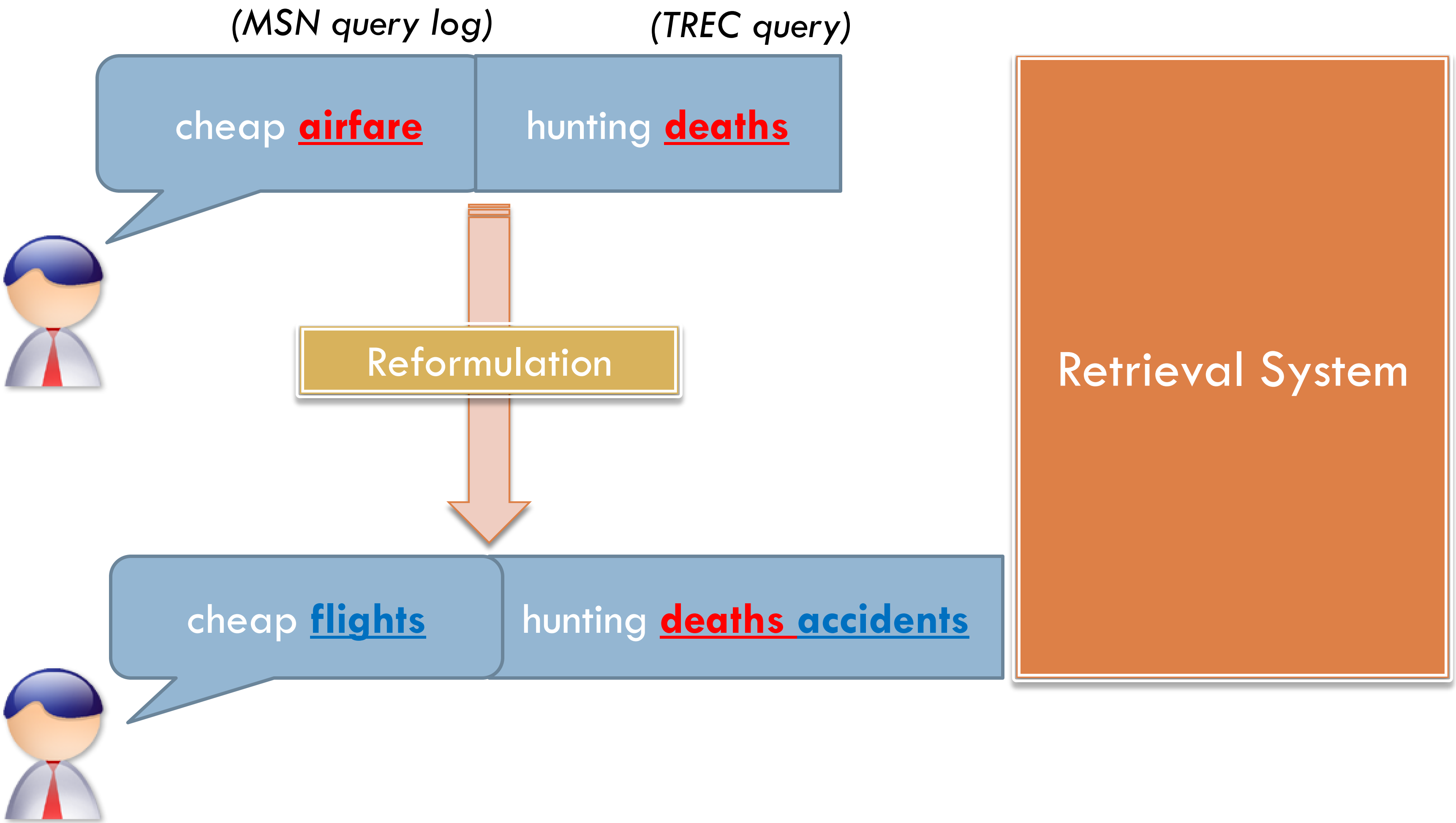
QUERY REFORMULATION USING ANCHOR TEXT

Van Dang and W. Bruce Croft

WSDM 2010

February 4th, 2010

Query Reformulation



(MSN query log)

(TREC query)

cheap **airfare**

hunting **deaths**

Reformulation

cheap **flights**

hunting **deaths** **accidents**

Retrieval System

Related work

- Relevance feedback
  - Well-known, not in the scope of this paper
- Recent reformulation techniques rely on query logs
  - [Jones et al., 06], [Wang and Zhai, 08]
  - These techniques have proven effective for real web queries
    - Many of these queries are badly formulated (“cheap airfare”)
  - What if queries are good? (e.g. “hunting deaths”)
    - Can these techniques still make them better?

