Abstract

Across the world, millions of users interact with search engines every day to satisfy their information needs. As the Web grows bigger over time, such information needs, manifested through user search queries, also become more complex. However, there has been no systematic study that quantifies the structural complexity of Web search queries. In this research, we make an attempt towards understanding and characterizing the syntactic complexity of search queries using a multi-pronged approach. We use traditional statistical language modeling techniques to quantify and compare the perplexity of queries with natural language (NL). We then use complex network analysis for a comparative analysis of the topological properties of queries issued by real Web users and those generated by statistical models. Finally, we conduct experiments to study whether search engine users are able to identify real queries, when presented along with model-generated ones. The three complementary studies show that the syntactic structure of Web queries is more complex than what \( n \)-grams can capture, but simpler than NL. Queries, thus, seem to represent an intermediate stage between syntactic and non-syntactic communication.

Keywords: Query complexity, Statistical language models, Word co-occurrence networks, Crowdsourcing

1. Introduction

Searching information on the World Wide Web by issuing queries to commercial search engines is one of the most common activities engaged in by almost every Web user [1]. The Web has grown extensively over the past two decades, and search engines have kept pace by incorporating progressively smarter algorithms to keep all the information at our fingertips [2, 3, 4]. This co-evolution of the Web and search engines have driven users to formulate progressively longer and more complex queries, as seen by a rise in mean lengths from 2.4 through 3.5 to about four words per unique query over the last twelve years [5, 6, 7]. Search queries represent a unique mode of interaction between humans and artificial systems, and they differ observably in syntax from that of the parent NL. This has led researchers to argue that probably queries are acquiring linguistic properties of their own [5, 6, 7, 8, 9, 10]. Arguing from the perspective of the function of queries, i.e., communication, and the factors that influence their self-organization, it can be fairly convincingly established that queries are indeed an evolving linguistic system [7]. Nevertheless, there is no systematic and comprehensive study of the syntactic properties of Web queries that can convincingly bring out the fact that queries are indeed a “language”. The challenge, of course, is to identify the unique syntactic features of an NL that make it different from any random or artificially generated sequence of symbols. Fortunately, there exist three lines of research that can address this fundamental question.

1.1. Background

One of the oldest statistical characterizations of NL comes from \( n \)-gram models that can be used for both generation of utterances or sentences to a certain degree of accuracy, and also
for quantifying the predictability (and hence the complexity) of a system of symbols \([12]\). Such a line of research in the past has been extensively used for analyzing ancient languages (like the Indus script hypothesis \([13]\)), studying languages with diverse typological properties \([14]\), and also for understanding non-linguistic systems such as music \([15]\) and the genetic code \([16]\). Over time, n-gram models have been appropriately generalized or restricted using more sophisticated linguistic features capturing various syntactic and semantic properties (see \([17]\) for a review). Collectively, these models are studied under the broad topic of statistical language modeling and extensively used in applications like automatic speech recognition \([18]\), machine translation \([19]\), spelling correction \([20]\) and information retrieval (IR) \([21]\).

A second and more recent line of investigation into linguistic systems is through complex network modeling of languages, where a language is modeled as a network of entities and their relations (see \([22]\) for a review). These studies were inspired by similar modeling techniques employed by physicists and biologists, which led to interesting insights into the systems being modeled. Such studies using network modeling have also revealed some interesting properties of languages \([23, 24]\).

A third approach to characterize a linguistic interaction is to study it from the perspective of the native speakers’ intuition, which says, to quote Noam Chomsky \([25]\): “The sentences generated will have to be acceptable to the native speaker”. Though the concept of a native speaker is debatable and eludes a clear definition \([26]\), in the context of queries it assumes an altogether new dimension, where it would refer to an average user of Web search engines.

1.2. Contributions

These three lines of investigation are, in fact, complementary, and therefore can be very well used for getting a more comprehensive picture of a linguistic system. The necessity of such a multi-pronged approach can also be appreciated in the context of a recent debate on the linguistic status of the Indus valley script: Based on the conditional entropy analysis of n-grams, Rao et al. \([13]\) had claimed that the script was indeed used for an ancient linguistic system; this work was later fiercely criticized and the claim was contested by the computational linguistics community, led by Richard Sproat, who argued that even simple and random generative models can lead to such statistical properties \([27]\). Subsequent work on Indus scripts used network modeling to further substantiate Rao et al.’s original claim \([28]\). However, as we shall see, such an analysis, by itself, is not sufficient. Therefore, if queries are indeed evolving linguistic system, then they should exhibit properties similar to NLs under statistical modeling, network modeling and cognitive analyses. Hence, in this work, we explore the syntactic properties of queries through these three different “lenses” and cross-validate our findings to come up with a holistic view. Specifically, we:

1. Build n-gram and n-term models \([29, 30]\) from our Web query log and analyze perplexity (or predictability) of the models and compare them with that of Standard English;
2. Build word co-occurrence networks (WCNs), the most popular and well-studied network modeling approach for NLs, for the real log and compare the topological properties of the networks with those built from artificial logs generated using the n-gram and n-term models; and finally,
3. Ask ordinary Internet users to rate the acceptability or “real”-ness of the search queries generated by the various language models.

Our study reveals that although queries seem to be more predictable (or less complex) than NL, n-gram models still fall short of generating a rich set of artificial queries. A typical user is able to tell apart a real query from an artificially generated one, even though a trigram-based generative model seems to overfit the data and is capable of confusing the user. The word order in queries seems to be the most important clue helping a user to differentiate between the real and artificially generated queries. Hence, the structure of current Web search queries indicates a linguistic system that has at least a rudimentary word ordering constraint, and several other syntactic and semantic constraints that lie beyond the scope of n-gram and n-term models. In short, queries represent a system which is in between a fully syntactic and a non-syntactic communication system.

We do not know of any previous systematic study on these aspects of Web search queries, even though features like autocomplete and query suggestions are indirect evidence that predictability of query terms is indeed exploited by Web search engines. On the other hand, this study can not only help in more systematic and principled techniques in relevant applications, but can also have an impact on the way we view query understanding today. The complete code for this research has been made available at https://github.com/tamrishiraj/QueryLMSStudy (accessed 02 February 2016).

1.3. Organization

The organization of the rest of this paper is as follows. First, we describe our dataset in Section 2. In Section 3 we present the generative language models for queries used in this work. Next, in Section 4 we discuss our complex network modeling technique and use the framework to compare the various generative models. We then present our experimental framework and results for crowdsourcing to measure human intuition of query syntax (Section 5). Subsequently, we discuss the implications of our results in Section 6 and make concluding remarks in the final section (Section 7).

2. Dataset

For all our experiments, we use a query log sampled from Bing Australia\(^{9}\) in May 2010. This raw data slice consisted of 16.7M (\(M = 10^6\)) queries. We subsequently extracted 11.9M queries from the raw data such that the queries were composed of ASCII characters only and were of length between two and ten words. The justification for imposing a filter

based on query length is as follows. One word queries do not contribute to network structures based on word co-occurrence and are not meaningful to be judged by humans (almost any word is potentially a search query). Very long queries (having more than ten words) are typically computer-generated messages or excerpts from NL text, and need separate query processing techniques (only 0.01% of our log). Out of the extracted queries, 4.7M are unique. In other words, the entire log contains 11.9M queries. However, there are duplicates in the log. If we remove the duplicates, we are left with 4.7M queries. But in order to preserve log properties arising out of the natural power law frequency distribution of queries [6], duplicates were retained for all experiments.

3. Statistical language modeling for queries

The n-gram language model (LM), which assumes that the probability of the nth word in a sentence depends only on the previous (n − 1) words [12], has been one of the most popularly used generative language models for NL with many applications [31, 32, 19]. Therefore, as the first steps, we evaluate query models based on n-grams (and their variants). An n-gram, in our context, is any continuous sequence of n words from a query. Mathematically, an n-gram LM over sequences of words is defined by a Markov chain of order (n − 1) [33], and the probability of a k-word query is given by Equation 1:

\[ P(w_1w_2 \ldots w_k) = \prod_{i=1}^{k} P(w_i|w_{i-1} \ldots w_{i-n+1}) \]  

(1)

Required probabilities can be initialized through training on the real query log from the relative counts, as shown in Equation 2 below:

\[ P(w_i|w_{i-1} \ldots w_{i-n+1}) = \frac{\text{count}(w_i \ldots w_{i-n+1})}{\text{count}(w_{i-1} \ldots w_{i-n+1})} \]  

(2)

Since queries are short (mean length was close to four words for distinct queries in our log), it is not practical to look beyond trigrams. Here distinct queries mean unique queries. Our query log, after removal of duplicate queries, contained 4.7M unique or distinct queries. The average length for these unique queries came out to be 3.98, very close to four words.

