



Microsoft®  
**Research**

# Unsupervised Approaches to Syntactic Analysis of Web Search Queries

**Rishiraj Saha Roy**

Ph.D. Student

## **Supervisors**

Prof. Niloy Ganguly (IIT Kharagpur)

Dr. Monojit Choudhury (Microsoft Research India)

PhD Defence Seminar  
Computer Science and Engineering  
IIT Kharagpur

# Collaborators



Niloy Ganguly  
IIT Kharagpur



Monojit Choudhury  
Microsoft Research  
India



Srivatsan Laxman  
Microsoft Research  
India



Kalika Bali  
Microsoft Research  
India



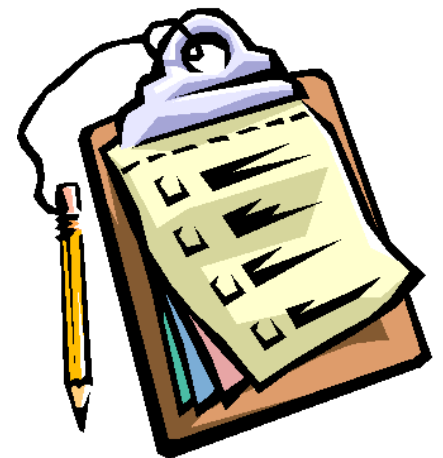
# Query Log Dataset

- 16.7 million search queries
- Each query accompanied by clicked URL and click count
- Other features like URL hash, page title also available
- Sampled from Bing Australia (<http://www.bing.com/?cc=au>)
- Collected in May 2010
- Acknowledgements: Anjana Das, Victor Das, Bhaskar Mitra  
(Microsoft India Development Center, Hyderabad)



# Overview

- Queries as a language – Idea and motivation
- Discovering syntactic units I: Flat segmentation
- Discovering syntactic units II: Nested segmentation
- Inducing roles for units: Content and intent
- Analyzing syntactic complexity
- Discussion on feedback
- Contributions and future directions



# Before we begin

- *“The tail end of unique terms is very long and warrants in itself a linguistic investigation. In fact, the whole area of query language needs further investigation. Such studies have potential to benefit IR system and Web site development.” – Jansen et al. (2000)*
- *“A small number of search terms are used with high frequency, and a great many terms are unique; the language of Web queries is distinctive.” – Spink et al. (2001)*



# Before we begin

- *“A modern expression of protolanguage can be observed in the use of search engines on the World Wide Web.” – Dessalles (2006)*
- *“It has been widely observed that search queries are composed in a very different style from that of the body or the title of a document... yet a large scale analysis on the extent of the language differences has been lacking.” – Huang et al. (2010)*



# Information needs to queries

- **How do I hide the network icon from the status bar?**
- **How many litres are there in a gallon?**
- **What are the available grants for setting up a business?**
- **What is the recipe for sweet green tomato pickles?**
- **Where can I buy an MS office guide book online?**



# Drop “unimportant” terms

- How do I hide the network icon from the status bar?
- How many litres are there in a gallon?
- What are the available grants for setting up a business?
- What is the recipe for sweet green tomato pickles?
- Where can I buy an MS office guide book online?





# Reorder words

- **hide** network icon status bar →

network **hide** icon status bar

- **sweet** green tomato pickles →

green tomato pickles **sweet**

- **buy** ms office guide book online →

ms office guide book **buy** online





ms office guide book buy online



26 Feb 2014

2020,00,000 RESULTS Narrow by language ▾ Narrow by region ▾

**Microsoft Office - Microsoft Word, Outlook & Excel ...**

[office.microsoft.com/en-in](http://office.microsoft.com/en-in)

Office Downloads · Try Office 365

Free **Microsoft Office** clip art, images, templates, how-to articles, downloads, help, and training for **Microsoft Office** Word, ... OFFICE ONLINE &nbsp;...

Non-relevant

**Microsoft® Office Specialist Study Guide Office 2003 ...**

[www.microsoft.com/learning/en-us/book.aspx?id=7389](http://www.microsoft.com/learning/en-us/book.aspx?id=7389)

23-06-2004 · Take an **online** skills test ... as well as dozens of **books** in the Quick Course ... **Microsoft Office Specialist Study Guide: 2007 Microsoft Office ...**

Relevant

**Office - Office.com**

[office.microsoft.com](http://office.microsoft.com)

Try or **buy Office** 365 for Home or Business, ... and on the web with **Office Online** for everywhere in between. ... **Microsoft Store**; Follow us. **Office Blogs**; Twitter;

Non-relevant

**Microsoft Press Books**

[www.microsoft.com/learning/en-us/microsoft-press-books.aspx](http://www.microsoft.com/learning/en-us/microsoft-press-books.aspx)

... and **online books** allow you to access the information you need to learn ... Training **Guide books**. ... Review our proposal guidelines in **Microsoft Office Word ...**

Relevant

**Buy MS Office 2010 Training Guide by : MS Office 2010 ...**

[www.infibeam.com/Books/office-2010-training-guide-s-jain/...](http://www.infibeam.com/Books/office-2010-training-guide-s-jain/)

**Buy MS Office 2010 Training Guide book by online. MS Office 2010 Training Guide book price, MS Office 2010 Training Guide reviews & ratings, ISBN: 8183334068, EAN ...**

Relevant

**Buy Microsoft Office 2013 suites and Office 365 ...**

[office.microsoft.com/en-us/buy](http://office.microsoft.com/en-us/buy)

OFFICE ONLINE &nbsp;... Choose ... **Buy Office** for Business . Compare **Office** suites; Common questions; Free **Office** trial; System requirements; **Office 365** University. At ...

Non-relevant



# Motivation for Research

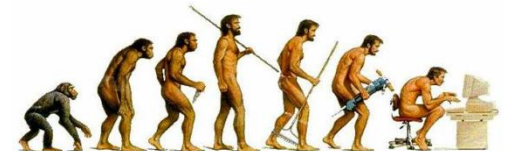
- Understanding queries as a language may add a fresh perspective to existing methods in query interpretation
- Suitable for improving IR on rare, long and “complex” queries
- Perfectly preserved dataset for studying language evolution
- Millions of global users, without direct interaction, developing a mode of communication with unique properties – Interesting!!



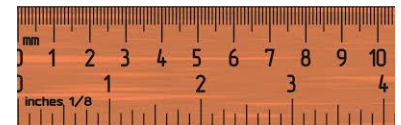
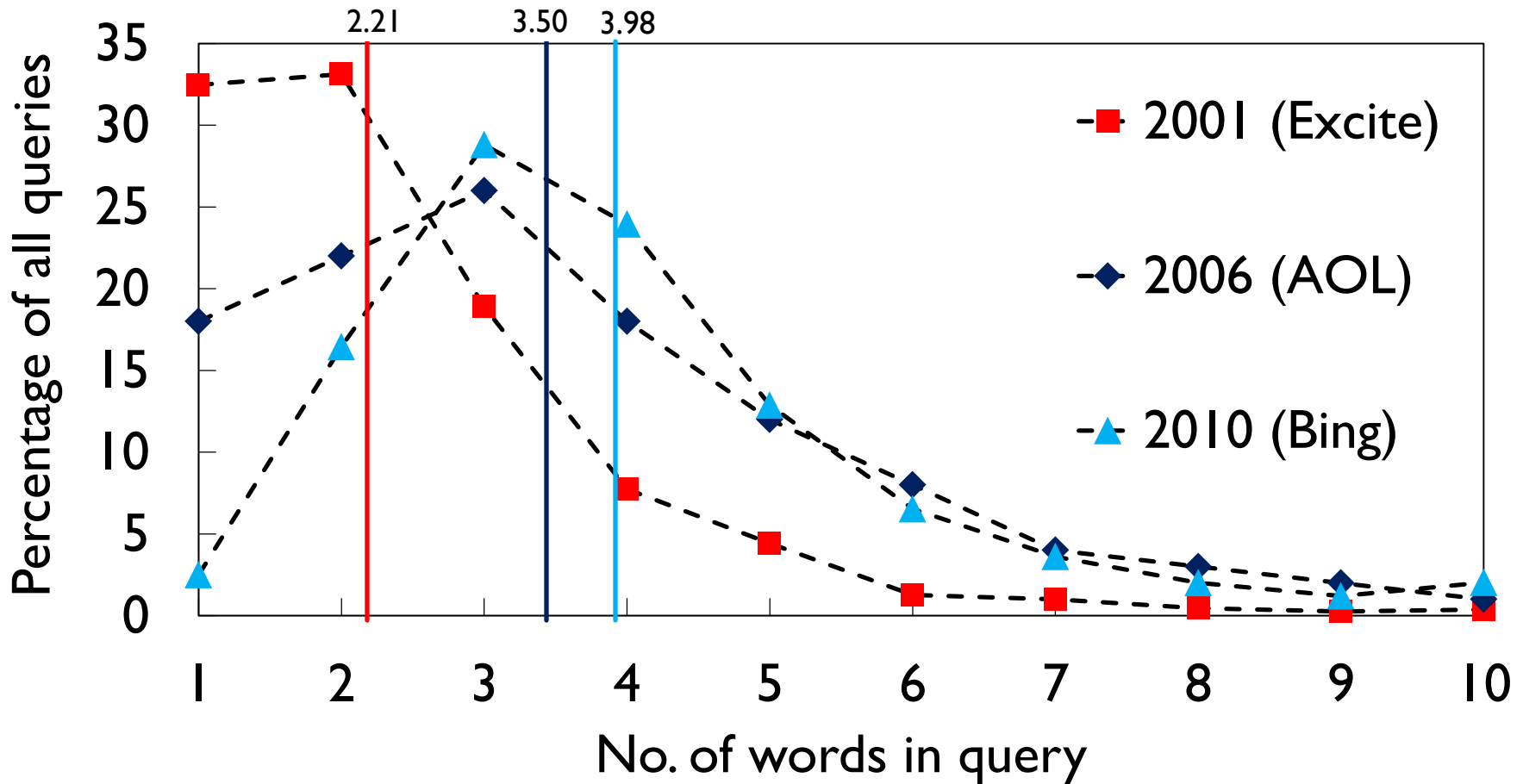
# A New Language?

- Three dimensions of analysis: Structure, function and dynamics
- Structure – Query syntax differs from parent NL
- Function – Satisfy eleven out of thirteen design features of spoken natural languages (Hockett, 1960)
- Dynamics – Continuous two-way interactions leading to more complex needs (user), model and algorithmic development (SE)

[BEST RESEARCH POSTER AWARD] R. Saha Roy, M. Choudhury and K. Bali, “Are Web Search Queries an Evolving Protolanguage?”, in *Proc. of the 9th International Conference on the Evolution of Language 2012 (Evolang IX)*, 13 – 16 March 2012, Kyoto, Japan, pp. 304-311.



# Query lengths



# Google introduces Hummingbird

- *“Google has overhauled its search algorithm to better cope with the **longer, more complex queries** it has been getting from Web users... need to match **concepts** and meanings in addition to words. ... The world has changed so much since then: billions of people have come online, the Web has grown exponentially.”*

– Reuters, 27 September 2013

<http://goo.gl/dgVK3l>



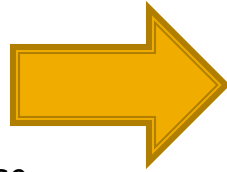
# Research Objectives – I

- How can we identify the syntactic units of queries?
- How can we utilize the knowledge of query structure to improve information retrieval?
- Identifying basic syntactic units from a corpus is the first step to understanding the language!!



