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)
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

Percentage of all queries vs. No. of words in query for 2001 (Excite), 2006 (AOL), and 2010 (Bing).
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?
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 appear 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

\[
P(w_n | w_1 \ldots w_{n-1}) = \frac{P(w_1 \ldots w_n)}{P(w_1 \ldots w_{n-1})}
\]
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
<table>
<thead>
<tr>
<th>1-gram</th>
<th>a dhcp ephemeral detailing</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-gram</td>
<td>create user account on roads</td>
</tr>
<tr>
<td>3-gram</td>
<td>a brief history of witchcraft europe</td>
</tr>
<tr>
<td>2-set</td>
<td>acer 5310 for flash player</td>
</tr>
<tr>
<td>3-set</td>
<td>adelaide entertainment explained seating plan</td>
</tr>
</tbody>
</table>
- 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

<table>
<thead>
<tr>
<th></th>
<th>3-disc.1</th>
<th>3-disc.2</th>
<th>4-disc.1</th>
<th>4-disc.2</th>
<th>4-disc.3</th>
<th>4-disc.4</th>
<th>4-disc.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-disc.1</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
</tr>
<tr>
<td>3-disc.2</td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
</tr>
<tr>
<td>4-disc.1</td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
<td><img src="image21.png" alt="Image" /></td>
</tr>
<tr>
<td>4-disc.2</td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td><img src="image24.png" alt="Image" /></td>
<td><img src="image25.png" alt="Image" /></td>
<td><img src="image26.png" alt="Image" /></td>
<td><img src="image27.png" alt="Image" /></td>
<td><img src="image28.png" alt="Image" /></td>
</tr>
<tr>
<td>4-disc.3</td>
<td><img src="image29.png" alt="Image" /></td>
<td><img src="image30.png" alt="Image" /></td>
<td><img src="image31.png" alt="Image" /></td>
<td><img src="image32.png" alt="Image" /></td>
<td><img src="image33.png" alt="Image" /></td>
<td><img src="image34.png" alt="Image" /></td>
<td><img src="image35.png" alt="Image" /></td>
</tr>
<tr>
<td>4-disc.4</td>
<td><img src="image36.png" alt="Image" /></td>
<td><img src="image37.png" alt="Image" /></td>
<td><img src="image38.png" alt="Image" /></td>
<td><img src="image39.png" alt="Image" /></td>
<td><img src="image40.png" alt="Image" /></td>
<td><img src="image41.png" alt="Image" /></td>
<td><img src="image42.png" alt="Image" /></td>
</tr>
<tr>
<td>4-disc.5</td>
<td><img src="image43.png" alt="Image" /></td>
<td><img src="image44.png" alt="Image" /></td>
<td><img src="image45.png" alt="Image" /></td>
<td><img src="image46.png" alt="Image" /></td>
<td><img src="image47.png" alt="Image" /></td>
<td><img src="image48.png" alt="Image" /></td>
<td><img src="image49.png" alt="Image" /></td>
</tr>
</tbody>
</table>

