Robust Pseudo Feedback & HMM Passage Extraction

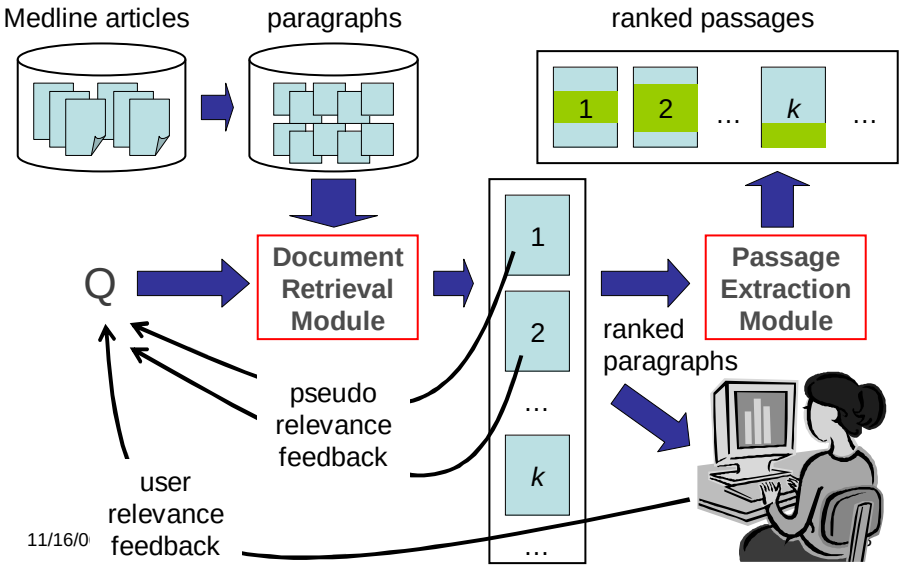
UIUC at TREC 2006 Genomics Track

Jing Jiang, Xin He, ChengXiang Zhai
University of Illinois at Urbana-Champaign

Goal of Participation

- To test the effectiveness of some recent language modeling methods for genomics retrieval
  - Robust pseudo feedback [Tao & Zhai 06]
  - HMM passage extraction [Jiang & Zhai 06]
- Task at 2006 genomics track
  - Document-level retrieval
  - Passage-level retrieval
  - Aspect-level retrieval

Overall Approach

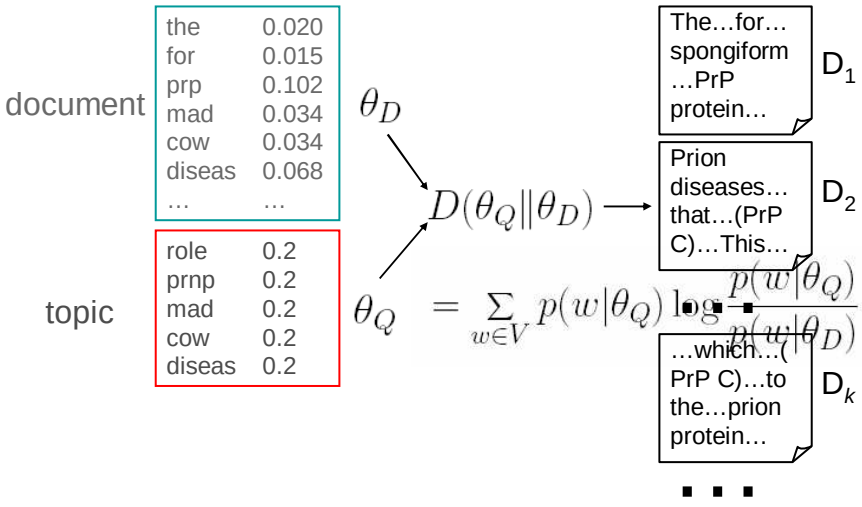


Goal of Participation

- To test the effectiveness of some recent language modeling methods for genomics retrieval
  - Robust pseudo feedback [Tao & Zhai 06]
  - HMM passage extraction [Jiang & Zhai 06]

