Microsoft® Research

Understanding the Linguistic Structure and Evolution of Web Search Queries

Rishiraj Saha Roy
Ph.D. Student, IIT Kharagpur

Dastagiri Reddy — IIT Kharagpur
Niloy Ganguly — IIT Kharagpur
Monojit Choudhury — Microsoft Research India

Evolang X
Vienna, Austria

Datasets

- Dataset 1: 12.8 million queries from AOL USA (2006)
- Dataset 2: 16.7 million queries Bing Australia (2010)
- Two to ten word queries only
- One word queries do not have structure
- Longer queries have different structural properties
- Newswire corpora for English

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)

Motivation

- Millions of global users, without direct interaction, developing a mode of communication with unique properties – Interesting!!
- Understanding queries as a language may add a linguistic perspective to existing methods in query interpretation
- Various sub-problems may prove directly useful for improving Information Retrieval performance
- Perfectly preserved dataset for studying language evolution

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**

An Evolving (Proto-)Language?

- Three dimensions of analysis: Structure, function and dynamics
- Structure – Query syntax differs from parent NL
- Function – Satisfy several design features, asymmetric communication, heterogeneous agents, click semantics
- Dynamics – Continuous two-way interactions leading to more complex needs (user) and algorithmic development (engine)

[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.

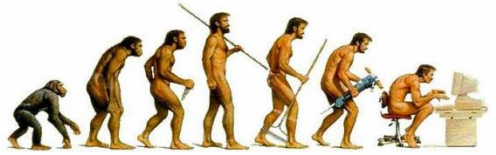
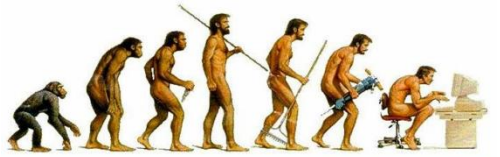


Image source: http://shishikwena.blogspot.in/

Query Structure

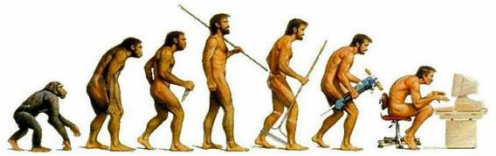
ms office guide book buy online

- More than bags-of-words but no formal grammar
- Identifying multiword expressions and intent expressions vital for information retrieval
- Flexible word order but some proximities and dependencies very important for query understanding



Why has query structure evolved?

- 15 years back
  - Web content limited → Simple information needs
  - Engines could not handle NL sentences – keywords only
- Now
  - Web content has exploded → Complex information needs
  - Search engines much smarter – Content types, user models
- Continuous two-way evolution!!