Do these methods work with good queries?

Related Work

- Recent reformulation techniques rely on query logs
  - [Jones, 06], [Wang and Zhai, 08]
- And so do many other tasks
  - Spelling correction: [Cucerzan et al., 04], [Ahmad et al., 05]
  - Stemming: [Peng et al., 07]
- Query logs might not be available to research community
  - Any alternatives?
- <anchor text, url> is just like <query, clicked doc>.

Can we use anchor text to simulate a query log?

Introduction

Do these methods work with good queries?

- Using TREC collections to evaluate the most recent log-based reformulation technique [Wang and Zhai, 08] on three tasks
  - Query Substitution
  - Query Expansion
  - Query Stemming

Can we use anchor text to simulate a query log?

- Uses anchor text in place of a query log

The Anchor Log

- Extract <anchor, url> pairs from the Gov-2 collection to create the *anchor log*.

	MSN Log	Anchor Log
# Total Queries	14 million	526 million
# Unique Queries	6 million	20 million
Avg. Query Length	2.68	2.62

- The anchor log is very noisy
  - "click here", "print version", … don't represent the linked page

Query Substitution

- A context of a word is the unigram preceding it
- Context distribution

$$P(c_i \mid w) = \frac{count_w(c_i)}{\sum_{c_j \in C(w)} count_w(c_j)}$$

The probability that the term **c_i** appears in ***w***'s context

- The translation model

$$t(s \mid w) = \frac{e^{-D(P(.|w) \| P(.|s))}}{Z}$$

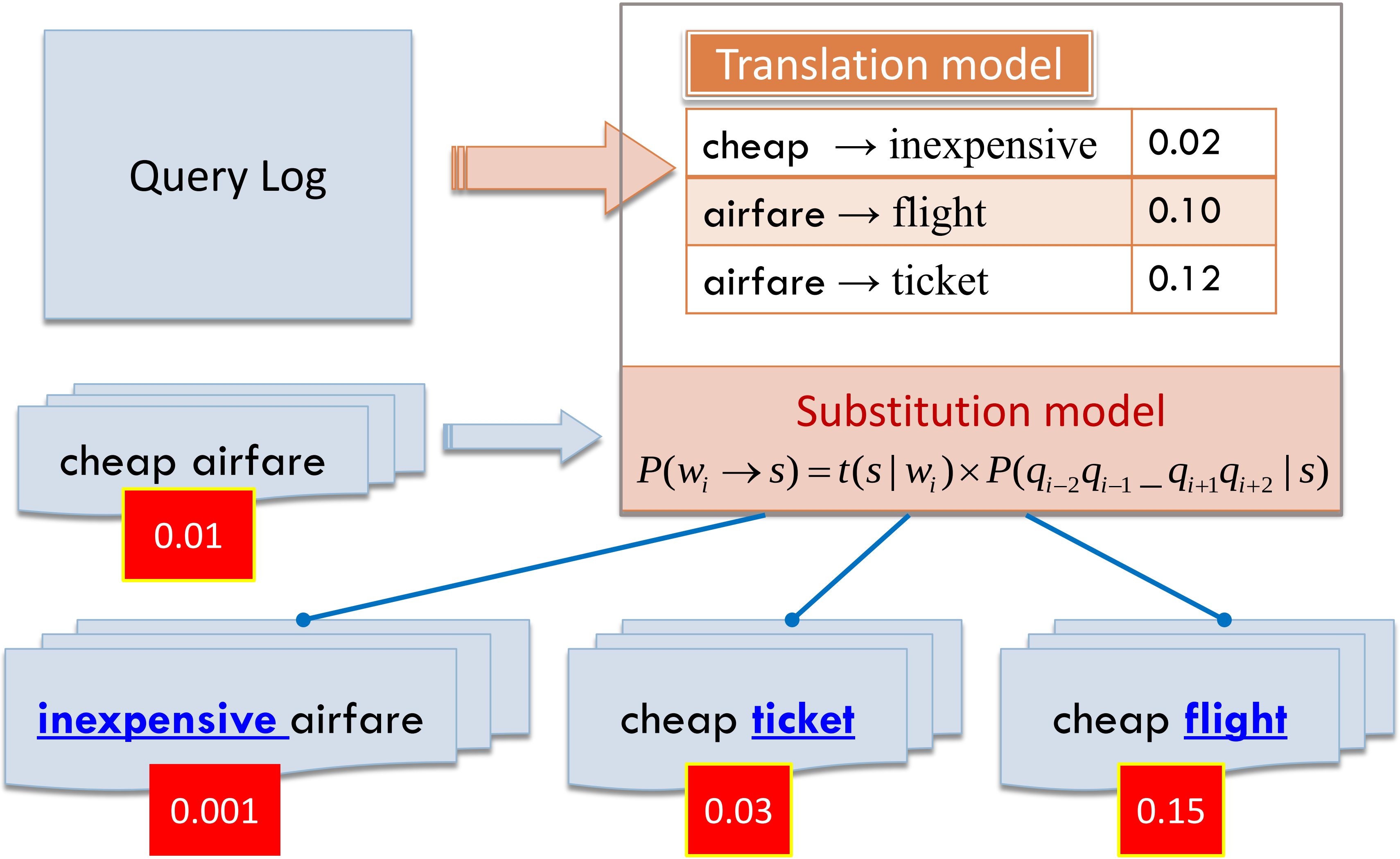
The KL divergence between the context distributions of **w** and **s**

