Longer Queries & Shorter Documents

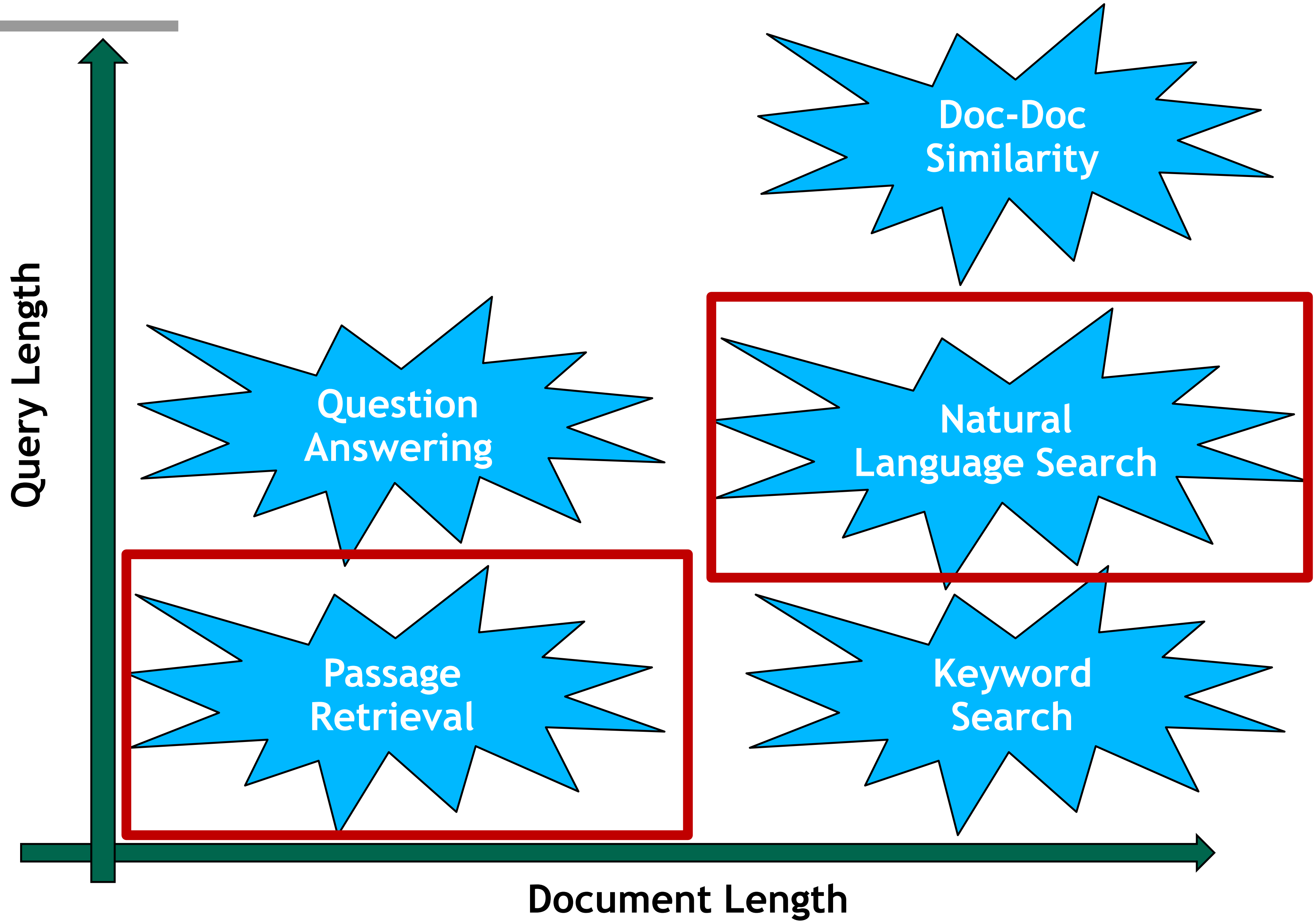
Michael Bendersky

Center for Intelligent Information Retrieval,

University of Massachusetts, Amherst

Joint Work with Bruce Croft, Oren Kurland

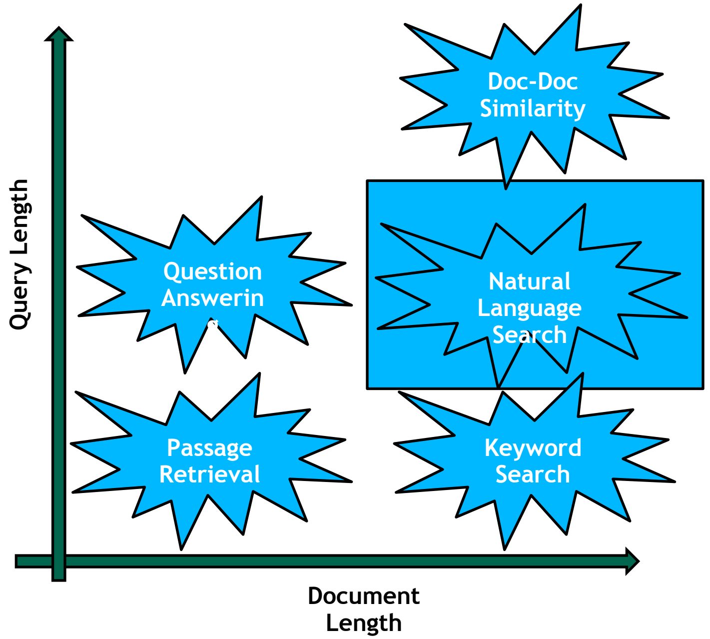
Doc & Query Length Continuum



Outline

- **Part I: Answering longer queries**