# Query Segmentation

2 *large* islands in the atlantic  
i need a crack for *windows 7*  
5 bedroom accommodation hobart *large* family  
*cant view* videos on youtube  
finding a word in *text files*  
american history in 1880 *large* disappearances  
increasing usable ram on *windows 7*  
convert *text files* to pdf format  
movie from rapidshare *cant view*  
improvements in *windows 7* official release  
3gs mobile phone bluetooth *large* screen  
dead space *text files* walkthrough ps3  
advantages of living in *large* cities  
*cant view* high resolution videos  
import multiple *text files* into access  
import video from camcorder *windows 7*  
facebook photos *cant view*  
index dat files in *windows 7*  
why *cant view* friends profiles  
how to compare 2 *text files*





**Candidates extracted and used to  
segment query:**

*cant view* | *large* | *text files* | *windows 7*





# Flat Query Segmentation

- Word level analysis is often insufficient
- Dividing a query into individual syntactic units (Li et al. 2011)
  - Non-overlapping sequences of words
- Example
  - *history of all saints church south australia* →
  - *history of | all saints church | south australia* 
  - *history of all | saints church south | australia* 



# Flat Query Segmentation

- Problem akin to NL chunking – but no POS and grammar
- Past methods relied on some form of document statistics (Tan and Peng 2008, Li et al. 2011, Hagen et al. 2012,)
- Only query logs: Discover distinct query syntax
- We do not wish to project NL document structure
- Goes beyond multiword named entity recognition
- Can improve IR precision, query suggestion



# Flat Query Segmentation Proposed Algorithm

- **Intuition:** If *leonardo da vinci* is a segment, then a query having *leonardo* and *da* and *vinci* will most likely contain *leonardo da vinci* together

*leonardo da vinci oil paintings* - FREQUENT

*leonardo di caprio in da vinci code* - RARE

- Differs from past MWE detection approaches – does not depend upon frequencies of constituent unigrams



# Flat Query Segmentation

## Proposed Algorithm

- Compute the expected probability of observing an  $n$ -gram in queries that contain all its words
- Compute the observed probability from query log
- Does observed probability significantly exceed expected probability?
- If yes,  $n$ -gram is a candidate MWE, associated with a score
- Build score lexicon of candidate segments



# Flat Query Segmentation

## Proposed Algorithm

- $X_i$  is the indicator variable for “ $n$ -gram  $M$  occurs in query  $q_i$ ”
- $P_i = P(X_i = 1) = \frac{(\ell_i - n + 1) \times (\ell_i - n)!}{\ell_i!} = \frac{(\ell_i - n + 1)!}{\ell_i!}$
- Assumes bag-of-words null model
- $X = \sum_i X_i$
- $Prob[X \geq N] \leq \exp\left(-\frac{2(N - E(X))^2}{k}\right) = \delta$
- $E(X) = \sum_i P_i$
- $Score(M) = -\log_e \delta$



# Flat Query Segmentation

## Proposed Algorithm

- Optimize total score over query – dynamic programming to search over all possible segmentations
- Finds segments like *gprs config, history of, how to, where do i, how do i, spot a fake, buy online*
- Problem of long named entities – enhance with Wikipedia titles

N. Mishra, R. Saha Roy, N. Ganguly, S. Laxman and M. Choudhury, "Unsupervised Query Segmentation Using only Query Logs", in *Posters of the 20th International World Wide Web Conference 2011 (WWW '11)*, 28 March - 1 April 2011, Hyderabad, India, pages 91 – 92 (companion).



# Flat Query Segmentation

## Past Evaluation

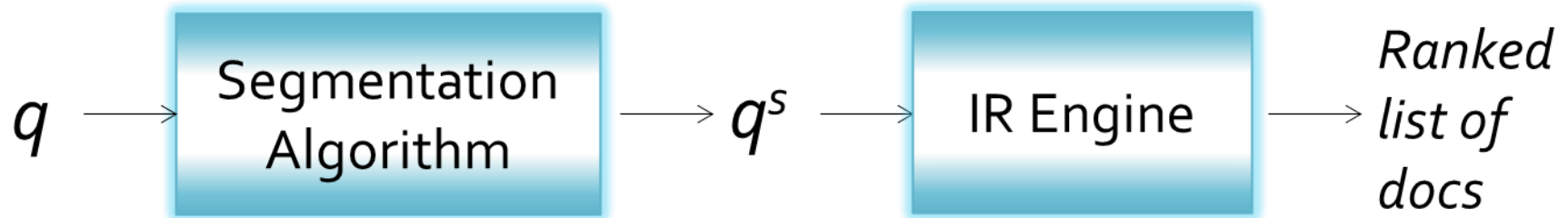
- An algorithm segments each query in test set
- A segmented query is matched against a human annotated query using five matching metrics (Bergsma and Wang 2007, Zhang 2009, Brenes 2010, Hagen et al. 2011)
- Low inter-annotator agreement on most metrics ( $\approx 70\%$ ) (Tan and Peng 2008)
- Not clear what should be the guidelines



# Flat Query Segmentation

## Proposed Evaluation

- Humans may not be the best judge as to which segments are best for IR – Humans are not the end users of segmentation!!



- End user of segmentation is the search engine
- An IR performance based evaluation
- How to use segmented query for retrieval??





# Flat Query Segmentation

## Proposed Evaluation

- Different segments need to be matched differently in documents  
*cannot view | word files | windows 7*
- Ordered (*windows 7*)
- Unordered (may have linguistic constraints) (*files in word*)
- Insertions, deletions, transpositions, substitutions (*cannot properly view*)
- MRF models of term dependence (Metzler and Croft, 2005)
- Some segments need not be matched (*view online, cheap, near*)



# Flat Query Segmentation

## Proposed Evaluation

- Current IR engines do not support these specifications
- Most retrieval systems support use of double quotes (exact match)
- However, simply putting double quotes around all query segments results in very poor retrieval performance!!
- Hagen et al. (2011) explore an evaluation with quotes around all segments, effective only for MWEs and negatively affecting overall results



# Flat Query Segmentation

## Proposed Evaluation

- We adopt a less constrained approach
- For each segmentation algorithm output, we generate all quoted versions of segmented query  $q^s$  (each segment can be quoted or unquoted)
- $2^k$  quoted versions for a  $k$ -segment query

SKIP TO RESULTS

R. Saha Roy, N. Ganguly, M. Choudhury and S. Laxman, "An IR-based Evaluation Framework for Web Search Query Segmentation", in *Proc. of the 35th Annual ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR '12)*, Portland, USA, 12 - 16 August 2012, pp. 881 – 890.



# Flat Query Segmentation

## Proposed Evaluation

Segmented query	Quoted versions
	<i>history of all saints church south australia</i>
	<i>history of all saints church “south australia”</i>
	<i>history of “all saints church” south australia</i>
<i>history of   all saints church   south australia</i>	<i>history of “all saints church” “south australia”</i>
	<i>“history of” all saints church south australia</i>
	<i>“history of” all saints church “south australia”</i>
	<i>“history of” “all saints church” south australia</i>
	<i>“history of” “all saints church” “south australia”</i>

# Flat Query Segmentation

## Proposed Evaluation

- Each version issued through IR engine (after query versions are deduplicated)
- IR system retrieves top  $k$  pages for each quoted version of a query
- Measure performance (eg. nDCG) of each quoted version (using human relevance judgments)



# Flat Query Segmentation

## Proposed Evaluation

Segmented query	Quoted versions	Score
	<i>history of all saints church south australia</i>	0.723
	<i>history of all saints church "south australia"</i>	0.788
	<i>history of "all saints church" south australia</i>	0.801
<i>history of   all saints church   south australia</i>	<i>history of "all saints church" "south australia"</i>	<b>0.852</b>
	<i>"history of" all saints church south australia</i>	0.632
	<i>"history of" all saints church "south australia"</i>	0.645
	<i>"history of" "all saints church" south australia</i>	0.652
	<i>"history of" "all saints church" "south australia"</i>	0.619

# Flat Query Segmentation

## Proposed Evaluation

- **Use of Oracle:** Highest nDCG from all quoted versions chosen as score achieved by  $q^s$
- Reflects “potential” of a segmented query
- Directly correlates to goodness of segmentation algorithm



# Flat Query Segmentation

## Proposed Evaluation

- For each algorithm, compute average oracle score over all queries
- **Find gold standard for IR performance:** Also perform brute force exhaustive search over all possible quoted versions of a query to find the one with the highest score
- Call it the best quoted version (BQV (BF)) of a query, irrespective of any segmentation algorithm
  - $2^{n-1}$  quoted versions for an  $n$ -word query





# Flat Query Segmentation

## Proposed Evaluation

- Search engine that supports double quotes (e.g. Lucene)
- Test set of queries
- Document pool
- Query relevance sets (*qrels*): For each query, human relevance judgments for the subset of documents in the pool possibly relevant to the query
- These resources are required for any IR-system evaluation



# Flat Query Segmentation

## Proposed Evaluation

- Query test set
- 500 test queries (5-8 words) sampled from Bing Australia
- Document collection (~15,000 in number)
- All possible quoted versions of a test query are issued through the Bing API 2.0
- Top 10 URLs retrieved are deduplicated and added to collection



# Flat Query Segmentation

## Proposed Evaluation

- Relevance judgments
- For each query, three sets of relevance judgments obtained for each URL retrieved for the query
- Much higher agreement on relevance judgments than human segment boundaries



# Flat Query Segmentation

## Proposed Evaluation

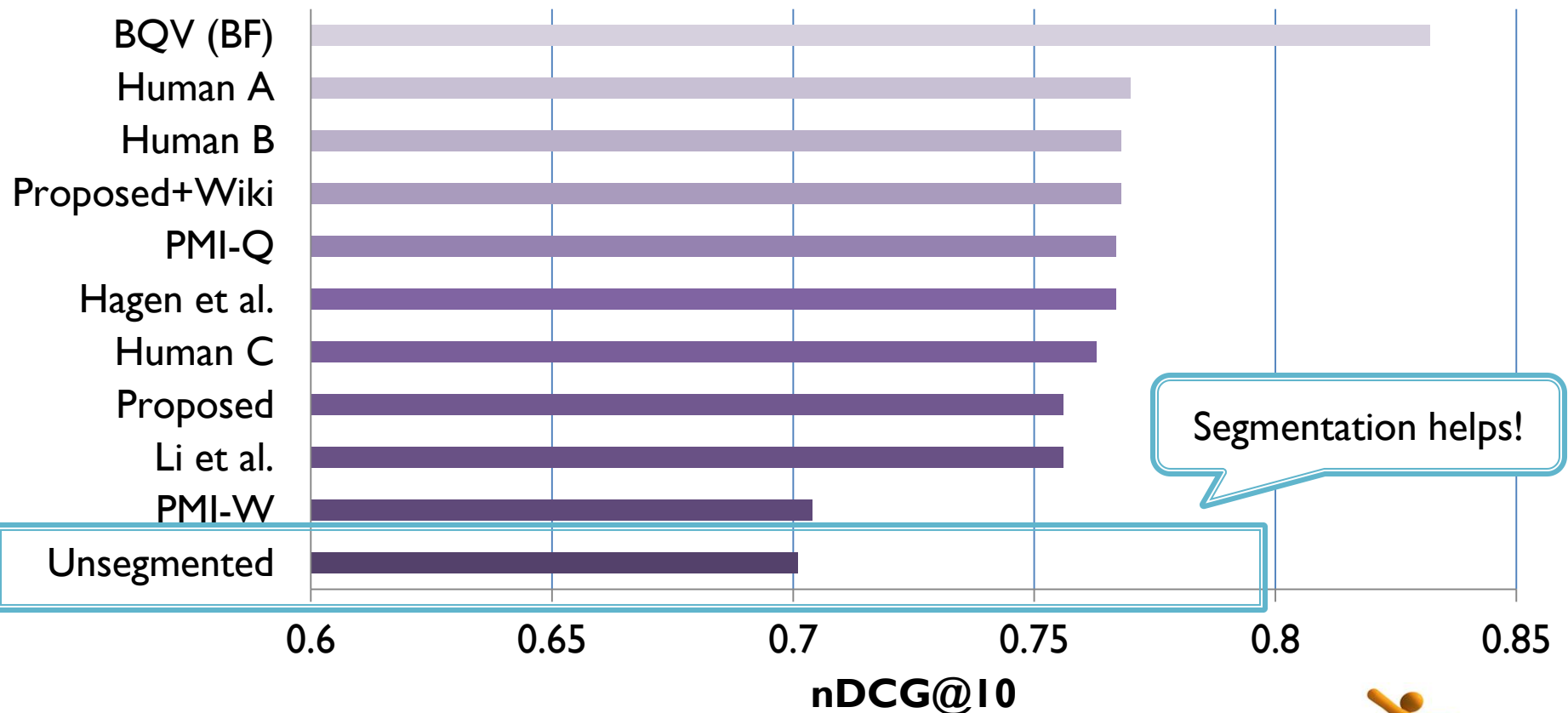
- Six segmentation strategies compared on our framework including (four state-of-the-art systems)
- Li et al. (SIGIR 2011), Hagen et al. (WWW 2011), Proposed (WWW 2011), Proposed + Wiki (SIGIR 2012)
- Baselines: PMI-W, PMI-Q
- Plus annotations by three human annotators A, B, C