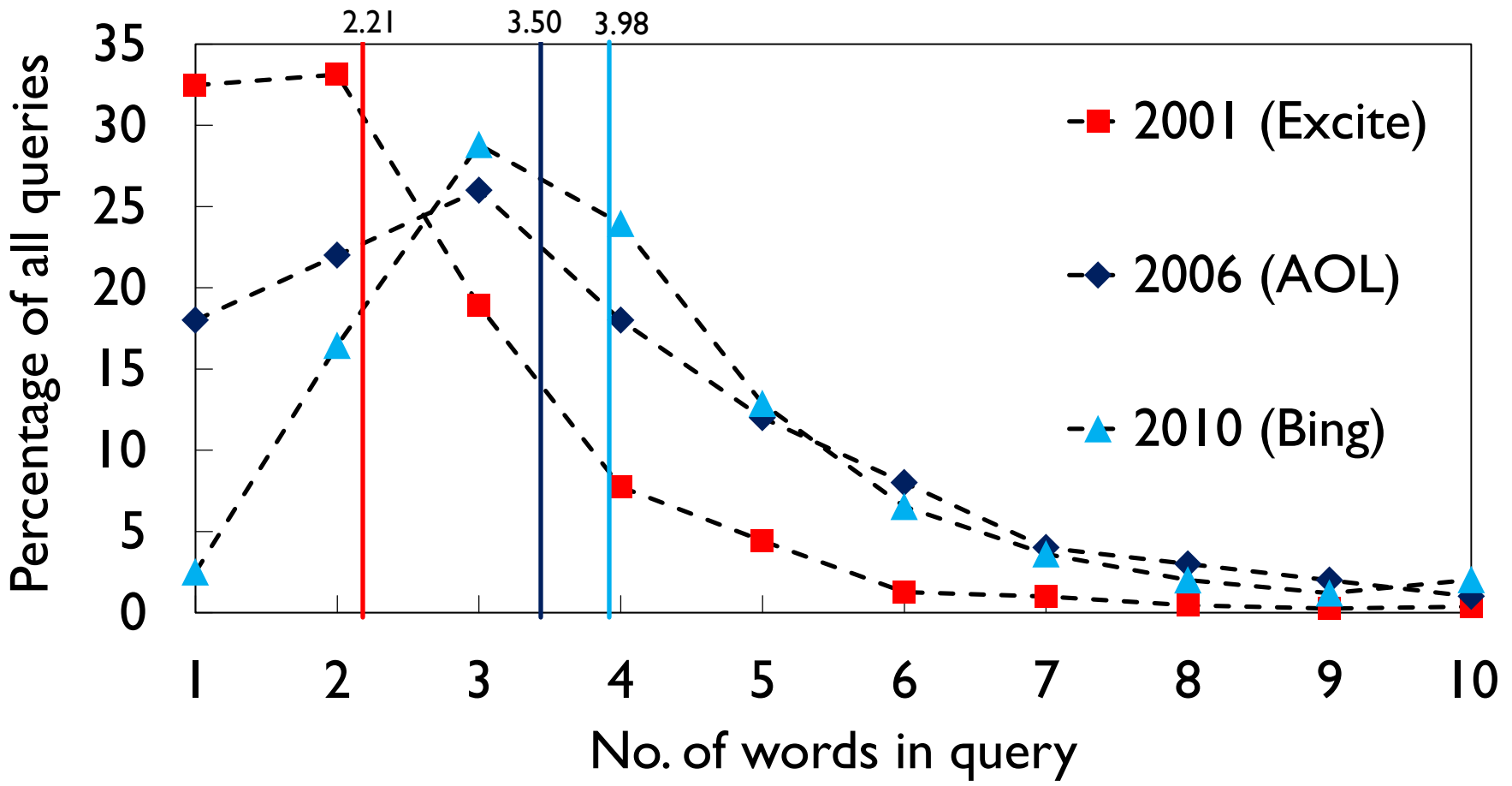
Google introduces Hummingbird

- *"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/dgVK3I

Query lengths

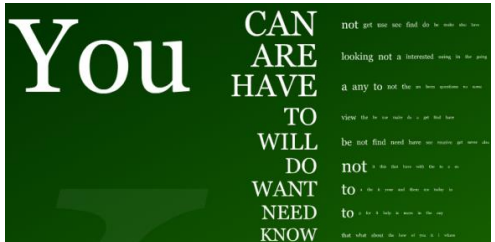


"Measuring" Evolution
Three Approaches

- Three approaches used in this study
- Statistical language modeling
- Complex network modeling
- Positional preferences of query segments
- Track changes using above dimensions
- Draw insights from observations

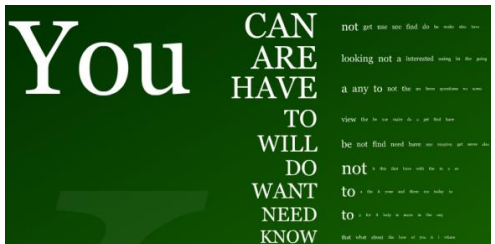
Statistical Language Models
Definition

- Probability distribution over the various possible strings that can be generated by the language
- Used *n*-gram and *n*-set based (generative) models
- Assumes probability of the *n*-th word in a sentence depends only on the previous (*n* -1) words
- Simple yet powerful in many real tasks



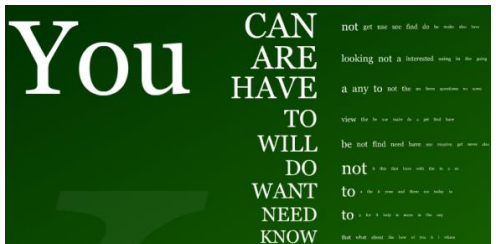
Statistical Language Models
Examples

- 1-grams: *apple, table, harry*
- 2-grams: *apple pie, table tennis, harry potter*
- 3-grams: *apple pie recipe, table tennis shots, harry potter games*
- 2-sets: {*apple, pie*}, {*table, tennis*}, {*harry, potter*}
- 3-sets: {*apple, pie, recipe*}, {*table, tennis, shots*}, {*harry, potter, games*}



Statistical Language Models
Measurement

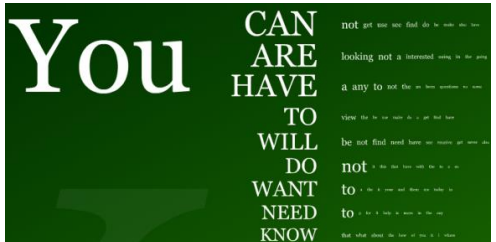
- Information theoretic measures for quantification
- Measured counts, entropy, perplexity, cross-entropy of probability distributions
- Entropy measures the randomness associated with a distribution
- Perplexity – number of choices that user has to predict *n*-th word, given the (*n* - 1) preceding words
- Cross-entropy measures the difference in information content between two probability distributions