### Connected 3- and 4-motifs

<table>
<thead>
<tr>
<th></th>
<th>3-line</th>
<th>3-clique</th>
<th>4-line</th>
<th>4-star</th>
<th>4-lineT</th>
<th>4-square</th>
<th>4-squareD</th>
<th>4-clique</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-line</td>
<td><img src="image50.png" alt="Image" /></td>
<td><img src="image51.png" alt="Image" /></td>
<td><img src="image52.png" alt="Image" /></td>
<td><img src="image53.png" alt="Image" /></td>
<td><img src="image54.png" alt="Image" /></td>
<td><img src="image55.png" alt="Image" /></td>
<td><img src="image56.png" alt="Image" /></td>
<td><img src="image57.png" alt="Image" /></td>
</tr>
<tr>
<td>3-clique</td>
<td><img src="image58.png" alt="Image" /></td>
<td><img src="image59.png" alt="Image" /></td>
<td><img src="image60.png" alt="Image" /></td>
<td><img src="image61.png" alt="Image" /></td>
<td><img src="image62.png" alt="Image" /></td>
<td><img src="image63.png" alt="Image" /></td>
<td><img src="image64.png" alt="Image" /></td>
<td><img src="image65.png" alt="Image" /></td>
</tr>
<tr>
<td>4-line</td>
<td><img src="image66.png" alt="Image" /></td>
<td><img src="image67.png" alt="Image" /></td>
<td><img src="image68.png" alt="Image" /></td>
<td><img src="image69.png" alt="Image" /></td>
<td><img src="image70.png" alt="Image" /></td>
<td><img src="image71.png" alt="Image" /></td>
<td><img src="image72.png" alt="Image" /></td>
<td><img src="image73.png" alt="Image" /></td>
</tr>
<tr>
<td>4-star</td>
<td><img src="image74.png" alt="Image" /></td>
<td><img src="image75.png" alt="Image" /></td>
<td><img src="image76.png" alt="Image" /></td>
<td><img src="image77.png" alt="Image" /></td>
<td><img src="image78.png" alt="Image" /></td>
<td><img src="image79.png" alt="Image" /></td>
<td><img src="image80.png" alt="Image" /></td>
<td><img src="image81.png" alt="Image" /></td>
</tr>
<tr>
<td>4-lineT</td>
<td><img src="image82.png" alt="Image" /></td>
<td><img src="image83.png" alt="Image" /></td>
<td><img src="image84.png" alt="Image" /></td>
<td><img src="image85.png" alt="Image" /></td>
<td><img src="image86.png" alt="Image" /></td>
<td><img src="image87.png" alt="Image" /></td>
<td><img src="image88.png" alt="Image" /></td>
<td><img src="image89.png" alt="Image" /></td>
</tr>
<tr>
<td>4-square</td>
<td><img src="image90.png" alt="Image" /></td>
<td><img src="image91.png" alt="Image" /></td>
<td><img src="image92.png" alt="Image" /></td>
<td><img src="image93.png" alt="Image" /></td>
<td><img src="image94.png" alt="Image" /></td>
<td><img src="image95.png" alt="Image" /></td>
<td><img src="image96.png" alt="Image" /></td>
<td><img src="image97.png" alt="Image" /></td>
</tr>
<tr>
<td>4-squareD</td>
<td><img src="image98.png" alt="Image" /></td>
<td><img src="image99.png" alt="Image" /></td>
<td><img src="image100.png" alt="Image" /></td>
<td><img src="image101.png" alt="Image" /></td>
<td><img src="image102.png" alt="Image" /></td>
<td><img src="image103.png" alt="Image" /></td>
<td><img src="image104.png" alt="Image" /></td>
<td><img src="image105.png" alt="Image" /></td>
</tr>
<tr>
<td>4-clique</td>
<td><img src="image106.png" alt="Image" /></td>
<td><img src="image107.png" alt="Image" /></td>
<td><img src="image108.png" alt="Image" /></td>
<td><img src="image109.png" alt="Image" /></td>
<td><img src="image110.png" alt="Image" /></td>
<td><img src="image111.png" alt="Image" /></td>
<td><img src="image112.png" alt="Image" /></td>
<td><img src="image113.png" alt="Image" /></td>
</tr>
</tbody>
</table>

- 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

\[ LNM\!\!C(\Psi^n_i) = \log_e \frac{\text{Actual count of } \Psi^n_i}{\text{Expected count of } \Psi^n_i \text{ in an E-R graph}} \]

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

\[
M-\text{Diff}(LM) = \sum_{k=3}^{4} \sum_i |LNM\!\!C(\Psi^k_i)_{\text{Real}} - LNM\!\!C(\Psi^k_i)_{LM}| \\
M-\text{Sum}(LM) = \sum_{k=3}^{4} \sum_i LNM\!\!C(\Psi^k_i)_{LM}
\]
Bigrams closest to real network

<table>
<thead>
<tr>
<th>Model</th>
<th>M-Diff</th>
<th>M-Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>0.000</td>
<td>74.758</td>
</tr>
<tr>
<td>1-gram</td>
<td>8.781</td>
<td>65.977</td>
</tr>
<tr>
<td>2-gram</td>
<td>1.590</td>
<td>76.348</td>
</tr>
<tr>
<td>3-gram</td>
<td>8.109</td>
<td>82.755</td>
</tr>
<tr>
<td>2-term</td>
<td>1.835</td>
<td>72.923</td>
</tr>
<tr>
<td>3-term</td>
<td>6.113</td>
<td>80.407</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Real percentage</th>
<th>Average rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>60.317</td>
<td>4.046</td>
</tr>
<tr>
<td>1-gram</td>
<td>9.645</td>
<td>2.406</td>
</tr>
<tr>
<td>2-gram</td>
<td>20.098</td>
<td>2.833</td>
</tr>
<tr>
<td>3-gram</td>
<td>28.095</td>
<td>3.276</td>
</tr>
<tr>
<td>2-term</td>
<td>18.072</td>
<td>2.880</td>
</tr>
<tr>
<td>3-term</td>
<td>15.625</td>
<td>2.875</td>
</tr>
</tbody>
</table>
Summary (1)

- 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