# Flat Query Segmentation

## Proposed Evaluation - Results

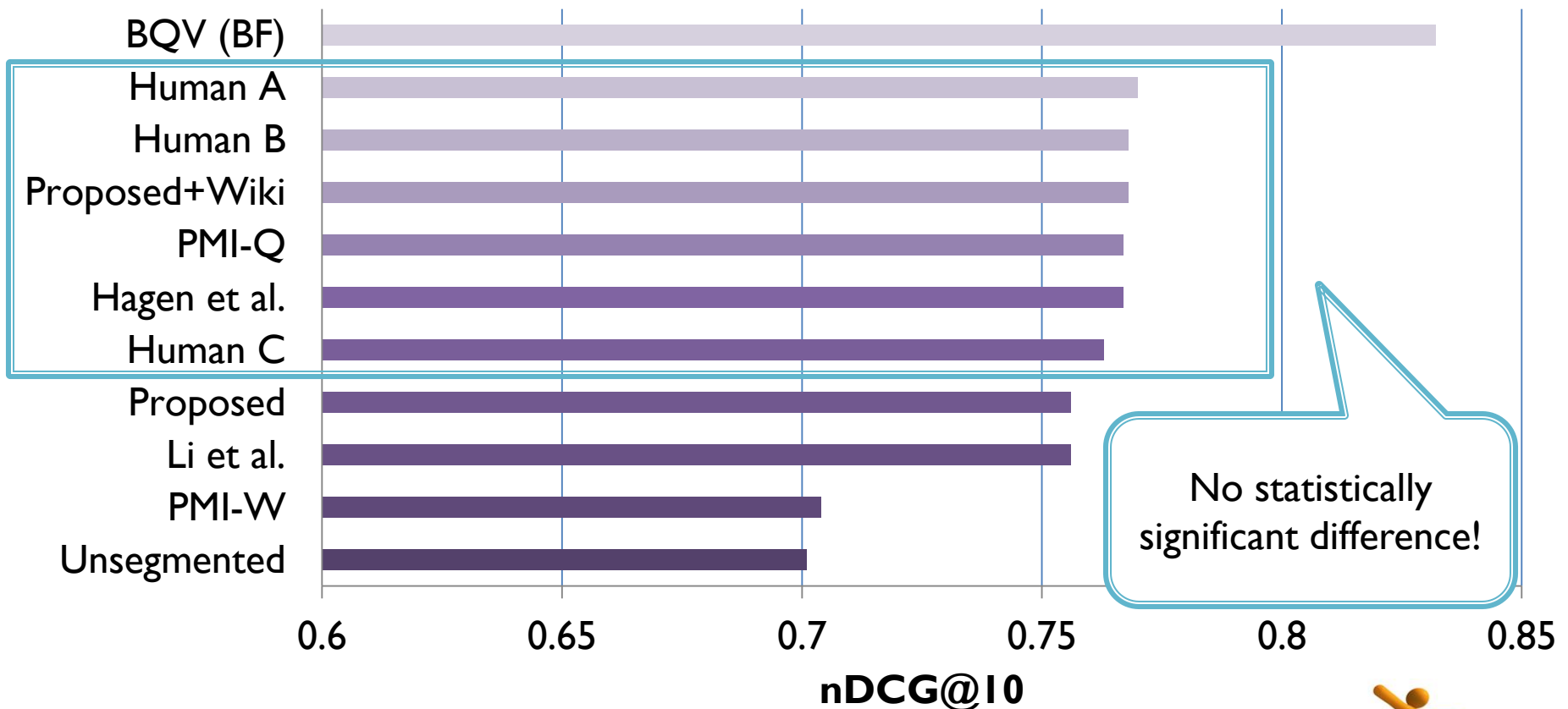
### IR Performance of Strategies



# Flat Query Segmentation

## Proposed Evaluation - Results

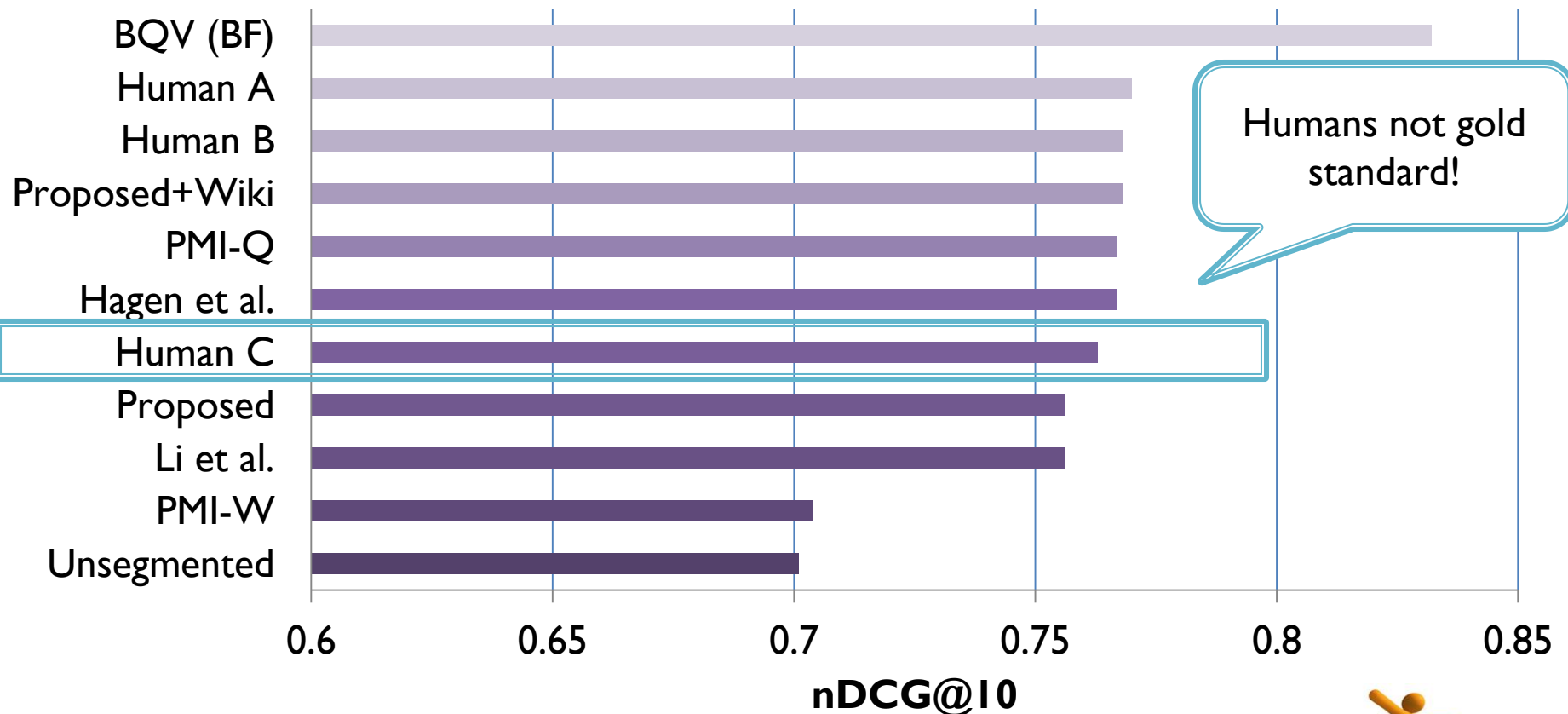
### IR Performance of Strategies



# Flat Query Segmentation

## Proposed Evaluation - Results

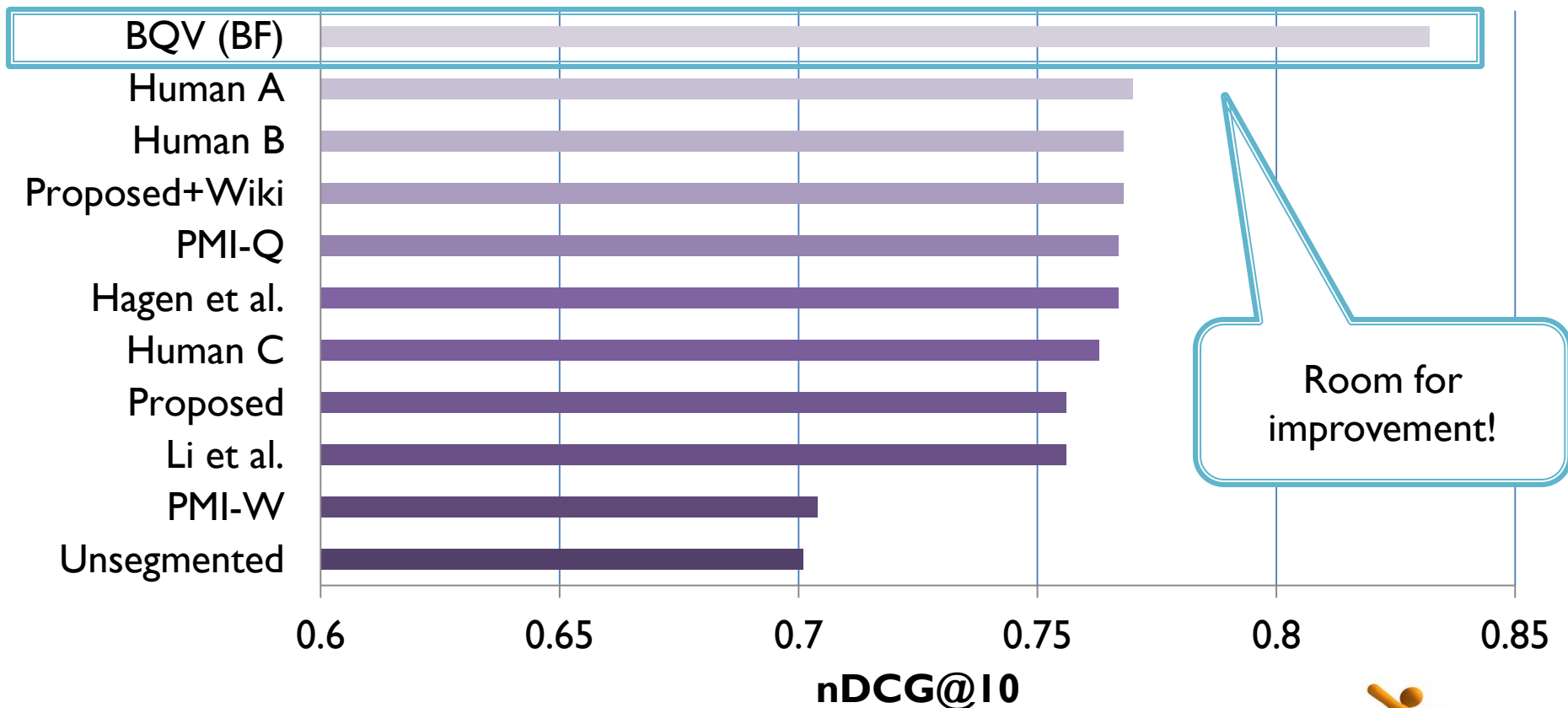
### IR Performance of Strategies



# Flat Query Segmentation

## Proposed Evaluation - Results

### IR Performance of Strategies





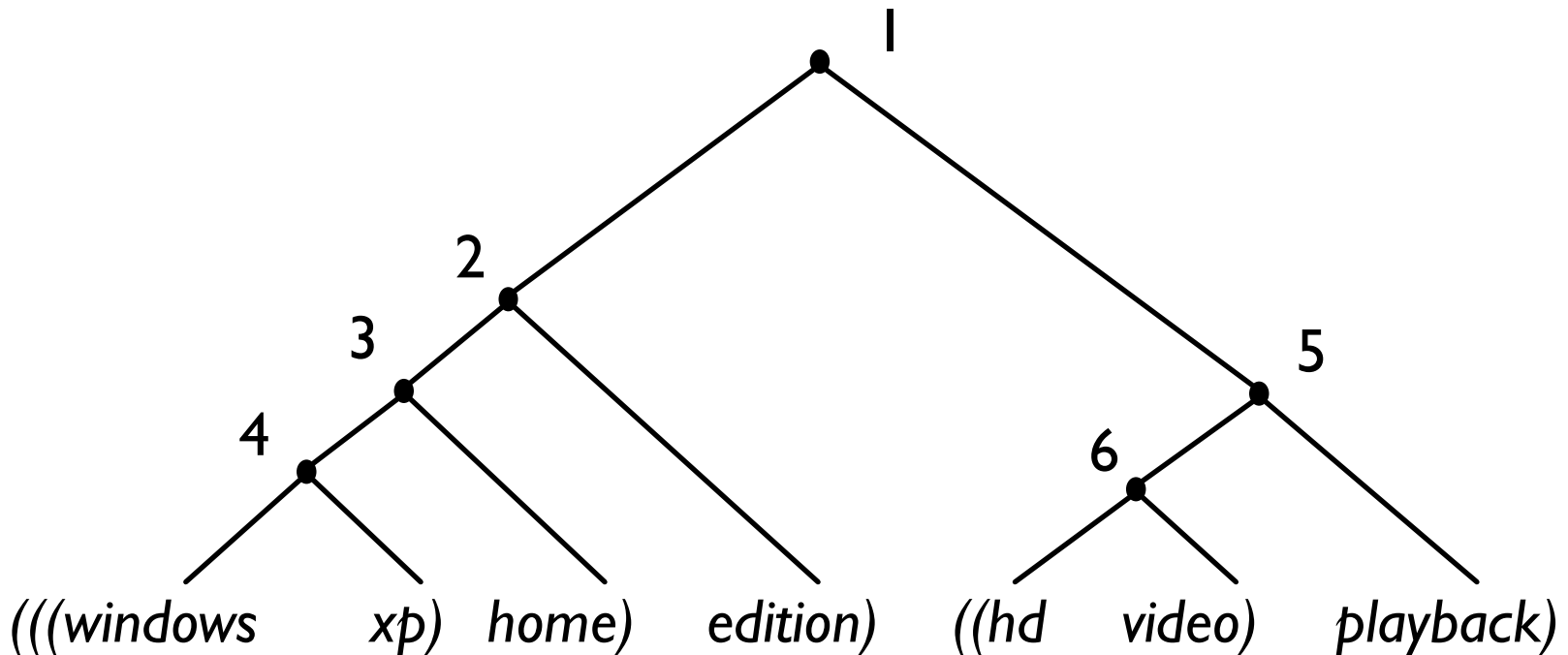
# Research Objectives - 2

- Issue of granularity – long or short segments – which is better?
- How to deterministically apply segmentation to improve document ranking?
- How to deal with situations when segments are not found in the exact order in documents?
- Find richer syntactic relationships!!