  - Background – Characteristics of long queries
  - Discovery of key concepts *(Bendersky & Croft, 2008)*

- **Part II: Leveraging shorter documents**
  - Background – Passage/Sentence Retrieval
  - Document–Passage Graphs *(Bendersky & Kurland, 2008)*



Part I: Answering Long Queries

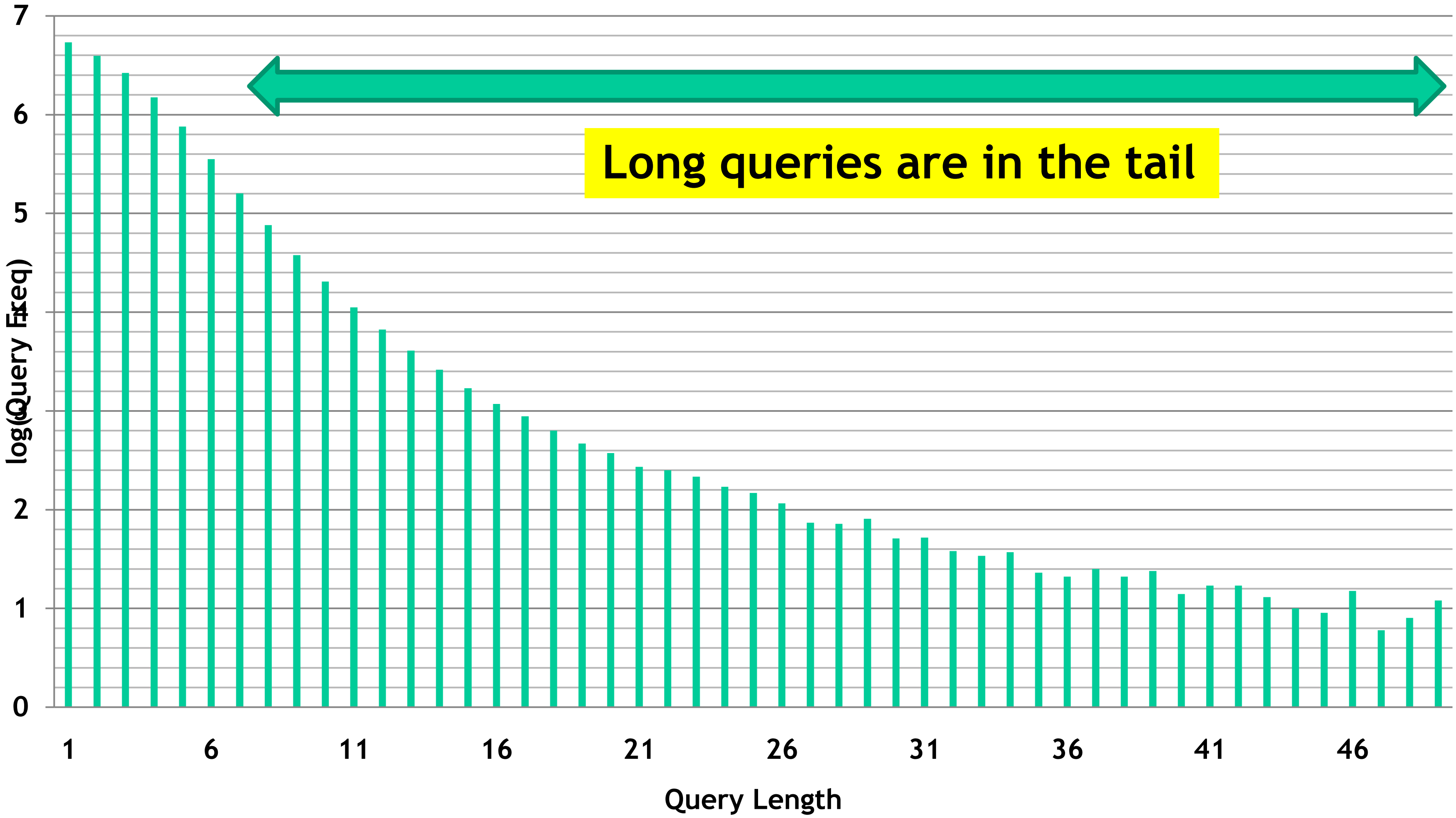
What is a long query?

