Language Models for Genomics Information Retrieval

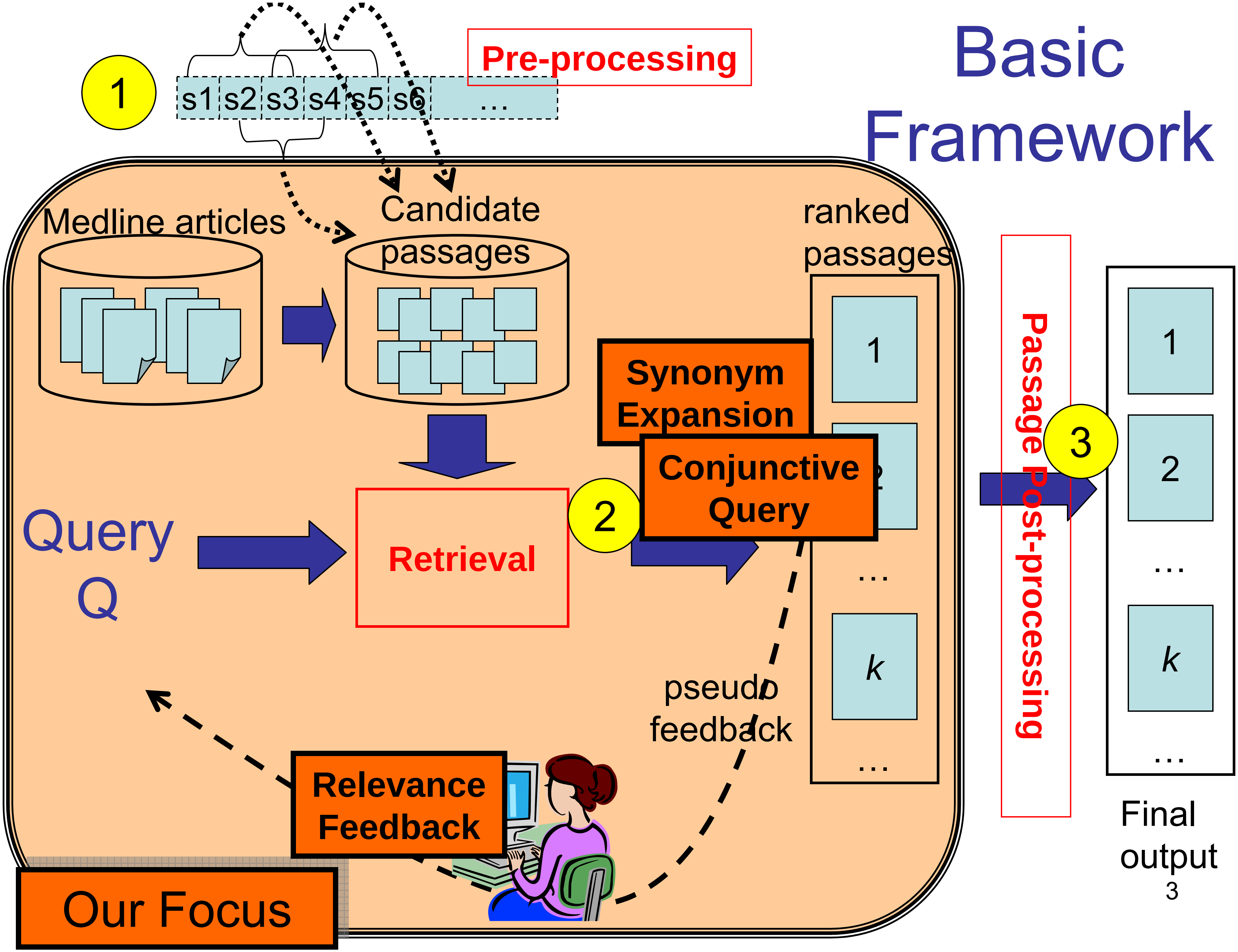
UIUC at TREC 2007 Genomics Track

Yue Lu, Jing Jiang, Xu Ling, Xin He, ChengXiang Zhai
University of Illinois at Urbana-Champaign

Goal of Participation

- Apply language models to genomics retrieval
- Extend standard language models for
  - gene synonym expansion
  - conjunctive query interpretation
- Experiment with relevance feedback

Basic Framework



Gene Synonym Expansion

"What [MOLECULAR FUNCTIONS] is LITAF involved in?"

Synonyms:

lps-induced tn factor

tbx 1

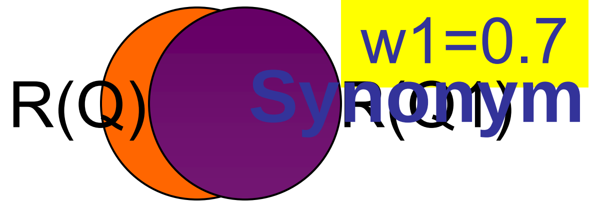
How to distinguish good synonyms from bad ones?

How to assign weights?

