Syntactic Complexity of Web Queries through the Lenses of Language Models, Networks and Users

Rishiraj Saha Roy
Max Planck Institute for Informatics, Saarbruecken, Germany

Smith Agarwal, Jadavpur University, Kolkata, India
Niloy Ganguly, Indian Institute of Technology (IIT) Kharagpur
Monojit Choudhury, Microsoft Research India

9th Satellite on Dynamics on and of Complex Networks (DOOCN 2016)
Amsterdam

Information needs to queries

- **How do I hide the network icon from the status bar?**
- **How many liters 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 →**

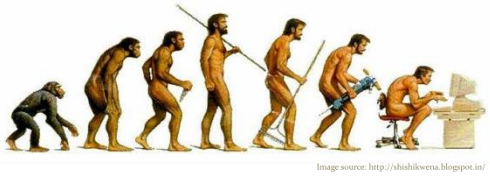
  green tomato pickles sweet

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

  ms office guide book buy online

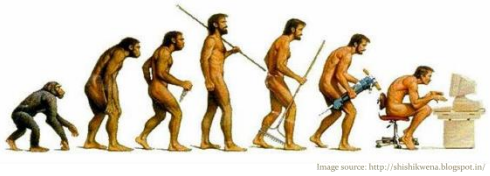
A new language?

- Three dimensions of analysis
  - 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)
- Discussion in Saha Roy et al. (2012, 2014)