- Natural language queries
- Questions from users in Q&A services
- Queries with more than one keyword or noun phrase from Web logs
- "Copy-Paste" queries: whole sentences or passages from documents

Query Length/Frequency

- **Length**
  - Q&A questions - more than 20 words
  - Web FAQs – about 9 words
  - TREC descriptions – 14-20 words average
- **Frequency**
  - Duplicates are rare
  - But, near duplicates or semantically similar queries are more common
    - In Q&A collections, 5-15 similar questions/query (found by pooling)

Long Tail



From a web query log excerpt (June, 2006)

Motivation

- Natural for some applications
  - e.g., Q&A, text reuse, professional/scholar
- May be the best way of expressing some information needs
  - i.e., perhaps selecting keywords is what is difficult for people (e.g., *SearchCloud.net*)
- Might become more widespread when search moves "out of the box"
  - e.g., speech recognition, search in context

Do Long Queries Work?

For people, yes; for search engines, no

- Long queries give generally poor, unpredictable results with current Web search engines
- Have sparser click-data than short queries
- TREC description queries don't work as well as title queries
- Searching Q&A archives is not very effective

Past Work on Long Queries

- **(Allan et al., 1997; Callan et al., 1995)**
  - Improving performance of long TREC queries
- **(Murdock & Croft, 2005; Balasubramanian et. al. 2007)**
  - Sentence Retrieval
- **(Kumaran & Allan, 2008)**
  - Interactive reduction/expansion of long queries

Discovering ***Key Concepts*** in ***Verbose Queries***

Michael Bendersky & Bruce Croft, SIGIR 2008

Introducing the problem

- A completely random TREC topic

<title> *Spanish Civil War Support*

<desc> *Provide information on all kinds of material international support provided to either side in the Spanish Civil War*

(Topic 829)

- How did three of the largest commercial web search engines do?

Introducing the problem [Cont.]

- For <*title*>
  - All results on the first results page refer to at least some aspect of the Spanish Civil War

<title>   Spanish Civil War Support

<desc>   Provide information on all kinds of material international support provided to either side in the Spanish Civil War

(Topic 829)

- For <*desc*>
  - *Six*, *three* and *one* results with at least some reference to ***Spanish Civil War*** in the top 10 results

Avoiding Morning Traffic in Seattle

Live Search™ how to avoid morning traffic in seattle

Web 1-10 of 1,090,000 results · Advanced
See also: Images, Video, News, Maps, More ▾

How to **Avoid Morning Traffic** to airport (Houston, West: travel, safe ...
I will be leaving Houston on a Friday **morning** during rush hour from Sam Houston Toll/Westpark Toll Westchase and I will be traveling to the airport (IA ...
www.city-data.com/forum/houston/390189-how-**avoid-morning-traffic**-airport.html · Cached page

Live Search™ "morning traffic" + seattle

Web 1-10 of 15,100 results · Advanced
See also: Images, Video, News, Maps, More ▾

Bus fire causes **morning traffic** jam | KOMO News - **Seattle** ...
SEATTLE -- A Metro Access bus that caught fire Thursday morning on Interstate 5 burned to the frame and caused a large traffic backup. The bus caught fire just before 8 a.m. in the ...
www.komonews.com/news/9899567.html · Cached page

Similar behavior for TREC data

	ROBUST04		W10g		GOV2	
	MAP	w/q	MAP	w/q	MAP	w/q
<title>	25.3	2.7	19.3	4.2	29.7	3.1
<desc>	24.5	8.3	18.6	6.4	25.3	6.1

- *<title> VS. <desc> queries on TREC corpora*
- *Mean Average Precision VS. Words Count Per Query*

Hypothesis

Identification of the key query concepts will have a (significant) positive impact on the retrieval performance for verbose queries

Hypothesis motivated

- Verbose queries tend to mix key (***Spanish Civil War***) and complementary (*material international support*) concepts

- Current retrieval techniques tend to treat these equally — potentially resulting in loss of focus on the main query topic(s)

Concept identification – *The ideal*

- Everything is a potential concept

(Bentivogli & Pianta, 2003)

  - Single words: *dog, cat*
  - Phrasal verbs: *catch up, come on*
  - Idioms: *break a leg, spend time*
  - Open compounds: *science fiction*
  - Named entities: *Spanish Civil War, Steve Jobs*
  - Free word combinations: *verbose queries*

Noun phrases as concepts

- In this work, we approximate concept identification by noun-phrase extraction
  - Reasonable approximation for the task at hand: nouns usually serve as query topics
  - Worked well in practice
  - Used in a previous work involving key phrases extraction
    - *Allan et al. (1997) – Core concepts in TREC queries*
    - *Hulth (2003) – Keywords in scientific abstracts*
    - *Yih et. al (2006) – Keywords for web advertisement*