Overlap-Based Synonym Weighting

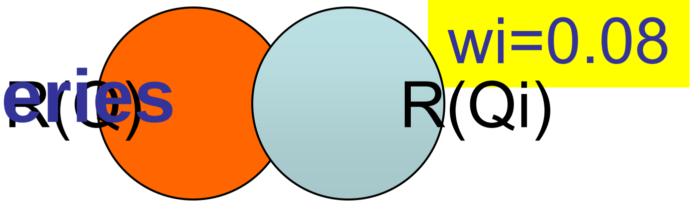
Q= “What [MOLECULAR FUNCTIONS] is LITAF involved in?”

Synonym Queries

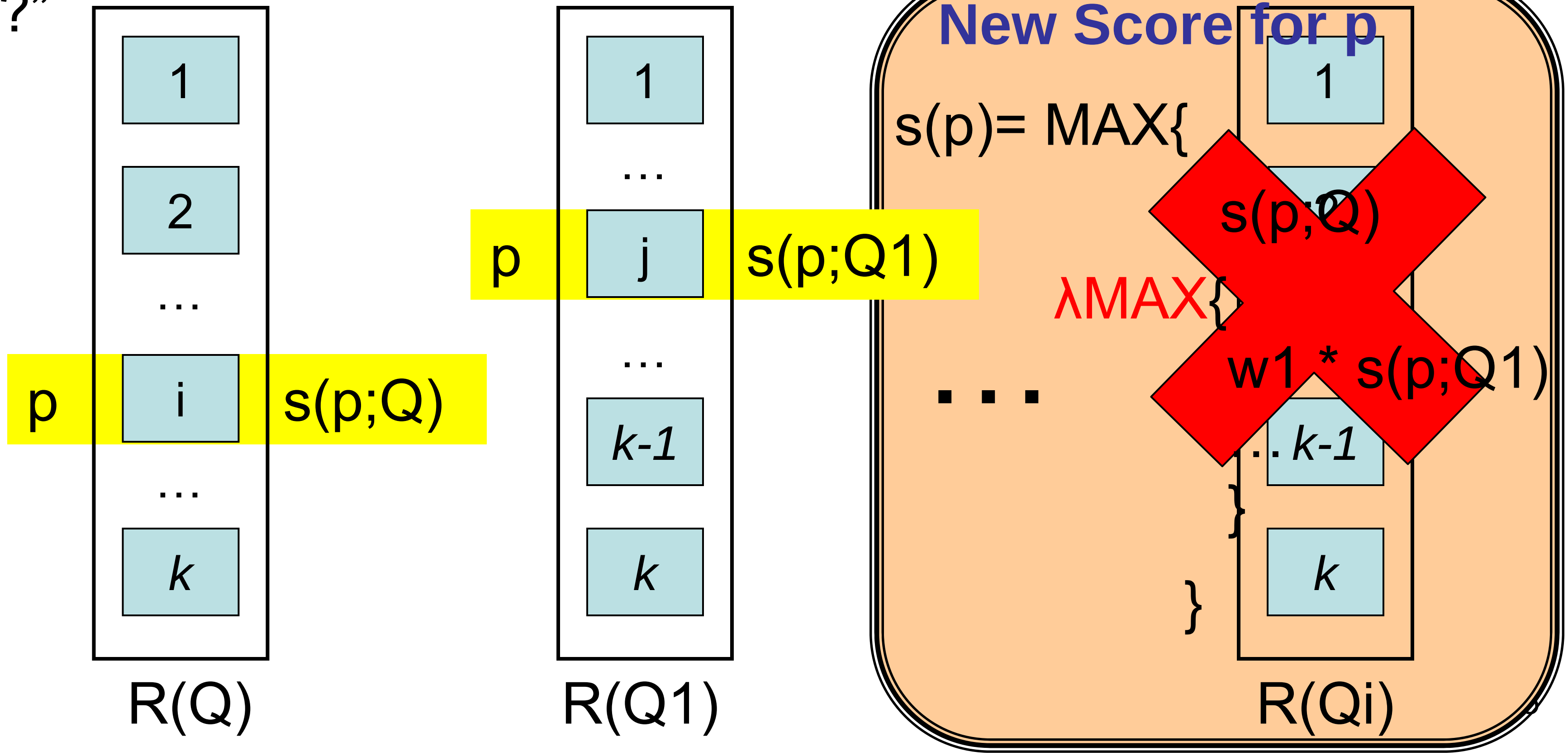


Q1=“… lps-induced tn factor…”

. . .



Qi=“… tbx 1…”



New Score for p

s(p)= MAX{

s(p;Q)

λMAX{

w1 * s(p;Q1)

. . .

…

}

}

Conjunctive Query Interpretation

“What [MOLECULAR FUNCTIONS] is LITAF involved in?”