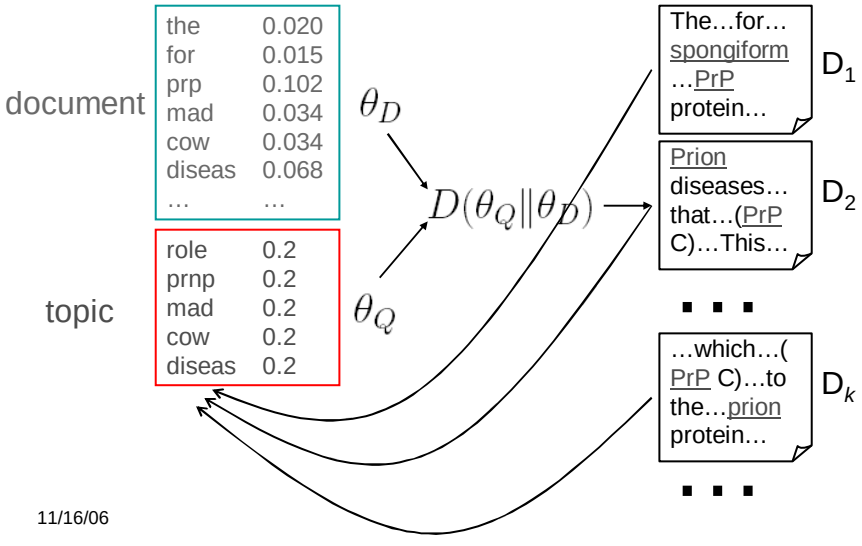
KL-Divergence Retrieval Model

[Lafferty & Zhai 01]



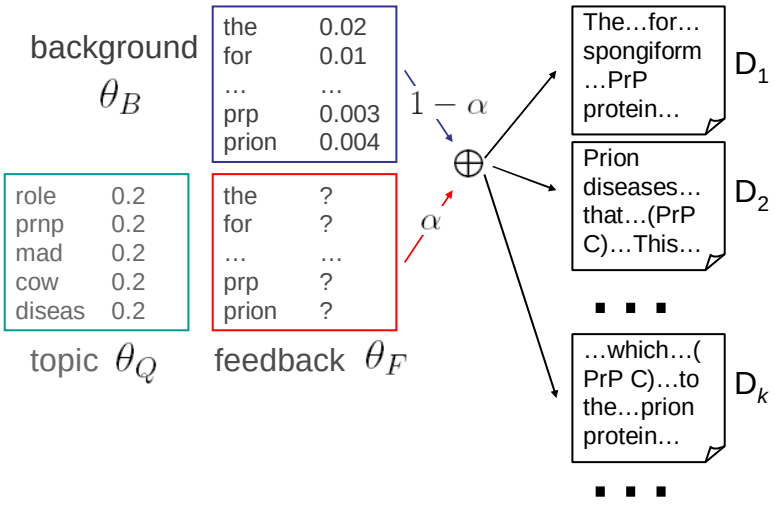
KL-Divergence Retrieval Model

[Lafferty & Zhai 01]



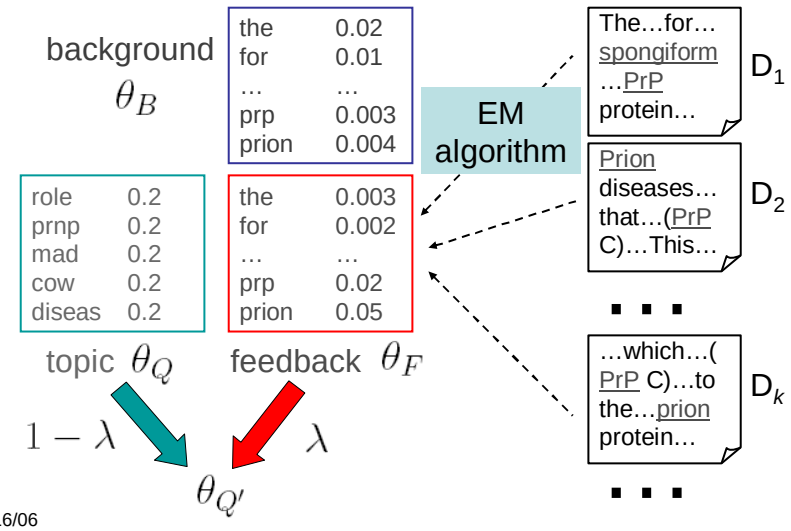
Model-Based Feedback

[Zhai & Lafferty 01]