Back to Topic 829

Provide information on all kinds of material international support provided to either side in the Spanish Civil War

Concept extraction

[information, kinds, material international support, side, Spanish Civil War]

Concept weighting principle

Assumption A

Each concept c_i can be assigned to one of the mutually exclusive classes

- ***KC*** *(key concepts class)*
- ***NKC*** *(non-key concepts class)*

Assumption B

A global function $h_k(c_i)$ indicates the confidence that concept c_i belongs to class ***KC***

Concept weighting principle

- Following assumptions, weight each query concept using the estimate

$$\hat{p}(c_i \mid q) = \frac{h_k(c_i)}{\sum_{c_i \in q} h_k(c_i)}$$

- That is, we ***rank query concepts***
- Concepts which have the highest confidence in membership in class ***KC*** are regarded as the best query representatives

Estimating $h_k(c_i)$

- As $h_k(c_i)$ is query-independent, we can
  a) Take an unsupervised approach to estimate it, e.g., use concept ***IDF***
  b) Try to "learn" it using a set of given concepts and their features

- What kind of features?
  - As $h_k(c_i)$ is query-independent, we can use any concept-related features

Collection-based features

cf(c_i) Concept frequency in the collection

idf(c_i) Concept IDF in the collection

ridf(c_i) Concept residual IDF in the collection

- *Actual IDF deviation from Poisson model prediction (Church & Gale, 1995)*

wig(c_i) Concept Weighted Information Gain

- *Information gain from a state where only average document is retrieved (Zhou & Croft, 2007)*

Collection-independent features

g_cf(c_i) Concept frequency in *Google n-grams*. Estimates concept frequency in a large web collection

l_qp(c_i) Number of times a concept was used as a part of a query, extracted from *Live Search* query logs

l_qe(c_i) Number of times a concept was used as an exact query, extracted from *Live Search* query logs

Query - Based

is_cap(c_i) Is concept capitalized in the query?

Collections

Collection	# Docs	# Topics
ROBUST04	528,155	250
W10g	1,692,096	100
GOV2	25,205,179	150

Concept classification task

- Task: identifying key concepts
- Simplifying assumption: single key concept per query is selected
- Train an *AdaBoost.M1* classifier on a set of labeled concept instances: $\boldsymbol{x_i} \in \{\boldsymbol{KC}, \boldsymbol{NKC}\}$
- Rank concepts for each query in the test-set according to their confidence in membership in class ***KC***