- The substitution model
  - Q= q$_1$, … q$_{i-2}$, q$_{i-1}$, **q$_i$**, q$_{i+1}$, q$_{i+2}$, … q candidate = **s**

$$\boxed{P(w_i \to s) = t(s \mid w_i) \times P(q_{i-2}q_{i-1} \_ q_{i+1}q_{i+2} \mid s)}$$

How fit the new term is to the context of the current query

Substitution: An example



Query Log

Translation model

cheap → inexpensive	0.02
airfare → flight	0.10
airfare → ticket	0.12

cheap airfare

0.01

Substitution model

$$P(w_i \rightarrow s) = t(s \mid w_i) \times P(q_{i-2} q_{i-1} \_ q_{i+1} q_{i+2} \mid s)$$

inexpensive airfare

0.001

cheap **ticket**

0.03

cheap **flight**

0.15

Query Expansion and Stemming

- Query Expansion is exactly the same as substitution
  - We add the new term and keep the original term

    substitution: "*cheap airfare*" → "*cheap flight*"

    expansion: "*cheap airfare*" → "*cheap airfare flight*"

- Stemming
  - New terms are restricted to Porter-stemmed root terms

    "drive direction" → "*drive driving direction*"

Experimental Setup

- Evaluation
  - Conducted on three TREC collections:
    - Robust-04 (news)
    - WT10G (web)
    - Gov-2 (web)

Collection	# Documents	# Queries
Robust-04	0.5 M	250
WT10G	1.5 M	100
Gov-2	25 M	150