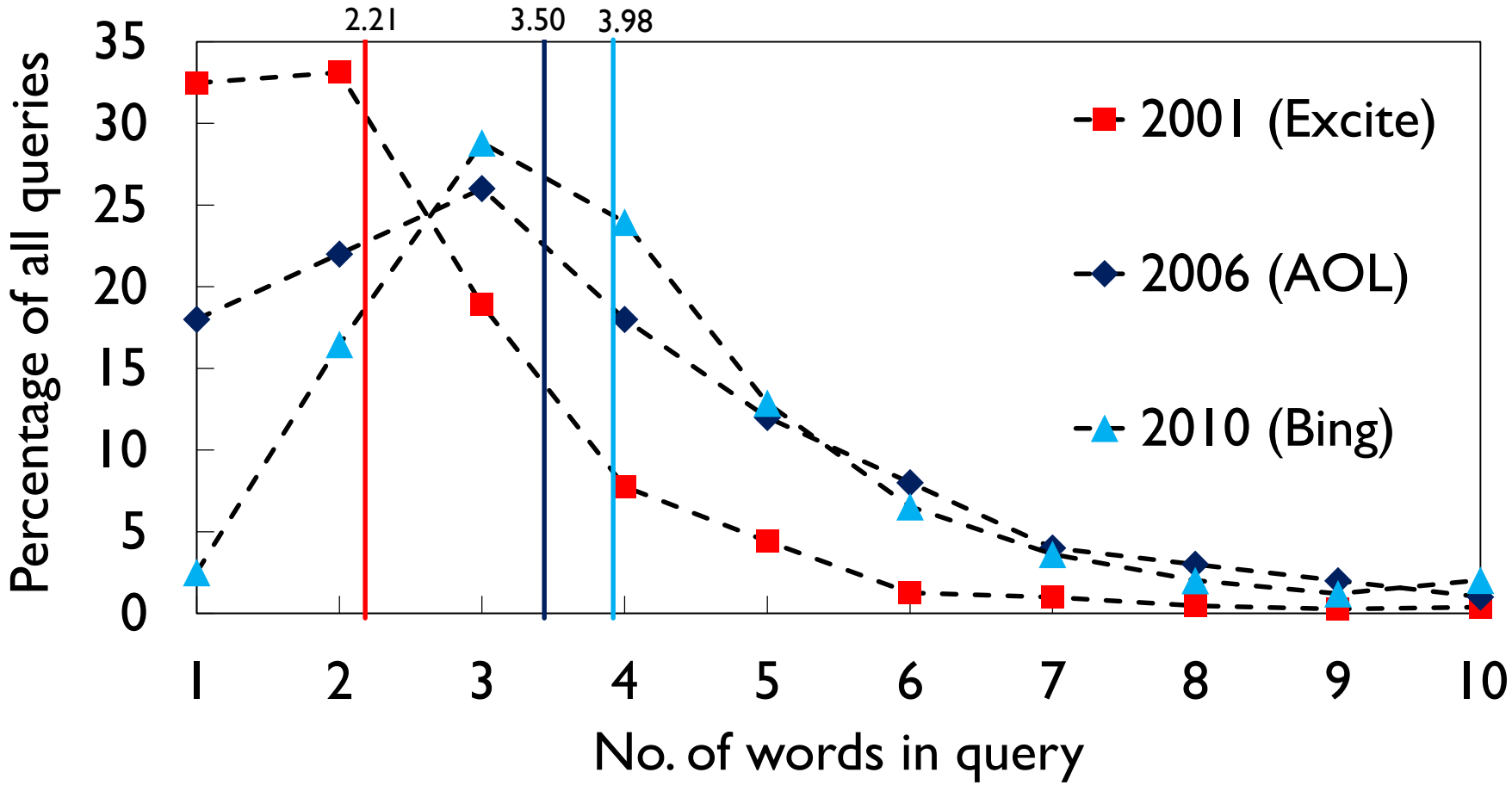


A new language?

- Three dimensions of analysis
  - **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)
- Discussion in Saha Roy et al. (2012, 2014)



Query lengths

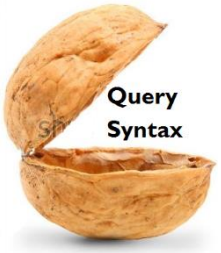


Research questions

- How simple or complex is query syntax with respect to random sequences and parent NL?
- How do we systematically quantify the syntactic complexity of Web queries?
- What inferences can we draw from multiple perspectives on query syntax?

Approach

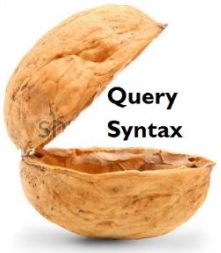
- Generative language models from query logs
- Simplest model: Based on word *n*-grams
- Two-pronged approach for evaluation
  - Macro-level: Complex network modeling
  - Micro-level: User intuition through crowdsourced studies
- Carefully interpret performance of generated queries when compared to real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



Generative language model

- Based on simple *n*-grams
- Consider query: *fifa world cup football 2014*
- *n*-grams: *n*-word strings that appears consecutively in a query
- Generate all 1-grams, 2-grams, 3-grams
- Example: 3-grams or trigrams are

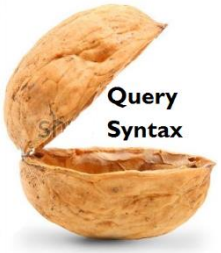
  <fifa world cup>, <world cup football>, <cup football 2014>



Generative language model

- *n*-terms: *n* words present in the query, but in any order
- Intuition from bag-of-words concept for queries
- 3-terms: {*fifa, world, football*}, {*fifa, cup, football*}, {*world, cup, 2014*}, {*fifa, cup, 2014*}, {*football, cup, 2014*}, …
- Conditional probabilities estimated from query log

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



Query generative process

- Select generative model
- Sample query length *L*
- 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

Model: 3-gram

L = 6

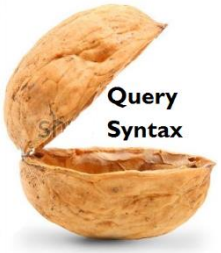
Begin: ***a brief***

Extend: *a brief* ***history***

Extend: *a brief history* ***of***

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

End: *a brief history of witchcraft* ***europe***



Examples of generated queries

1-gram	*a dhcp ephemeral detailing*
2-gram	*create user account on roads*
3-gram	*a brief history of witchcraft europe*
2-set	*acer 5310 for flash player*
3-set	*adelaide entertainment explained seating plan*