An n-term [29, 30] is an unordered set of n words, all of which occur in a query, but not necessarily next to each other. Since queries have been considered to be bags-of-words in several contexts [34, 35, 21, 36], a systematic exploration of n-terms becomes necessary. The meaning of an n-term is equivalent to an unordered skipgram [37]. We also note that n-term models are akin to the full dependence model [38], which can capture long-range dependencies between terms but ignore their ordering and proximity.

3.1. Query generative models

We explore five generative models for queries, viz. 1-gram, 2-gram, 3-gram, 2-term, and 3-term. Example queries generated by each of these models are shown in Table 1. 1-gram queries typically do not make much sense; they are just some random words thrown in to create a query, albeit maintaining the real word probability distribution. The notion of local coherence immediately becomes clear as one moves to 2-grams. The crossovers from one bigram to the next seem smooth, but as a whole the queries are generally not meaningful. Clearly, the chances of a query being meaningful in its entirety diminishes with increasing query length. The same thing can be observed for trigrams; but since several queries only have three, four or five words (can be generated with only one, two or three trigrams, respectively), such queries are often realistic. For 2-terms and 3-terms, we have similar observations but it is also apparent that the sequencing of the words is not very natural. Another point to note is that such models often contain word repetitions. Both these aspects are fall-outs of the relaxation of the strict ordering constraint.

3.1.1. Query generation process

We now explain our process of generating artificial queries using n-grams or n-terms. We denote the query length (in words) as L, a random variable that can take integral values between two and ten. Query length distribution refers to the probability distribution of L, which is empirically estimated from the log. All the generation models described here are made to follow this distribution. The process for generating artificial queries using n-grams is as follows (the process is exactly analogous for n-terms): first a value l of L is stochastically sampled from the query length distribution. Then, an n-gram is sampled in proportion to its probability, forming the initial query. An extra word is added to this initial query trying to extend the previous string of (n − 1) words. This new word is chosen probabilistically from all the n-grams which have their first (n − 1) words as the string concerned. This process is continued till the desired query length, i.e., L, is reached. When a query generation process is unable to add a new word to a partially generated query using an n-gram model, then it backs off to an (n − 1)-gram generative model to generate the new word.

Example of an n-gram module query generation. We ex-

<table>
<thead>
<tr>
<th>Model</th>
<th>Example queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram</td>
<td>a dhcp ephemeral detailing</td>
</tr>
<tr>
<td></td>
<td>create user account on roads</td>
</tr>
<tr>
<td>2-gram</td>
<td>access 2003 not working with info</td>
</tr>
<tr>
<td></td>
<td>shaolin kung fu panda</td>
</tr>
<tr>
<td>3-gram</td>
<td>a brief history of witchcraft</td>
</tr>
<tr>
<td></td>
<td>a thousand miles sheet music for websliders</td>
</tr>
<tr>
<td>2-term</td>
<td>adobe acrobat free download windows</td>
</tr>
<tr>
<td></td>
<td>adelaide entertainment explained seating plan area</td>
</tr>
<tr>
<td>3-term</td>
<td>anti software virus windows vista</td>
</tr>
<tr>
<td></td>
<td>adobe photoshop 7 for mac</td>
</tr>
</tbody>
</table>
plain the generation process with one of the example queries for 3-grams in Table 1: a thousand miles sheet music for websliders. First, we stochastically sample a query length seven from the real log query length distribution, which implies that the generated query will have seven words. Next, the 3-gram a thousand miles is sampled from the list of all real trigrams based on its occurrence probability. We now try to extend the query by looking at all 3-grams that start with the 2-gram thousand miles. Using similar frequency-biased sampling as above, we obtain the trigram thousand miles sheet. Next, a search for 3-grams beginning with miles sheet fails, and hence we back off to the 2-gram model and sample for 2-grams starting with sheet, which produces sheet music. We resume our search for trigrams using the rightmost 2-gram and add the words for and websliders in consecutive iterations. No backing off was required in the last two steps. Since we have now reached the desired query length of seven, the query generation process stops.

We do not choose a fixed length for all the queries in the artificial logs — we maintain the query length distribution of the real log. In other words, the proportion of queries of a particular length (in words) is approximately the same in an artificial log as in the real log. In our algorithm for generating artificial logs, we first stochastically pick a query length \( l \) and continue the query generation process till the query reaches \( l \) words, and then stop. Since our query generation process (for any language model) works by adding one word at a time, it is not difficult to stop the query at a particular number of words. For more discussion on this topic, see Section 3.1.2.

**Example of an \( n \)-term model query generation.** We explain the query generation process for an \( n \)-term model with one of the example queries in Table 1: a thousand miles sheet music explained seating. Next, we resume the search for 3-terms containing any two of the words from the following set with four items: \{adelaide, entertainment, explained, seating\}. We sample \{entertainment, seating, plan\} and the new word plan is appended to the end of the query. In the next step, we sample \{adelaide, seating, area\} and so add area to the query. No backing off was required in the last two steps. Since we have now reached the desired query length of six, the query generation process stops.

We create the \( n \)-terms by the following process: we make one pass on the log and identify the unique \( n \)-terms (each query with \( k \) unique words contributes \( C(k, n) \) unique \( n \)-terms, \( k \geq n \)). In the same pass, we can also compute the frequencies of each of these \( n \)-terms, by incrementing an \( n \)-term count by one each time we encounter a duplicate. We then normalize the \( n \)-term frequencies with the number of queries in the log to obtain the \( n \)-term probabilities. Now, using these \( n \)-term probabilities, we follow the algorithm described earlier to generate the query. In short, we first stochastically sample a query length (in words) \( l \) from the real length distribution. Next, we choose an \( n \)-term in proportion to its occurrence probability. We then try to extend the query by one word by selecting an \( n \)-term which has \((n-1)\) words (any \((n-1)\) words, not necessarily the last) in common with the current (partial) query. This process is repeated using similar frequency-biased sampling of \( n \)-terms till the desired query length \( l \) is reached. In case an \( n \)-term is not found with the required words in common, we back off to an \( n \)-term model which is one order less than the current model.

Words are often repeated in \( n \)-term queries because of the following reason. When an \( n \)-term is selected for a query, it usually has a relatively high probability. Next, its words do not go out of contention for being reselected for the query. For example, if the 3-term \{anti, software, virus\} (anti software virus in the query) is selected, the next step is to look for a relatively high-probability 3-term with any two words of the query in common. It is quite probable that the 3-term \{anti, virus, software\} is selected again, with the query now becoming anti software virus anti or anti software virus software or anti software virus. In our model, we did not explicitly disallow the original \( n \)-term not to be reconsidered for query extension. Moreover, while extending the query in an \( n \)-term model, it is not only the last \((n-1)\) words that are considered. Instead, the new \( n \)-term to be selected may have any \((n-1)\) words of the query in common. For example, in the query housing thailand what is cost of housing, once we have housing thailand what is cost of, the next 3-term can have any of the two words from the set \{housing, thailand, what, is, cost, of\}. So, since a frequent \( n \)-term always has the possibility of being “re-selected”, we often see duplicate words in the \( n \)-term model-generated queries. The initial intuition behind our exploring the \( n \)-term models was to examine the hypothesis that queries were relatively free-word order, i.e., word order did not matter much for the user while formulating the query. We observed that \( n \)-term models failed to generate realistic queries, which gave us...
evidence that word order was indeed important (Section 5).

The algorithms for generating artificial logs from real logs are presented in Algorithms 1 (n-gram-Log-Gen) and 2 (n-term-Log-Gen). An expression refers to any n-gram or any n-term in the two algorithms, respectively. The backoff step is shown in bold in both algorithms. The steps unique to the n-term model are marked with comments in boldface.

