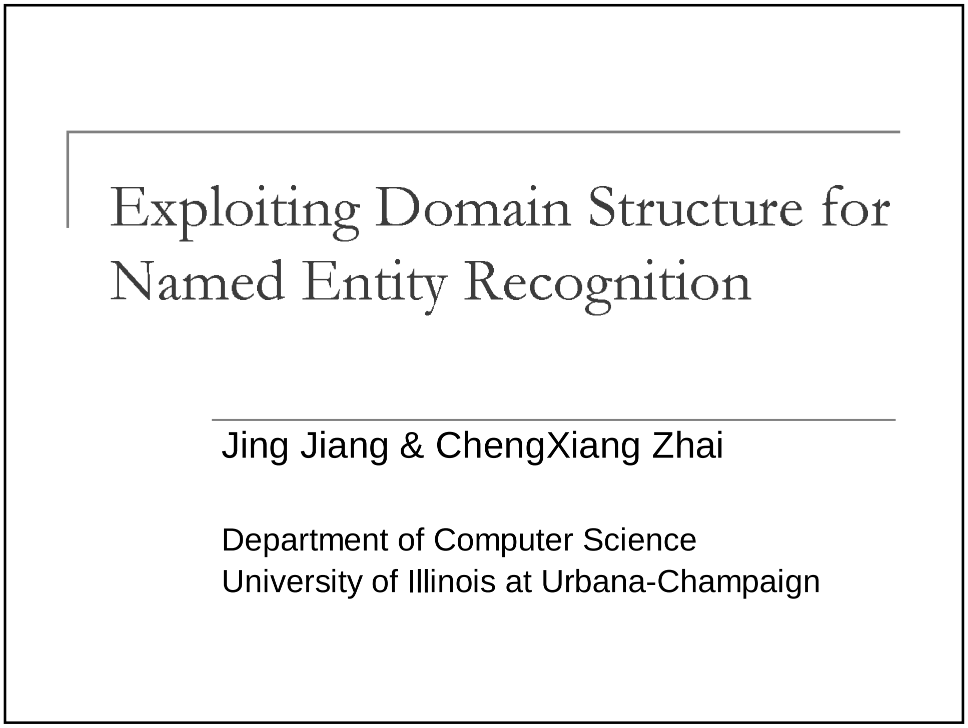
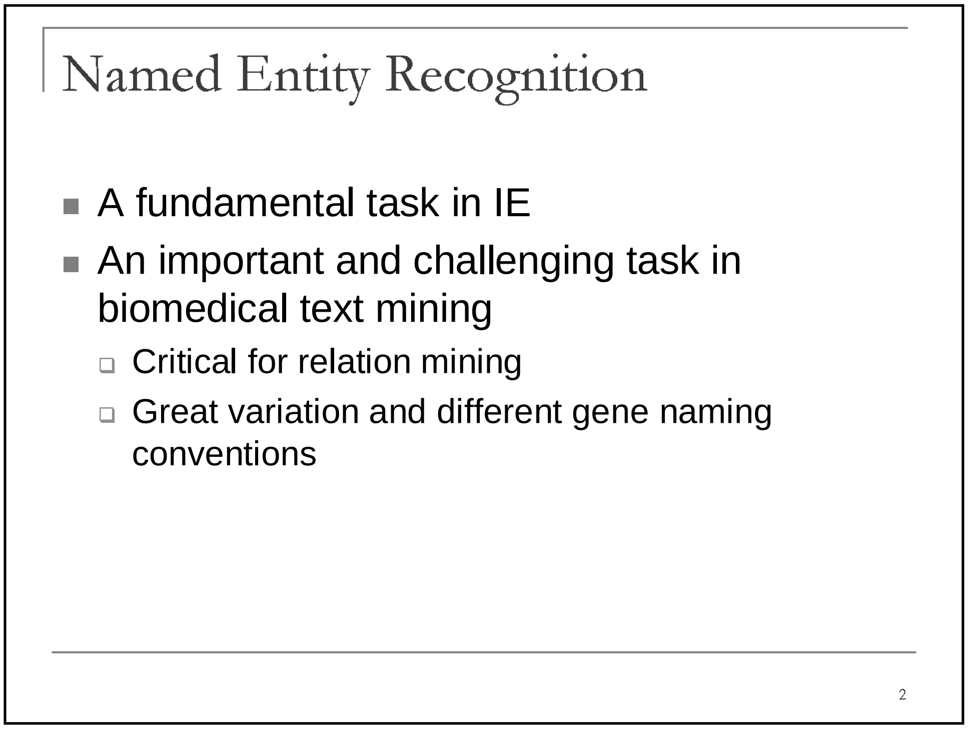


