

Concept Recognition, Information Retrieval, and Machine Learning in Genomics Question Answering

J. Gregory Caporaso

Center for Computational Pharmacology
Department of Biochemistry and Molecular Genetics
University of Colorado Health Sciences Center

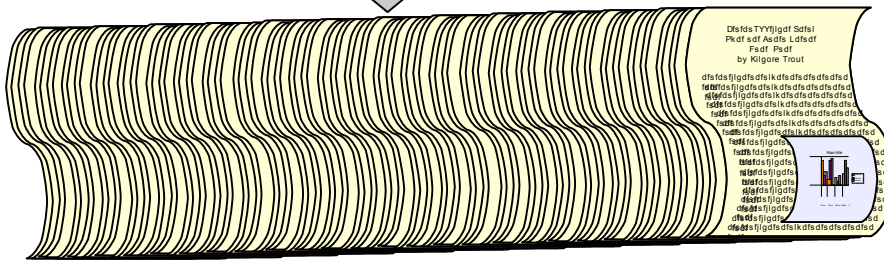
gregcaporaso@gmail.com
<http://hsc-turing.ucbsc>

TREC Genomics 2006

Q: What is the role of huntingtin in Huntington's Disease?

28 topic questions

162,259 full-text journal articles



A: Repeats in the huntingtin gene cause protein aggregation and likely cause HD.

Up to 1000 passages per topic questions

Mean average precision at document, passage and aspect levels

Relevance judged by human domain experts

UCHSC approach

● uchsc1 ● uchsc2 ● uchsc3

📁 Document processing: HTML parsing, paragraph splitting, document zoning ● ● ●

📄 Section filtering ● ● ●

📄 In dri indexing ● ● ●

📄 Query expansion

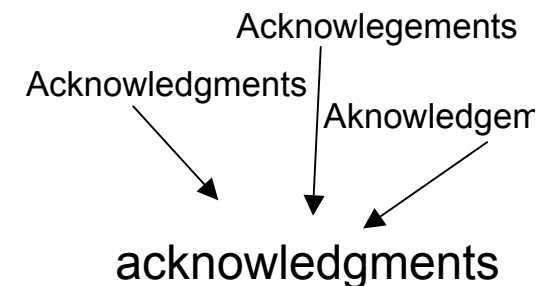
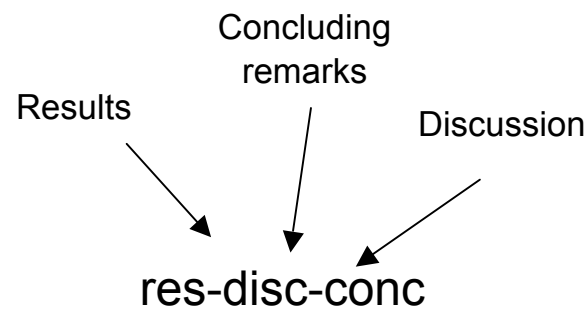
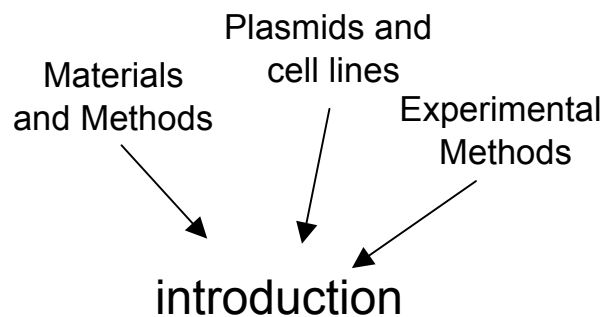
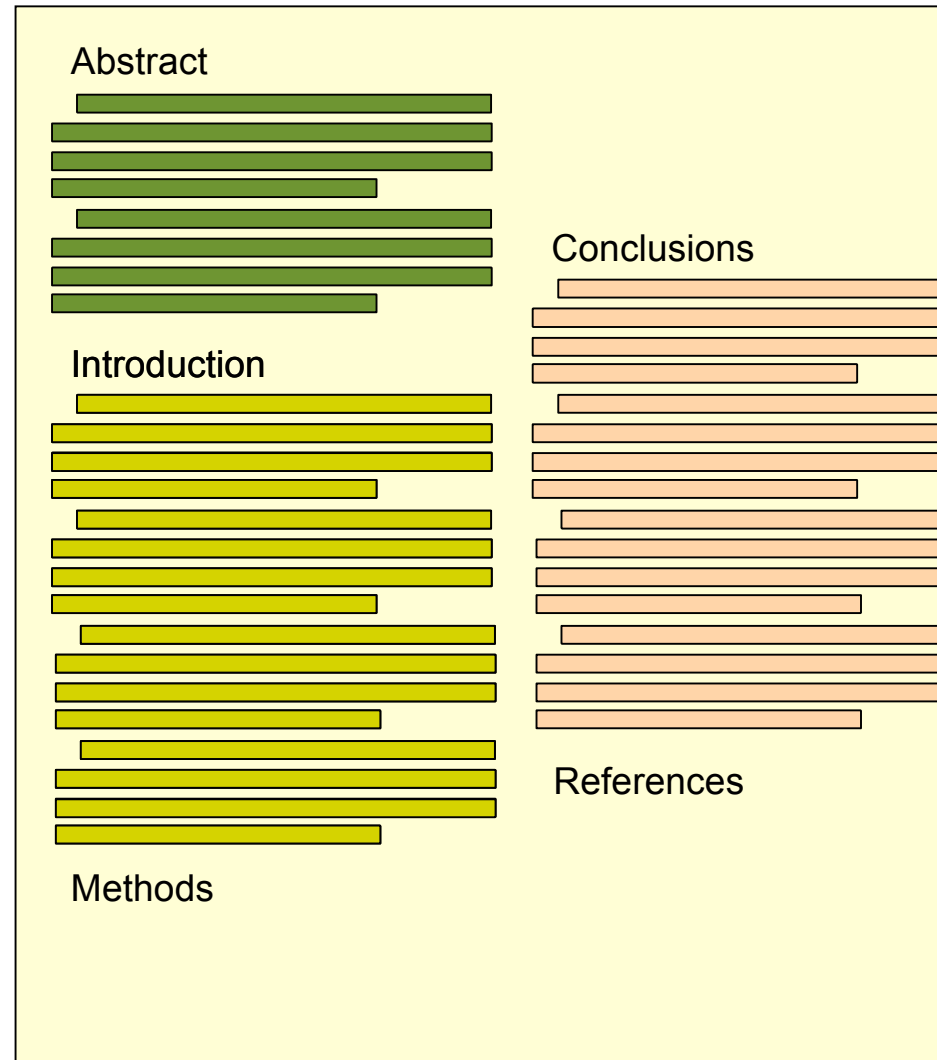
⊗ Strict ● ●