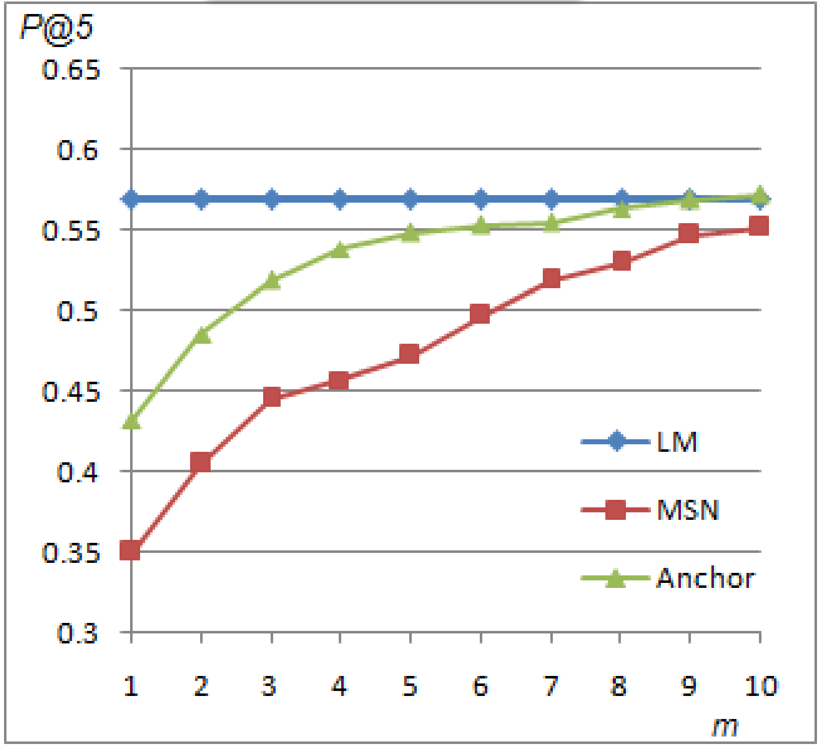
- Title queries vs. Description queries

Evaluation of Reformulated Query

Original Queries	MSN-Log Substitution	Anchor-Log Substitution
Query 1	Substitution 1 Substitution 2 … Substitution m	Substitution 1 Substitution 2 … Substitution m
…	…	…
Query n	Substitution 1 Substitution 2 … Substitution m	Substitution 1 Substitution 2 … Substitution m
P@5	P@5	P@5

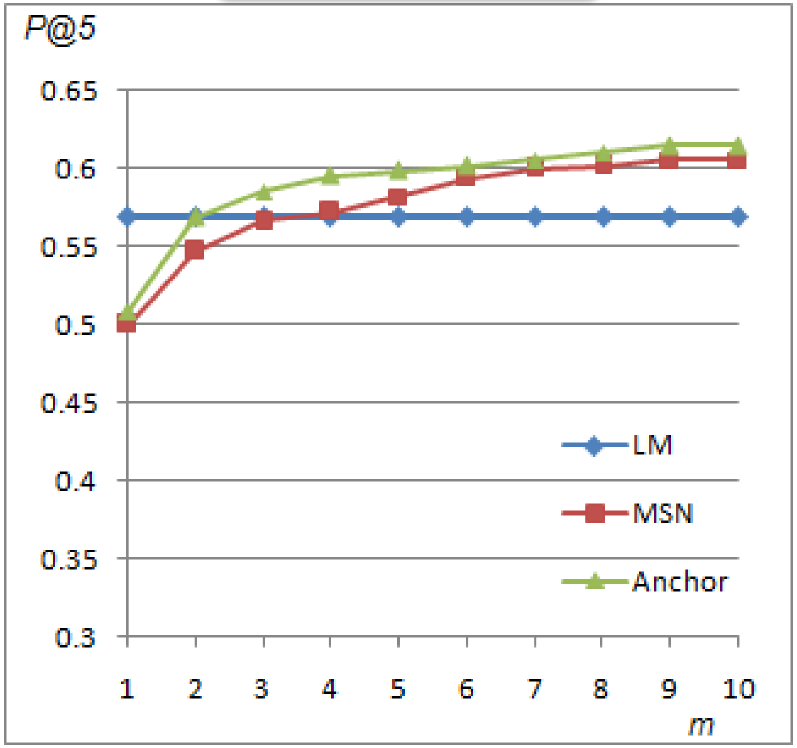
Substitution vs. Expansion (Title Q.)

Substitution



Does NOT help

Expansion



HELPS

The *Anchor log* is comparable to the *MSN Log*

“Chance” vs. “Risk”

- Substitution works for web queries [Wang and Zhai, 08]
  - Does not work here
  - Expansion is much better
  - **Why?**
- Both Substitution and Expansion
  - Introduce a new term to the query
    - “chance”: it brings more relevant documents
    - “risk”: it brings more non-relevant documents

“Chance” vs. “Risk”

- Results
  - Among 99 queries that were reformulated

	# Queries	P@5 change
Substitution helps	34	**+110.94%**
Expansion helps	32	+88.72%
Substitution hurts	**32**	-55.29%
Expansion hurts	14	-53.85%