Exploiting Domain Structure for Named Entity Recognition

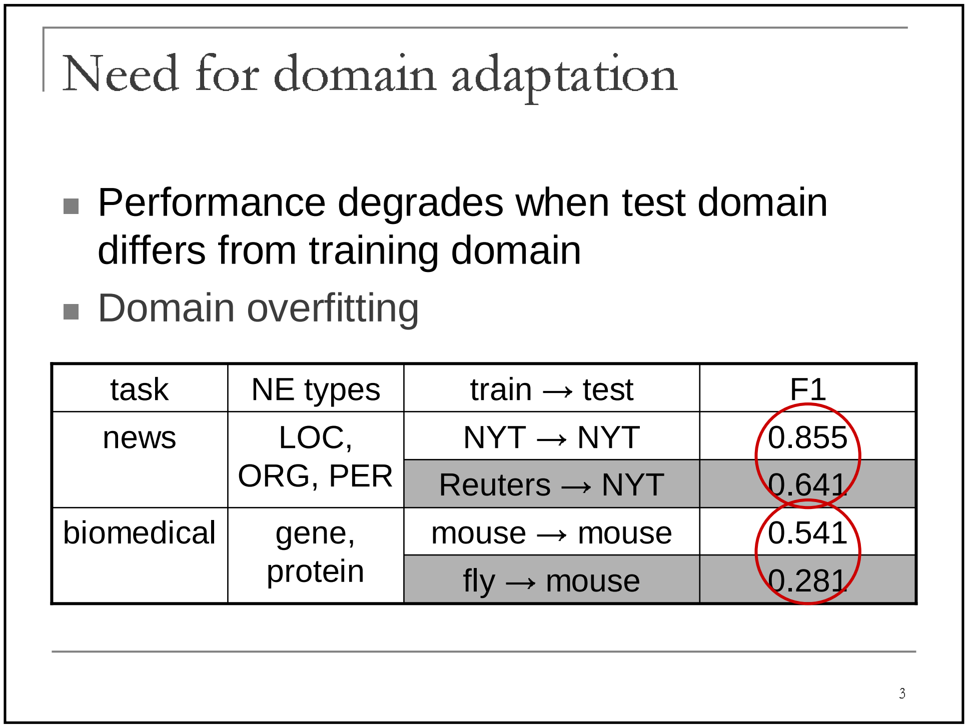
Jing Jiang & ChengXiang Zhai

Department of Computer Science
University of Illinois at Urbana-Champaign



Named Entity Recognition

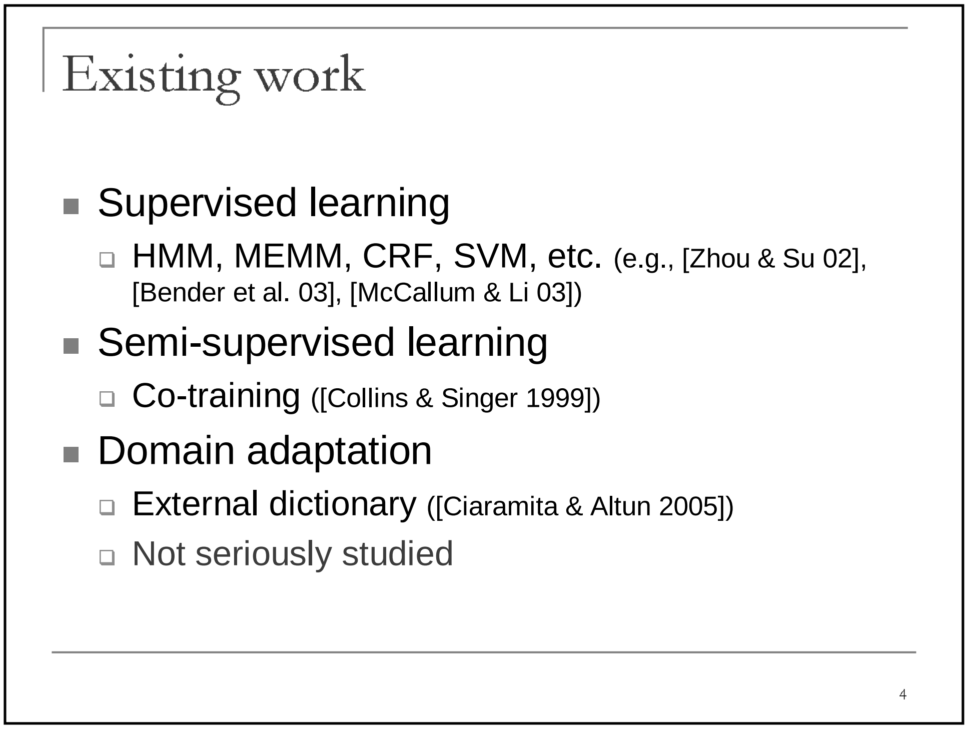
- A fundamental task in IE
- An important and challenging task in biomedical text mining
  - Critical for relation mining
  - Great variation and different gene naming conventions



Need for domain adaptation

- Performance degrades when test domain differs from training domain
- Domain overfitting

task	NE types	train → test	F1
news	LOC, ORG, PER	NYT → NYT	0.855
		Reuters → NYT	0.641
biomedical	gene, protein	mouse → mouse	0.541
		fly → mouse	0.281