Statistical Language Models
Results

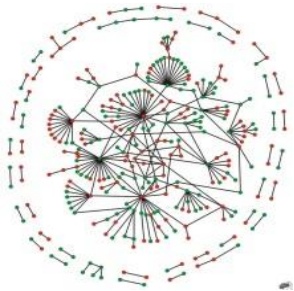
Model	Counts			Perplexity		
	Queries (06)	Queries (10)	NL	Queries (06)	Queries (10)	NL
1-gram	0.4M	0.7M	0.3M	7,869	8,481	2,143
2-gram	3.5M	4.7M	4.4M	75	109	188
3-gram	2.1M	2.8M	11.7M	5	6	12
2-set	6.9M	9.0M	37.5M	128	179	815
3-set	15.3M	20.6M	N.A.	11	13	N.A.

- 0.2M words common, 0.5M words added, 0.2M words deleted
- Unigram perplexity much greater for queries
- But more predictable for higher levels
- Cross-entropy differs markedly from entropy

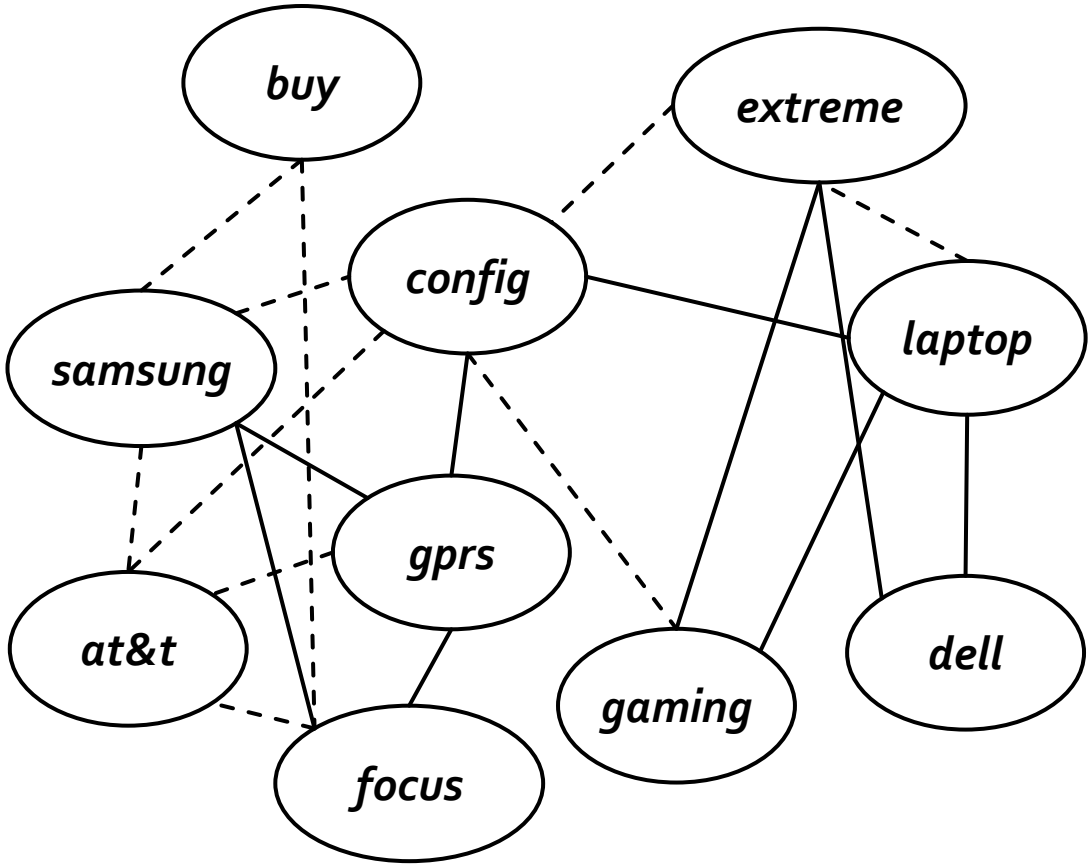


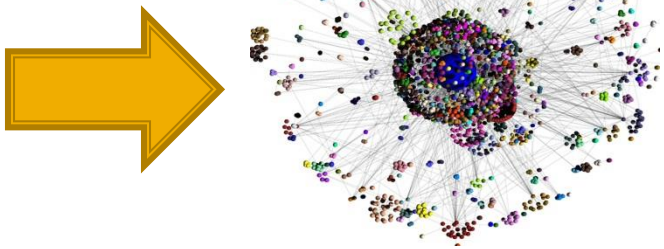
Complex Network Models
Definition

- Complex networks – powerful mathematical framework for modeling complex systems, also useful for visualization
- Use word co-occurrence networks in this study (Cancho and Solé 2001)
- Each word forms a node in the network
- Words co-occurring in queries have an edge between them
- Joint probability measures used to prune random collocations