Substitution

Helps substantially

Hurts drastically

Does NOT help in general

Expansion

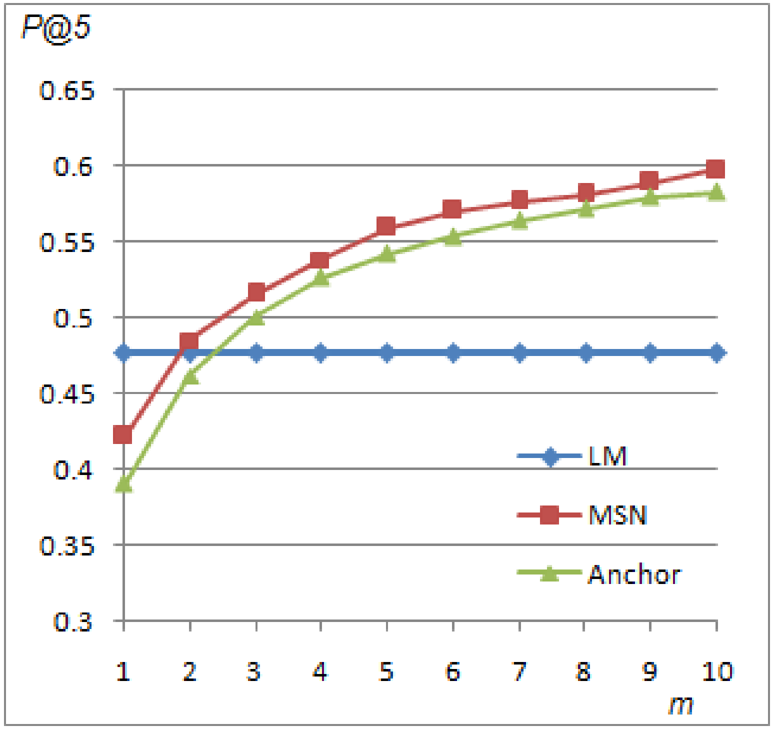
Helps more than it hurts, thus better

“Chance” vs. “Risk”

- Translation model does NOT provide « synonyms »
  - {women, men, children}
  - {diamond, gold, necklace, watches}
- It is undesirable to
  - “diamond smuggling” → “watches smuggling”
- TREC queries have good quality
  - Complete substitution is too risky

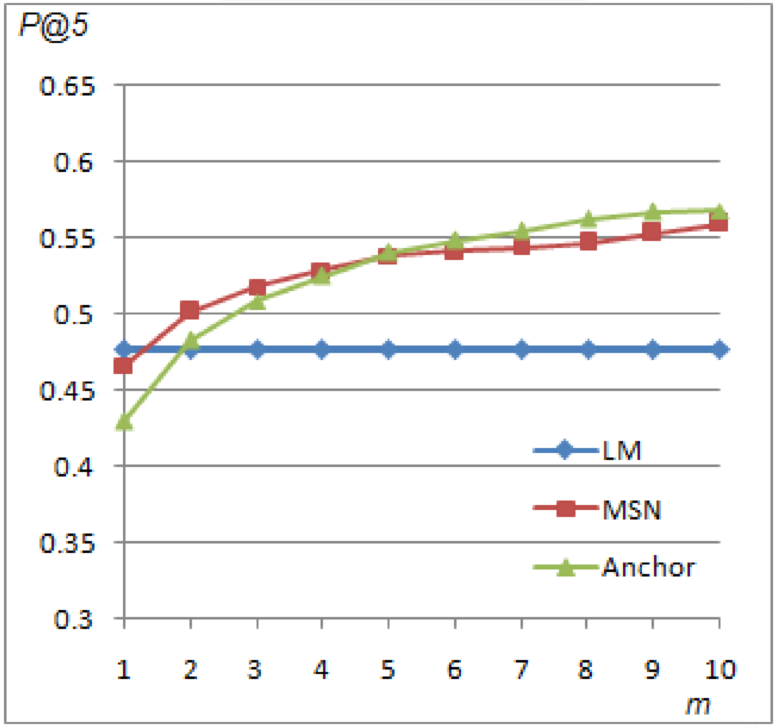
Substitution vs. Expansion (Desc Q.)