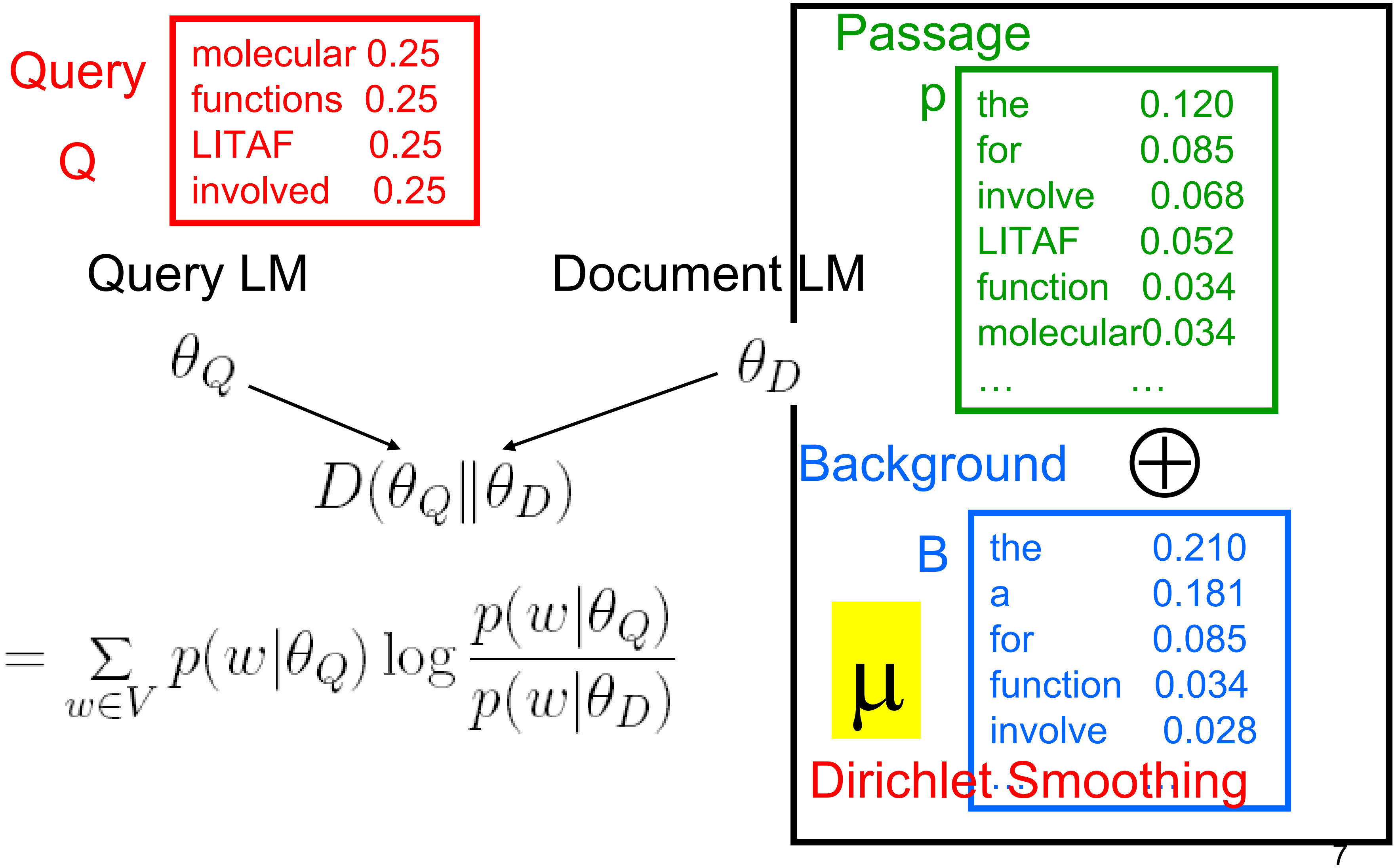
p1 = “LITAF …involve … LITAF… involved … LITAF …”

Missing “Molecular Function”

p2 = “… LITAF … involve … molecular function …”