# Nested Query Segmentation

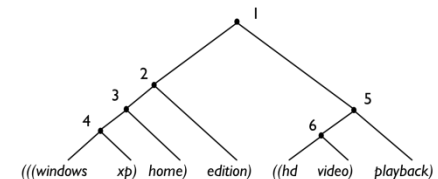
- First issue resolved by nested or hierarchical segmentation
- **Nested segmentation tree** shows richer syntactic structure



# Nested Query Segmentation

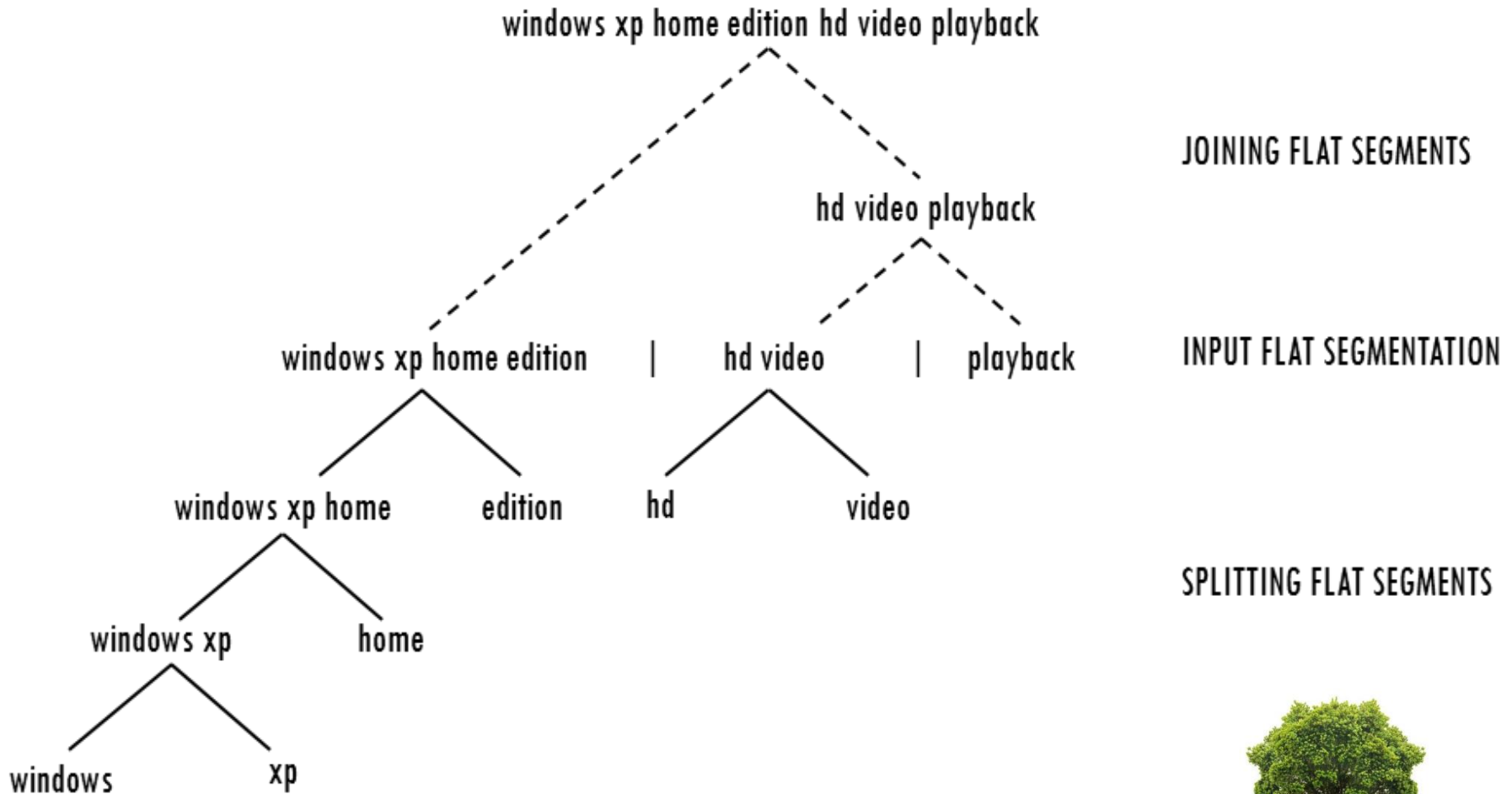
## Proposed Algorithm

- Concept originally proposed by Huang et al. (2010)
- Do not relate nested segmentation to IR
- Approach involved top down recursive binary partitioning
- Similar by NL parse trees, but relevant methods not applicable
- Hierarchical tree structure induced by simple low order n-gram statistics and strengths of flat segment boundaries
- Flat segments involve optimization over query, and we do not want to discard information



# Nested Query Segmentation

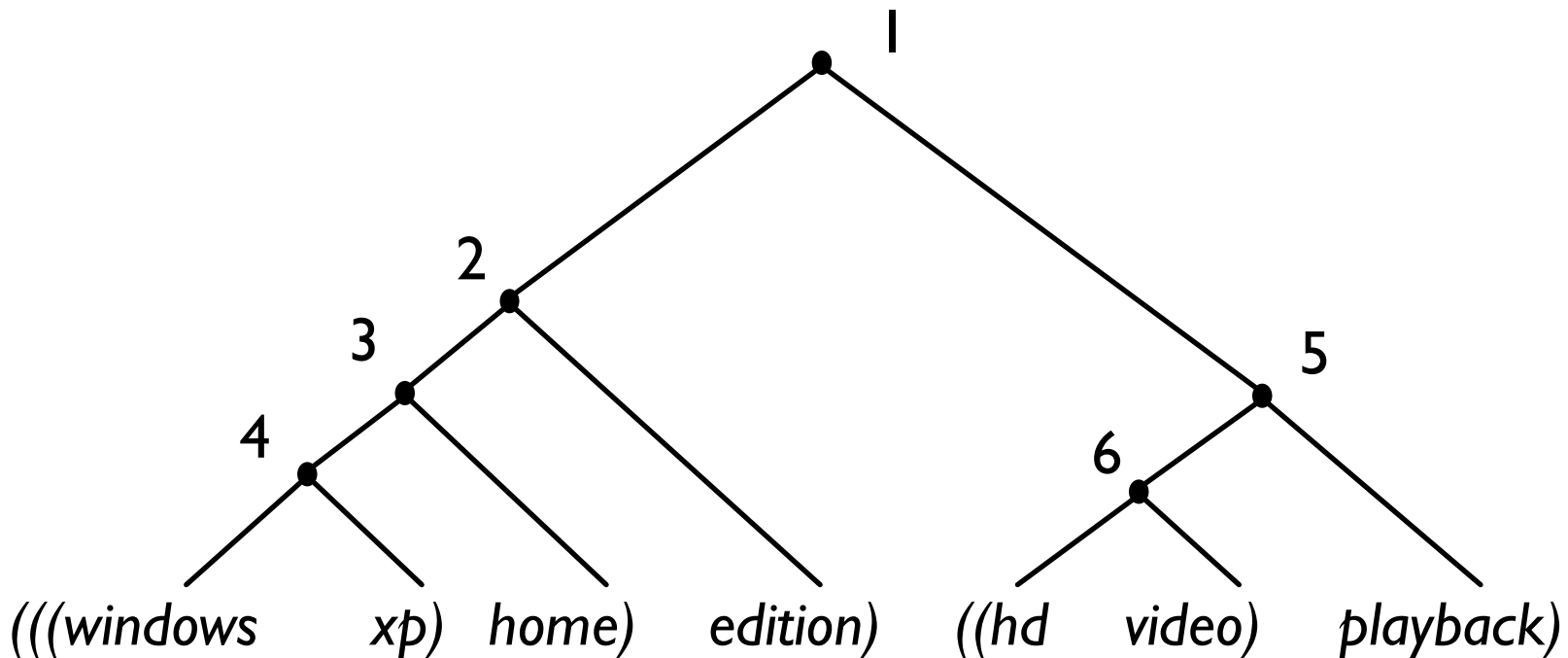
## Proposed Algorithm



# Nested Query Segmentation

## Application to IR: Intuition

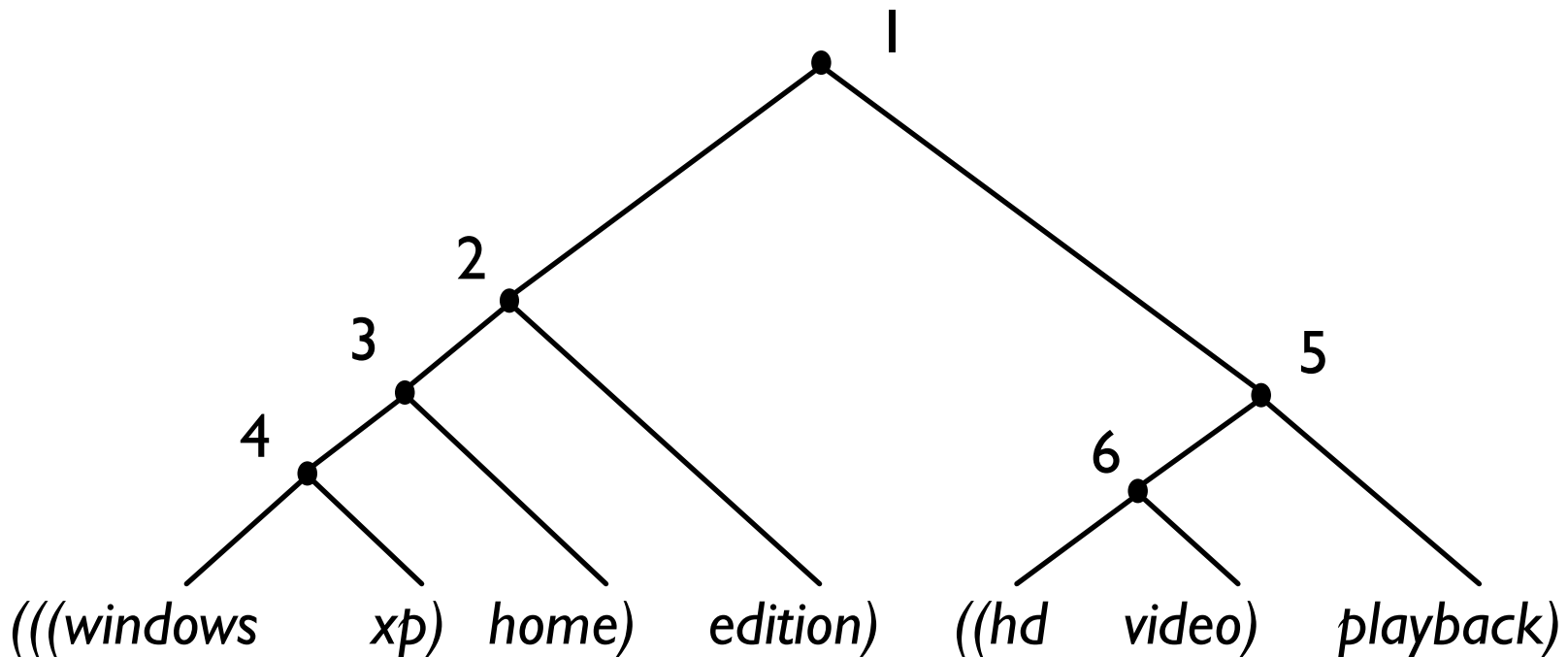
- Pair of words that have a low tree distance should not have a high document distance



# Nested Query Segmentation

## Application to IR: Intuition

- Query: windows xp home edition hd video playback
- Document fragment: ... **windows 7** compatibility of office **xp**  
**home edition...** → High penalty