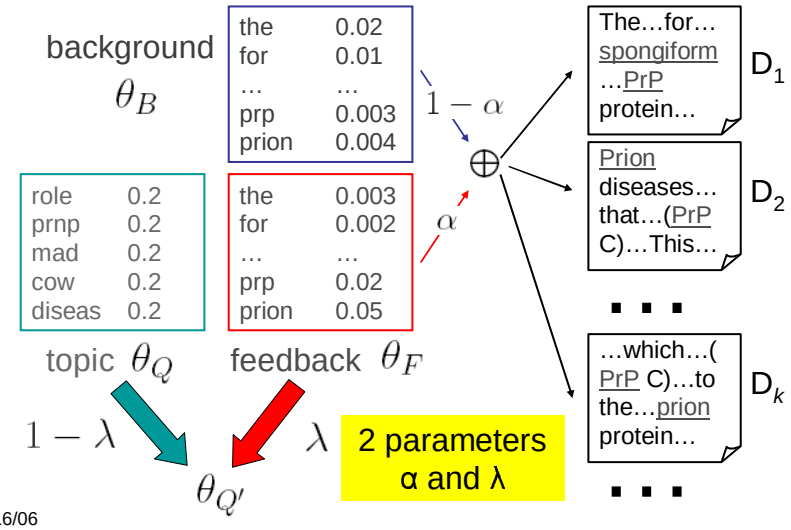
Model-Based Feedback

[Zhai & Lafferty 01]



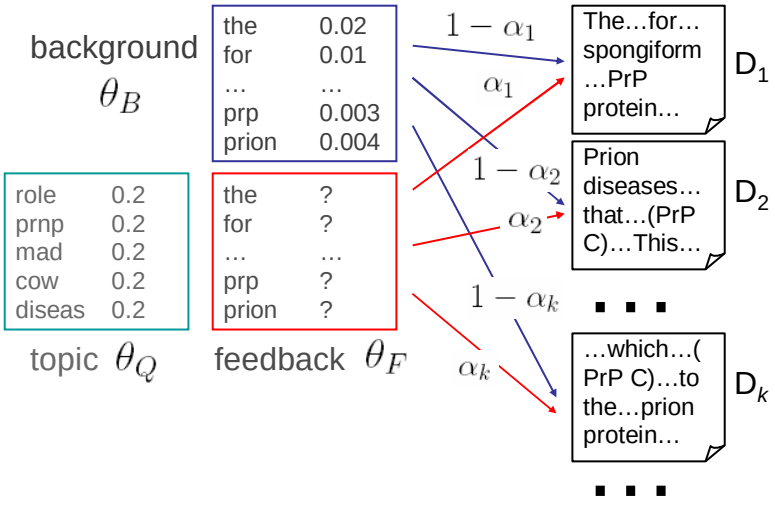
Model-Based Feedback

[Zhai & Lafferty 01]



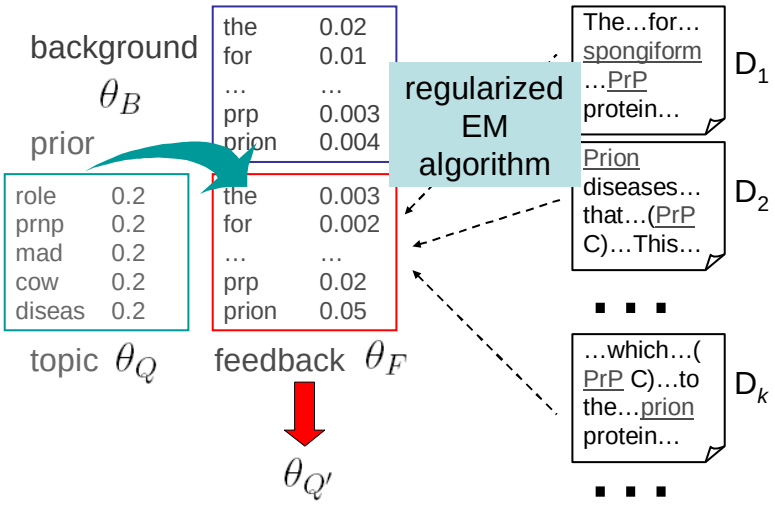
Regularized Estimation

[Tao & Zhai 06]