Substitution



Helps even more

Expansion

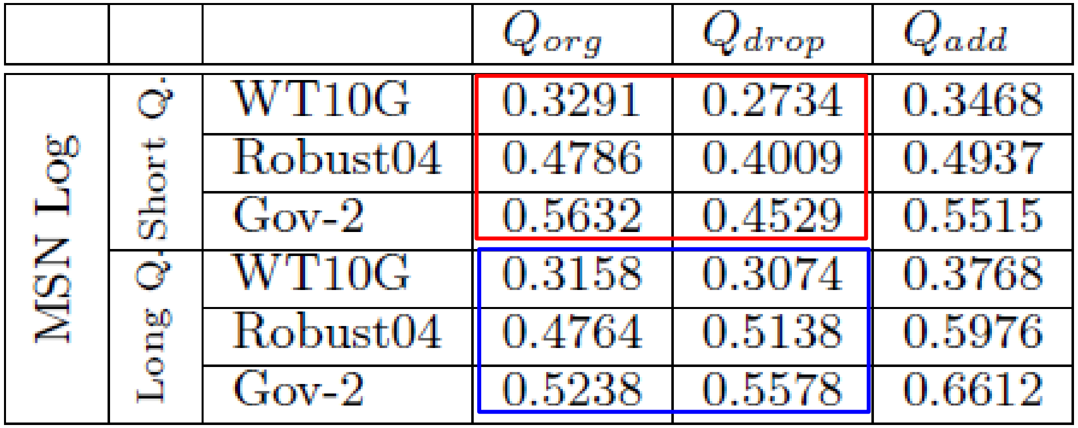


Helps

The *Anchor log* is comparable to the *MSN Log*

Substitution good for Long Query?

- Substitute w for s = drop w + add s
  - Q_{org}: original query
  - Q_{drop}: drop the target word
  - Q_{add}: add the substitution candidate



			Q_{org}	Q_{drop}	Q_{add}
MSN Log	Short Q.	WT10G	0.3291	0.2734	0.3468
		Robust04	0.4786	0.4009	0.4937
		Gov-2	0.5632	0.4529	0.5515
	Long Q.	WT10G	0.3158	0.3074	0.3768
		Robust04	0.4764	0.5138	0.5976
		Gov-2	0.5238	0.5578	0.6612

Dropping hurts

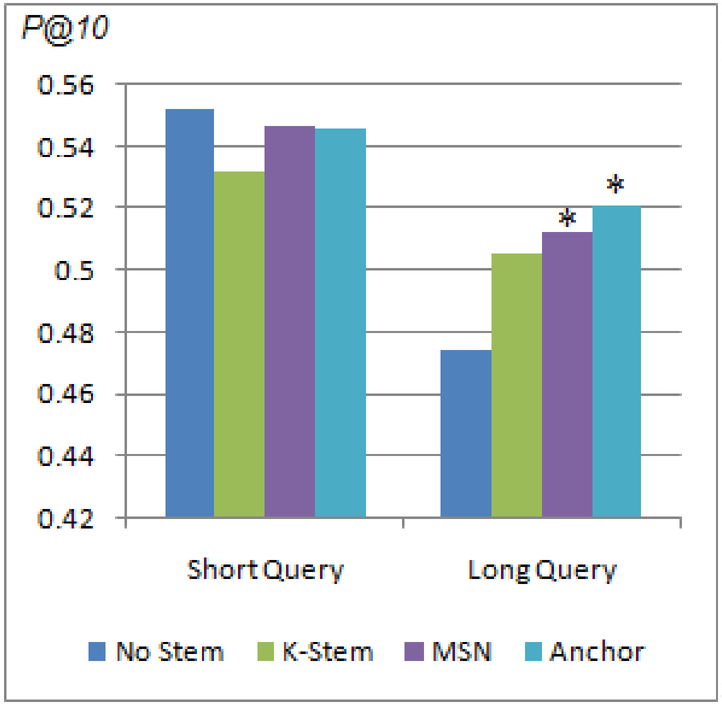
Dropping helps
[Kumaran et al., 09]

Similar improvement

It is the dropping that helps

Stemming

- We compare using P@10 queries
  - Unstemmed
  - Krovetz
  - Log-based (**MSN** vs. **Anchor** Log)



The *Anchor log* is comparable to the *MSN Log*

Conclusions

- **Anchor text gives comparable performance to MSN log** on
  - Substitution
  - Expansion
  - Stemming
- Expansion is more reliable than substitution
- Substituion helps with long (desc) queries
  - It is the dropping that helps
- Log-based stemming is promising