Match all query terms

KL-Divergence Retrieval Model



Query

Q

molecular	0.25
functions	0.25
LITAF	0.25
involved	0.25

Query LM

Document LM

θ_Q

θ_D

$$D(\theta_Q||\theta_D)$$

$$= \sum_{w\in V} p(w|\theta_Q)\log\frac{p(w|\theta_Q)}{p(w|\theta_D)}$$

Passage

p

the	0.120
for	0.085
involve	0.068
LITAF	0.052
function	0.034
molecular	0.034
…	…

Background $\oplus$

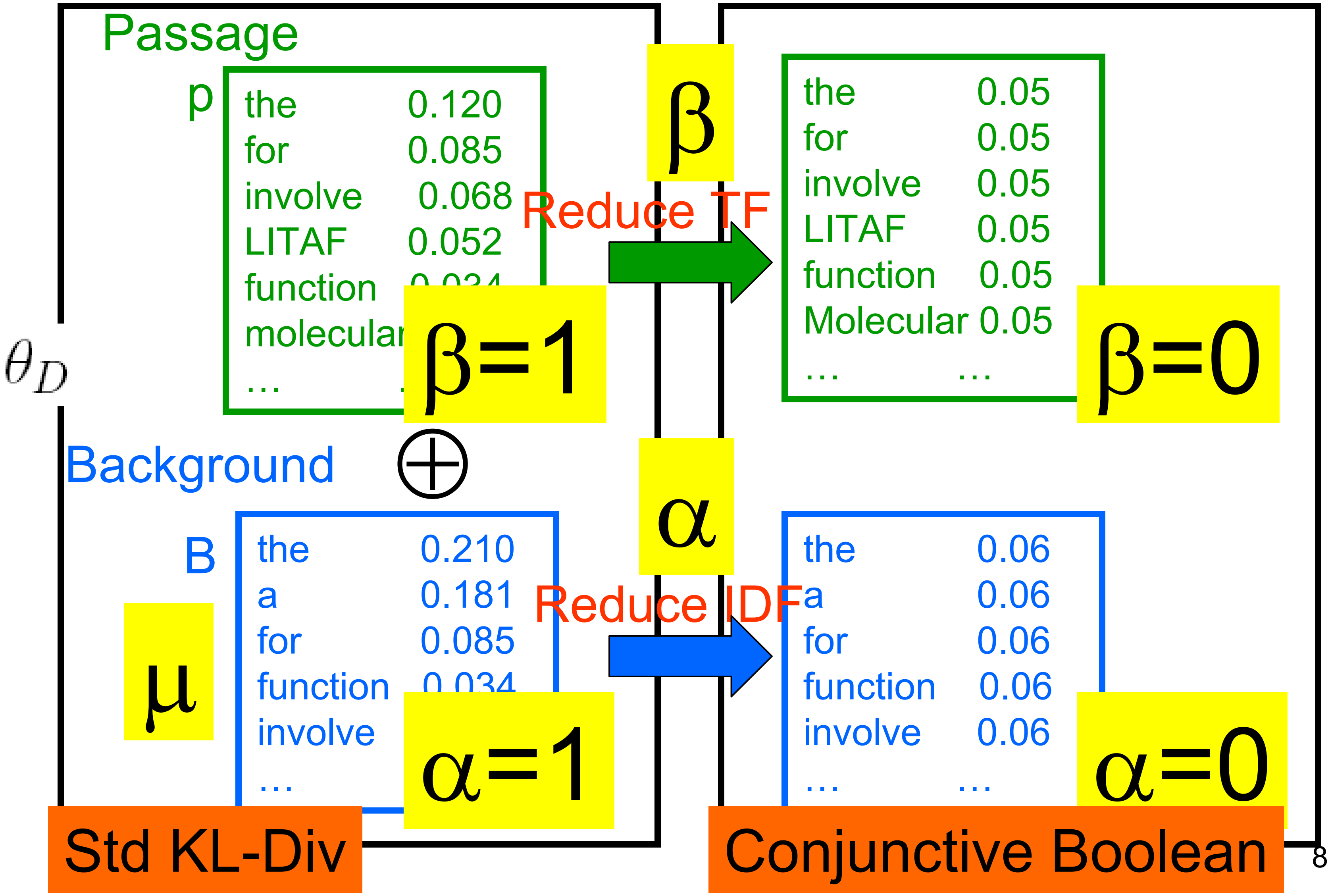
B

the	0.210
a	0.181
for	0.085
function	0.034
involve	0.028
…	…

μ

Dirichlet Smoothing

Conjunctive Scoring in LM



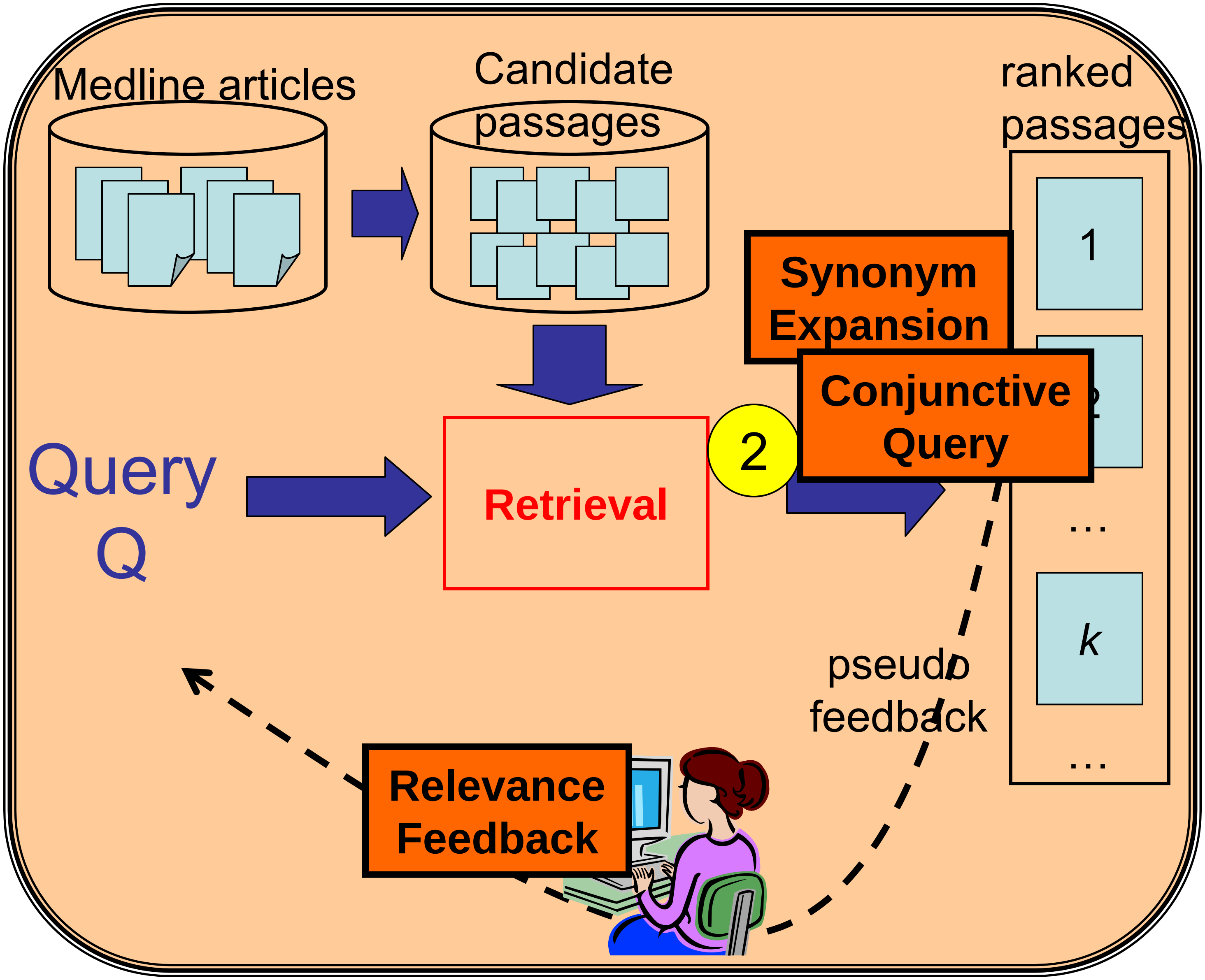
the	0.120
for	0.085
involve	0.068
LITAF	0.052
function	0.034
molecular	
…	…