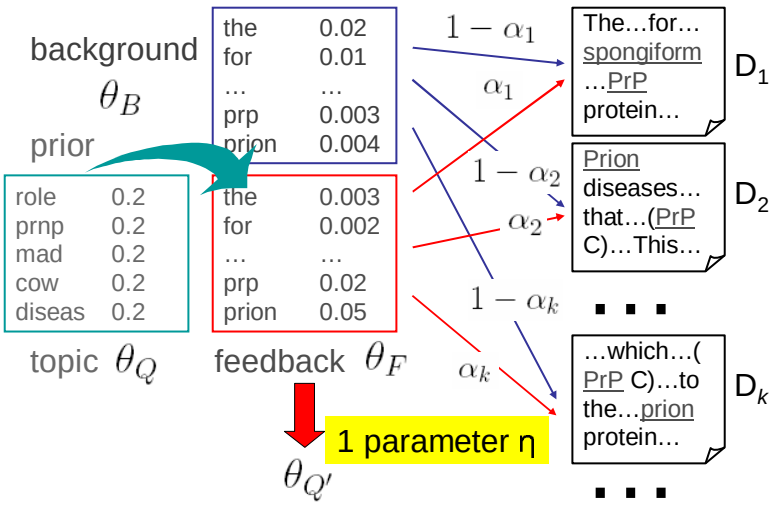
Regularized Estimation

[Tao & Zhai 06]

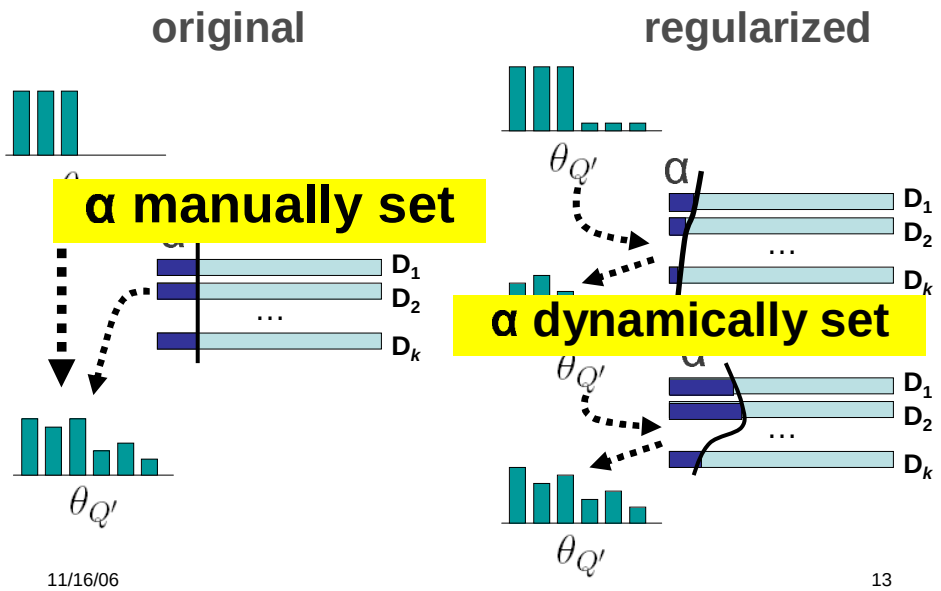


Regularized Estimation

[Tao & Zhai 06]