# Nested Query Segmentation

## Application to IR: Intuition

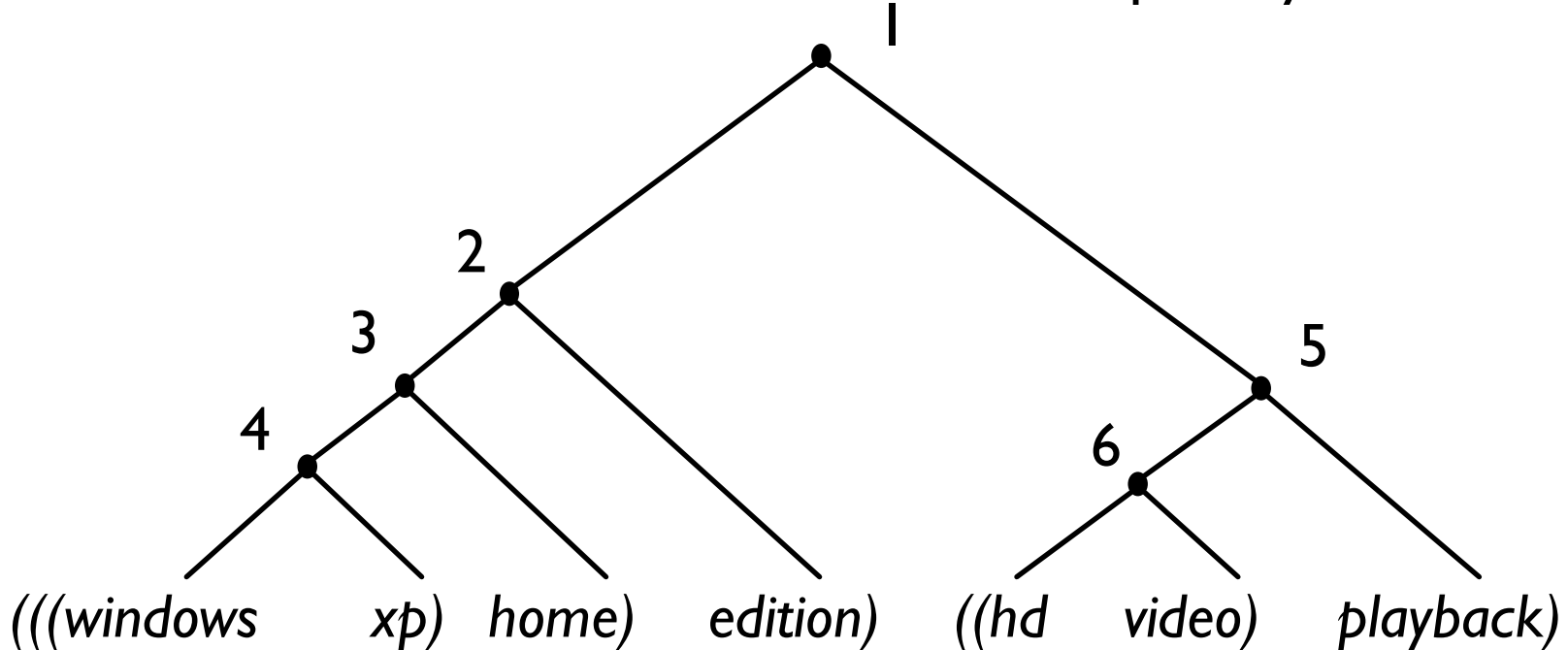
- Query: **windows xp** home edition hd video playback
- Document fragment: ... **windows 7** compatibility of office **xp**  
**home edition...** → High penalty
- Document fragment: ... **has several new tools for hd** audio  
**playback and standard video** ... → High penalty



# Nested Query Segmentation

## Application to IR: Intuition

- Segment: windows xp home **edition** hd video playback
- Document fragment: ... windows xp home **edition** has various tools for **hd** video ... → Lower penalty





# Nested Query Segmentation

## Application to IR: Formulation

- Define term proximity measure in document (pairwise minimum distances, Tao and Zhai 2007, Cummins and O’Riordian 2009)

$$AIDD(a, b; D)_{a \neq b} = \frac{1}{d_1} + \frac{1}{d_2} + \frac{1}{d_3} + \dots + \frac{1}{d_k}$$

- Re-rank documents by scaling document distance by tree distance

$$RrSV_{\mathcal{D}} = \sum_{\substack{t_i, t_j \in q \cap \mathcal{D} \\ t_i \neq t_j}} \frac{AIDD(t_i, t_j; \mathcal{D})}{td(t_i, t_j; n(q))}$$

- Fuse original and new ranks (Agichtein et al. 2006)



# Nested Query Segmentation

## Application to IR

- Baseline: Flat segmentation
- Terms within a flat segment are expected to appear close to each other in the document (Bendersky et al. 2009)
- Re-rank using document distance restricted to term pairs within flat segment

$$RrSV_D = \sum_{k=1}^p \sum_{t_i, t_j \in S_k \cap D} AIDD(t_i, t_j; D)$$

# Nested Query Segmentation

## Application to IR

- Results with flat segmentation baseline
- Input flat segmentation: Hagen et al. (2011)

Metric	Unsegmented	Flat Segmentation	Nested segmentation
nDCG@10	0.6997	0.7081	<b>0.7262</b>
MAP	0.8337	0.8406	<b>0.8468</b>

- Input flat segmentation: Proposed algorithm (Query logs+Wiki)

Metric	Unsegmented	Flat Segmentation	Nested segmentation
nDCG@10	0.6997	0.7044	<b>0.7268</b>
MAP	0.8337	0.8423	<b>0.8477</b>

# Nested Query Segmentation

## Application to IR

- Most improvements statistically significant
- Improvement of other baselines like document distances only, document distance and *query distances*
- Improvement over algorithm by Huang et al. (2010)

Metric	Doc Dist	Query Dist	Tree Dist (Proposed)	Huang et al.
nDCG@10	0.7193	0.7255	<b>0.7268</b>	0.7224

- Robust to parameter variations



# Research Objectives - 3

- What roles do units play in Web search queries?
- Can such “roles” be used to improve result quality?
- Role induction of syntactic units is the next step in a linguistic analysis!!



# Initial Observations

*list of | jim carrey | movies*

*hotels | near | gold coast | au*

*cheap | flights | from | brisbane | to | vienna*

*very hungry caterpillar | download | pdf*

*barack obama | elections | latest news*

*how to | gprs config | nokia n96*

# Initial Observations

*list of | jim carrey | movies*

*hotels | near | gold coast | au*

*cheap | flights | from | brisbane | to | vienna*

*very hungry caterpillar | download | pdf*

*barack obama | elections | latest news*

*how to | gprs config | nokia n96*



# Content and Intent Segments

- Content segments
  - Carry core information need within query (topic)
  - Removing these units makes query lose its central idea
  - Need to be matched within documents for effective retrieval
- Intent segments (or Content-Free)
  - Specify user intent
  - **Need not** match in documents for relevance
  - Intelligent techniques increase the relevance of result pages





# Content and Intent Segments

- Earlier notion of intent associated with query as a whole
- Informational, navigational and transactional (Broder 2002)
- Here intent associated at word level
- Aligned to entity-attribute model (Pasca 2005)
- Aligned to intent heads and intent modifiers (Li 2009)
- Content segments need to be matched; extensive taxonomies
- *Analysis of intent segments more interesting for understanding user requirements!!*



# Intuitions for Separation

- Rigorous manual analysis of query logs reveals that content and intent segments have distributional properties similar to content and function words of natural language
- Content words carry lexical meaning
- Function words specify important relationships between content words
- Content and intent words perform similar roles and are broad lexical categories of the *query language*



# Parallels with NL

- *Postulate I*: Function words co-occur with more distinct words
  - Co-occurrence counts expected to be a better indicator
- *Postulate II*: Co-occurrence distribution of function words have much less bias towards specific words
  - Co-occurrence entropy expected to be a better indicator

$$Entropy(w) = - \sum_{t_i \in context(w)} p_{t_i|w} \log_2 p_{t_i|w}$$



# Parallels with NL

- Variants: Left, right and total counts and entropies (LCC, LCE, RCC, RCE, TCC, TCE)
- AP@500 values for function word detection (prepositions, conjunctions, determiners, pronouns, and interjections)

Lang	Adp Typ	Fr	LCC	LCE	TCC	TCE	RCC	RCE
English	Pre-	0.453	0.477	<b>0.493</b>	0.468	0.464	0.439	0.365
French	Pre-	0.390	0.430	<b>0.438</b>	0.405	0.398	0.357	0.313
Hindi	Post-	0.497	0.458	0.394	0.511	0.505	<b>0.523</b>	0.521
Bangla	Post-	0.522	0.543	0.537	0.579	0.599	0.589	<b>0.603</b>

# Extension to queries

## Ranks by Total Co-occurrence Entropy

Ranks 1-10	Ranks 11-20	Ranks 21-30	Ranks 51-60	Ranks 91-100
<i>in</i>	<i>with</i>	<b><i>for sale</i></b>	<i>home</i>	<i>time</i>
<i>the</i>	<b><i>lyrics</i></b>	<i>is</i>	<i>de</i>	<i>your</i>
<i>and</i>	<i>by</i>	<i>what is</i>	<b><i>pictures of</i></b>	<b><i>book</i></b>
<i>for</i>	<i>from</i>	<i>best</i>	<b><i>music</i></b>	<b><i>show</i></b>
<i>of</i>	<i>2010</i>	<b><i>vs</i></b>	<i>uk</i>	<i>la</i>
<b><i>free</i></b>	<b><i>online</i></b>	<b><i>video</i></b>	<i>jobs</i>	<i>myspace</i>
<i>To</i>	<i>new</i>	<i>2009</i>	<i>black</i>	<i>Baby</i>
<i>on</i>	<i>at</i>	<i>my</i>	<b><i>song</i></b>	<i>james</i>
<b><i>how to</i></b>	<i>2008</i>	<b><i>pictures</i></b>	<b><i>news</i></b>	<b><i>cheap</i></b>
<i>a</i>	<b><i>download</i></b>	<i>school</i>	<i>about</i>	<i>does</i>

# Extension to queries

## Ranks by Total Co-occurrence Entropy

- Effect of total co-occurrence entropy on ranks

Segment	Rank by frequency	Rank by TCE
<i>blog</i>	490	127
<i>biography</i>	824	171
<i>wedding</i>	138	438
<i>define</i>	503	51

- Notion of content-intent context sensitive

*roger federer wikipedia* vs *what is wikipedia*



# Extension to queries

## Validation against human annotations

- Three human annotators asked to label pooled segments
- AP@500 values for intent segment detection
- Performance of TCC, TCE best
- Annotator C strictest, mean pairwise Kappa  $\approx 0.6$

Anno	Fr	LCC	LCE	TCC	TCE	RCC	RCE
A	0.462	0.495	0.498	<b>0.548</b>	<b>0.519</b>	0.513	0.479
B	0.528	0.617	0.631	<b>0.665</b>	<b>0.674</b>	0.590	0.567
C	0.338	0.361	0.359	<b>0.401</b>	0.385	<b>0.392</b>	0.381



# In-query labeling

## Intuition

- Content-intent labeling useful only if done within query
- Intentness score (IS) defined by a simple log-linear combination

$$\text{IS}(s) = \log_2(\text{Fr}(s)) + \log_2(\text{LCC}(s)) + \text{LCE}(s) + \log_2(\text{TCC}(s)) + \text{TCE}(s) \\ + \log_2(\text{RCC}(s)) + \text{RCE}(s)$$

- Intent segments expected to have higher intentness score
- A query always has at least one content segment, zero or more intent segments
- In case of doubt, always safer to label as content!!





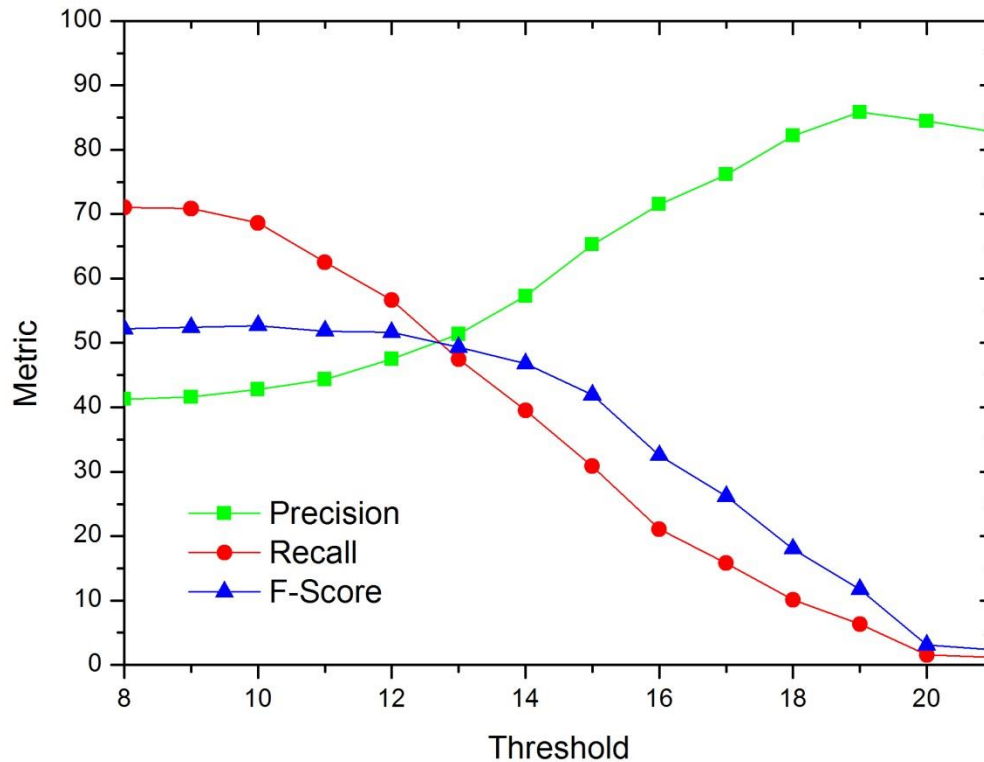
# In-query labeling Algorithm

- Analysis restricted to two-segment queries ( $\approx 44\%$  of log)
- Labelling algorithm
  - Label segment with lower intentness score as content
  - If score of other segment greater than threshold, label intent
  - Else also label other segment as content
- Ask same annotators to label content and intent in 1000 queries
- Measure precision, recall, and F-Score