Existing work

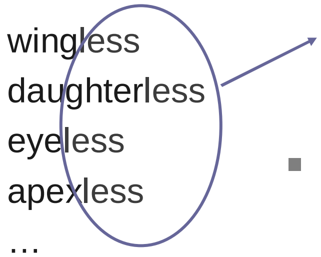
- Supervised learning
  - HMM, MEMM, CRF, SVM, etc. (e.g., [Zhou & Su 02], [Bender et al. 03], [McCallum & Li 03])
- Semi-supervised learning
  - Co-training ([Collins & Singer 1999])
- Domain adaptation
  - External dictionary ([Ciaramita & Altun 2005])
  - Not seriously studied

Outline

- Observations
- Method
  - Generalizability-based feature ranking
  - Rank-based prior
- Experiments
- Conclusions and future work

Observation I

- Overemphasis on domain-specific features in the trained model



wingless
daughterless
eyeless
apexless
…

fly

“suffix –less” weighted high in the model trained from fly data

- Useful for other organisms?

  in general NO!

- May cause generalizable features to be downweighted

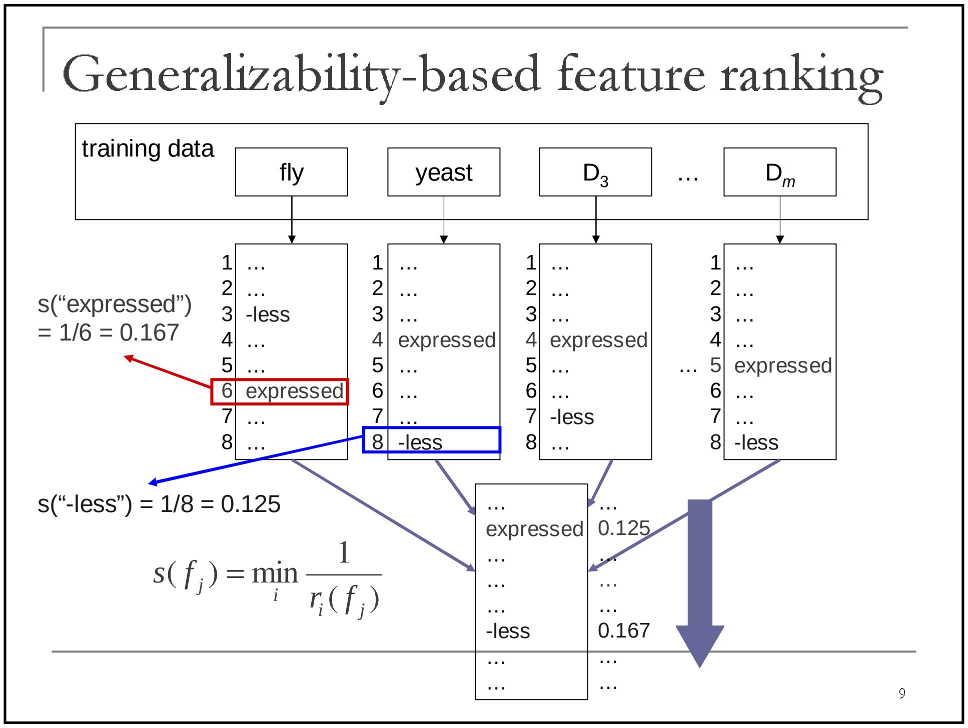
Observation II

- Generalizable features: generalize well in all domains
  - …**decapentaplegic** and **wingless** are expressed in analogous patterns in each primordium of… (fly)
  - …that **CD38** is expressed by both neurons and glial cells…that **PABPC5** is expressed in fetal brain and in a range of adult tissues. (mouse)

Observation II

- Generalizable features: generalize well in all domains
  - …**decapentaplegic** and **wingless** are expressed in analogous patterns in each primordium of… (fly)
  - …that **CD38** is expressed by both neurons and glial cells…that **PABPC5** is expressed in fetal brain and in a range of adult tissues. (mouse)

“w_{i+2} = expressed” is generalizable



Generalizability-based feature ranking

training data

	fly	yeast	D$_3$	…	D$_m$
1	…	…	…		…
2	…	…	…		…
3	-less	…	…		…
4	…	expressed	expressed		…
5	…	…	…	…	expressed
6	expressed	…	…		…
7	…	…	-less		…
8	…	-less	…		-less