the	0.210
a	0.181
for	0.085
function	0.034
involve	
…	

the	0.05
for	0.05
involve	0.05
LITAF	0.05
function	0.05
Molecular	0.05
…	…

the	0.06
a	0.06
for	0.06
function	0.06
involve	0.06
…	…

Experiments



Gene Synonym Expansion

Method		DocMAP	Psg2MAP
No expansion	Baseline1	0.1777	0.0391
Gene Synonym Expansion	UIUCsyn	0.1926	0.0392

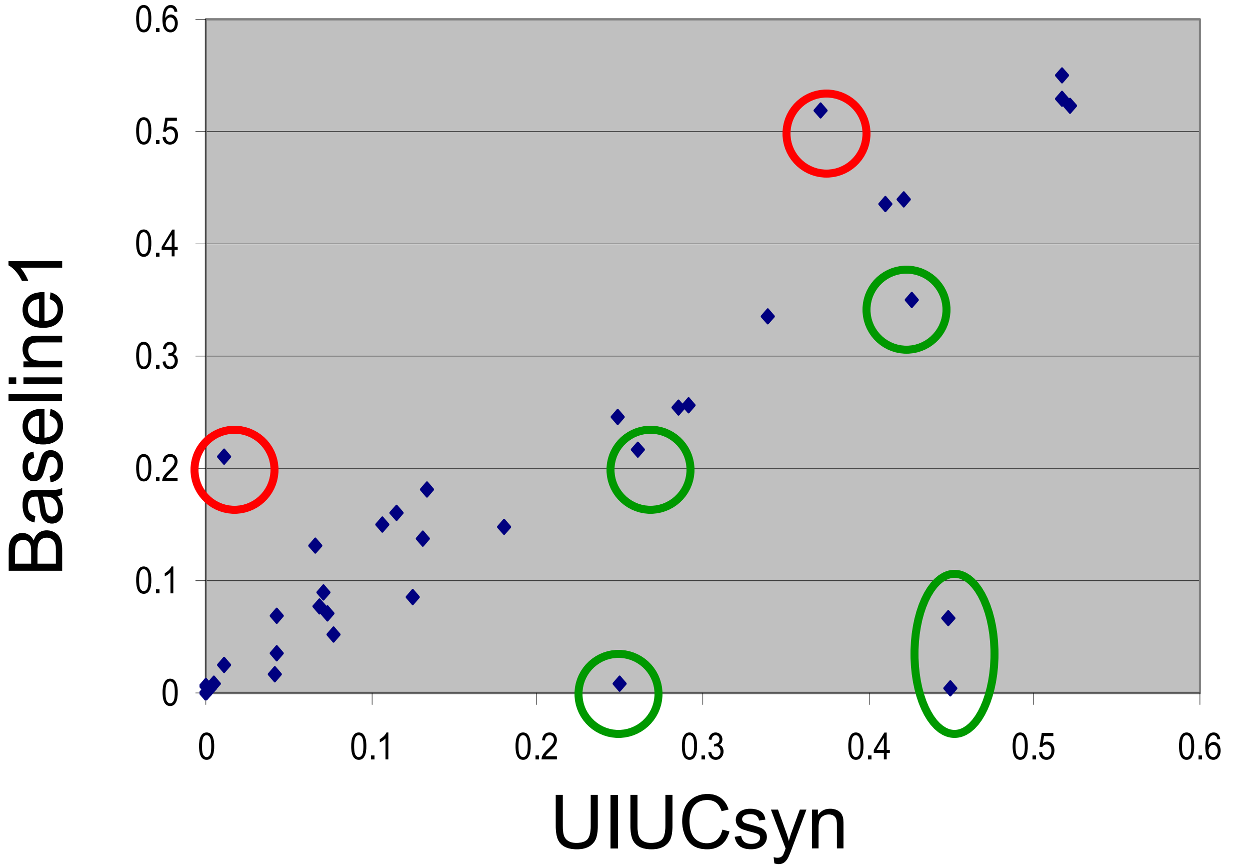
Gene Synonym Expansion

Method		DocMAP	Psg2MAP
No expansion	Baseline1	0.1777	0.0391
Gene Synonym Expansion	UIUCsyn	0.1926	0.0392
Improvement over Baseline1		+8.38%	

