Complex Network Analysis of Word Co-occurrence Networks

Natural Language
|K| = 5000 units
|P| = 84,000 units
Sentences formed by units from K and P, or only K
Intra-K edges dominate
K more tightly coupled

Queries
|K| = 1000 units
|P| = 1,200,000 units
Queries mostly formed by units from K and P, or only P
K to P edges dominate
K less tightly coupled

Property
Degree Distribution
Clustering coefficient
Average shortest path

NL
2-regime
0.437
2.670

Queries
2-regime
0.630
3.305

Structural Complexity of Web Search Queries
through the Lenses of Positionality, Language Models and Networks

Rishiraj Saha Roy and Niloy Ganguly, (IIT Kharagpur, India) and Monojit Choudhury (Microsoft Research India)

Change in Segment Positions

- For 2006 and 2010 logs, segments with the highest co-occurrence counts are labeled intent, and the rest as content
- For each segment, query beginning probability \(P_b\), ending probability \(P_e\) and that of occurring in the middle \(P_m\) are computed
- Navigational queries like imdb and youtube are now appended as intent \(P_b\), drops
- Intent segments (how to, news) stabilizing towards ends of query \(P_b\) rises or \(P_e\) rises
- Stacking of intent segments gradually making search queries longer

Perplexity of Language Models

<table>
<thead>
<tr>
<th>Model</th>
<th>NL (Perplexity)</th>
<th>Queries (Perplexity)</th>
<th>NL (Counts)</th>
<th>Queries (Counts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram</td>
<td>1,406.59</td>
<td>6,417.28</td>
<td>0.3M</td>
<td>0.2M</td>
</tr>
<tr>
<td>2-gram</td>
<td>193.722</td>
<td>104.337</td>
<td>3.5M</td>
<td>1M</td>
</tr>
<tr>
<td>3-gram</td>
<td>17.663</td>
<td>5.43</td>
<td>9.7M</td>
<td>1.1M</td>
</tr>
<tr>
<td>2-set</td>
<td>893.851</td>
<td>384.945</td>
<td>48.1M</td>
<td>4.2M</td>
</tr>
<tr>
<td>3-set</td>
<td>N.A.</td>
<td>23.36</td>
<td>N.A.</td>
<td>24.8M</td>
</tr>
</tbody>
</table>

- Perplexity is an information theoretic measure of how perplexed a user is in predicting the \(n^{th}\) word
- Perplexity of unigram model much larger for queries
- In contrast, bigram and trigram perplexity much lower for queries

Conclusions

- Web search queries provide a very interesting case of a complex self-organizing communication system which has its unique characteristics
- Queries structurally simpler than NL, but more complex than bags-of-words model
- Several similarities with NL that make this system interesting to study from a language evolution perspective

Contacts: rishiraj.saharoy@gmail.com
niloy@cse.iitkgp.ernet.in
monojitc@microsoft.com