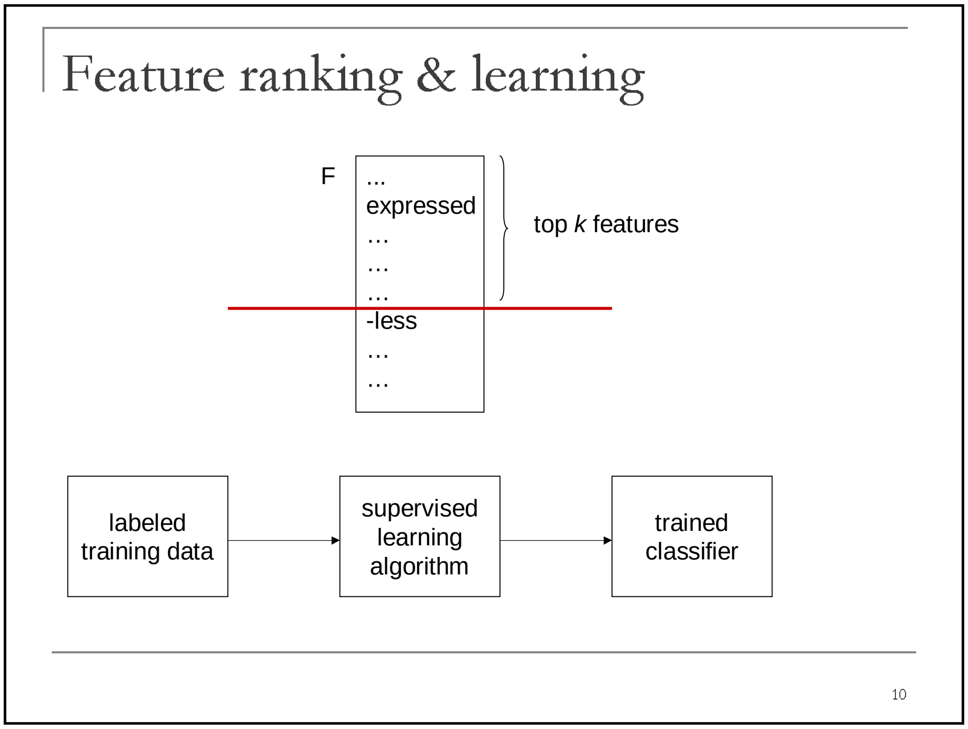
s("expressed") = 1/6 = 0.167

s("-less") = 1/8 = 0.125

$$s(f_j) = \min_i \frac{1}{r_i(f_j)}$$

…	…
expressed	0.125
…	…
…	…
…	…
-less	0.167
…	…
…	…

9



Feature ranking & learning

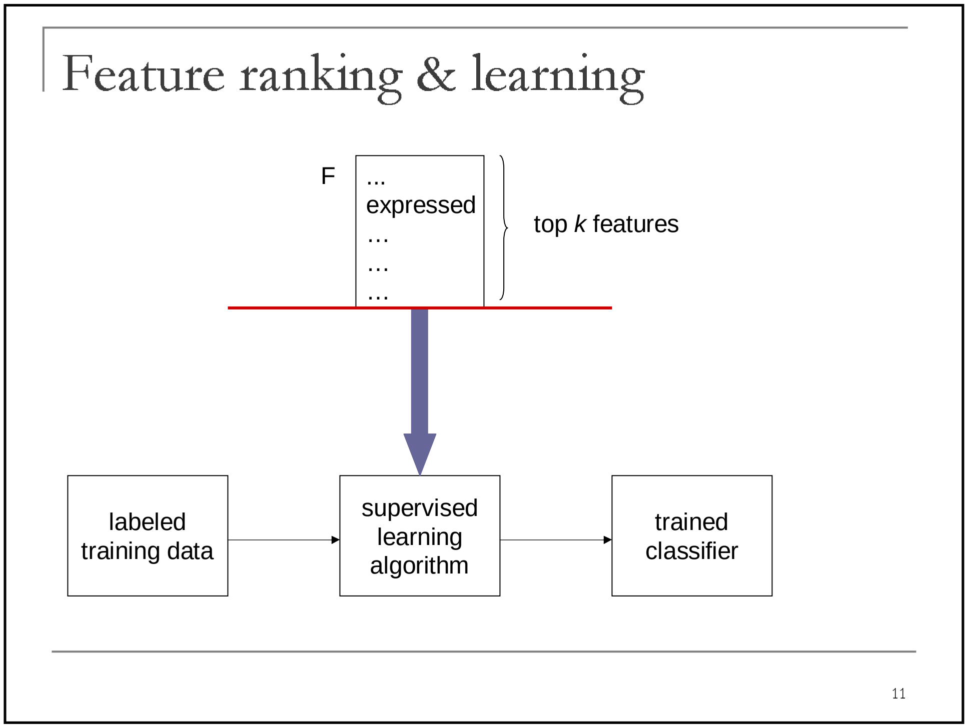
F

…
expressed
…
…
…
-less
…
…

top *k* features

labeled training data → supervised learning algorithm → trained classifier

10



Feature ranking & learning

F

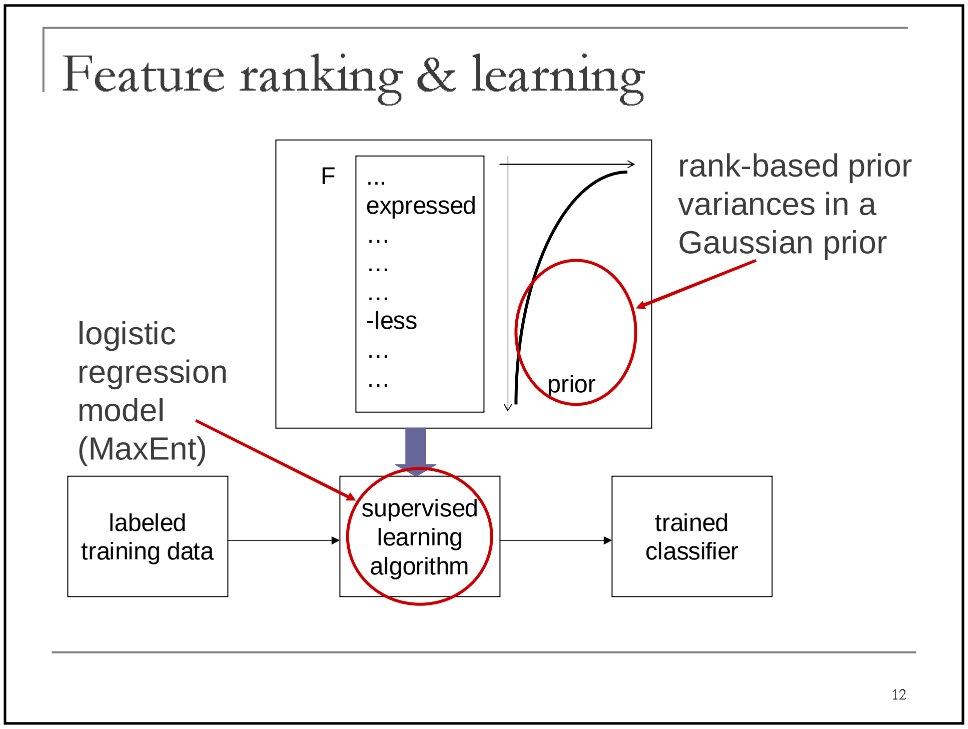
...
expressed
…
…
…

top *k* features

labeled training data

supervised learning algorithm

trained classifier



Feature ranking & learning

F

...
expressed
…
…
…
-less
…
…

prior