Gene Synonym Expansion

Method		DocMAP	Psg2MAP
No expansion	Baseline1	0.1777	0.0391
Gene Synonym Expansion	UIUCsyn	0.1926	0.0392
Improvement over Baseline1			≈0

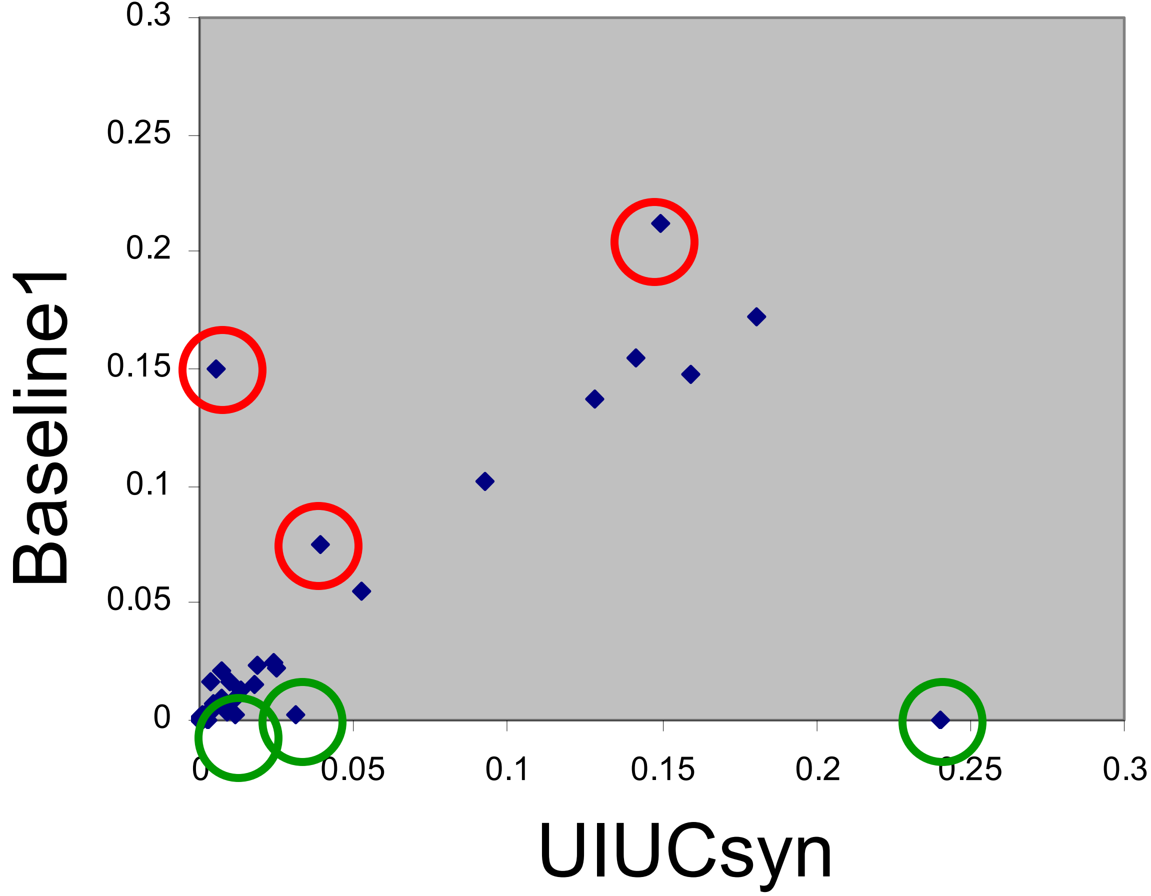
Scatter Plot of DocMAP



UIUCsyn improves DocMAP on many topics

UIUCsyn decreases DocMAP on a few topics

Scatter Plot of Psg2MAP



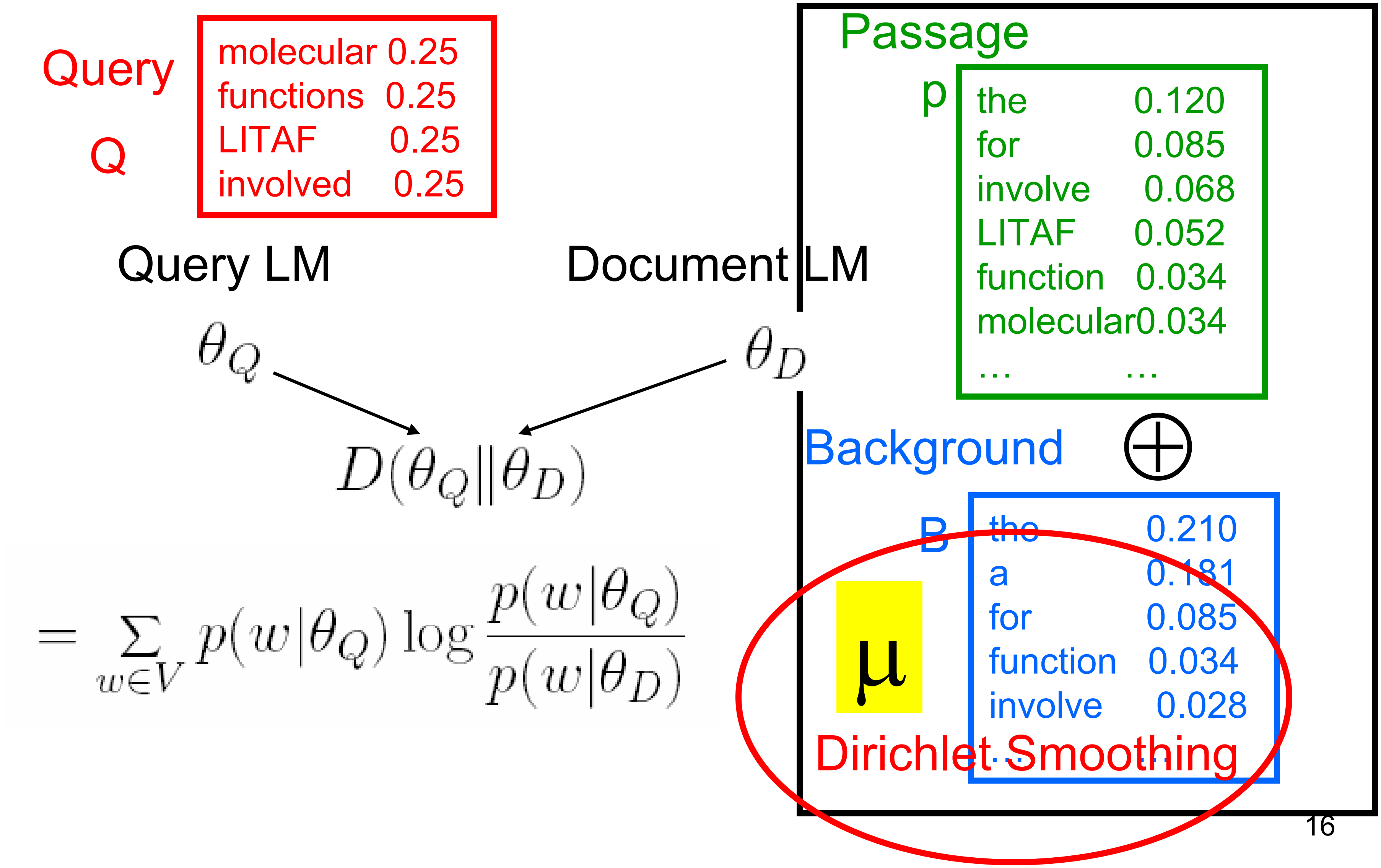
UIUCsyn improves Psg2MAP on some topics

UIUCsyn decreases Psg2MAP on some topics

Conjunctive Query Interpretation

Method		DocMAP	Psg2MAP
Std KL-Div.+fb	Baseline2	0.1918	0.0422
~~Official run~~	~~UIUCconj~~	~~0.1495~~	~~0.0296~~
Strict Conj. Boolean	UIUCconj2	0.1688	0.0351
Partly discount IDF	UIUCconj3	0.1932	0.0424
Partly discount TF	UIUCconj4	0.1931	0.0423

