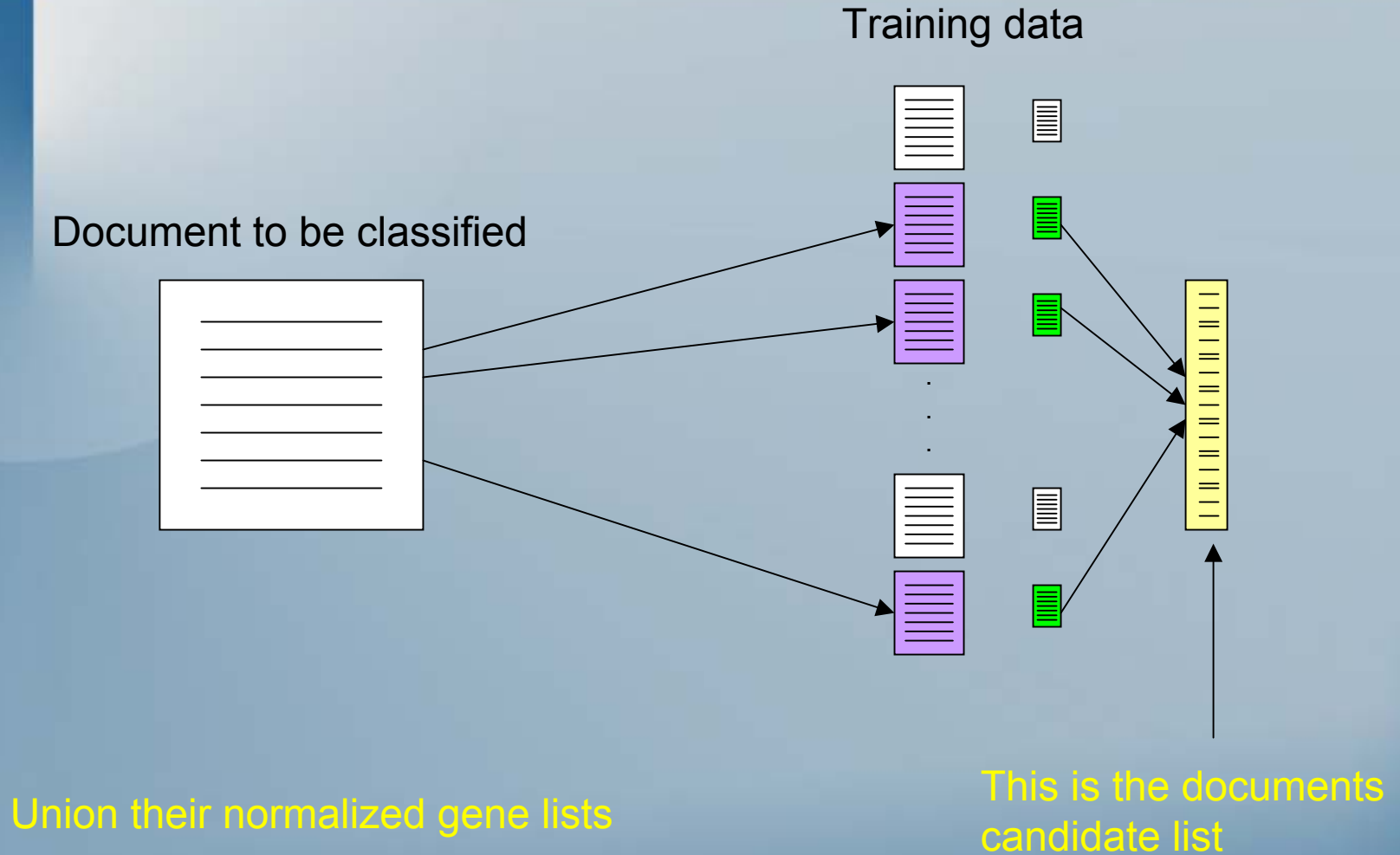


System 1 – Candidate List



Find k-nearest-neighbours (using cosine distance)

System 1 – Candidate List



System 1 - stemming

- Stem synonym list & document before matching
- We used the Porter Stemmer ('80)

Fly

System	Precision	Recall	F-meas
Simple	0.033	0.861	0.063
Simple+SL	0.458	0.727	0.562
Simple+SL+CL	0.709	0.667	0.687
Simple+SL+CL+Stem	0.713	0.690	0.701