rank-based prior variances in a Gaussian prior

logistic regression model (MaxEnt)

labeled training data

supervised learning algorithm

trained classifier

Prior variances

- Logistic regression model

$$p(y_k \mid \vec{x}, \vec{\beta}) = \frac{\exp(\vec{x} \cdot \vec{\beta}_k)}{\sum_l \exp(\vec{x} \cdot \vec{\beta}_l)}$$

- MAP parameter estimation

$$\hat{\vec{\beta}} = \arg\max_{\vec{\beta}} p(\vec{\beta}) \prod_{i=1}^{n} p(y_i \mid \vec{x}_i, \vec{\beta})$$

prior for the parameters

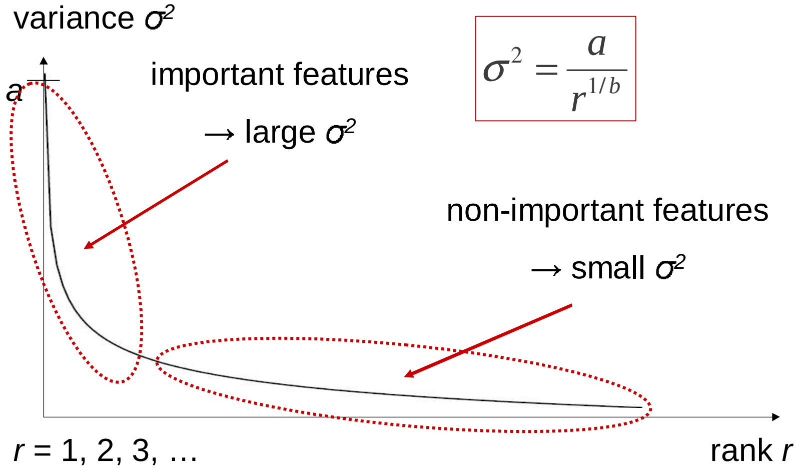
$$p(\vec{\beta}) = \prod_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp(-\frac{\vec{\beta}_j^2}{2\sigma_j^2})$$