Original vs. Regularized EM

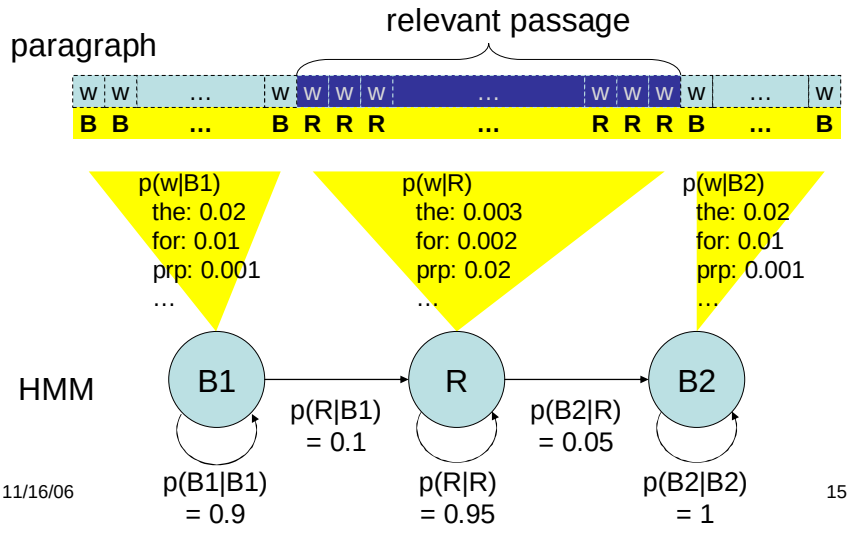


Goal of Participation

- To test the effectiveness of some recent language modeling methods for genomics retrieval
  - Robust pseudo feedback [Tao & Zhai 06]
  - HMM passage extraction [Jiang & Zhai 06]

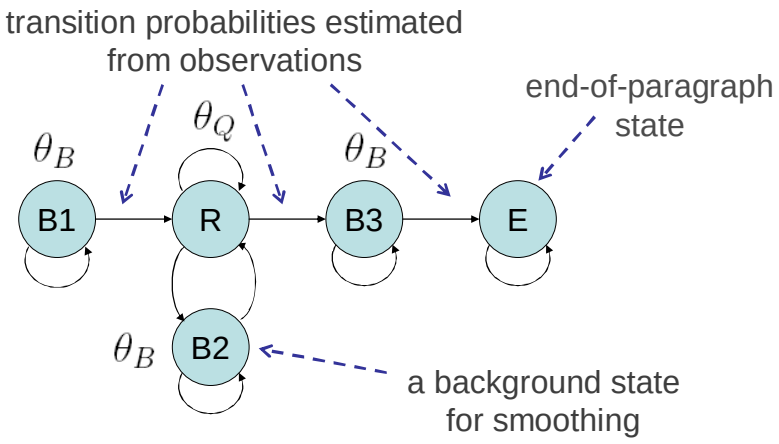
HMM Passage Extraction

[Jiang & Zhai 06]



HMM Passage Extraction

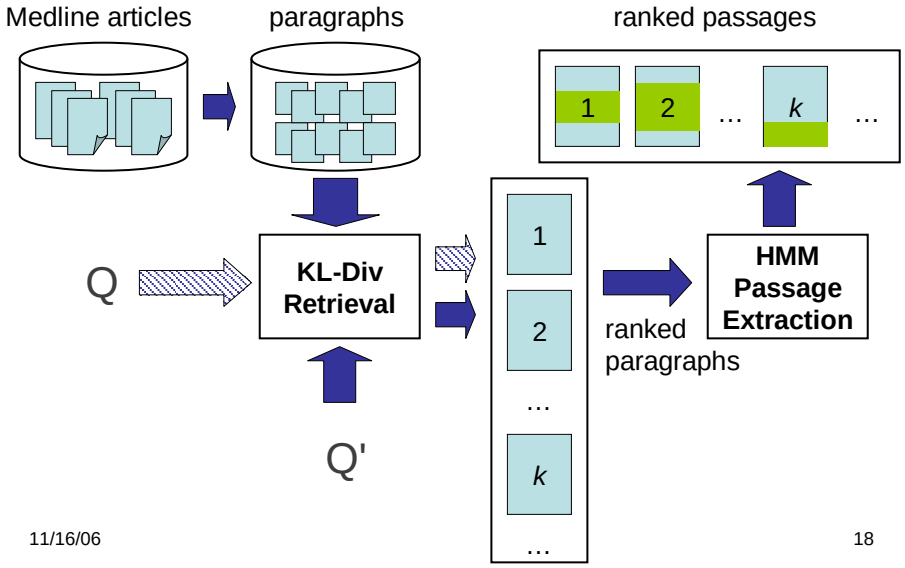
[Jiang & Zhai 06]