3.1.2. Query termination

Deciding the final length of a query based on the partial query generated so far is an interesting possibility. However, it is non-trivial, or rather extremely hard to decide the length in the n-term model based on the current partial query, using syntactic properties only. So we have taken a more natural and non-trivial, or rather extremely hard to decide the length in the generated so far is an interesting possibility. However, it is

non-trivial, or rather extremely hard to decide the length in the

n-term model based on the current partial query, using syntactic properties only. So we have taken a more natural and non-trivial, or rather extremely hard to decide the length in the generated so far is an interesting possibility. However, it is

3.2. Measuring model perplexity

Perplexity is one of the most common metrics used for evaluating n-gram systems. It can be intuitively thought of as the weighted average number of values that a random variable can take. Thus, perplexity of an n-gram model tells us that if, on an average, a string of \((n-1)\) words of the language are known, how many words are likely to occur in the next position. This indicates the number of words that can follow a given \((n-1)\)-gram in the language. A higher value means that the \(n^\text{th}\) word is less predictable from the previous \((n-1)\)-word context. A higher perplexity value for a language model thus implies less certainty in the user’s mind about its predictability. The perplexity of a probability distribution \(p(x)\) of a random variable \(X\)

is defined as \(2^{H(X)}\), where \(H(X)\) is the entropy of \(X\) and is given by Equation 5

\[
H(X) = -\sum_{x \in X} p(x) \log_2 p(x)
\]

(3)

**Computational method.** For \(n = 1\), the computations are straightforward, i.e. the probability distribution of unigrams is the one whose entropy (and thus perplexity) is calculated. For \(n\)-grams and \(n\)-terms where \(n > 1\), the weighted average of entropies over all \((n-1)\)-grams is considered as the entropy of the model. The weight is the probability of occurrence of the corresponding \((n-1)\)-gram in the corpus. The entropy of a particular \((n-1)\)-gram is the entropy of the probability distribution over all words that can appear in the \(n^\text{th}\) position given that the first \((n-1)\) words are fixed. This entropy, raised to the power of two, gives the corresponding perplexity. Specifically, every \(n\)-gram slot’s appearance is referred to by some random variable \(X\), with probability distribution \(p(x)\). The individual \(p(x)\)-s refer to the points in the probability distribution \(p(x)\), or instantiations in which the slot can be filled. For an example with the 2-gram model, let \(X\) represent the 2-gram slot \(<\text{apple} \ldots>\), where ... can be filled by juice, tart or pie. Then, apple juice, apple tart and apple pie would be represented by the \(x_i\)-s \((x_1, x_2, x_3)\), their occurrence probabilities would be denoted by the individual \(p(x)\)-s \((p(x_1), p(x_2), p(x_3))\). Then, we would compute an entropy for \(<\text{apple} \ldots>\) using Equation 3 and the final entropy of the 2-gram model would be the weighted sum of the entropies of all such 2-gram slots, the individual weights being the occurrence probabilities of the respective 1-grams in the first positions of the 2-grams (like apple). The individual \(p(x)\)-s \(<1\) and the individual \(\log_2 p(x)\)-s \(<0\), but the negative of the resultant summation, i.e., \(H(X)\), is not less than zero. It varies between 0 and \(\log_2 N\), where \(N\) is the number of points in the probability distribution \(p(x)\). Thus, the perplexity, which is \(2^{H(X)}\), varies between 1 and \(N\).

3.3. Experimental results

Table 2 reports the perplexities and counts of the different models for NL and Web search queries. The corresponding perplexity values can be obtained simply by raising 2 to the power of the entropy value \(H(X)\) (as obtained from Equation 3). For NL, the corpus used contained 1M randomly sampled sentences

<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>1,406.593</td>
<td>6,417.283</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>5.340</td>
<td>9.7M</td>
<td>1.1M</td>
</tr>
<tr>
<td>2-term</td>
<td>893.851</td>
<td>384.945</td>
<td>48.1M</td>
<td>4.2M</td>
</tr>
<tr>
<td>3-term</td>
<td>N.A.</td>
<td>23.360</td>
<td>N.A.</td>
<td>24.8M</td>
</tr>
</tbody>
</table>

* Dictionary runs out of memory even with 64 GB of RAM.
from newswire data in 2010. Newswire text was chosen because, in general, they contain cleaner NL sentences than random Web data. For comparability of NL values with queries, we kept the dataset size similar for the latter by randomly sampling 1M queries from our dataset. To preserve the natural frequency distribution, duplicates were not removed from either dataset. The NL text was case-folded and only alphanumeric characters (and whitespace) were retained. Perplexity values for Standard English reported in Brown et al. [45] are obtained from corresponding cross-entropy values and hence are not directly comparable to those in Table 2.

### 3.4. Interpretation

It is quite interesting to note that while the perplexity of the unigram model for queries is much higher than that of NLs, the perplexity of bigrams and trigrams show just the opposite trend. The explanation for this surprising trend is as follows. We observed in our data that the rate of encountering a new word for queries is much higher (about one per 20 words) than NL (about one per 58 words). Hence, the unigram distribution of queries is more diverse than NL. In fact, queries have a much larger specialized or peripheral vocabulary and very small core vocabulary as compared to NLs [23, 47]. This unique feature makes the perplexity of the unigram model very high for queries.

On the other hand, queries are also repeated and their repetition frequency is known to follow a power-law distribution [6]. In NL, sentences are rarely repeated exactly, except for phrases
Algorithm 2 n-term-Log-Gen(Q, n, numQueries)

Require: Real query log Q, model number n (n = 3 for 3-term model), no. of queries to be generated numQueries
1: E ← ∅ \[→ E is the list of all expressions in Q (n-terms, (n-1)-terms, . . . , 1-terms), initialized to NULL\]
2: for all q ∈ Q do \[→ For each query q in input log Q\]
3: Extract all expressions E_q from q \[→ n-terms required for basic model, lower order expressions required for backoff\]
4: for all expressions e ∈ E_q do \[→ For every expression e in the query q\]
5: if e ∈ E then \[→ Expression e not found in list E\]
6: E ← E ∪ {e} \[→ Expression e added to list E\]
7: count(e, Q) ← 1 \[→ Frequency of e in Q is set to 1 when encountered for the first time in Q\]
8: else \[→ Expression e has been encountered in Q earlier\]
9: count(e, Q) ← count(e, Q) + 1 \[→ Increment count of e in Q by one\]
10: end if
11: end for
12: end for
13: for all e ∈ E do \[→ For every expression e in Q\]
14: prob(e, Q) ← count(e, Q)/|Q| \[→ Compute probabilities of expressions e in Q\]
15: end for
16: Partition E into E(3), E(2), . . . , E(0) by no. of words in expression \[→ 1-terms go to E(3), 2-terms to E(2), . . . , n-terms to E(0)\]
17: Q ← ∅ \[→ List of generated queries Q initialized to NULL\]
18: while |Q| < numQueries do \[→ Generate queries till we reach the desired number numQueries\]
19: qgen ← stochastic-sample-fn(lendist) \[→ lendist = Real-len-dist(Q), we use a stochastic sampling function stochastic-sample-fn(·) to sample a query length qlen from the word length distribution function of the real queries Real-len-dist(·)\]
20: q ← while \[→ Query to be generated q initialized to empty string\]
21: e ← freq-biased-sample-fn(E_q) \[→ We use a frequency-biased sample function freq-biased-sample-fn(·) to sample an expression e’ from a list in proportion to its occurrence probability in Q\]
22: e’ ← permute(e) \[→ Create an n-gram e’permute by an arbitrary permutation of the words in e’\]
23: q’ ← concat(q, e’permute) \[→ Concatenate e’permute to q to form q’\]
24: while len(q’)/qlen do \[→ Continue till we reach the desired query length\]
25: select-flag ← FALSE \[→ select-flag is a Boolean variable indicating whether we have found a new e’ to extend q’\]
26: while select-flag = FALSE do \[→ Try to select the next expression e’ for extending q’, x acts as a model selector\]
27: x ← 1 \[→ x must be the same as any (n − i) of an expression in E(n−i+1), the set\]
28: while x ≤ n − 1 do \[→ Non-zero candidate expressions found\]
29: E’ ← ∅ \[→ Break loop\]
30: if x < n − 1 then \[→ No expressions found\]
31: for all E’(x−1) ∈ E(n−x−1) \[→ Back off to lower order model\]
32: if [set(q’)/x] ∈ E(n−x−1) + n − x then \[→ Sample from E’ in proportion to frequency\]
33: function converts a string into a bag-of-words set form, n-terms in E(n) are already in set form \[→ An expression has been picked for query extension\]
34: E’ ← E’ ∪ E’(1) \[→ End while select-flag = FALSE\]
35: end if
36: else \[→ End while\]
37: E’ ← E’(1) \[→ End while select-flag = FALSE\]
38: end if
39: if E’ ≠ ∅ then \[→ Non-zero candidate expressions found\]
40: break \[→ No expressions found\]
41: else \[→ Break loop\]
42: x ← x + 1 \[→ Non-zero candidate expressions found\]
43: if x < n − 1 then \[→ Non-zero candidate expressions found\]
44: E’ ← freq-biased-sample(E’) \[→ Sample from E’ in proportion to frequency\]
45: select-flag ← TRUE \[→ An expression has been picked for query extension\]
46: end while \[→ End while\]
47: new-word ← e’ \& set(q’) \[→ The new word to be added is the set difference between the words in the query so far q’ and the new expression e’\]
48: q’ ← concat(q’, new-word) \[→ Concatenate the new word to the query so far\]
49: end while \[→ End while\]
50: Q ← Q ∪ q’ \[→ End while\]
51: numQueries ← numQueries + 1 \[→ End while\]
52: end while \[→ End while\]
53: return Q \[→ Return list of generated queries\]