⊗ Loose ●

📄 In dri query, pseudo-relevance feedback ● ● ●

🕒 Result expansion with Latent Semantic Analysis ● ●

📄 Naïve Bayes-based result filtering ●



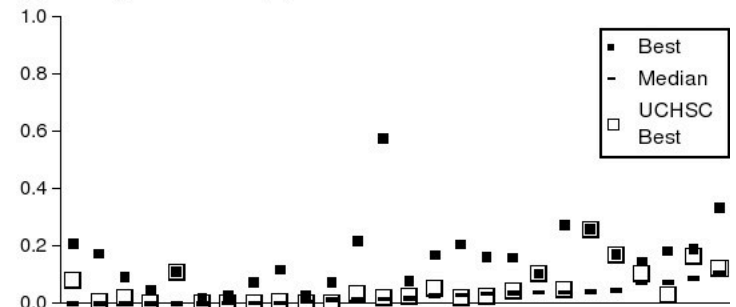
UCHSC results and conclusions

Above mean on each performance metric

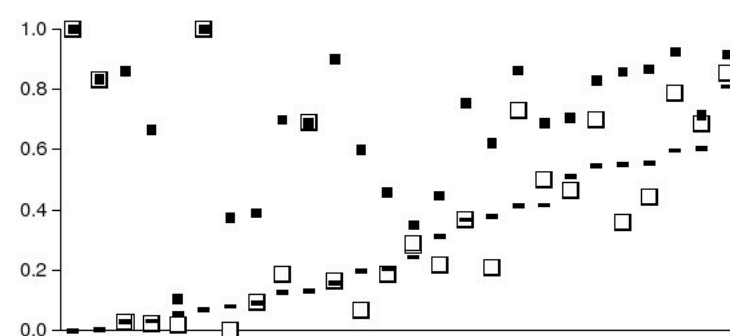
Performed well where median was low

Run	Aspect MAP	Document MAP	Passage MAP
uchsc1	0.250	0.406	0.055
uchsc2	0.247	0.419	0.056
uchsc3	0.247	0.404	0.054
overall mean	0.161	0.289	0.039