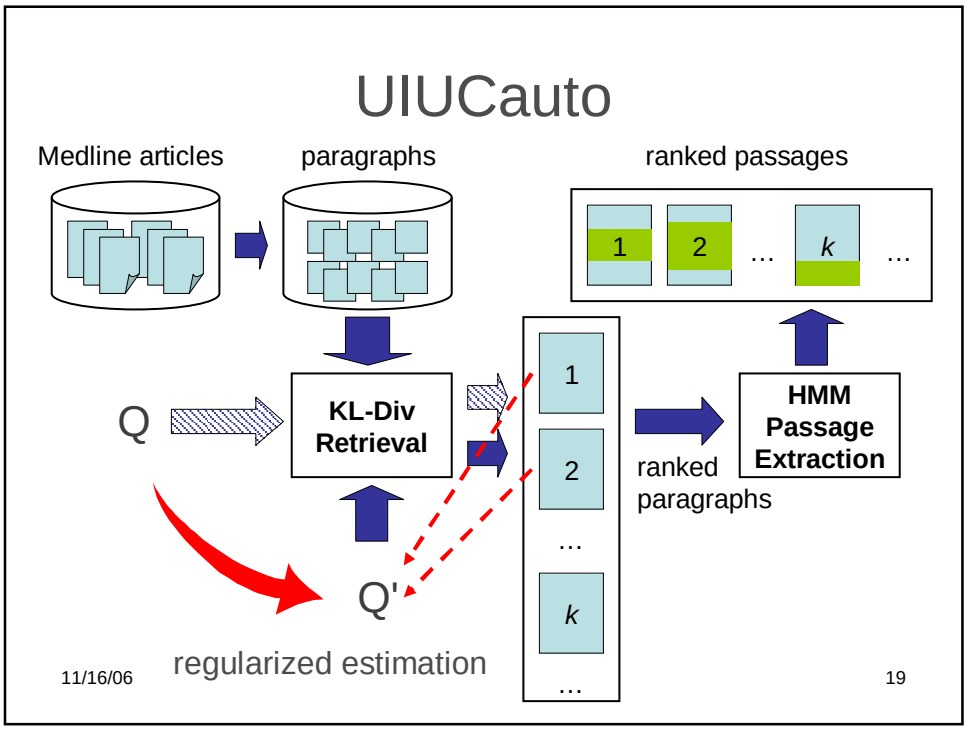


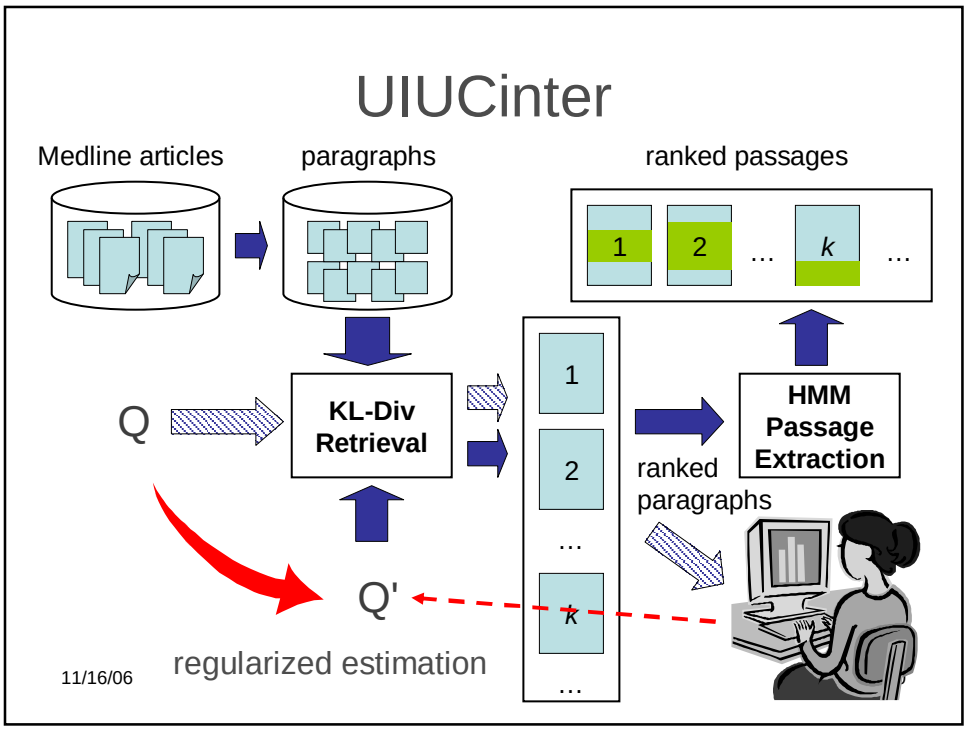
Experiment Design

- Pre-processing
  - HTML parsing
  - paragraph boundaries
  - Tokenization
- User relevance feedback