σ_j^2 is a function of r_j

13

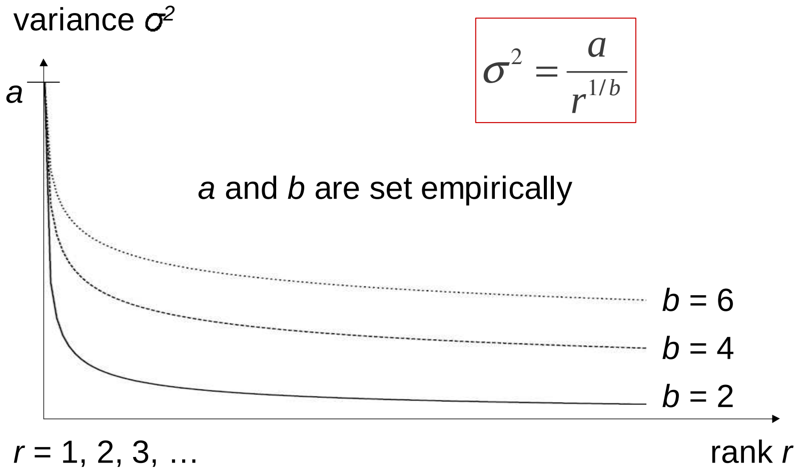
Rank-based prior

$$\boxed{\sigma^2 = \frac{a}{r^{1/b}}}$$



14

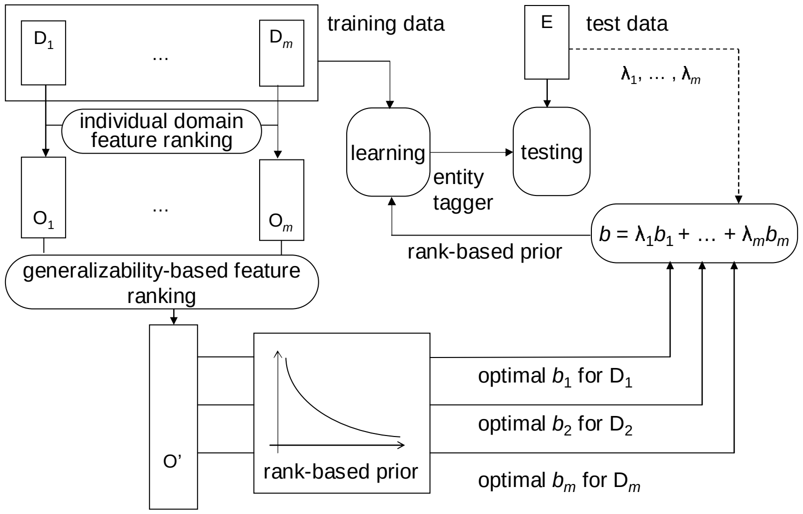
Rank-based prior



$$\sigma^2 = \frac{a}{r^{1/b}}$$

a and *b* are set empirically

Summary


D$_1$ … D$_m$

training data

E

test data

$\lambda_1, \ldots, \lambda_m$

individual domain feature ranking

O$_1$ … O$_m$

learning

testing

entity tagger

rank-based prior

$b = \lambda_1 b_1 + \ldots + \lambda_m b_m$

generalizability-based feature ranking

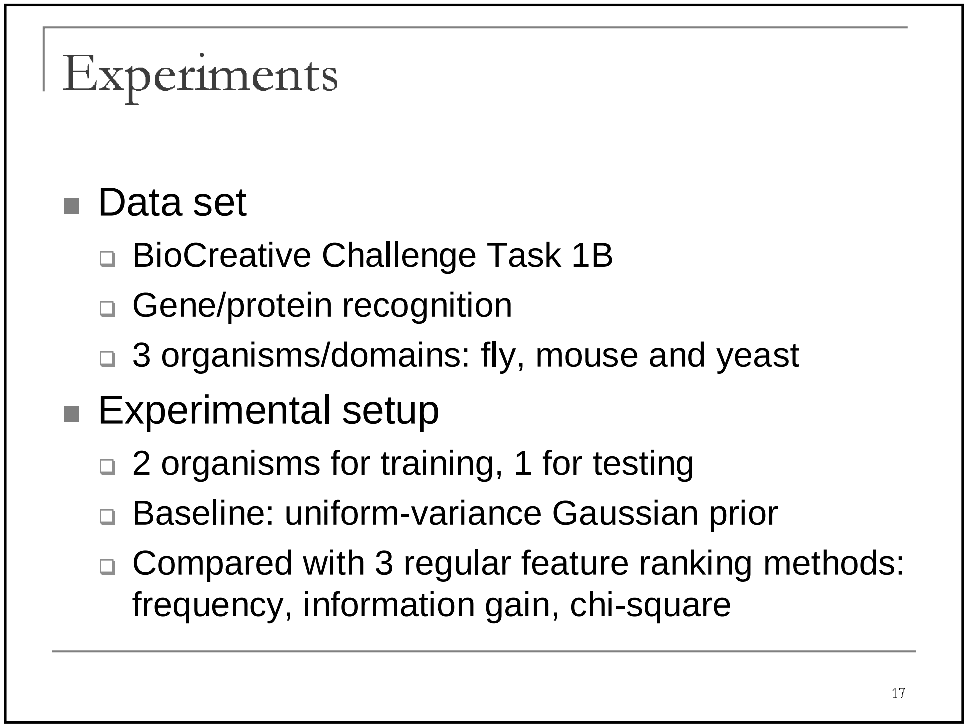
O'