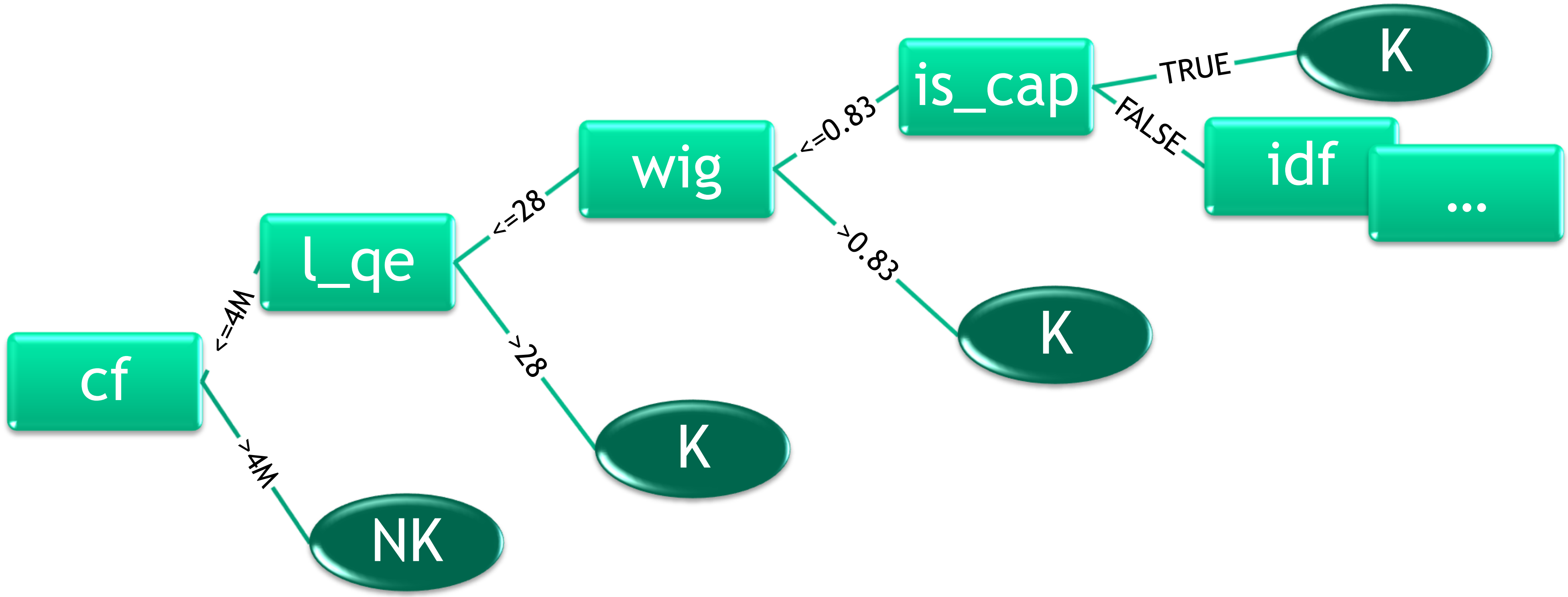
Concept classification results

	AdaBoost.M1		$idf(c_i)$	
	Accuracy	MRR	Accuracy	MRR
ROBUST04	**76.4**	**84.5**	56.4	74.2
W10g	**81.0**	**85.3**	66.0	78.6
GOV2	**84.0**	**88.9**	74.7	85.7

Accuracy and MRR results for cross-validation learning VS. using IDF estimate for $h_k(c_i)$ ***(with optimal features selection)***

CIIR Center for Intelligent Information Retrieval

What Makes a Key Concept?



[Example]:
- A high-weight decision tree for key concept classification in GOV2 collection

<u>So what does this say about retrieval</u>?

Does identifying key concepts (with a reasonable accuracy) help at all?

Does the concept weighting help?

Concept weighting for ranking

- Having estimated ***p(c$_i$|q)*** we may use a linear combination of query and all weighted concepts for ranking

Concept Weight

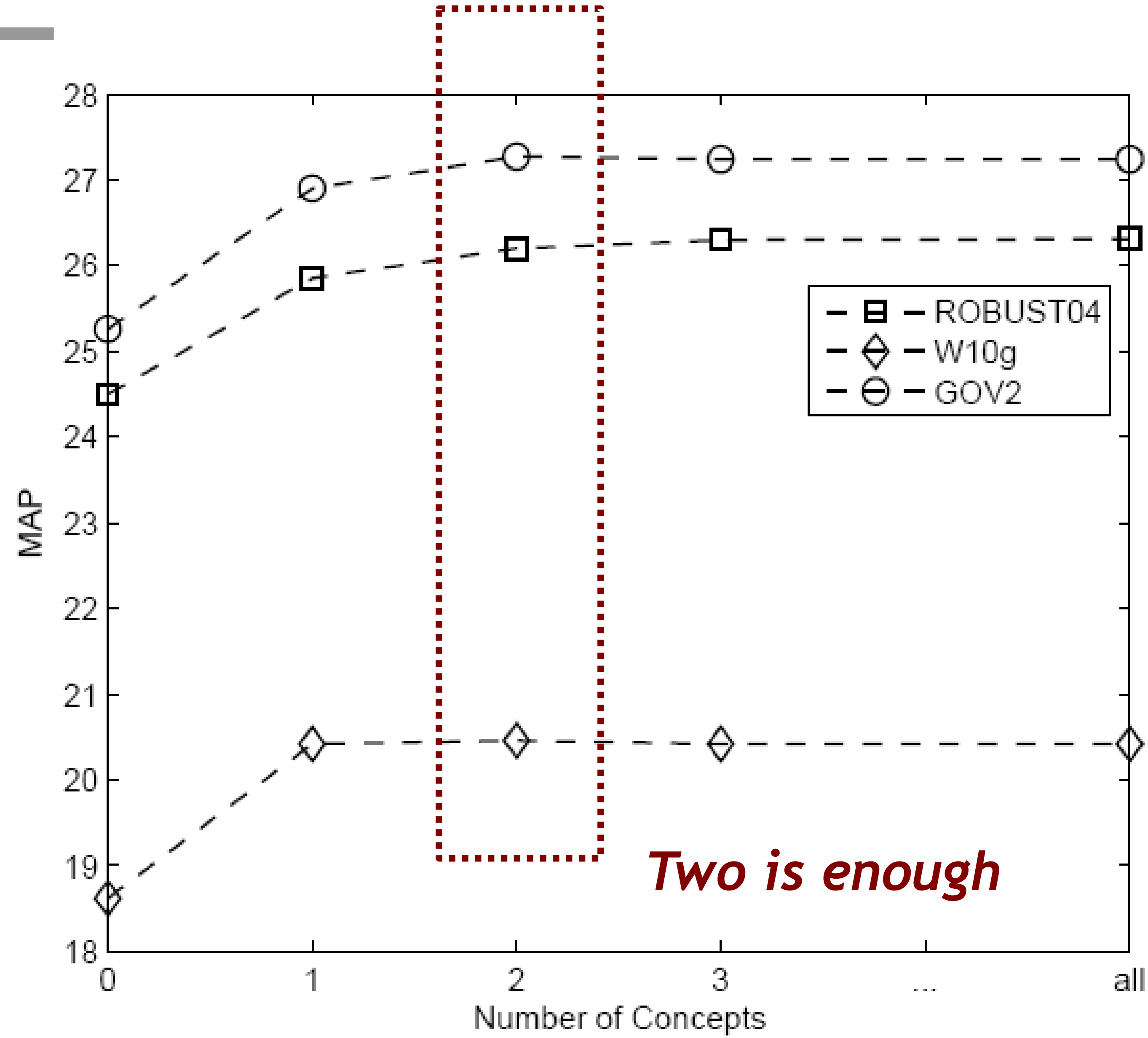
$$rank(d) \propto \lambda \log p(q \mid d) + (1-\lambda) \sum_{c_i \in q} \log p(c_i \mid d) p(c_i \mid q)$$