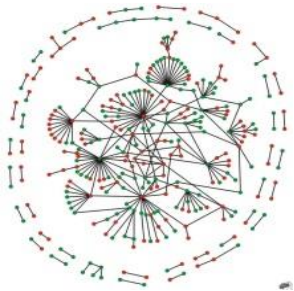
Complex Network Models
Example





Complex Network Models
Measurement

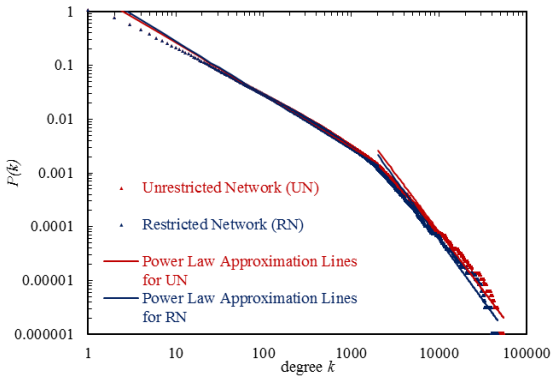
- Find largest connected component of network
- Numbers of nodes and edges
- Average degree (number of connections) of nodes
- Cumulative degree distribution – probability distribution of nodes having degree greater than or equal to value
- Clustering coefficient – Degree of triadic closure in network
- Average shortest path length between every pair of nodes



Complex Network Models
Results

Model	Queries (06)	Queries (10)	NL
|Nodes|	83,525	136,555	460,902
|Edges|	1.1M	1.4M	16.1M
Average degree	25.404	20.660	69.863
Degree distribution	2-regime	2-regime	2-regime
Clustering coefficient	0.592	0.630	0.437
Average shortest path length	3.193	3.305	2.670

- Two-regime power law in queries and NL
- Reflects kernel-periphery structure in both
- Interesting differences in graph structure!!
- Less tight kernel, more kernel-periphery edges



Positional Preferences of Segments
Definition

- Structural units of queries identified through query segmentation (Saha Roy et al. 2012)

harry potter | online videos | youtube

- Content and intent segments labeled using co-occurrence statistics (Yu and Ren 2012)

harry potter | online videos | youtube

- Segment positions in queries labeled as beginning, middle or end

harry potter (beg) | online videos (mid) | youtube (end)

Positional Preferences of Segments
Measurement

- Compute class-wise positional probabilities for each log
- For each (content or intent) segment in each log, measure
  - Probability of being at beginning of query
  - Probability of being at the middle of query
  - Probability of being at end of query
- Interpret changes for possible implications

Positional Preferences of Segments

Results

Content in 2006, Intent in 2010

Beginning probability drops	393/445
Dominant Trend	#Support Segments

Segment	Example query in 2006	Example query in 2010
youtube	*youtube videos*	*new visions 31 youtube*
xbox	*xbox logo*	*sonic the hedgehog xbox*

Positional Preferences of Segments
Results

- 2006: Segments mostly issued as navigational queries for internal searches or as informational queries
- 2010: Appended with content words as search engines can handle direct queries now
- Manifold increase in frequency leads to user-guided search engine rules
- Positional dynamics vital to query intent detection!

Positional Preferences of Segments
Results

Intent in 2006, Content in 2010

Ending probability drops	481/576
Dominant Trend	#Support Segments

Segment	Example query in 2006	Example query in 2010
yellow pages	*granger indiana yellow pages*	*yellow pages wikipedia*
motels	*maryland motels*	*motels for sale brisbane*

Positional Preferences of Segments
Results

- 2006: High popularity leads to intent labels
- 2010: Usage becoming obsolete, esoteric interests
- Error analysis: *what to do*, *cheat codes* and *official site* labeled as content in 2010 due to drops in co-occurrence counts
- But relative positions still indicate "intent"-ness!

Positional Preferences of Segments
Results

- Segments whose roles have not have changed show stabilizations in their relative positions (content at beginnings, intent at ends)
- Users conceptualize and type in content segments first
- Add intent segments to specify user intention explicitly
- Stacking of intent segments main contributor to increasing query lengths over years
- Example: *titanic* → *titanic movie review imdb*

Conclusions and Future Work

- Analyzed evolution of structural properties of queries over four years through three approaches
- Web search queries structurally simpler than NL, but more complex than the bags-of-words model
- SLM measures show that queries approaching NL-like properties
- CNM based analysis shows reverse trend of divergence from NL
- Observations underline uniqueness of linguistic evolution of Web search queries

Conclusions and Future Work

- Web search queries are a very interesting case of a self-organizing communication system
- System has its unique characteristics, but also has several similarities with NL that make this system interesting to study from a language evolution perspective
- Can significantly enrich our knowledge of NL evolution, utilizing large volumes of well-preserved query logs
- **Questions please!!**

Thank you!!