Shortcomings of Pattern Matching

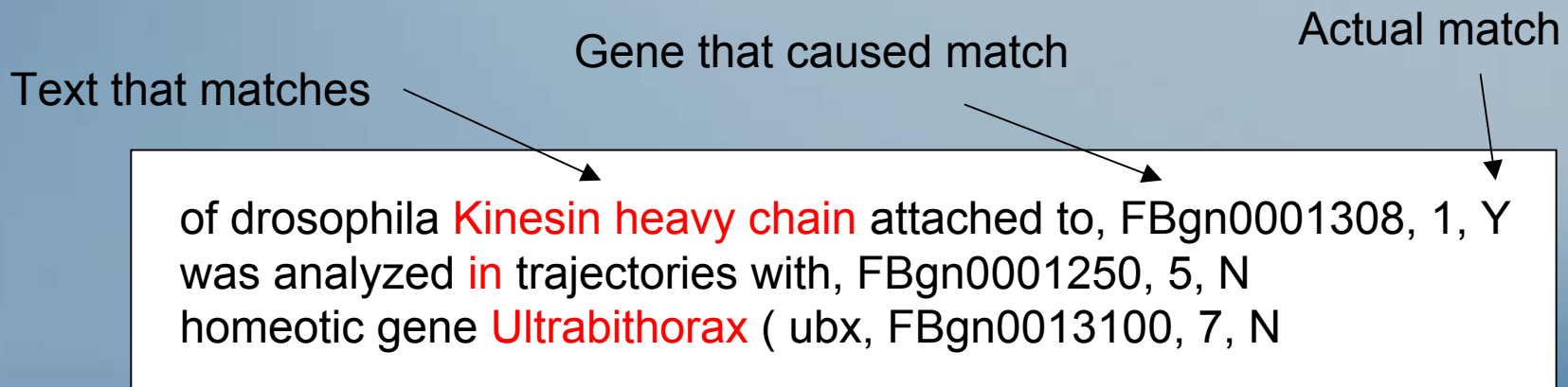
- Variables tuned on devel data
 - δ , k-neighbours (cause over-fitting)
- Complex: data transformed often
 - Stemming, useful syns, candidates
- Different systems for each organisms
 - Candidate list different for all organisms
 - Yeast does not use stemming
 - Method will not generalize to other organisms!!

Observation

- Matching can have up to 0.92% recall
- Need to identify good matches vs. bad matches to increase precision
- Can train a classifier to do just this
- Collect from the training data
 - Good matches: matches where gene in normalized gene list
 - Bad matches: matches where gene is not in normalized gene list

Creating the training data

- For each synonym match collect:
 - The text that matched
 - 2 words before and after the match
 - The norm of the gene causing match
 - # of other genes matching this text



System 2 - MaxEnt Classifier

- Trained a MaxEnt classifier to recognize good matches vs. bad
- Features:
 - matched text
 - previous words & next words
 - gene causing match
 - # genes matching text

Final Results - testing

Fly

System	Precision	Recall	F-meas
Pattern Matching	0.64	0.7	0.67
Max Ent	0.70	0.78	0.74

Mouse

System	Precision	Recall	F-meas
Pattern Matching	0.83	0.67	0.74
Max Ent	0.79	0.73	0.76

Yeast

System	Precision	Recall	F-meas
Pattern Matching	0.95	0.89	0.92
Max Ent	0.96	0.88	0.92

Maximum Entropy Advantages

- One system for all organisms
- All parameters trained on training data
- Extensible
 - Can easily add more expert knowledge to maximum entropy models
 - Right now features are all contextual
- Does equally well or better than pattern matching

Thanks

Acknowledgements:

Seth Kulick, Mark Liberman, Mark Mandel, Andy Schein, Scott Winters, Pete White

University of Pennsylvania: Biomedical Data Mining Project

Work on statistical IE part of a larger project at University of Pennsylvania

Annotating large set of data on drug development & pediatric oncology

Entities: chemicals, genes, variation events, malignancies

Relations: I.e. gene variation causing malignancy

Treebanking: Adding syntactic structure over entity structure

Propbanking: Adding semantic role information

Have prototype disks available containing:

Data: entity tagged and POS tagged

Tools: annotating tool and POS tagger

Info: Info on the project

Talk with me or Jay if you want a copy