# In-query labeling Results

- Precision increases with threshold, reverse for recall
- Increasing threshold → more segments get labeled as content



# In-query labeling

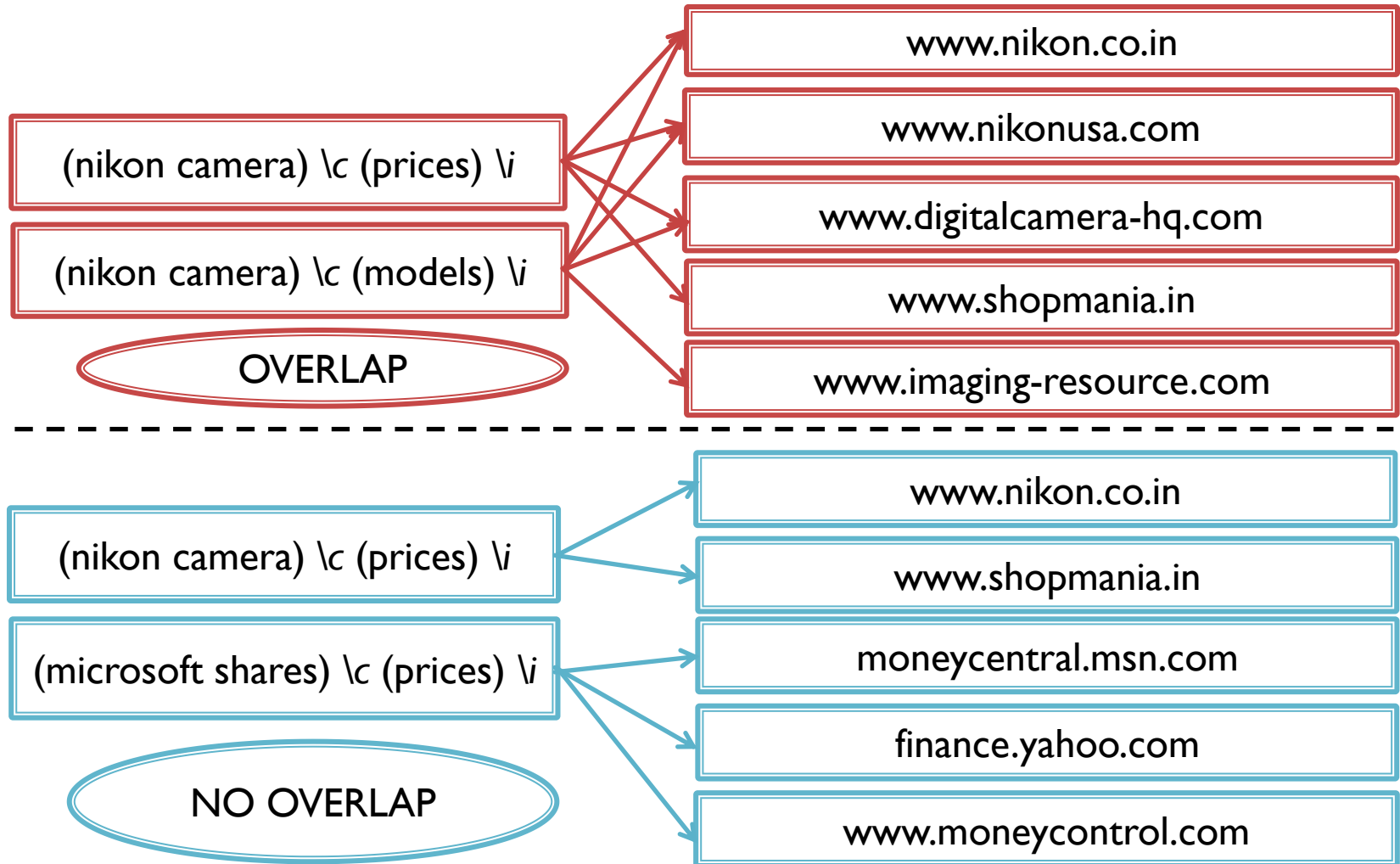
## Alternative Evaluation Using Click Data

- Not good to be tied to manual annotations for evaluation
- Content and intent differences are reflected through clickthrough information
- Intuition: More URL overlap in queries with the same content segment than with those with the same intent segment



# In-query labeling

## Alternative Evaluation Using Click Data



# In-query labeling

## Alternative Evaluation Using Click Data

- Modeling click overlap in set of URLs
- First formulate pairwise URL overlap score – non-trivial

*en.wikipedia.org/wiki/fox* vs *en.wikipedia.org/wiki/guitar*

*cse.iitkgp.ac.in/faculty* vs *cse.iitkgp.ac.in/pg*

- Then compute mean overlap over set



# In-query labeling

## Alternative Evaluation Using Click Data

- Overlap  $o$  between two URLs  $x$  and  $y$  is a function of
  - Number of common strings in path starting from beginning
  - Numbers of strings in paths
  - Inverse URL frequency of URL stem  $st$ ,  $IUF(st) = \log_{10} \frac{1+|U|}{|U_{st}|}$
  - Click counts  $c_x$  and  $c_y$  on  $x$  and  $y$  for the same query

$$o(x, y) = IUF(s_{x_1}) \times \frac{2k}{n_1 + n_2} \times \min(c_x, c_y)$$

# In-query labeling

## Alternative Evaluation Using Click Data

- For both segments in query separately, collect set of clicked URLs and compute overlap
- Label segment with higher overlap as content, other as intent
- Produce (pseudo)-gold standard
- Compute precision, recall and F-score similar to validation with human annotations

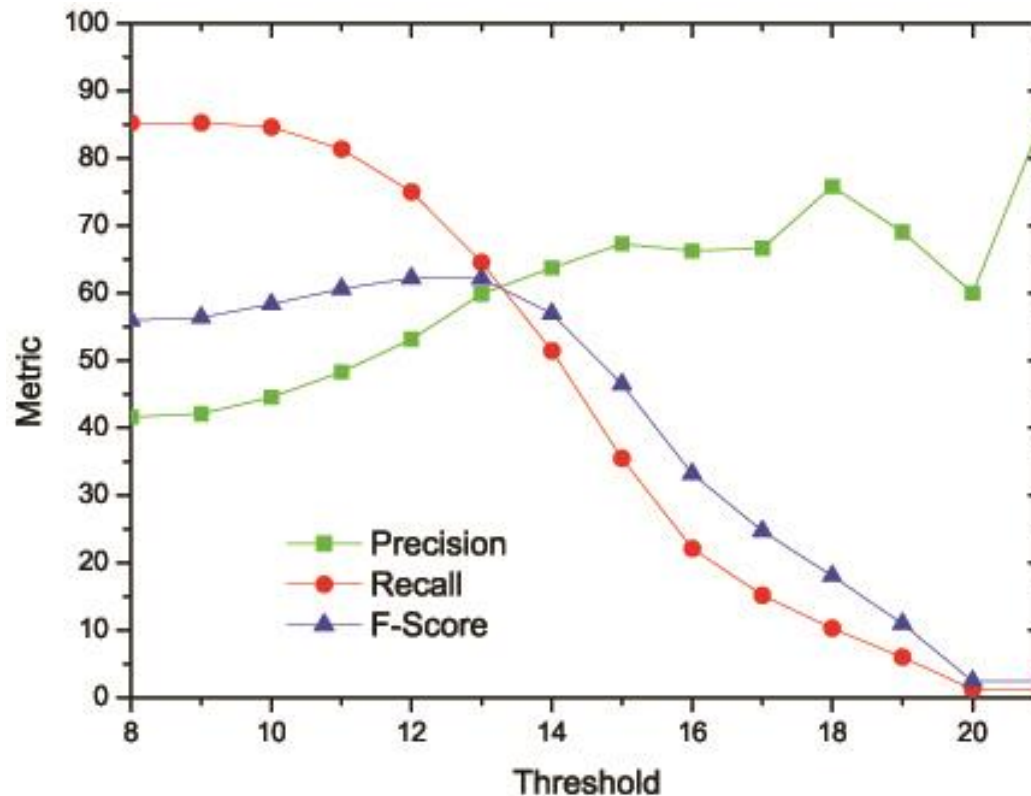


# In-query labeling

## Alternative Evaluation Using Click Data

- Exactly similar trends as with human annotations!
- Optimum threshold also similar – Click data reliable for validation

(or labeling!)





# In-query labeling

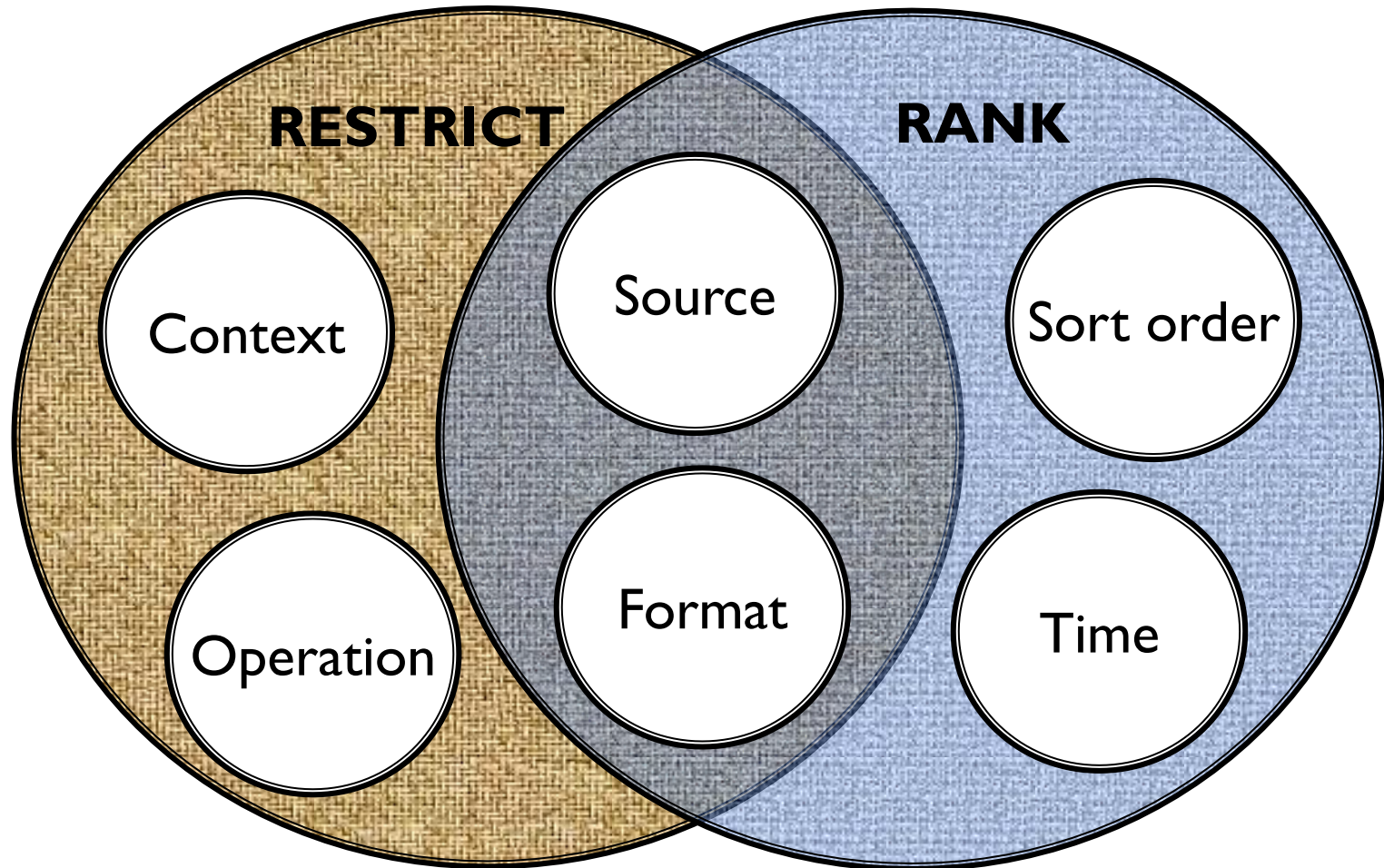
## Effect on IR

- Hypothesis behind operational definitions
- Content units must match exactly in documents, while intent units need not do so

Metric	Both units quoted	Content quoted; Intent unquoted or deleted
nDCG@10	0.772	<b>0.890</b>
MAP	0.359	<b>0.397</b>

Rishiraj Saha Roy, Rahul Katare, Niloy Ganguly, Srivatsan Laxman and Monojit Choudhury, “Discovering and understanding word level user intent in Web search queries”, in *Web Semantics: Science, Services and Agents on the World Wide Web*, Elsevier, 2014 (in press).