rank-based prior

optimal b_1 for D$_1$

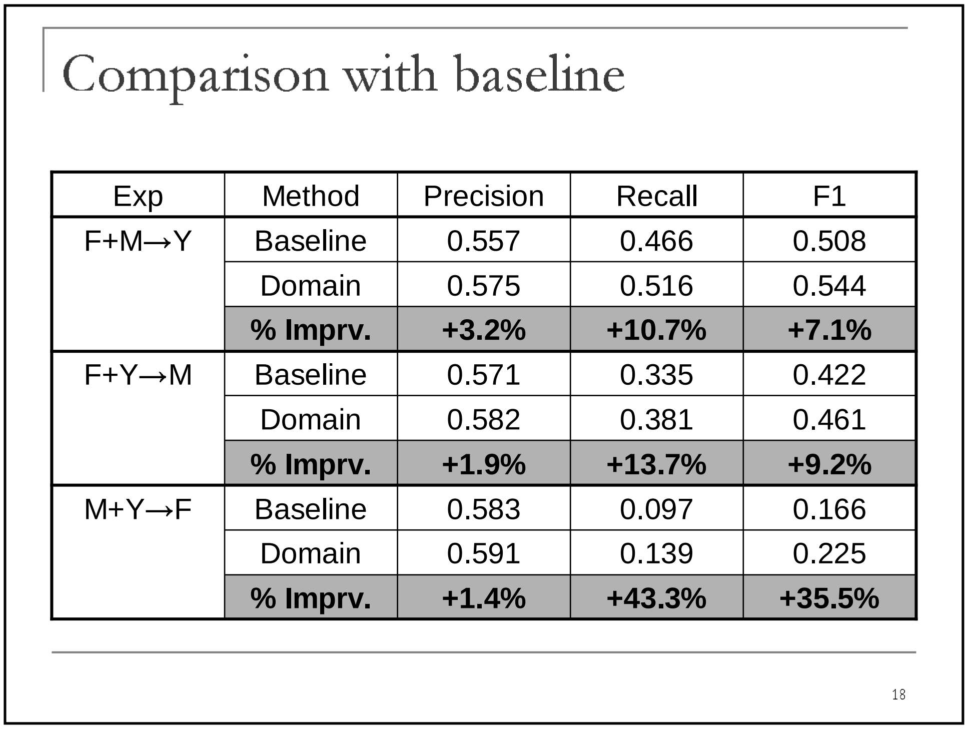
optimal b_2 for D$_2$

optimal b_m for D$_m$



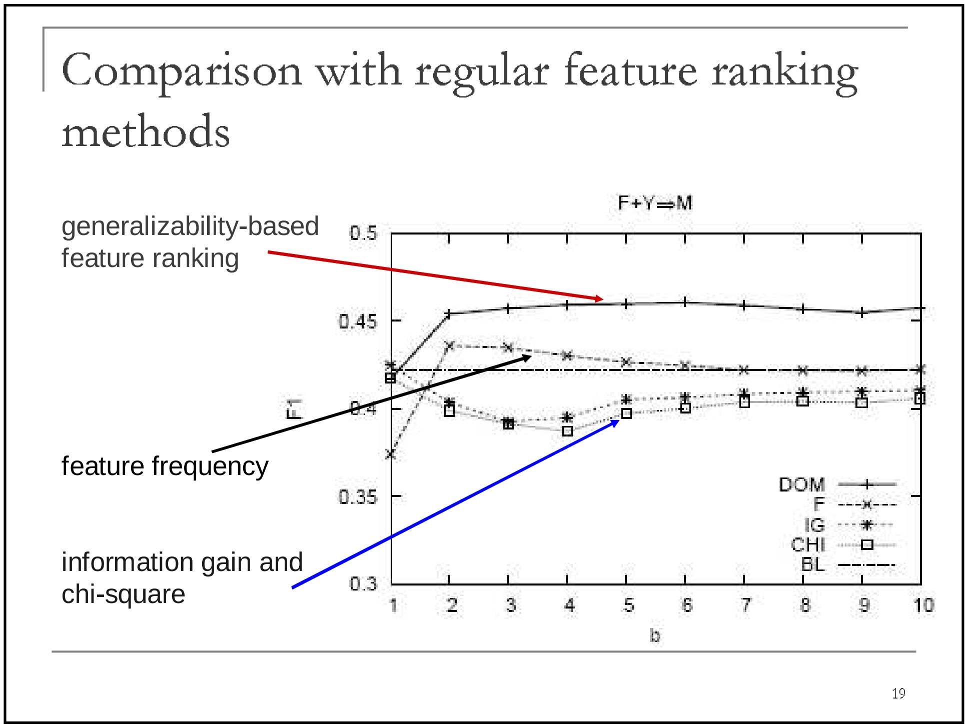
Experiments

- Data set
  - BioCreative Challenge Task 1B
  - Gene/protein recognition
  - 3 organisms/domains: fly, mouse and yeast
- Experimental setup
  - 2 organisms for training, 1 for testing
  - Baseline: uniform-variance Gaussian prior
  - Compared with 3 regular feature ranking methods: frequency, information gain, chi-square



Comparison with baseline

Exp	Method	Precision	Recall	F1
F+M→Y	Baseline	0.557	0.466	0.508
	Domain	0.575	0.516	0.544
	% Imprv.	**+3.2%**	**+10.7%**	**+7.1%**
F+Y→M	Baseline	0.571	0.335	0.422
	Domain	0.582	0.381	0.461
	% Imprv.	**+1.9%**	**+13.7%**	**+9.2%**
M+Y→F	Baseline	0.583	0.097	0.166
	Domain	0.591	0.139	0.225
	% Imprv.	**+1.4%**	**+43.3%**	**+35.5%**