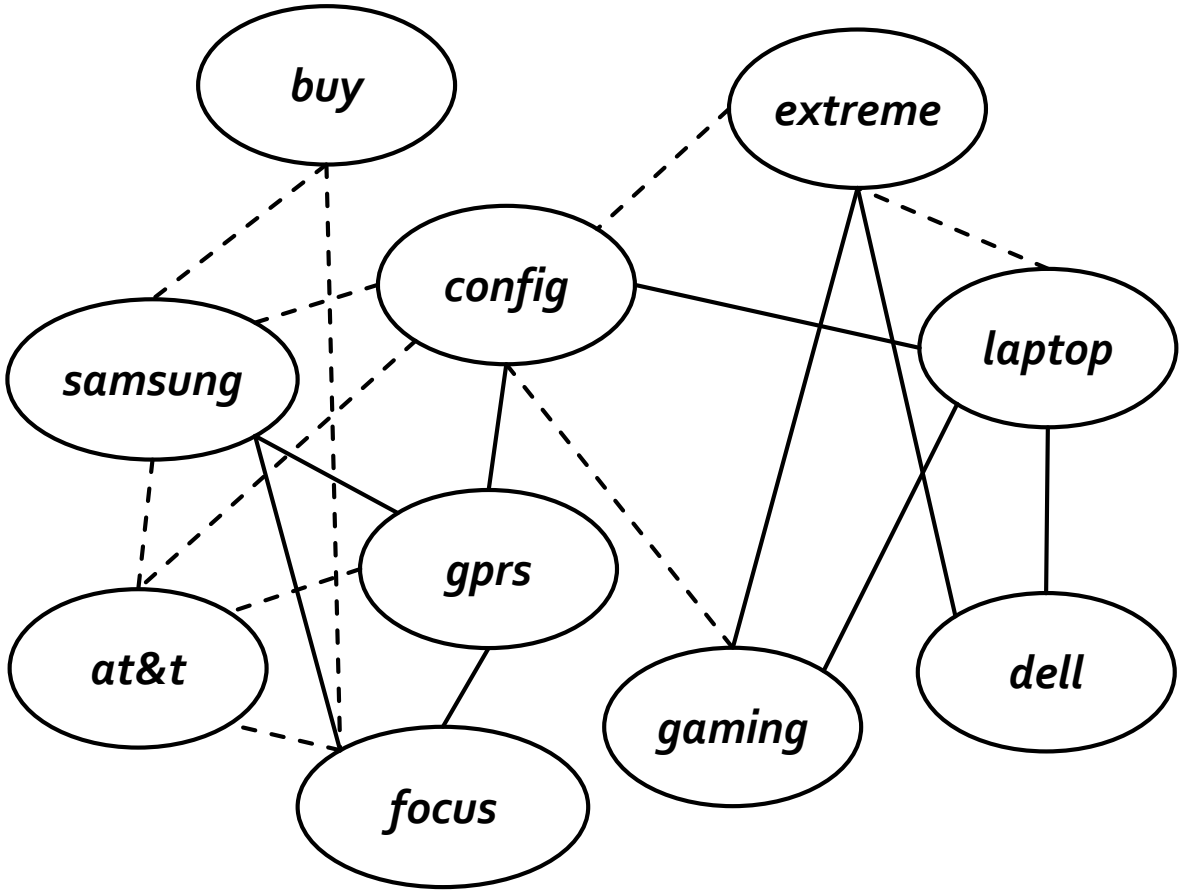
Query log dataset

- 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 in May 2010
- 100 samples of one million queries each created for each of the five generative models

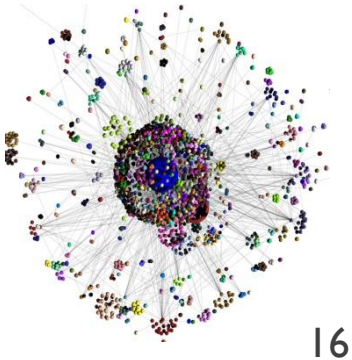
Macro-level evaluation

- Networks an elegant framework for modeling complex systems
- Very useful for modeling large text corpora [Dorogovtsev and Mendes 2001, Ferrer-i-Cancho and Sole 2001]
- Captures both local and aggregate-level properties of system
- Word co-occurrence networks created from each sample log
- Insignificant edges pruned using joint probabilities ($p_{ij} > p_i p_j$) [Ferrer-i-Cancho and Sole, 2001]

Word co-occurrence networks

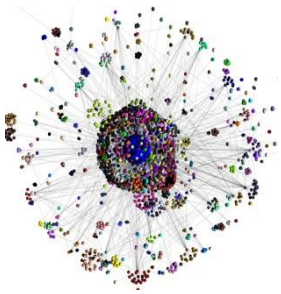


samsung focus gprs config
dell laptop extreme gaming config
extreme gaming dell laptop config
buy samsung focus at&t
gprs config at&t samsung focus
samsung focus gprs config at&t



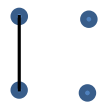
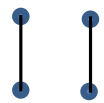
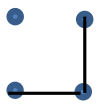
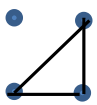
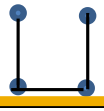
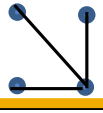
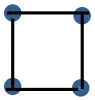
Macro-level evaluation

- Largest connected component analyzed
- Network properties to be compared between real and generated logs
- Degree distribution, clustering coefficient, average shortest path lengths shown to be easy to replicate using simple generative models
- Need network property that better captures relations and non-relations among words better
- *Hard to manipulate or replicate through the generative model*



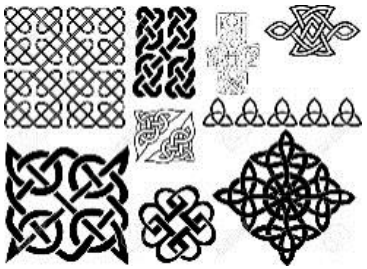
Network motifs

- Small subgraphs occurring much more frequently in real networks than random graphs [Wernicke et al. 2005]
- We analyze *undirected* and *connected* 3 and 4-node motifs only
- Motif signatures unique properties of networks [Biemann et al. 2012]