KL-Divergence Retrieval Model



Query

Q

molecular	0.25
functions	0.25
LITAF	0.25
involved	0.25

Query LM

Document LM

θ_Q

θ_D

$$D(\theta_Q\|\theta_D)$$

$$= \sum_{w\in V} p(w|\theta_Q)\log\frac{p(w|\theta_Q)}{p(w|\theta_D)}$$

Passage

p

the	0.120
for	0.085
involve	0.068
LITAF	0.052
function	0.034
molecular	0.034
…	…

Background ⊕

B

the	0.210
a	0.181
for	0.085
function	0.034
involve	0.028
…	…

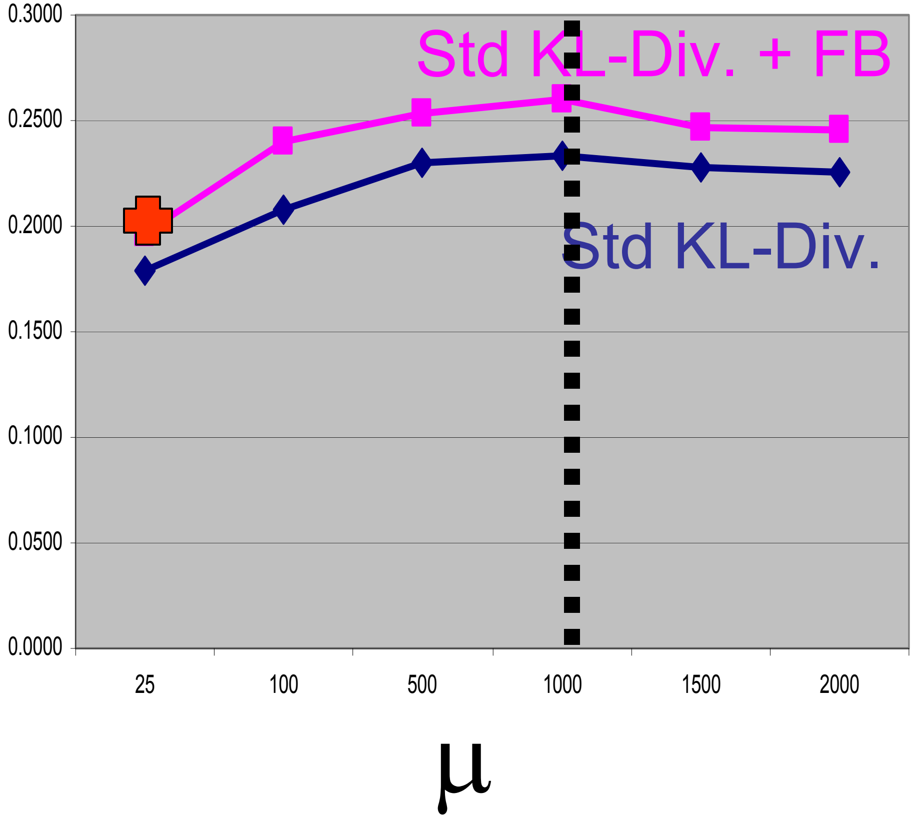
μ

Dirichlet Smoothing

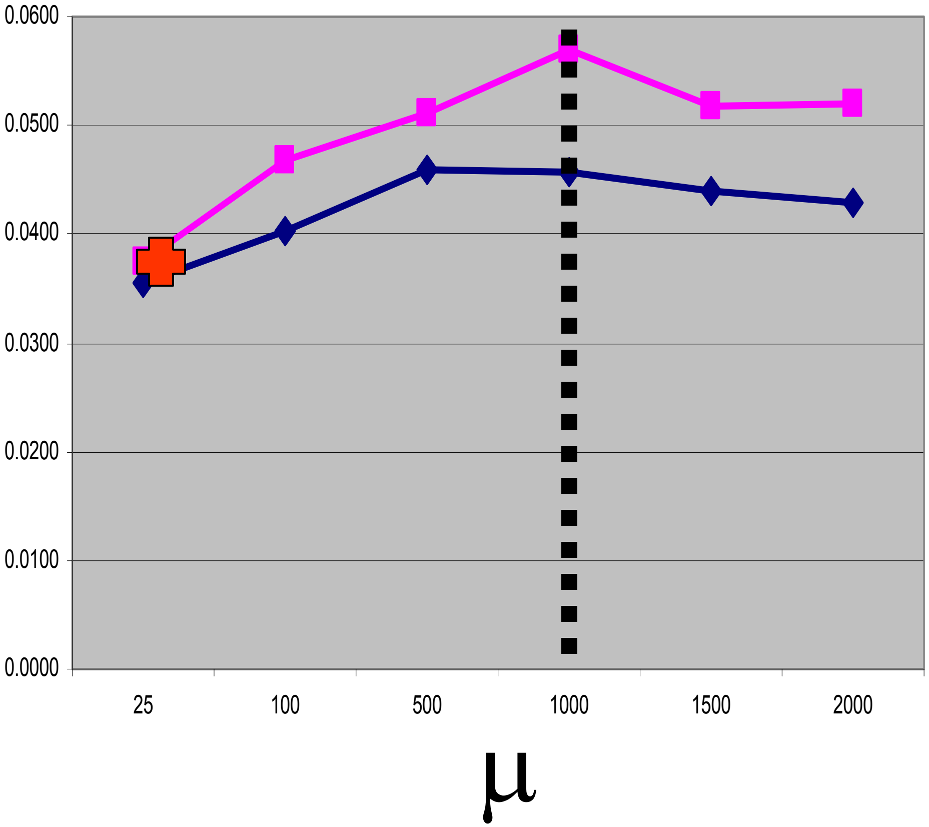
μ - Dirichlet Smoothing

✚ Performance of our best official run

DocMAP



Psg2MAP



Conjuctive Scoring over Optimum μ

Optimum: μ=1000

Method		DocMAP	Psg2MAP
Std KL-Div.+fb	Baseline2	0.2598	0.0570
Partly discount IDF	UIUCconj3	0.2660	0.0680
Improvement		+2.4%	19.3%

Relevance Feedback

Method		DocMAP	Psg2MAP
No Feedback	Baseline1	0.1777	0.0391
Pseudo Feedback	Baseline2	0.1918	0.0422
Relevance Feedback	UIUCrelfb	0.1940	0.0364

Both feedback methods improve DocMAP,
but NOT necessarily Psg2MAP

Conclusions and Future Work

- Standard KL-Div. retrieval method are effective but also sensitive to Dirichlet smoothing μ
- Conjunctive scoring improves performance based on optimum μ
- Synonym expansion and User relevance feedback tend to improve DocMAP but not Psg2MAP
- Future work
  - Automatically set optimum Dirichlet smoothing
  - More aggressive synonym expansion

Questions?