like Thank you and Good morning. In our dataset, the number of duplicates (repeated at least once) in NL and queries were found to be 447 and 164, 185 respectively. Furthermore, the mean sentence lengths for NL and queries were found to be 18.159 and 3.980 words, respectively. These two factors play a crucial role in bringing down the bigram and trigram perplexities for queries. One interesting conclusion from these findings on perplexity is that for a random sentence or query, a native speaker (or search engine user in case of queries) will be able to predict a random word present in an NL sentence much more certainly than for a query. On the other hand, if one or more words are shown, it is much easier to predict the rest of the words in a query than in an NL sentence. This is precisely why an autocomplete feature can work much better for Web search engines than for a word editor. An alternative perspective is as follows: Web search queries generated using bigrams or trigrams will look much more realistic than NL sentences generated using the same LMs.

The perplexities of n-terms are greater than their corresponding n-grams due to the manifold increase in the number of possibilities in the former (Table 2). The number of 3-grams is comparable to the number of the 2-grams for queries because of the presence of a substantial number of 2-word queries, that do not contribute to 3-gram counts.

The perplexity experiments were conducted to examine the hypothesis of queries evolving into a distinct linguistic system, different from the mother NL from which the words of the “query language” are borrowed. Had the n-gram model
perplexity values between NL and queries been almost similar, the above hypothesis could have been directly questioned. However, that was not the case and the values differed notably. Nevertheless, just having entropy or perplexity similar to or lower than NL need not, by itself, be indicative of an underlying language system \([27]\), and hence further experimentation was deemed necessary. The low perplexity of trigram models in query logs can be related to the fact that trigram models are restrictive in generating synthetic queries and probably overfit the data.

### 4. Complex network modeling for queries

Network analysis provides an elegant mathematical framework to study various complex systems \([48, 49, 50]\). The success of such network-based techniques in the last couple of decades is primarily due to the fact that a network can capture aggregate properties of a system, while considering both local and long range (global) interactions present in a system. Of special interest to us here is the application of network models to linguistics and corpus studies \([51, 22]\). The most popular and well-studied representation of a language corpus is the word co-occurrence network (WCN) \([23, 24, 52, 39]\), which we have applied here to study the statistical properties of query logs. In the recent past, such WCNs have also been used for term weighting in IR \([53]\). The acronyms used in this section are expanded in Table 3.

### 4.1. Network definition and construction

A WCN for any given text corpus is defined as a network \(\mathcal{N} : (N, E)\), where \(N\) is the set of nodes each labeled by a unique word and \(E\) is the set of edges. Co-occurrence can be defined on a query-level, sentence-level, paragraph-level, document-level, word and corpus-level or in a window of \(|N| \approx 180,000\) (both for local and global models), while \(|E| \approx 1.5M\) (local) and \(|E| \approx 2.0M\) (global), on an average. Thus, these are very sparse networks with average edge density \(i.e., p_{ij} = |E|/(|N|^2)\) of the order of \(10^{-4}\).

### 4.2. Topological properties of WCNs

We now explain the topological properties of word co-occurrence networks that we use for characterizing real and generated query logs, namely, degree distribution, clustering coefficient, average shortest path length, and network motifs.

#### 4.2.1. Degree distribution

The degree of a node in a network is the number of nodes that it is connected to. The degree distribution (DD) of a network is the probability distribution \(\pi_k\) of a node having a degree.
k. A cumulative degree distribution (CDD) \( \hat{P}_k \) (probability of a node having degree \( \geq k \)) is more robust to noisy data points and is preferred for visualization. A representative CDD for a query WCN built from \( 1M \) randomly sampled queries is shown in Figure \( 2 \) (local model; global model is almost exactly similar) and is found to resemble a two-regime power law. This is also found for NLs \([23]\) and indicates the presence of a kernel-periphery structure. A kernel is a small subgraph of the network where all nodes have very high degrees. The nodes in the periphery have relatively lower degrees than the nodes in the kernel. The majority of the nodes in a query WCN are observed to form very small peripheral clusters which are all connected to the kernel.

### 4.2.2. Clustering coefficient

Let a node \( r \) in the network have \( k \) neighbors. Then, \( \binom{k}{2} \) edges are possible among its neighbors. The clustering coefficient (CC) of \( r \), \( CC_r \), is the fraction of these edges that actually exist in the network \([54]\). The \( CC_N \) of the entire network \( N \) is defined as the average of \( CC_r \) over all \( r \in N \). A high CC indicates that the network consists of one or more dense subgraphs or clusters. The average CCs for the real WCNs (built from \( 1M \) random queries) are 0.429 (local) and 0.521 (global). The CC for the global network is slightly higher because of the higher edge density. These values are quite high as compared to the CC of an E-R random graph \([55, 23]\), which is of the order of \( 10^{-5} \). CC for similar networks for NL has been reported to be 0.437 \([23]\). This result shows lower density for NL WCNs than those for queries.

### 4.2.3. Average shortest path length

The shortest path length between a pair of nodes is the minimum number of edges that one must traverse to reach one node from the other. The average shortest path length (ASPL) is defined as the mean of the shortest path lengths between all pairs of nodes in the network that can be reached in a finite number of steps from each other. The ASPL for the WCN built from a \( 1M \) real query sample is 3.519 (local) and 3.004 (global). The ASPL for the global network is slightly lower because of the higher edge density. These values are quite small for a network with close to 180,000 nodes, and near to the expected ASPL for an E-R random graph of similar size and density (4.253 (local) and 3.830 (global)). ASPL for an E-R random graph is given by \( \ln|N|/\ln(k) \) \([54]\), where \( k \) is the average degree of the graph, given by \( 2 \times |E|/|N| \). It has been argued that the low ASPL for similarly constructed NL WCNs (2.67) is an outcome of an optimization of language structure necessary for fast recognition and retrieval of words \([23]\). As an aside, we also note that a network with high CC, low ASPL and low edge density (all with respect to random graphs) is known as a small world network \([50]\). Hence, like social networks and WCN for NLs, WCNs for queries are also examples of small world networks.

### 4.2.4. Network motifs

Network motifs are small subnetworks that are found to occur in significantly higher numbers in real networks than in random networks \([50, 57, 58, 59, 60, 61]\). For example, cliques with four nodes have an occurrence probability of \( 10^{-11} \) in a real WCN, while their expected probability in an E-R random graph \([55]\) with the same number of nodes and edge density, is only \( 10^{-20} \). A \( \Psi^m \) motif is defined as a subgraph of \( n \) distinct nodes in the network, unique to structural isomorphism. Counting motifs for large graphs is computationally expensive, because beyond \( \Psi^4 \), motifs typically have a very large number of possible isomorphisms. In this study, we only consider connected \( \Psi^3 \) and \( \Psi^4 \) motifs. The motif detection has attracted attention as a useful technique to uncover structural design principles of networks in various domains like biochemistry, neurobiology, ecology, and engineering \([57, 60]\). Here, we use the algorithm FANMOD \([60]\) to detect \( \Psi^3 \) and \( \Psi^4 \) motifs.