Query Score

Concept Score

CIIR Center for Intelligent Information Retrieval

How many concepts do we need?



```
#combine( Spanish Civil War support )
```

```
#combine( information kinds material international support provided side Spanish Civil War )
```

<title> and <desc> - #combine query

```
#weight(
  0.85 #combine( information kinds material international support provided side Spanish Civil War )
  0.10 #combine( #1(information kinds) #1(kinds material) #1(material international)
                 #1(international support) #1(support provided) #1(provided side)
                 #1(side Spanish) #1(Spanish Civil) #1(Civil War) )
  0.05 #combine( #uw8(information kinds) #uw8(kinds material) #uw8(material international)
                 #uw8(international support) #uw8(support provided) #uw8(provided side)
                 #uw8(side Spanish) #uw8(Spanish Civil) #uw8(Civil War) ))
```

<desc> - sequential dependency model query

(Metzler & Croft, 2005)

```
#weight(
  0.8 #combine( information kinds material international support
                provided side Spanish Civil War  )
  0.2 #weight( 0.99994 #combine ( Spanish Civil War )
               0.00006 #combine ( material international support )))
```

<desc> - key concepts-expanded query

Retrieval results

	ROBUST04	W10g	GOV2
	MAP	*MAP*	*MAP*
<title>	25.28	19.31	**$\underline{29.67}_d$**
<desc>	24.50	18.62	25.27^t
SeqDep<dep>	25.69_d	19.28	27.53^t_d
KeyConcept[2]<desc>	**$\underline{26.20}_d$**	**$\underline{20.46}^t_d$**	27.27^t_d

Query expansion by key concepts

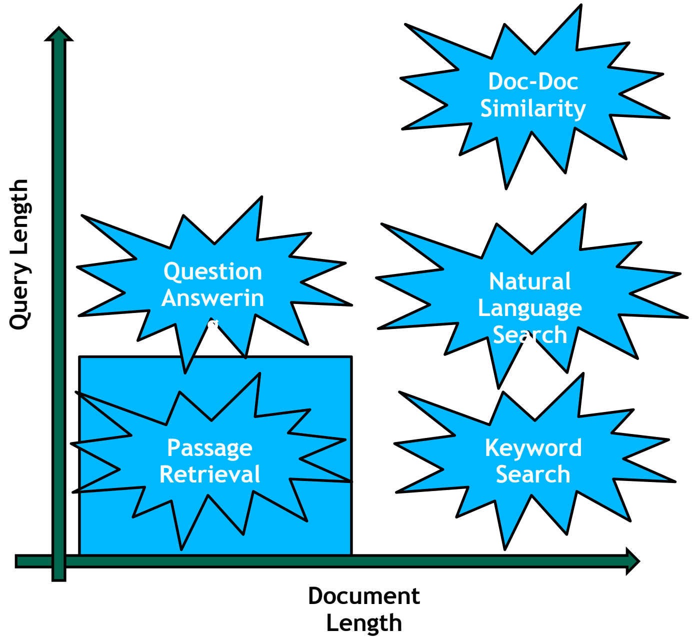
a) always outperforms the original description queries

b) comparable performance to SeqDep model

c) more efficient than SeqDep model

Conclusions

- Identifying key concepts in queries can be done with reasonable accuracy using supervised learning with very limited training data
- Query expansion by weighted concepts improves retrieval performance for verbose queries
- Resulting queries are efficient: on average, no more than 2 concepts are needed, even for very long queries



Part II: Leveraging Short Documents

Passage Retrieval

- The ***ad hoc document retrieval task*** is to rank documents in a corpus in response to a query
- ***Large and/or heterogeneous*** documents may pose a challenge for ad hoc retrieval
  - e.g., books, news-feeds, blogs
- ***Passage retrieval***
  - rank and return only a part of the document considered to be relevant

Challenges in Passage Retrieval

- **Higher number of judgments**
  - Judge each passage for relevance
- **Passage segmentation is a hard problem**
- **Retrieval of smaller text units**
  - Less statistical evidence to estimate relevance
  - Higher term mismatch