# Taxonomy of Intent Segments



# Taxonomy of Intent Segments

	<b>RESTRICT</b>		<b>RESTRICT</b>	<b>+ RANK</b>		<b>RANK</b>	
<b>Context</b>	<b>Operation</b>	<b>Other aspects</b>	<b>Source</b>	<b>Format</b>	<b>Sort order</b>	<b>Time</b>	<b>Other preferences</b>
<i>book</i>	<i>how to</i>	<i>recipe</i>	<i>wikipedia</i>	<i>pdf</i>	<i>near</i>	<i>latest</i>	<i>online</i>
<i>movie</i>	<i>what is</i>	<i>benefits</i>	<i>youtube</i>	<i>mp3</i>	<i>cheap</i>	<i>recent</i>	<i>free</i>
<i>game</i>	<i>where are</i>	<i>reviews</i>	<i>india</i>	<i>slides</i>	<i>fast</i>	<i>2012</i>	<i>printable</i>
<i>tv show</i>	<i>download</i>	<i>biography</i>	<i>ebay</i>	<i>videos</i>	<i>large</i>	<i>new</i>	<i>public</i>
<i>ps2</i>	<i>compare</i>	<i>obituary</i>	<i>bestbuy</i>	<i>pictures</i>	<i>close to</i>	<i>current</i>	<i>exclusive</i>
<i>soap</i>	<i>difference between</i>	<i>history</i>	<i>facebook</i>	<i>photos</i>	<i>high-res</i>	<i>last 24 hours</i>	<i>private</i>
<i>windows</i>	<i>buy</i>	<i>lyrics</i>	<i>linkedin</i>	<i>images</i>	<i>shortest</i>	<i>today</i>	<i>black</i>
<i>scientist</i>	<i>upload</i>	<i>recipe</i>	<i>australia</i>	<i>ppt</i>	<i>budget</i>	<i>now</i>	<i>best</i>

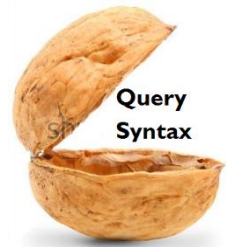
# Research Objectives - 4

- How simple or complex is query syntax with respect to random sequences and parent NL?
- How do we systematically quantify syntactic complexity of queries?
- How do we measure user intuition?
- Multiple independent perspectives are needed to understand complexity of query syntax



# Analyzing Query Syntactic Complexity Approach

- Generative language models from query logs
- Simplest model: Based on word  $n$ -grams
- Two-pronged approach for evaluation
- Complex network modeling
- Native speaker intuition: User studies
- Carefully interpret performance of generated queries when compared to real queries

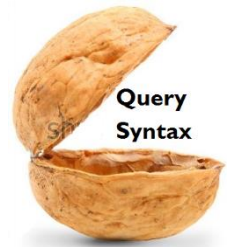


# Analyzing Query Syntactic Complexity

## Generative language model

- Consider query *fifa world cup football 2014*
- *n*-grams: *n*-word strings that appears consecutively in a query
- 3-grams: *<fifa world cup>*, *<world cup football>*, *<cup football 2014>*
- *n*-terms: *n* words present in the query, but order does not matter
- Intuition from bag-of-words concept for queries
- 3-terms: *{fifa, world, football}*, *{fifa, cup, football}*, *{world, cup, 2014}*...
- Conditional probabilities estimated from query log

$$P(w_n | w_1 \dots w_{n-1}) = \frac{P(w_1 \dots w_n)}{P(w_1 \dots w_{n-1})}$$



# Analyzing Query Syntactic Complexity

## Query generation process

- Compute probabilities for each  $n$ -gram in log ( $n = 1, 2, 3$ )
- Sample query length  $L$  from real distribution
- Generate initial  $(n-1)$ -gram
- Try to extend query using last  $(n-1)$ -gram, one word at a time
- Iterate till  $L$  is reached
- Back-off to  $(n-1)$ -gram if needed
- Similarly for  $n$ -terms

$L = 6$ ; 3-gram model

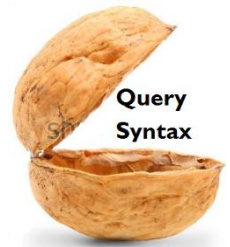
Begin: *a brief*

Extend: *a brief history*

Extend: *a brief history of*

Back-off: *a brief history of **witchcraft***

End: *a brief history of witchcraft europe*



# Analyzing Query Syntactic Complexity

## Examples of generated queries

1-gram	<i>a dhcp ephemeral detailing</i>
	<i>map rc2 western pacific kennedy</i>
2-gram	<i>create user account on roads</i>
	<i>access 2003 not working with info</i>
3-gram	<i>a brief history of witchcraft</i>
	<i>a thousand miles sheet music for webservers</i>
2-set	<i>acer 5310 for flash player</i>
	<i>housing thailand what is cost of housing</i>
3-set	<i>adelaide entertainment explained seating plan</i>
	<i>anti software virus windows vista</i>
2-set-GR	<i>flash player for acer 5310</i>
	<i>what is cost of housing thailand</i>
3-set-GR	<i>adelaide entertainment seating plan explained</i>
	<i>anti virus software windows vista</i>



# Analyzing Query Syntactic Complexity

## Perplexity

Model	NL	Queries	NL	Queries
	(Perplexity)	(Perplexity)	(Counts)	(Counts)
1-gram	1406.593	<b>6417.283</b>	0.3M	0.2M
2-gram	193.722	104.337	3.5M	1M
3-gram	17.663	<b>5.43</b>	9.7M	1.1M
2-term	893.851	384.945	48.1M	4.2M
3-term	N.A.*	23.36	N.A.*	24.8M

$$H(X) = \sum_{x \in X} p(x) \log_2 p(x)$$

$$\text{Perplexity}(X) = 2^{H(X)}$$

- Rate of encountering new word much higher for queries
- Queries more predictable than NL

Fill in the

# Analyzing Query Syntactic Complexity

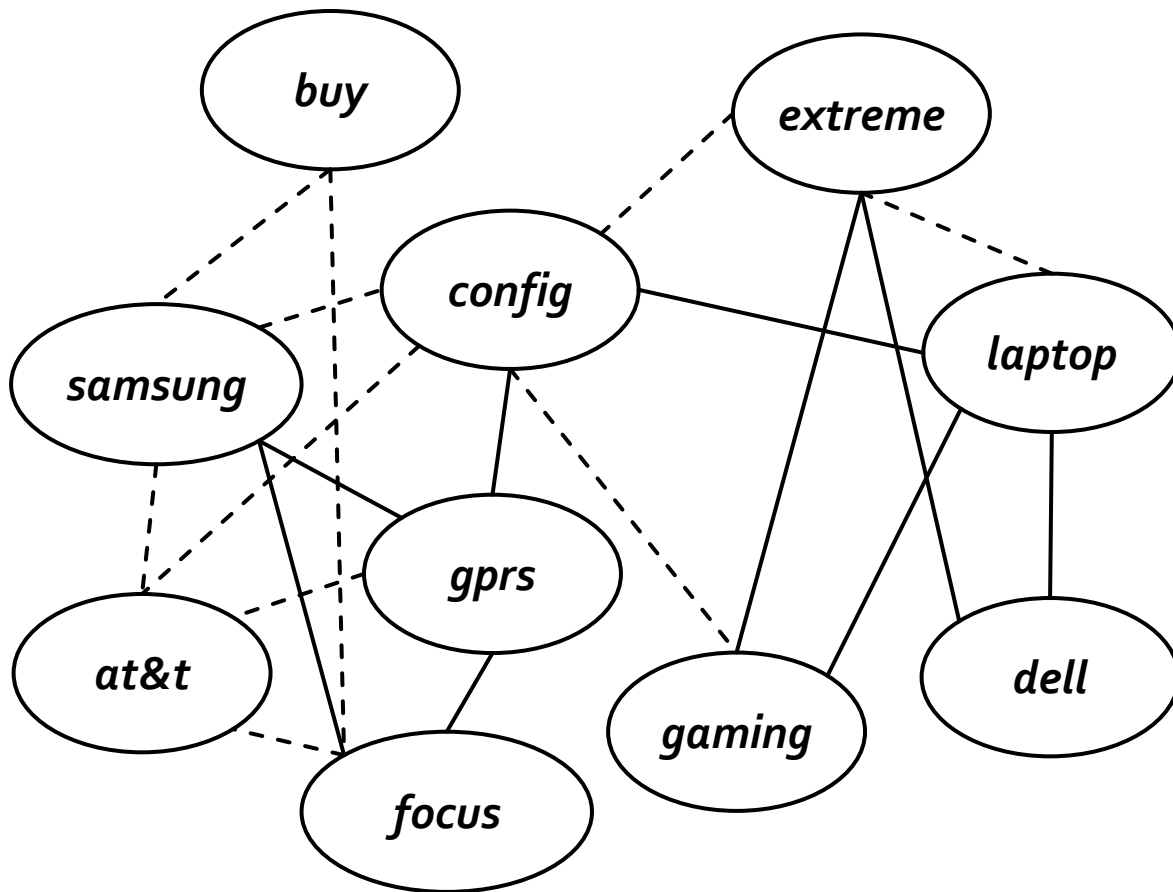
## Macro-level Evaluation

- Networks an elegant framework for modeling complex systems
- Captures local and aggregate-level properties of system
- 100 logs of a million queries each generated for all the seven models
- Word co-occurrence networks formed created from each sample log
- Insignificant edges pruned using joint probabilities  $[p_{ij} > p_i p_j]$
- Largest connected component analyzed
- Network properties compared between real and generated logs
  - Degree distribution, clustering coefficient, average shortest path length
  - **Network motifs**



# Analyzing Query Syntactic Complexity

## Word co-occurrence networks



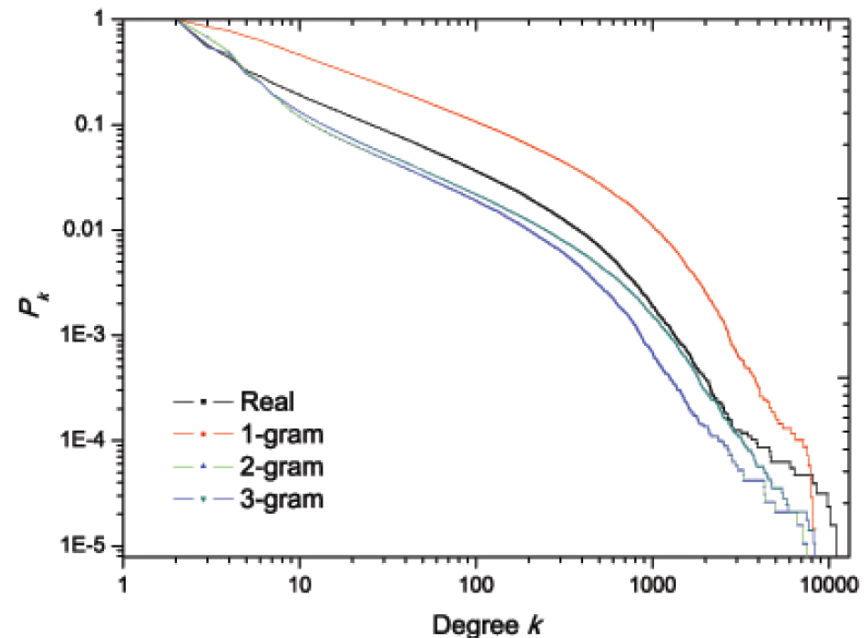
samsung focus gprs config  
dell laptop extreme gaming  
config  
extreme gaming dell laptop  
config  
buy samsung focus at&t  
gprs config at&t samsung focus  
samsung focus gprs config at&t



# Analyzing Query Syntactic Complexity

## Basic network properties

- Network properties observed to be robust to size variation
- Degree distribution, clustering coefficient and average shortest path length reasonably to replicate using n-gram models
- $CC(2\text{-gram}) = 0.619$   
 $CC(\text{Real}) = 0.623$
- $ASPL(3\text{-gram}) = 3.302$   
 $ASPL(\text{Real}) = 3.472$
- Focus on network motifs



# Analyzing Query Syntactic Complexity

## Network motifs

- Small connected subgraphs occurring more frequently in real networks than random graphs [Biemann et al. 2012]
- We analyze *undirected* and *connected* 3 and 4-node motifs only
- Motif signatures definitive properties of networks

Disconnected 3- and 4-motifs							
3-disc.1	3-disc.2	4-disc.1	4-disc.2	4-disc.3	4-disc.4	4-disc.5	
Connected 3- and 4-motifs							
3-line	3-clique	4-line	4-star	4-lineT	4-square	4-squareD	4-clique
2.941	7.185	4.101	6.057	10.315	7.212	15.463	21.484
motherboard, mercury, element	mercury, element, compound	planet, mercury, motherboard, lan	mercury, messenger, planet, motherboard	mercury, motherboard, lan, card	mercury, mining, iron, element	mercury, water, system, requirements	mercury, water, system, steel

# Analyzing Query Syntactic Complexity

## Comparing motifs

$$LNMC(\Psi_i^n) = \log_e \frac{\text{Actual count of } \Psi_i^n}{\text{Expected count of } \Psi_i^n \text{ in an E-R graph}}$$