Figure \( 3 \) enumerates all the connected \( \Psi^3 \) and \( \Psi^4 \) motifs. Since motif counts are dependent on the size of a network, they must be suitably normalized by their corresponding expected counts in an E-R random graph with the same number of nodes and edge density \([60]\). Since the ratios of the probabilities can be very skewed, we take the natural logarithms of these quantities for meaningful comparisons of values, which we shall refer to as the Log Normalized Motif Counts (LNMC), defined in Equation \( 4 \) below:

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

where, \( \Psi^n_i \) is the \( i^{th} \) \( n \)-sized motif. For example, following Figure \( 3 \), \( \Psi_3^3 \) and \( \Psi_4^3 \) would be the 3-clique and the 4-loop-out respectively. Figure \( 3 \) (third row) reports the LNMC values for \( \Psi^3 \) and \( \Psi^4 \) motifs for the real WCN (local co-occurrence model). This vector of eight elements for motifs is referred to as the motif signature of the network \([59]\). These ratios indicate that the probabilities of occurrence of all the connected motifs in a real WCN are several orders of magnitude higher than probabilities in an E-R random graph. Figure \( 3 \) also contains...
Figure 3: Connected $\Psi^3$ and $\Psi^4$ motifs with examples and LNMCs from the real log. LNMC represents log normalized motif counts, i.e., logarithms of motif counts normalized by expected counts (or, corresponding probabilities) in an E-R random graph.

<table>
<thead>
<tr>
<th>$3$-chain ($\Psi^3$)</th>
<th>$3$-clique ($\Psi^3$)</th>
<th>$4$-chain ($\Psi^4$)</th>
<th>$4$-star ($\Psi^4$)</th>
<th>$4$-loop-out ($\Psi^4$)</th>
<th>$4$-box ($\Psi^4$)</th>
<th>$4$-semi-clique ($\Psi^4$)</th>
<th>$4$-clique ($\Psi^4$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mercury motherboard</td>
<td>element</td>
<td>planet</td>
<td>log</td>
<td>element motherboard</td>
<td>mercury element</td>
<td>water requirements</td>
<td>water steel</td>
</tr>
</tbody>
</table>

real examples of each type of motif from the network. Motifs in WCNs capture semantic relatedness between the words, and certain $\Psi^4$ motifs like boxes and chains are representative of semantic concepts like synonymy and polysemy, respectively [39]. We found motifs to offer insights into query semantics as well (Section 6).

Basic network properties like the numbers of nodes and edges, the degree distribution, the clustering coefficient and the average shortest path length are indeed very commonly used metrics for comparison. However, it is true that they cannot completely capture the complexity of the model generating the network, and can be replicated to a good extent using simple models. On the other hand, using motif signatures is a fairly new technique for this purpose, being proposed by Biemann et al. in 2012 [39]. While certain motifs like $4$-cliques or $4$-semi-cliques do represent strong connections among nodes in the network when present in substantially high counts, other motifs like chains, boxes, loop-outs and stars represent interesting semantic relationships among its nodes. Non-existent edges in the motifs represent non-connections between pairs of nodes, which may often convey non-trivial interpretations. Exactly replicating the motif signature of a network appears to be fairly hard, and thus the proximity of the motif signature of a network built from model-generated query logs (or NL sentences) to the signature of the network built from real logs can indeed shed light on the complexity of the model [39]. A motif signature is a more effective property for comparing networks at a macroscopic or aggregate level, than the standard characteristics like degree distribution, clustering coefficient and average shortest path length.

4.3. Stability of WCNs

The trends in the network statistics described above are useful only if they remain reasonably fixed when the size of the log is varied. Hence, to analyze the stability of the WCN statistics, we varied the network size by controlling the number of queries from which the network is created. Let $W (= 1M$ here) be the number of sample queries used to build a large WCN. We construct smaller query logs consisting of $W/10$, $W/20$, $W/50$, and $W/100$ queries by random subsampling of the entire log and computing the network statistics. To minimize sampling bias, for each $W/s$-sized log, the experiments were repeated $s$ times and statistics were averaged over these $s$ instances (not applicable for DD). Table [4] reports $|V|$, $|E|$, CC and ASPL for each of these sizes. Subscripts $l$ and $g$ represent local and global models of co-occurrence. Figure [4] shows the CDD plots for the respective networks (one specific sample from each network size; local co-occurrence model). In this figure, from the top left, we show degree distributions for networks constructed from query logs of sizes $W$, $W/10$, $W/20$, $W/50$, and $W/100$. We do not reduce the network size beyond $W/100$ because $W/1000 \approx 1000$ queries which is too small for any reliable network analysis.

The results (mean values over $s$ experiments) in Table [4] show that the statistics are extremely robust to network size variation. Similar stability for WCNs for NL text has been reported in in Biemann et al. [39]. Importantly, the standard deviations for the $s$ experiments were found to be very low in all cases, further indicating network stability (of the orders of $10^{-3}$ for CC and $10^{-2}$ for ASPL respectively). Even when the dataset size is decreased by two orders of magnitude ($W$ to $W/100$), the CC and ASPL change only by $\simeq 17\%$ $(g)$ to $19\%$ $(l)$ and $\simeq 5\%$ $(g)$ to $6\%$ $(l)$, respectively. We also note an interesting trend here – both CC and ASPL increase as the network size decreases. This is slightly counter-intuitive, because a large network with a small ASPL is expected to have a high CC. However, the trend is explained as follows. As the network grows in size due to an increase in the number of queries, (a) new nodes join the network with edges connecting them to existing nodes; and (b) new edges form between existing nodes. Since new nodes do not immediately have several new connections, the first event decreases the CC and increases the ASPL. The second event decreases the ASPL and increases the CC. However, these rates of increase and decrease of CC and ASPL caused by the two events are not the same. For Web queries, with lots of new and rare words (arising from various proper nouns) continuously joining the network, it is the first event that mostly dominates and is responsible for the drop in CC and increase in ASPL.

Examining the degree distribution, it is evident that the network stabilizes to its final form marked by a two-regime power law, at about $W/10$, which is $100,000 = 0.1M$ queries. The motif statistics for the local network indicate similar stabilization trends and have been shown in Table [5]. Motif results for the global network also show similar behavior and hence are not reported here. From these results, we infer that for dependable query WCN analysis, one must have at least $0.1M$ queries in their sample.

Proximity of local and global co-occurrence networks. An important observation from these experiments on network stability is the relative invariance in the properties of local and global co-occurrence networks for queries. Even though the global networks have more edges, differences in the CC and ASPL values are very small. Moreover, trends observed in degree and motif distributions are also quite similar. Thus, in the rest of this paper, results will be reported only on local WCNs.
Table 4: Basic network statistics for various network sizes (local and global models indicated by subscripts \(l\) and \(g\) respectively). The properties include the numbers of nodes and edges, the clustering coefficient and the average shortest path length, of the largest connected component in the network. \(W\) represents 1M queries.

| Log size | \(|N|_l\)  | \(|N|_g\)  | \(|E|_l\)  | \(|E|_g\)  | \(CC_l\)  | \(CC_g\)  | \(ASPL_l\)  | \(ASPL_g\)  |
|----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| \(W\)    | 177,900   | 177,975   | 1,526,043 | 2,089,729 | 0.429     | 0.521     | 3.519     | 3.320     |
| \(W/10\) | 46,528    | 46,523    | 299,496   | 412,218   | 0.463     | 0.555     | 3.536     | 3.328     |
| \(W/20\) | 30,398    | 30,399    | 169,670   | 233,932   | 0.474     | 0.566     | 3.567     | 3.355     |
| \(W/50\) | 17,160    | 17,162    | 77,052    | 106,306   | 0.493     | 0.588     | 3.644     | 3.410     |
| \(W/100\)| 10,991    | 10,990    | 41,270    | 56,992    | 0.512     | 0.611     | 3.741     | 3.481     |

Table 5: \(Ψ^3\) and \(Ψ^4\) Motif signatures at various network sizes (local co-occurrence). The vector composed of the LNMC values of the two 3-motifs and the six 4-motifs is referred to as the motif signature of a network.