Leveraging Passages for Document Retrieval

- *(Callan, 1994; Cai et al., 2004; Ogilvie & Callan, 2004, Dang et al., 2007; Bendersky & Kurland, 2008)*
  - Interpolation of document and passage scores

$$Score_d = \lambda p(q \mid d) + (1 - \lambda) \max_{g_i \in d} p(q \mid g_i)$$

$$\underbrace{\lambda p(q \mid d)}_{\text{Document Score}} \qquad \underbrace{\max_{g_i \in d} p(q \mid g_i)}_{\text{Max-Scoring Passage}}$$

- *Higher scores for documents with a single relevant passage*
- *No need to judge each individual passage*

Re-ranking Search Results Using Document-Passage Graphs

Michael Bendersky & Oren Kurland, SIGIR 2008

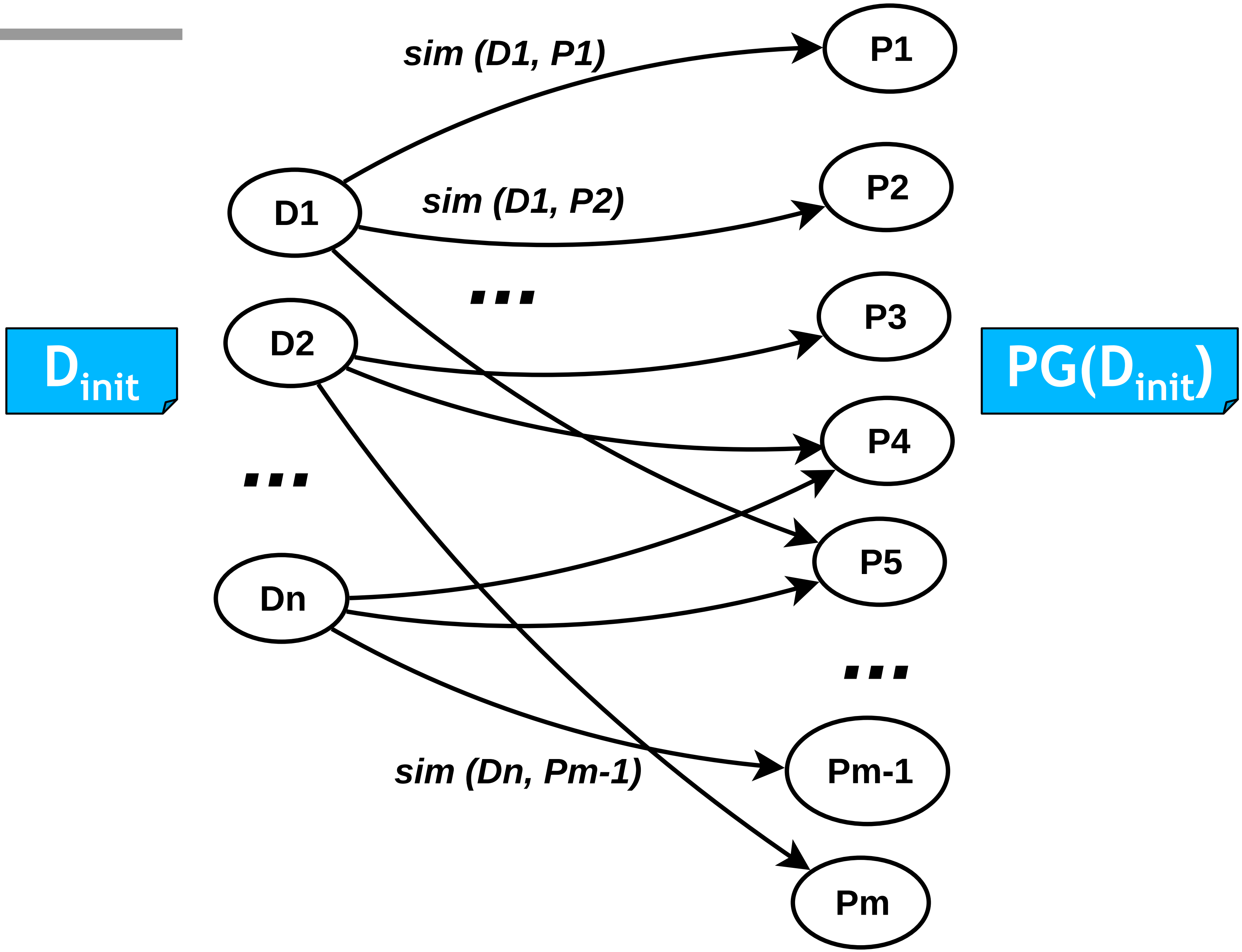
Motivation

- Existing techniques usually assume document independence in the ranking process
- But, context is important
  - Relevant **documents** tend to be similar to other relevant documents (“Cluster Hypothesis”)
  - Relevant **passages** tend to be similar to other relevant passages
  - Mutually reinforcing relation between the documents and the passages in the retrieved list