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

Network motifs

- Network motifs shown to be very useful when comparing generative models for natural languages (Biemann et al. 2012)
- Has also found wide application to biological networks



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

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

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

Bigrams closest to real network

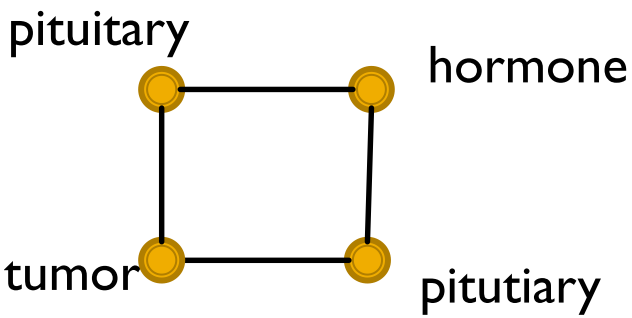
Model	M-Diff	M-Sum
Real	0.000	74.758
1-gram	8.781	65.977
2-gram	**1.590**	76.348
3-gram	8.109	**82.755**
2-term	**1.835**	72.923
3-term	6.113	**80.407**

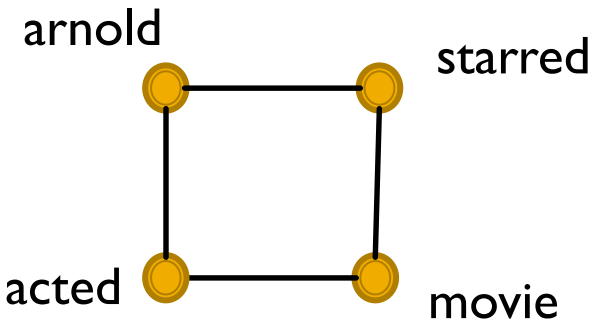
- Bigram-based models within striking distance of real network
- 3-gram-based models have an abundance of connected motifs

Network motifs

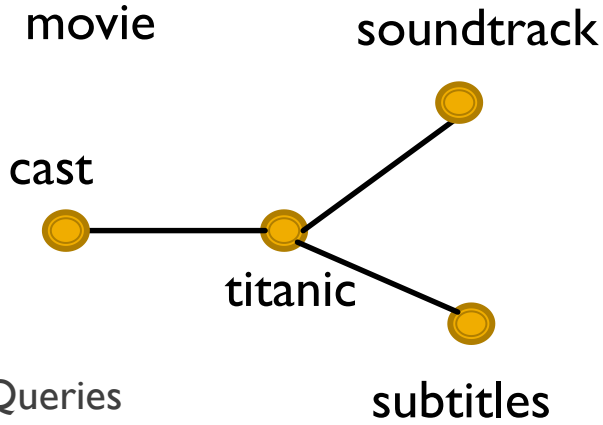
- Polysemy in **chain motifs**

- Synonymy and typographical errors in **box motifs**





- Entity-attribute relationships in **star motifs**



Micro-level evaluation

- Micro- or individual query evaluation is necessary for commenting on quality of generative model
- Capturing native speaker intuition is vital
- Who are the native speakers here?
- Do searchers have a notion of how well-formed a query is?
- Asking users to rate standalone query strings not meaningful!!

Capturing user intuition

- Users were given a triplet having one real query and two model generated queries on Amazon Mechanical Turk
- 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
- About 700 such triplets were annotated

Results of user experiments

Model	Real percentage	Average rating
Real	**60.317**	**4.046**
1-gram	9.645	2.406
2-gram	20.098	2.833
3-gram	**28.095**	**3.276**
2-term	18.072	2.880
3-term	15.625	2.875

Summary (I)

- Ideal generative model between 2- and 3-grams
- Queries may be explained by more principled generative models that incorporate both better syntax and semantics
- Motif analysis insightful
- Often capture semantics
- Motif signatures can be used to compare semi-structured text in other domains

Summary (2)

- Web users possess a cognitive model for queries
- More complex than what random sequences and *n*-grams can capture, but simpler and more predictable than NL
- Insights towards generating and evaluating **synthetic search query logs**

Rishiraj Saha Roy, Smith Agarwal, Niloy Ganguly and Monojit Choudhury, "Syntactic Complexity of Web Queries through the Lenses of Language Models, Networks and Users", *Elsevier Information Processing and Management (IP&M), Volume 52 (5), pp. 923 – 948, September 2016.*