<table>
<thead>
<tr>
<th>Log size</th>
<th>3-chain</th>
<th>3-clique</th>
<th>4-chain</th>
<th>4-star</th>
<th>4-loop-out</th>
<th>4-box</th>
<th>4-semi-clique</th>
<th>4-clique</th>
</tr>
</thead>
<tbody>
<tr>
<td>(W/20)</td>
<td>2.873</td>
<td>7.172</td>
<td>4.066</td>
<td>5.890</td>
<td>10.251</td>
<td>7.189</td>
<td>15.427</td>
<td>21.460</td>
</tr>
</tbody>
</table>

Table 6: Mean network statistics for the query LMs. Since DD cannot be summarized by a single average value, the Kullback-Leibler Divergence (KLD) values are computed between the DDs (after applying add-one Laplace smoothing) of the real networks and the DDs of the model-generated networks. The CC, ASPL and KLD values for the models with minimum deviations from those for the real network are shown in boldface.

| Model | \(|N|\)  | \(|E|\)  | CC   | ASPL | KLD  |
|-------|---------|---------|------|------|------|
| Real  | 34,209  | 242,680 | 0.623| 3.302| 0.000|
| 1-gram| 28,748  | 311,955 | 0.280| 2.823| 0.349|
| 2-gram| 33,257  | 209,947 | 0.619| 5.011| 0.077|
| 3-gram| 60,594  | 292,210 | 0.591| 3.472| 0.068|
| 2-term| 28,978  | 227,146 | 0.630| 4.968| 0.054|
| 3-term| 47,292  | 249,966 | 0.634| 3.538| 0.031|

4.4. Comparison of real and model-generated query WCNs

We have seen that the statistical properties of the networks are stable as long as the log consists of at least 0.1M queries. Therefore, for reliable results, we decided to conduct all our experiments on logs having 1M queries. We sampled the entire real query log to construct 100 subsamples of 1M queries each, each subsample preserving the query length distribution by words. These will serve as our real logs for all the following experiments. Similarly, we stochastically generated 100 logs, each consisting of 1M queries, for each of the five generative models. We constructed the WCNs for these 6 (real and five models) \(\times 100 = 600\) logs and computed the DD, CC, ASPL and motif signatures for each network. We observed negligible variance in the network statistics across the 100 samples generated from the same model, which further demonstrates the robustness of network modeling. Thus, we report only the average values for \(|N|\), \(|E|\), CC and ASPL in Table 6 and the average LNMC values for the connected \(Ψ^3\) and \(Ψ^4\) motifs in Table 7. We note that the values reported here for the real log do not necessarily match those corresponding to \(W\) in the previous sub-section because now the sampling is done preserving the query length distribution by words; the sampling was intentionally random during the experiments on network stability for emphasizing the idea of stability. Henceforth, we will refer to the WCNs generated from the real query logs as real networks and the WCNs generated from the model-based query logs as model-generated networks.

Since DD cannot be summarized by a single average value, the Kullback-Leibler Divergence (KLD) \(\text{KL}(\cdot)\) is computed between the DDs (after applying add-one Laplace smoothing \(\text{KL}(\cdot)\)) of the real networks and the DDs of the model-generated networks. These values are also reported in Table 6. The smaller the KLD, the closer is the DD of the network to that of the real WCN. Figure 5 shows the CDDs for one of the subsamples each from the real and the model-generated networks. The degree \(k\) is the number of nodes that a particular node is connected to, in the network. The plots have been split into two groups, to somewhat mitigate the problem of almost completely overlapping curves.

**Smoothing for KLD.** We smooth the graph DDs to make them suitable for KLD computation in the following way. The KLD between two probability distributions \(P\) and \(Q\) is defined as shown in Equation 5.

\[
D_{\text{KL}}(P \parallel Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}
\]
avoid such cases, we must smooth the DD of $M$ at all points where $Q(k)$ is zero. First, if the maximum degree for the real network ($k_{\text{max}}(R)$) is greater than that for $M$ ($k_{\text{max}}(M)$), we set the maximum data point for $Q$ to be at $k_{\text{max}}(R)$. Now for $M$, we first compute smoothed frequencies (number of nodes having a particular degree) and then compute the probabilities. Say, all the degree points $(a+i)$ in $M$ between degree $a$ and degree $b$ are missing (zero), where $i$ is an integer and $i_{\text{min}} = 1$, where $a < b$ and $\text{freq}(a) < \text{freq}(b)$. Then, we replace all these missing values with $\text{freq}(b) - \text{freq}(a)/(b - a)$, and this value is rounded off to the nearest integer. If it becomes zero, we add one to all frequencies in $M$ (add-one Laplacian correction) [63]. We can similarly handle cases where $\text{freq}(a) > \text{freq}(b)$. If the last $t$ degree points for $M$ are missing (zero) and the last degree that has a non-zero frequency is $m$, then the frequencies of the last $t$ points are replaced by $\text{freq}(m)/t$ and this value is rounded off to the nearest integer, and the add-one Laplacian correction is applied if necessary. Finally, the corresponding probabilities are computed by dividing the updated frequencies by the sum of the updated frequencies of all degrees of nodes in the network.

We can observe that even the DD of the 1-gram model is like a two-regime power law (Figure 5), which means that DD is the easiest of the network statistics to replicate. For other generated networks, the DD is almost identical to the real network, a fact also apparent from the KLD values in Table 6. However, the 1-gram model has much lower CC and ASPL than the real network. The CC matches (is close to the real network) for all models where $n \geq 2$, and the ASPL matches only for the 3-gram and 3-term LMs. Thus, ASPL is a harder statistic to replicate artificially than CC.

A general observation from the motif signatures is that like NLA [59], it is possible to be close to real networks on $\Psi^3$ motif counts with 2- and 3-grams. To examine deeper, we computed two aggregate motif statistics for each model: $M-$Diff, the motif-wise (pointwise) sum of the absolute LNMC differences between real and generated networks, and $M-$Sum, the sum of the LNMC values for all connected motifs. Equations 6 and 7 show computations of $M-$Diff and $M-$Sum for a model $M$:

$M-$Diff($M$) = $\sum_{k=3}^{4} \sum_{i} |\text{LNMC}(\Psi_i^k)_{\text{Real}} - \text{LNMC}(\Psi_i^k)_{\text{LM}}|$ (6)

$M-$Sum($M$) = $\sum_{k=3}^{4} \sum_{i} \text{LNMC}(\Psi_i^k)_{\text{LM}}$ (7)

The existence of a large number of connected motifs in a network is an indication of its non-randomness [60]. We argue that the larger the sum of LNMC values for connected motifs, the more structured the network is. However, if this sum exceeds that of the real WCN, it implies that the network is becoming more restrictive than what is naturally expected. With $M-$Diff values as low as 1.590 and 1.835, the 2-gram and 2-term models have almost the same motif signatures as the real network. On the other hand, going by the abundance of connected motifs, the 3-gram and 3-term models seem to be the most restrictive (>
Table 7: \( \Psi^3 \) and \( \Psi^4 \) motif signatures for the query LMs. M-Diff represents the sum of the absolute differences between a model network’s LNMC values with corresponding values from the real network. M-Sum is simply the sum of the individual values in a network’s motif signature. The two minimum M-Diff-s and the two maximum M-Sum-s are shown in boldface.

<table>
<thead>
<tr>
<th>Model</th>
<th>3-chain</th>
<th>3-clique</th>
<th>4-chain</th>
<th>4-star</th>
<th>4-loop-out</th>
<th>4-box</th>
<th>4-semi-clique</th>
<th>4-clique</th>
<th>M-Diff</th>
<th>M-Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>2.941</td>
<td>7.185</td>
<td>4.101</td>
<td>6.057</td>
<td>10.315</td>
<td>7.212</td>
<td>15.463</td>
<td>21.484</td>
<td>0.000</td>
<td>74.758</td>
</tr>
</tbody>
</table>

The three lowest and the highest values in the M-Diff and M-Sum columns, respectively, are marked in boldface.

Figure 5: (Color online) Sample CDDs for query LMs. It shows CDDs for one of the subsamples each from the real and the model-generated networks. The plots have been split into two groups, to somewhat mitigate the problem of almost completely overlapping curves.

| Log     | \(|N|\) | \(|E|\) | CC    | ASPL  | KLD   |
|---------|--------|--------|-------|-------|-------|
| Real    | 34,209 | 242,680| 0.623 | 3.302 | 0.000 |
| 2-gram-w| 28,748 | 311,955| 0.280 | 2.823 | 0.349 |
| 2-gram-seg | 270,687 | 1,642,641 | 0.572 | 3.519 | 0.523 |

Table 8: Basic network statistics for word (w) and segment (seg) co-occurrence networks for the same generative model 2-gram.