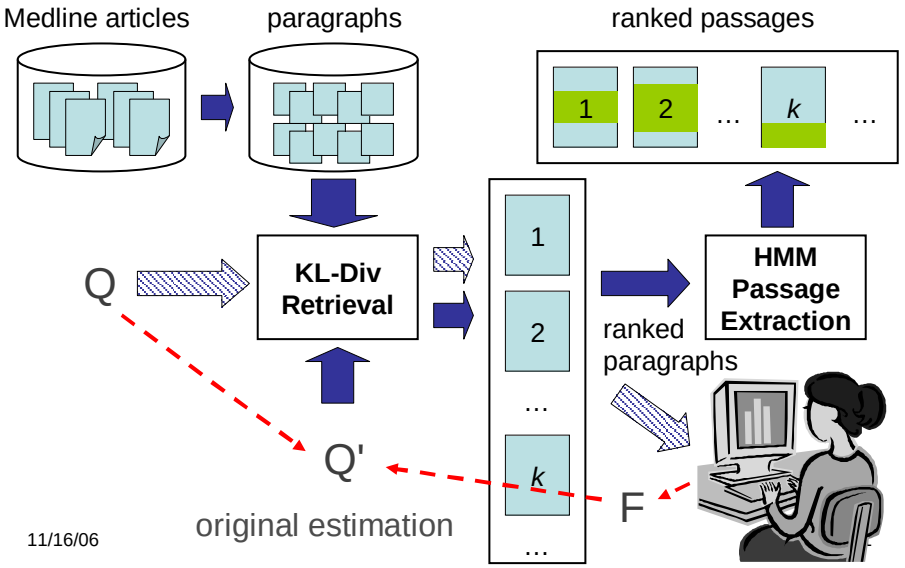
Official Runs




UIUCauto
Medline articles
paragraphs
ranked passages
1
2
…
k
…
Q
KL-Div Retrieval
1
2
…
k
…
HMM Passage Extraction
ranked paragraphs
Q'
regularized estimation
11/16/06
19


UIUCinter
Medline articles
paragraphs
ranked passages
1
2
…
k
…
Q
KL-Div Retrieval
1
2
…
k
…
HMM Passage Extraction
ranked paragraphs
Q'
regularized estimation
11/16/06

UIUCinter2



11/16/06

Pseudo Relevance Feedback (k = 10)

Method		Doc MAP	Rel. Impr.
Baseline (no feedback)		0.3484	N/A
Original Estimation	Def	0.3606	**+3.50%**
	Opt	0.3943	**+13.2%**
Regularized Estimation	Def	0.3842 (UIUCauto)	**+10.3%**
	Opt	0.3952	**+13.4%**

η is similar to λ / (1 − λ)

11/16/06 22

Pseudo Relevance Feedback (k = 10)

Method		Doc MAP	Rel. Impr.
Baseline (no feedback)		0.3484	N/A
Original Estimation	Def	0.3606	**+3.50%**
	Opt	0.3943	**+13.2%**
Regularized Estimation	Def	0.3842 (UIUCauto)	**+10.3%**
	Opt	0.3952	**+13.4%**

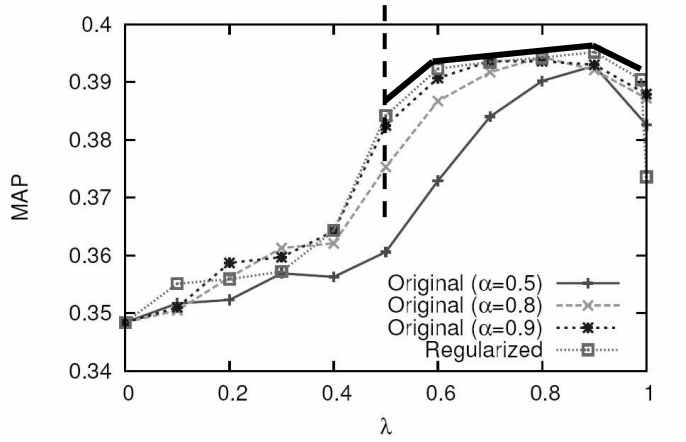
η is similar to λ / (1 − λ)

Pseudo Relevance Feedback (k = 10)