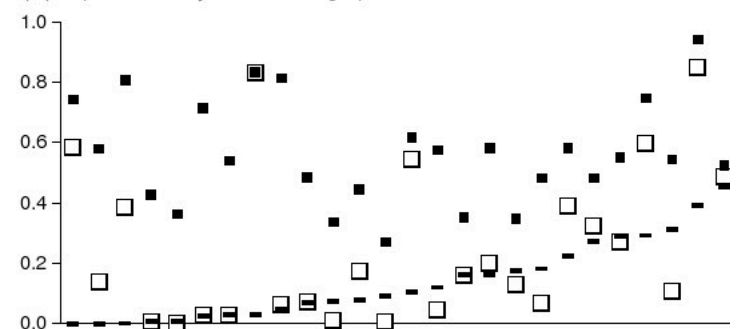
(A) Passage mean average precision



(B) Document mean average precision



(C) Aspect diversity mean average precision

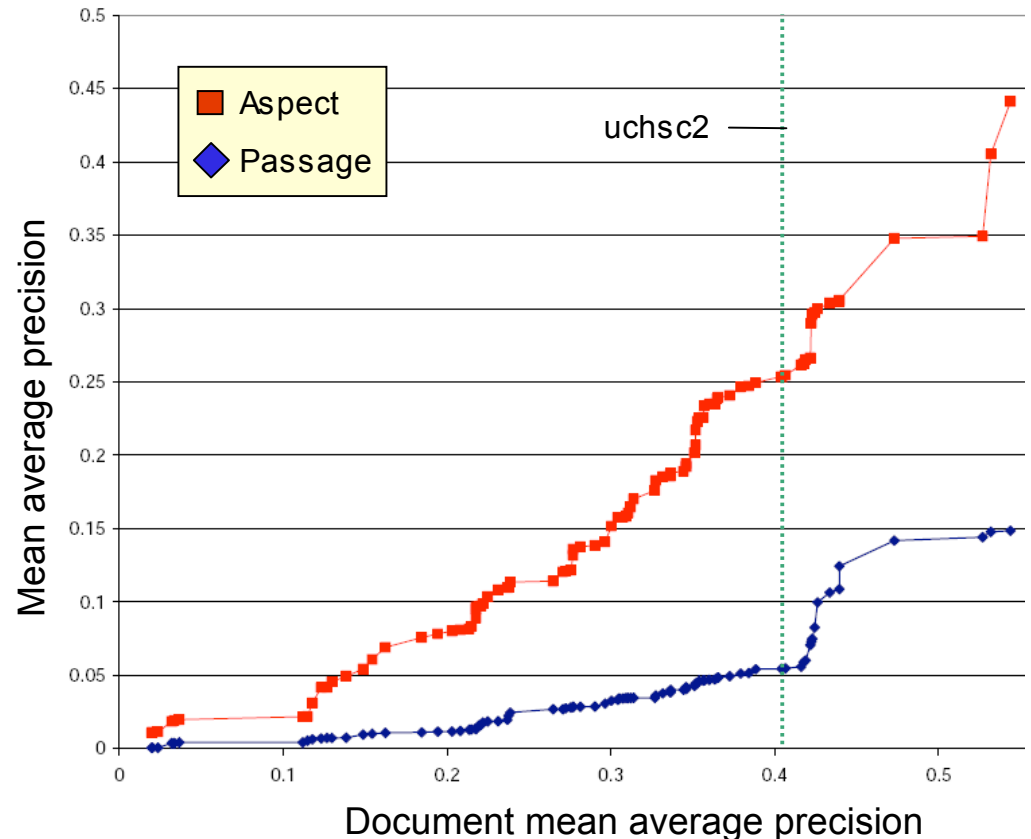


Genomics track results and conclusions

Best systems:
indexed paragraphs,
expanded queries,
selected sentences

Successful techniques
included concept-based
query expansion, acronym
expansion

Unsuccessful techniques
included cluster-based
reranking



Graph modified from: TREC 2006 Genomics Track Overview, W. He Cohen, P. Roberts, H.K. Rekapalli, TREC 2006 Notebook (68), Noven 2006.

Future Directions

TREC Genomics 2007: same corpus, very similar or identical task

Gold-standard will be released allowing for in-depth comparison of methods

TREC 2006 co-authors

Larry Hunter	Hyunmin Kim
Bill Baumgartner	Anna Lindemann
Kevin Cohen	Zhiyong Lu
Lynne Fox	Olga Medvedeva
Helen Johnson	Elizabeth White

Concept Recognition, Information Retrieval, and Machine Learning in Genomics
Question-Answering, J. G. Caporaso, W.A. Baumgartner, Jr., H. Kim, Z. Lu, H.L.
Johnson, O. Medvedeva, A. Lindemann, L. Fox, E.K. White, K. B. Cohen, L. Hunter.
TREC 2006 Proceedings (723), November, 2006.

gregcaporaso@gmail.com
<http://hsc.turing.uchsc>