Real M-Sum). As supporting evidence, we note that the 3-gram model has lower perplexity than the 1-gram or 2-gram model (Table 7). We also note that the n-gram models have motif signatures closer to the real network than the corresponding n-term LMs. Thus, relative word ordering in queries is important. For example, in the query convert pdf to word windows, the intent of the query would be altered if the order of pdf and word was interchanged to convert word to pdf windows.

4.4.1. Segment co-occurrence networks

In continuation to the discussion in Section 3.1.2, we now evaluate how segment co-occurrence networks behave in the current framework. The basic network properties were computed on the LCCs and the results are presented in Table 8 and Figure 6. We observe that the segment 2-gram model (2-gram-seg) has matched up quite close to real logs (CC and ASPL); it is slightly nearer to real logs on ASPL and slightly farther away on CC and KLD, than the word-level 2-gram model 2-gram-w. The property values are quite close to real logs and to 2-gram-w in spite of the very large network size for segments than words (almost eight times larger both in terms of nodes and edges than the word level 2-gram-w). The number of nodes in the segment model are so high because segments are created by permutations of individual words, which results in a very large number of potential segments (although only a small fraction of the possible segments are actually present in the real query log). All values are averaged over 100 samples of 1M queries each. While it is intuitive that using segment statistics could ultimately produce better quality logs, the degree distribution shows a slightly higher deviation from the real distribution (the 2-gram-seg degree distribution plot is almost a one-regime instead of a two-regime power law) which needs a more thorough investigation. Segment-based query generative models remains an important avenue to explore.

4.5. Variation of the significance threshold

In our co-occurrence model, we consider a co-occurrence between a pair of words only if they occur together more often...
than expected by random chance. This is ensured by having an edge in the WCN only if the joint probability of occurrence \( p_{ij} \) of words \( i \) and \( j \) is greater than the product of their individual occurrence probabilities \( p_i \) and \( p_j \), i.e., \( p_{ij} > p_i \times p_j \) (edge restriction condition). This condition is imposed because if occurrences of \( i \) and \( j \) are independent, then \( p_{ij} = p_i \times p_j \). Thus, edges are not created simply by a binary indicator, i.e., if the two words are co-occurring in queries or not.

However, weighted networks capture more information than unweighted graphs, and we have used the latter in our experimental model. Nevertheless, computing and comparing motif signatures between real and model-generated query logs are an integral part of our experimental framework, and counting motifs has not yet been defined for a weighted graph in the literature. However, we did perform a set of experiments on unweighted graphs that, in a way, simulates the effect of weighting the edges. We do this by having a parameter \( \alpha \) as the exponent of the \( p_{ij} \) term in the edge restriction condition (Section 2.1), and varying \( \alpha \) between 0.0 and 2.0 in steps of 0.1. Thus, our edge restriction condition becomes \( p_{ij}^\alpha > p_i p_j \). When \( \alpha = 0 \), \( p_{ij}^\alpha = 1 \), and since \( p_i \times p_j \) is always less 1 if probabilities \( p_i \) and \( p_j \) are both non-zero, we have a network with all the possible edges being present, irrespective of \( i \) and \( j \) co-occurring in a query. This is a boundary condition. As \( \alpha \) is increased from 0.0 to 1.0 in small steps, the number of edges in the network decreases. Consider a specific pair of words \( i \) and \( j \). Here a positive fraction is raised to the power of a positive fraction, and while the base of the quantity \( p_{ij} \) remains constant, the exponent \( (\alpha) \) is increasing, and thus the value of the expression decreases. But the right-hand side of the condition remains constant \((p_i \times p_j)\). Thus, it gradually becomes harder to have edges in the network. At \( \alpha = 1.0 \), we reach our original edge restriction condition. As \( \alpha \) is gradually increased from 1.0 to 2.0, the value of left-hand side further decreases due to the nature of the values of the base and the exponent resulting in a continued decrease in the number of edges in the network. Thus, the number of the edges steadily decreases in the network as \( \alpha \) is increased from 0.0 to 2.0. We found that at \( \alpha = 2.0 \), the network ceases to exist as none of the word pairs satisfy the edge restriction condition. This framework with unweighted graphs has essentially the same effect on network properties as gradually increasing the weights of edges in a weighted-graph model.

Specifically, we performed this \( \alpha \) variation experiment on the logs generated by our 2-gram model. We report the basic network statistics and the motif signatures in Tables 9 and 10, respectively. As usual, all computations are performed on the LCC of the graph. The numbers reported are the average values of 100 random samples. There is negligible standard deviation among the properties of the random samples.

The number of edges falls sharply from \( \alpha = 1.6 \), with almost no edges being created from then on. The numbers of nodes in the LCC become 11, 6 and 0 for \( \alpha = 1.6, 1.7, 1.8 \) respectively, and the numbers of edges in the LCC become 17, 8 and 0, respectively. So we do not consider \( \alpha > 1.5 \) in our analysis.

We observe that with decreasing \( \alpha \), the number of edges decreases, making the network sparser. This is shown by the steady decrease of CC from 0.656 to 0.562. Similarly, as the edge density decreases, the shortest paths become longer, from 3.739 up to 7.807. We also observed that there is no change in the LCC from \( \alpha = 0.0 \) to 0.6, with the first noticeable decrease in edges in the LCC happening from \( \alpha = 0.7 \). The real network’s CC and ASPL are 0.623 and 3.302 respectively. Since closeness to the real values is preferable, we find the best \( \alpha \) to be 1.0 (CC) and \( <= 0.7 \) (ASPL). However, we see that there is no substantial deviation in CC or ASPL till about 1.0, after which the network starts becoming sparse at a reasonably high rate. The minimum KLD with the real degree distribution is again found to be the lowest for \( \alpha = 1.0 \), thus underlying the choice of \( \alpha = 1.0 \) to be close to optimum (closeness to real logs is what we try to achieve). CDDs at some sample \( \alpha \) values are shown in Figure 7. The plots are almost exactly overlapping.

| \( \alpha \) | \( |V| \) | \( |E| \) | CC | ASPL | KLD |
|---|---|---|---|---|---|
| 0.0 | 33,213 | 225,370 | 0.656 | 3.739 | 0.084 |
| 0.1 | 33,213 | 225,370 | 0.656 | 3.739 | 0.084 |
| 0.2 | 33,213 | 225,370 | 0.656 | 3.739 | 0.084 |
| 0.3 | 33,213 | 225,370 | 0.656 | 3.739 | 0.084 |
| 0.4 | 33,213 | 225,370 | 0.656 | 3.739 | 0.084 |
| 0.5 | 33,213 | 225,370 | 0.656 | 3.739 | 0.084 |
| 0.6 | 33,213 | 225,370 | 0.656 | 3.739 | 0.084 |
| 0.7 | 33,213 | 225,366 | 0.656 | 3.739 | 0.085 |
| 0.8 | 33,213 | 225,229 | 0.655 | 3.741 | 0.084 |
| 0.9 | 33,213 | 222,777 | 0.644 | 3.739 | 0.083 |
| 1.0 | 33,213 | 210,047 | 0.619 | 3.887 | 0.079 |
| 1.1 | 33,213 | 182,645 | 0.595 | 4.137 | 0.081 |
| 1.2 | 33,210 | 147,534 | 0.579 | 4.466 | 0.091 |
| 1.3 | 32,962 | 112,668 | 0.569 | 4.942 | 0.128 |
| 1.4 | 32,564 | 82,731 | 0.562 | 5.865 | 0.202 |
| 1.5 | 30,828 | 57,438 | 0.562 | 7.807 | 0.329 |

The row corresponding to the original condition with \( \alpha = 1.0 \) is shown in boldface.
from \(\alpha = 0.0\) to 1.0, and after \(\alpha = 1.2\) the two-regime power law clearly breaks down, indicating a drastic change in network structure due to the rapid removal of edges.