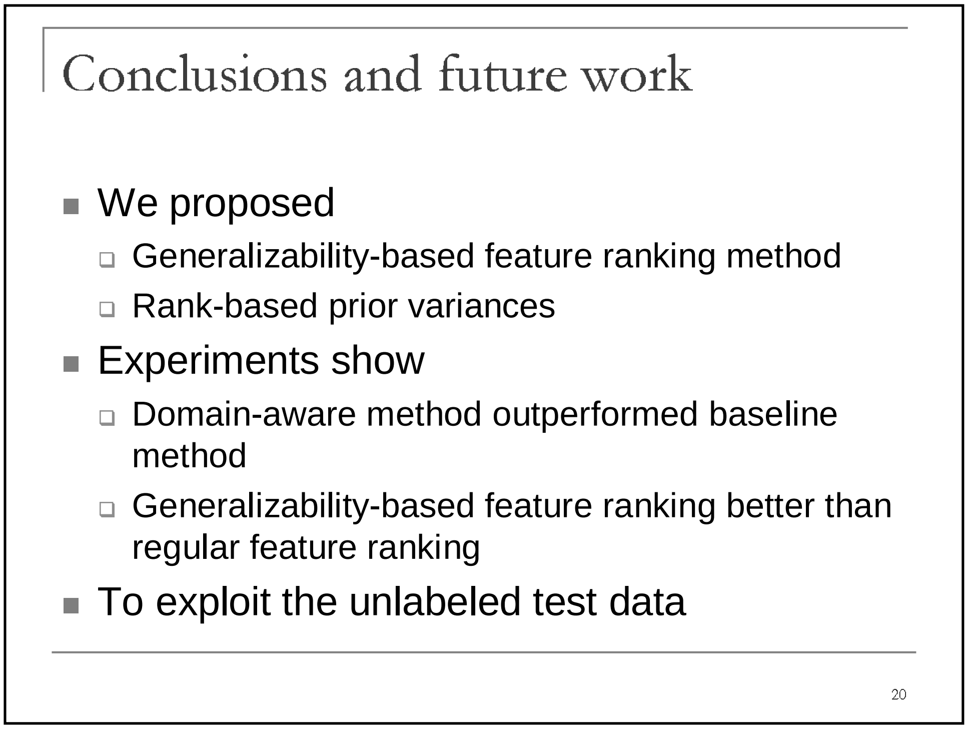
Comparison with regular feature ranking methods

generalizability-based feature ranking

feature frequency

information gain and chi-square

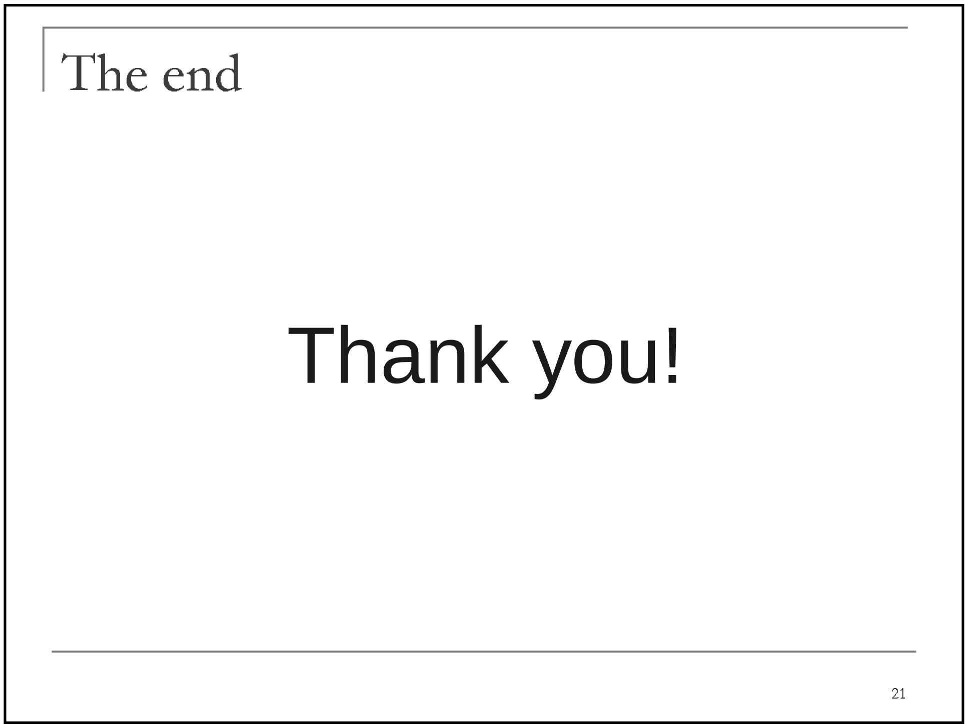
19



Conclusions and future work

- We proposed
  - Generalizability-based feature ranking method
  - Rank-based prior variances
- Experiments show
  - Domain-aware method outperformed baseline method
  - Generalizability-based feature ranking better than regular feature ranking
- To exploit the unlabeled test data

20



The end

Thank you!