Graph-Based Re-ranking

- Retrieve initial list of documents D_{init}
- Extract a list of passages $PG(D_{init})$
- Build a weighted directed bipartite graph **G**
  - Weight of the edge ***wt(d,g)*** is a similarity between document ***d*** and passage ***g***
  - Retain only ***k*** edges with the highest weight, for each document
- Re-rank documents in the graph based on the **centrality** of their constituent passages

Document-Passage Graph



Re-ranking Algorithms

Method	Document Score
DocBase	$sim(q,d)$
PsgBase	$max_{gi \in d}\, sim(q,g_i)$
InterPsgDoc	$\lambda sim(q,d) + (1-\lambda) max_{gi \in d}\, sim(q,g_i)$
MultPsgDoc	$sim(q,d) max_{gi \in d}\, sim(q,g_i)$
Centrality	$\boldsymbol{sim(q,d) max_{gi \in d}\, Cent(g_i)}$
	influx $\quad Cent(g) = \Sigma_d\, sim(d,g)$
	authority $\quad Cent(g) = authority(g)$

Document Retrieval Evaluation

	AP		TREC8		WSJ	
	p@5	*p@10*	*p@5*	*p@10*	*p@5*	*p@10*
DocBase	45.7	43.2	50.0	45.6	53.6	48.4
PsgBase	46.1	41.7	44.8^{d}	43.0	48.8^{d}	44.6
InterPsgDoc	46.1	41.7	50.4^{p}	46.0^{p}	54.0^{p}	48.8^{p}
MultPsgDoc	45.3	43.4	49.6	46.4^{p}	52.8^{p}	47.8^{p}
influx	**50.7^{d}**	46.7^{dp}_{im}	55.2^{dp}_{im}	47.8^{p}	**55.6^{p}**	**50.8^{p}**
authority	50.3	47.3^{dp}_{i}	**55.6^{p}**	**48.2**	53.2	49.2

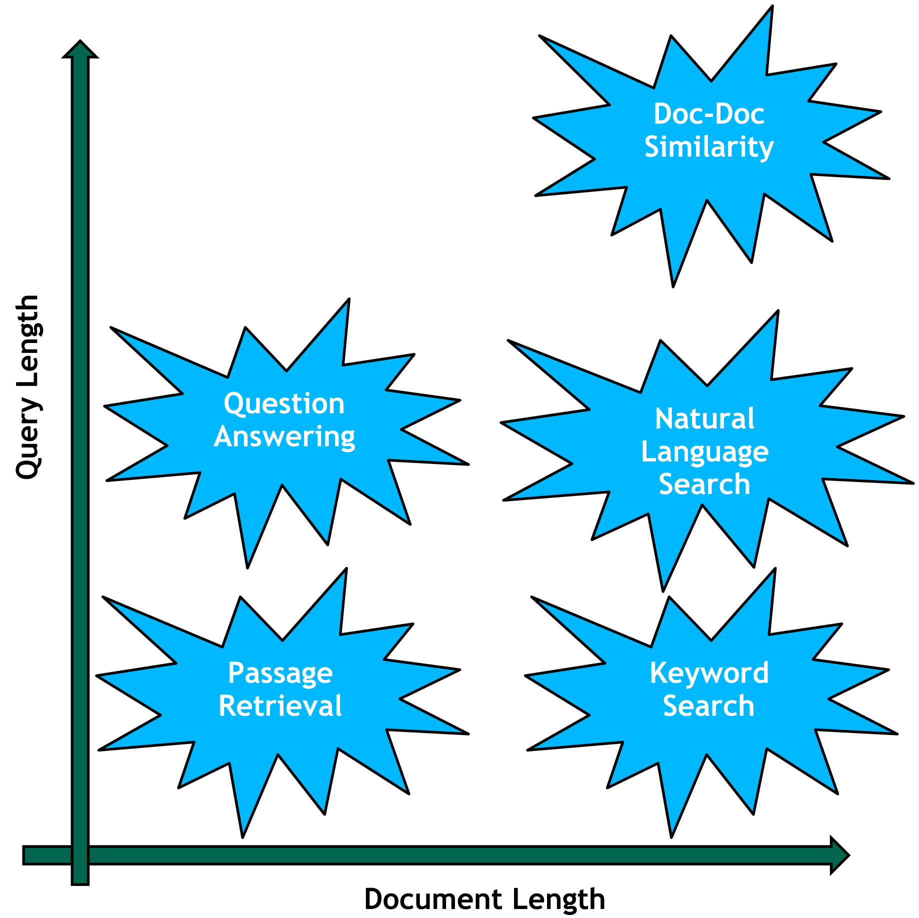
- **Centrality**
  - Always better than the initial ranking
  - Superior to other re-ranking techniques

Other possible benefits

- Harder to evaluate, but
  - Highest ranked passages are good snippets
  - Most central documents are good "diversity" results
  - Most central passages are good summaries for the retrieved list

Talk Summary

- A technique for discovering key concepts in verbose queries
- A technique for graph-based re-ranking using document-passage graph
- General introduction of various retrieval tasks on document-query continuum



THANK YOU