Method		Doc MAP	Rel. Impr.
Baseline (no feedback)		0.3484	N/A
Original Estimation	Def	0.3606	**+3.50%**
	Opt	0.3943	**+13.2%**
Regularized Estimation	Def	0.3842 (UIUCauto)	**+10.3%**
	Opt	0.3952	**+13.4%**

η is similar to λ / (1 − λ)

Parameter Sensitivity (pseudo feedback, k = 10)



User Relevance Feedback

Method		Doc MAP		
		Pseudo Feedback	**User Feedback**	**Rel. Impr.**
Original Estimation	Def	0.3606	0.3986	**+10.5%**
	Opt	0.3943	0.4511	**+14.4%**
Regularized Estimation	Def	0.3842 (UIUCauto)	0.4261 (UIUCinter)	**+10.9%**
	Opt	0.3952	0.4515	**+14.2%**

User Relevance Feedback

Method		Doc MAP		
		Pseudo Feedback	**User Feedback**	**Rel. Impr.**
Original Estimation	Def	0.3606	0.3986	**+10.5%**
	Opt	0.3943	0.4511	**+14.4%**
Regularized Estimation	Def	0.3842 (UIUCauto)	0.4261 (UIUCinter)	**+10.9%**
	Opt	0.3952	0.4515	**+14.2%**

User Relevance Feedback

Method		Doc MAP		
		Pseudo Feedback	**User Feedback**	**Rel. Impr.**
Original Estimation	Def	0.3606	0.3986	**+10.5%**
	Opt	0.3943	0.4511	**+14.4%**
Regularized Estimation	Def	0.3842 (UIUCauto)	0.4261 (UIUCinter)	**+10.9%**
	Opt	0.3952	0.4515	**+14.2%**

HMM Passage Extraction

Method		Psg MAP
UIUCauto	Paragraph	0.03753
	HMM Passage	0.04864
	Rel. Impr.	+29.6%
UIUCinter	Paragraph	0.04481
	HMM Passage	0.05906
	Rel. Impr.	+31.8%
UIUCinter2	Paragraph	0.04580
	HMM Passage	0.06038
	Rel. Impr.	+31.8%

Passage Length (In Bytes)

	Max	Min	Avg	Std
True Passages	6928	27	**399.8**	489.4
HMM Passages	6955	34	**1525.8**	949.7
Paragraph	8670	60	**2105.4**	1136.8

HMM passages are generally too long!

Example Passage

Prion diseases, which include Creutzfeldt-Jacob disease in humans, mad cow disease in cattle, and scrapie in sheep, involve the misfolding of the benign cellular prion protein (PrP C) 1 to the infectious disease-causing scrapie isoform PrP Sc. The prion protein (PrP C) is a copper-binding cell surface glycoprotein. The role of copper in the normal function of PrP, as well as in prion diseases, has been the subject of a number of excellent reviews. The mature cellular form of PrP consists of residues 23 to 231 and is tethered to the cell surface via a glycosylphosphatidylinositol anchor at the C terminus. There are now a number of NMR solution structures of copper-free mammalian PrPs. A crystal structure of PrP C has also been published; this structure is dimeric involving domain swapping of the monomeric form.

Example Passage

Prion diseases, which include Creutzfeldt-Jacob disease in humans, mad cow disease in cattle, and scrapie in sheep, involve the misfolding of the benign cellular prion protein (PrP C) 1 to the infectious disease-causing scrapie isoform PrP Sc. The **prion** protein (**PrP** C) is a copper-binding cell surface glycoprotein. The role of copper in the normal function of **PrP**, as well as in **prion diseases**, has been the subject of a number of excellent reviews. The mature cellular form of PrP consists of residues 23 to 231 and is tethered to the cell surface via a glycosylphosphatidylinositol anchor at the C terminus. There are now a number of NMR solution structures of copper-free mammalian PrPs. A crystal structure of PrP C has also been published; this structure is dimeric involving domain swapping of the monomeric form.

Conclusions and Future Work

- The two language modeling methods in general works well in genomics domain
  - Regularized feedback estimation can effectively eliminates parameter α
  - HMM passages improves over paragraphs
- User relevance feedback is effective
- Limitations and future work
  - Regularized feedback estimation still has parameter η to tune
    - How to eliminate η?
  - The inherent coherence property of HMM passages may not suit the task well
    - Different/better HMM architecture?

The End

- Questions?