- Log normalized motif counts more comparable across networks  
[Wernicke et al. 2005]
- Signatures difficult to compare directly, need aggregate statistics

$$M-Diff(LM) = \sum_{k=3}^4 \sum_i |\text{LNMC}(\Psi_i^k)_{Real} - \text{LNMC}(\Psi_i^k)_{LM}|$$

$$M-Sum(LM) = \sum_{k=3}^4 \sum_i \text{LNMC}(\Psi_i^k)_{LM}$$

# Analyzing Query Syntactic Complexity

## Bigrams closest to real network

Model	M-Diff	M-Sum
Real	0.000	74.758
1-gram	8.781	65.977
2-gram	<b>1.590</b>	76.348
3-gram	8.109	<b>82.755</b>
2-term	<b>1.835</b>	72.923
3-term	6.113	<b>80.407</b>
2-term-GR	<b>2.363</b>	72.395
3-term-GR	6.132	<b>80.318</b>

- Bigram-based models within striking distance of real network
- 3-gram-based models have an abundance of connected motifs

# Analyzing Query Syntactic Complexity

## Capturing user intuition

- Do searchers have a notion of how well-formed a query is?
- Asking users to rate standalone query strings not meaningful
- Users were given a triplet having one real query and two model generated queries
- They were asked to identify the real query in this triplet
- Remaining two queries were to be rated on a five-point scale
- Triplets had several words in common
- 19 combinations, 665 triplets,  
1995 queries to be judged





# Analyzing Query Syntactic Complexity

## Capturing user intuition

- Conducted experiments on AMT, cost structuring [Alonso and Baeza-Yates 2011]
- Articulated guidelines; solved examples provided

1. map weather moreton island all	<input checked="" type="radio"/> 0	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5
a map victoria australia home kit	<input type="radio"/> 0	<input checked="" type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5
map of east timor and surrounding islands	<input type="radio"/> 0	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input checked="" type="radio"/> 5
2. fun relay races for kids	<input type="radio"/> 0	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input checked="" type="radio"/> 5
for kids fun easter gifts buy where to	<input type="radio"/> 0	<input type="radio"/> 1	<input checked="" type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5
musical relay races for kids	<input type="radio"/> 0	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input checked="" type="radio"/> 4	<input type="radio"/> 5
3. white fish and potatoes recipes	<input type="radio"/> 0	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input checked="" type="radio"/> 5
fish potato big computer child	<input checked="" type="radio"/> 0	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5
a recipes for cook white fish	<input type="radio"/> 0	<input type="radio"/> 1	<input type="radio"/> 2	<input checked="" type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5

# Analyzing Query Syntactic Complexity

## Results of user experiments

Model	#Total triplets	#Consistent triplets	Judged “Real”	Real percentage	Average rating
Real	665	630	380	<b>60.317</b>	<b>4.046</b>
1-gram	210	197	19	9.645	2.406
2-gram	210	204	41	20.098	2.833
3-gram	210	210	59	<b>28.095</b>	<b>3.276</b>
2-term	175	166	30	18.072	2.88
3-term	175	160	25	15.625	2.875
2-term-GR	175	158	35	22.152	3.076
3-term-GR	175	172	38	22.093	3.163

# Analyzing Query Syntactic Complexity Summary

- Motif analysis insightful, certain motifs reflect semantic structures
- Trigram-based models overfit the data
- Scope for better generative models
- Relative word ordering is important
- Users have a cognitive model for queries
- More complex than random sequences and what  $n$ -grams can capture, more predictable than NL



Rishiraj Saha Roy, Smith Agarwal, Niloy Ganguly and Monojit Choudhury, "Syntactic Complexity of Web Queries through the Lenses of Language Models, Networks and Users", *communicated to PLOS ONE, Public Library of Science (under review)*.

# Discussion on Feedback

**Prof. Vasudeva Varma**

- Would you be able to make similar progress, if Bing query logs are not available? What would have been a viable alternative?
- What if you had access to the query logs and other relevant data from major search engines such as Google, Yahoo in addition to Bing? What more could be done?
- What is the application of this work for Indian languages, when the query logs are not available in some quantity?
- What are the language dependent elements used in your work? Is it possible to reduce them significantly, if not completely eliminate them?



# Discussion on Feedback

Prof. Vasudeva Varma

- Literature review on model-generated queries
- Discussion on synthetic search queries
- Test set and query frequency
- Performance of TF-IDF in function word detection
- Typographical errors, minor clarifications and editorial corrections addressed



# Discussion on Feedback

## Prof. Josiane Mothe

- What is the motivation of using Wikipedia ? What would have been the results/impacts if the document collection or part of it has been used?
- With regard to IR evaluation, the author decides to consider each MWE as a quoted expression in the search engine. In that case and contrary to the theoretical framework he defined, the order of terms is considered during the query-document matching. The impact such a modification may have on the results is not discussed. What type of impact this choice can have had on the results? How could the experimental framework be considered in the theoretical framework?



# Discussion on Feedback

## Prof. Josiane Mothe

- With regards to the segmentation methods proposed in chapter 3: Are the results consistent over queries or are there queries that are improved and other that are better when non-segmented?
- How would the candidate integrate the segmentation results in the analysis of the query language? (chapter 6)
- Has the candidate some suggestions to use the indicators he suggested in chapter 4 for other purposes in IR?



# Discussion on Feedback

## Prof. Josiane Mothe

- Choice of parameter  $\beta$  in segmentation algorithm
- Explanation of equations in nested segmentation
- Details of test collections provided
- Baselines in evaluation of nested query segmentation
- Typographical errors, minor clarifications and editorial corrections addressed





# Takeaways

- Segmentation algorithm generally applicable for *mining new word associations* and syntactic analysis of ungrammatical texts
- Simple statistics can discover hierarchical syntactic structure
- Co-occurrence entropy often marks up an interesting zone for detailed analysis (sponsored or enterprise search)
- Modelling click overlaps among query sets often useful
- Motif profiling can be a useful tool in several text mining tasks



# Contributions of this Thesis

- First flat segmentation algorithm primarily based on query logs
- First IR-based evaluation framework for flat query segmentation
- First nested segmentation algorithm only based on query logs
- First deterministic IR-application for nested query segmentation
- Co-occurrence entropy as reliable indicator for intent units
- Corpus-level and query-level frameworks for understanding syntactic complexity of search queries
- Unsupervised and lightweight techniques



# Future Directions

- Develop ways of incorporating notions of content and intent in the nested segmentation framework
- Develop automatic classifiers for intent taxonomy
- Develop ways to integrate intelligent techniques into search interface for leveraging knowledge of intent
- More sophisticated generative language models for queries, using constraints based on content and intent



# Publications from the Thesis (I)

## Journals

- Rishiraj Saha Roy, Rahul Katare, Niloy Ganguly, Srivatsan Laxman and Monojit Choudhury, “Discovering and understanding word level user intent in Web search queries”, in **Web Semantics: Science, Services and Agents on the World Wide Web**, Elsevier, 2014 (*in press*). (Long Paper)
- Rishiraj Saha Roy, Smith Agarwal, Niloy Ganguly and Monojit Choudhury, “Syntactic Complexity of Web Queries through the Lenses of Language Models, Networks and Users”, communicated to **PLOS ONE**, Public Library of Science (*under review*). (Long Paper)
- Rishiraj Saha Roy, Anusha Suresh, Niloy Ganguly and Monojit Choudhury, “Nested Query Segmentation for Information Retrieval”, communicated to **ACM Transactions on Information Systems** (*under review*). (Long Paper)



# Publications from the Thesis (2)

## Conference Long Papers

- Rishiraj Saha Roy, M. Dastagiri Reddy, Niloy Ganguly and Monojit Choudhury, "Understanding the Linguistic Structure and Evolution of Web Search Queries", in Proceedings of the 10th International Conference on the Evolution of Language (**Evolang X**), 14 - 17 April 2014, Vienna, Austria, pages 286 – 293. (Long Paper)
- Rishiraj Saha Roy, Rahul Katare, Niloy Ganguly and Monojit Choudhury, "Automatic Discovery of Adposition Typology", in Proceedings of the 25th International Conference on Computational Linguistics (**Coling '14**), 23 – 29 August 2014, Dublin, Ireland, pages 1037 - 1046.
- Rohan Ramanath, Monojit Choudhury, Kalika Bali and Rishiraj Saha Roy, "Crowd Prefers the Middle Path: A New IAA Metric for Crowdsourcing Reveals Turker Biases in Query Segmentation", in **ACL 2013**, pages 1713 – 1722. (Long Paper)
- Rishiraj Saha Roy, Niloy Ganguly, Monojit Choudhury and Srivatsan Laxman, "An IR-based Evaluation Framework for Web Search Query Segmentation", in **SIGIR 2012**, pages 881 – 890. (Long Paper)
- Rishiraj Saha Roy, Monojit Choudhury and Kalika Bali, "Are Web Search Queries an Evolving Protolanguage?", in **Evolang 2012**, pages 304 – 311. (**Best Research Poster Award**) (Long Paper)

# Publications from the Thesis (3)

## Conference Short Papers

- Rishiraj Saha Roy, Yogarshi Vyas, Niloy Ganguly and Monojit Choudhury, "Improving Unsupervised Query Segmentation using Parts-of-Speech Sequence Information", in Proceedings of the 37th Annual ACM SIGIR Conference on Research and Development on Information Retrieval (**SIGIR '14**), 6 - 11 July 2014, Gold Coast, Australia, pages 935 - 938. (*Short Paper*)
- Rishiraj Saha Roy, Anusha Suresh, Niloy Ganguly and Monojit Choudhury, "Place Value: Word Position Shifts Vital to Search Dynamics", in Posters of **WWW 2013**, 13 – 17 May 2013, Rio de Janeiro, Brazil, pages 153 – 154 (companion). (*Short Paper*)
- Nikita Mishra, Rishiraj Saha Roy, Niloy Ganguly, Srivatsan Laxman and Monojit Choudhury, "Unsupervised Query Segmentation Using only Query Logs", in Posters of **WWW 2011**, pages 91 – 92 (companion). (*Short Paper*)



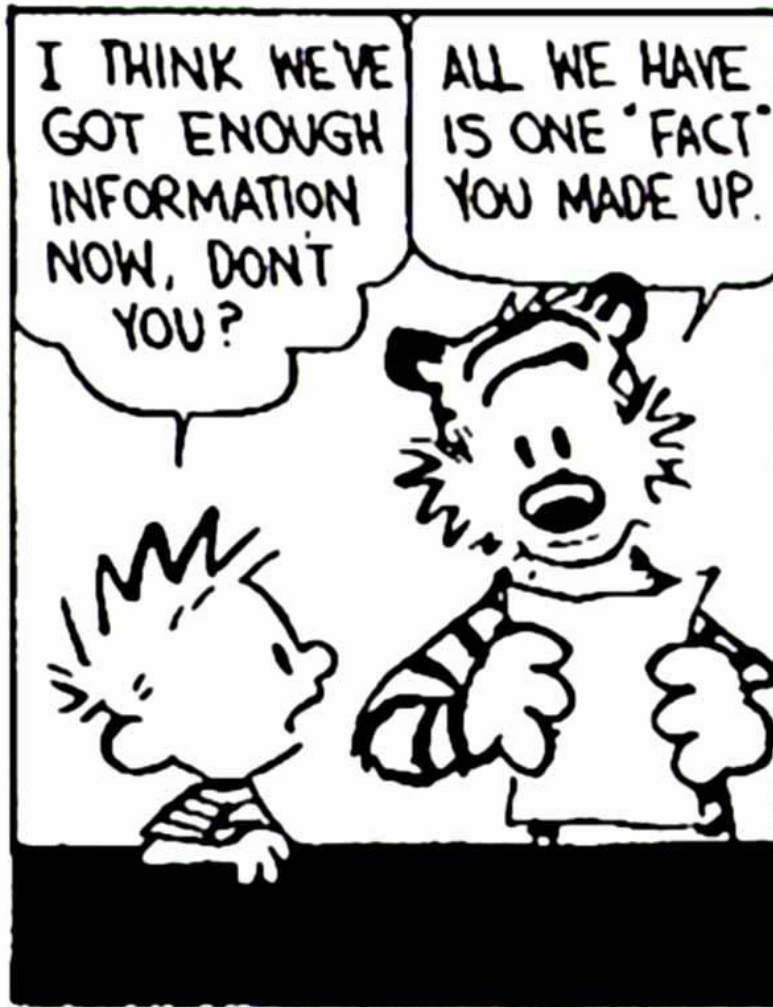
# Publications from the Thesis (4)

## Workshops and Doctoral Consortiums

- Rishiraj Saha Roy, “Analyzing Linguistic Structure of Web Search Queries”, in Doctoral Consortium of the 22nd International World Wide Web Conference (**WWW '13**), Rio de Janeiro, Brazil, 13 – 17 May 2013, pages 395 – 399 (companion).
- Rishiraj Saha Roy, Niloy Ganguly, Monojit Choudhury and Naveen Kumar Singh, “Complex Network Analysis Reveals Kernel-Periphery Structure in Web Search Queries”, in Proceedings of the 2nd International ACM SIGIR Workshop on **Query Representation and Understanding 2011 (QRU '11)**, 28 July 2011, Beijing, China, pages 5 – 8.



# Thank you



THAT'S PLENTY. BY THE TIME WE ADD AN INTRODUCTION, A FEW ILLUSTRATIONS, AND A CONCLUSION, IT WILL LOOK LIKE A GRADUATE THESIS.