We observed that M-Diff initially decreased from 12.711 (\(\alpha = 0.0\) through 0.7) to 5.329 (\(\alpha = 1.0\), original). Then it increased steadily to 25.655 with increase in \(\alpha\). Since lower M-Diff is an indicator of proximity to the real network (Section 4.4), incidentally it turns out that \(\alpha = 1.0\) is indeed the best setting. The M-Sum values, however, continue to decrease monotonically with increase in \(\alpha\) from 0.0 to 1.5, through-out. However, as discussed within the text, the increase in the LNMC values only implies that the network is getting more restractive and probably trying to overfit the training data. Thus, from these experiments, we find that with an exponent value of 1.0 for \(\alpha\), we get the closest proximity to the real network (the first row in Table 7).

### 4.6. Effect of stopword removal

We wanted to study the effect of stopword removal. Again, we chose the 2-gram model as the representative for experimentation. We removed the stopwords from the logs generated by the 2-gram model and constructed the new WCN. We then compared the network statistics with those obtained from the original log containing the stopwords. Basic network properties are presented in Table 11 and the motif signatures in Table 12. Sample degree distributions are shown in Figure 8. From both the tables and the figure, we found that removing stopwords does not have a substantial effect on the network statistics. This is perhaps because users already omit most of the stopwords while formulating Web search queries.

### 5. User intuition of real queries

One of the important aspects of any NL is the grammatical correctness and coherence of the sentences, which is typically verified through native speakers’ judgments [64][65]. Native speakers can also predict the next word in a sentence.
Table 12: Effect of stopword removal on motif signature of 2-gram model. Removing stopwords does not appear to substantially affect the motif signature.

<table>
<thead>
<tr>
<th>Stopwords</th>
<th>3-chain</th>
<th>3-clique</th>
<th>4-chain</th>
<th>4-star</th>
<th>4-loop-out</th>
<th>4-box</th>
<th>4-semi-clique</th>
<th>4-clique</th>
</tr>
</thead>
</table>

Figure 8: (Color online) Effect of stopword removal on the cumulative degree distribution of the 2-gram model. Removing stopwords does not seem to have a substantial effect on the cumulative degree distribution.

given the previous \((n - 1)\) words with a reasonable degree of accuracy [69], which makes them a good point of comparison against \(n\)-gram models. Therefore, statistical and network modeling-based analyses of query syntax would not be complete without a native speaker evaluation on acceptability of the generated queries. The challenge, however, is to redefine the concept of native speakers in the context of queries, and to design the corresponding query-acceptability task.

If queries are considered as a language, then clearly anybody generating a query can be considered a native speaker of the language. Thus, for our experiments, we deem an average search-engine user as the native speaker of the query language. However, asking a user whether a query is acceptable or not seems quite a meaningless task – any random sequence of keywords could constitute a query that has been issued by a real user, because as such there is no consensus on grammatical constraints on queries. To get around this problem we carefully designed our experiment in the following way:

1. Users were given a triplet having one real query and two generated queries.
2. They were asked to identify the real query in this triplet.
3. The remaining two search queries were to be rated on a five-point scale.

To make the comparison meaningful, the three queries in the triplet had some words (at least one word to at most three words) in common. The selection of query triplets was automated by a program; the program conditionally selected triplets from thousands of query sets if the three chosen queries of a triplet had at least one to at most three words in common. The philosophy behind this evaluation strategy is that if a generated query is sufficiently realistic, the user will have to make a random choice between the generated and the real query. Moreover, the rating points awarded to the queries in the triplet will give us information about the relative quality of the underlying generative models. We did not consider preference judgments [67] for the models as these are useful if one is only interested in the relative performance of the models. Here, on the other hand, we would like to find out the absolute goodness of a model with respect to real queries. This would not be captured through preference-based judgments. Moreover, after selecting the real query, since the users had to score the two remaining queries, the ranking within a triplet can be easily inferred.

On a related note, Li et al. [68] describe an experiment where artificial queries were rejected if they were not acceptable to human judges. However, in their setup, new queries were created by string transformation methods from a real query. Users only had to judge if the generated queries had the same intent as the corresponding original ones, and therefore, their experimental framework is not applicable to our problem.

5.1. Method

We use crowdsourcing through Amazon’s Mechanical Turk (AMT) for our user experiments. Apart from being a cheap and fast method for gathering large data [69, 70, 71], a Turker (AMT task participant) is expected to be as good as an average search-engine user because AMT experiments are done online and often require one to really conduct Web searches. Hence, crowdsourcing is amenable to our experimental setup. We designed the Human Intelligence Task or HIT (a unit task in AMT) as follows. The five LMs can be combined with real queries to create \(\binom{5}{3} = 10\) triplets. This gives us ten combinations of \{Real query-Model i-Model j\} triplets. For each of these ten combinations, we randomly selected 35 triplets such that individual queries had some words in common (at least one word to at most three words) – making a total of \(10 \times 35 = 350\) triplets containing \(1,050\) queries to be judged. Some solved examples are shown in Figure 9. The choice of assigning a 5 appears difficult for question 2, and one of the realistic queries is model-generated. We assume that the annotator makes a random choice in such cases, and sufficiently realistic artificial queries will generally accumulate high scores in the end. To reduce annotator bias, we tightened our guidelines as far as possible, which are specified below:

1. Each question in the datasheet contains three queries – one is issued by a real human user, and the other two are generated by an algorithm.

Solved Examples

1. map weather moreton island all
2. a map victoria australia home kit
3. map of east timor and surrounding islands
4. fun relay races for kids
5. for kids fun easter gifts buy where musical relay races for kids
6. white fish and potatoes recipes
7. fish potato big computer
8. a recipes for cook white fish

Figure 9: Some solved examples for the crowdsourced query annotation task posted on AMT. To reduce annotator bias, we tightened our guidelines as far as possible, which are detailed in the text.

2. The task is to find the query that is most likely to be issued by a human user and mark it 5. For example, how to play a cd on my computer.
3. The remaining queries are to be scored on a scale of 0 – 4 according to the following guidelines.
4. Mark 4 if you think that the query is generated by an algorithm, but could almost be the real query (some external factor may make the query unrealistic, like the lack of the context in the example shown). For example, queries like ancient rome exchange limit.
5. Mark 3 if you think that the query is generated by an algorithm, and it makes sense, but has incorrect grammar or spelling. For example, queries like i fit to get blood transfussion.
6. Mark 2 if you think that the query is generated by an algorithm, and represents incomplete information needs or jumbled units, but could be meaningful if completed or reordered. For example, queries like creating pdf a file share tab security.
7. Mark 1 if you think that the query is generated by an algorithm, and parts of the query are coherent, but not as a whole. For example, queries like how to slaves how did.
8. Mark 0 if you think that the query is generated by an algorithm, and it is totally nonsensical. For example, queries like and anzac to jungle characters 101.

Task details are presented in Table 13. An individual HIT was defined as the task of rating all the queries in five triplets. In other words, each HIT was defined by us as identifying the real query and scoring the other queries in the triplet for a set of five query triplets, i.e., a human annotator had to examine a set of five (independent) triplets and provide his annotations. In other words, (s)he may not choose to solve only three or four triplets in the task. Additionally, for a task to be approved by us, the annotator had to provide ratings to all the fifteen queries in the five triplets. Approval was subject to the condition of having exactly one score of five (“Real”) in a triplet. We had requested and had obtained five annotations for every HIT. This means that we procured scorings by five different people for every query triplet (every query in every triplet). This was done to compute inter-annotator agreement (which turned out to be quite good) and allowed us to report aggregated ratings which are often more robust to noise than individual ratings. Aspects of the AMT experiment setup like cost, allowed time for each HIT and task descriptions are crucial to receiving quick and reliable responses. For our research, we followed the general guidelines presented in Alonso and Baeza-Yates [69].

Participants for our study needed to have at least fifty tasks approved in order to be eligible. Apart from this, no other restriction was placed on attributes of participants. Amazon allows any person to sign up for participation in AMT tasks, irrespective of his/her demographic properties. Consent of participants was obtained using Amazon. All relevant information can be found at Amazon’s participation agreement[3] and the privacy notice[4]. None of the authors had any access to identifying information for the participants. AMT only provides user-ids to task publishers which lack any personally identifiable information.

5.2 Results and observations

Table 14 reports the average rating assigned to a query within a triplet by the five annotators, averaged over all queries from a model (reported in the “Average Rating” column). The “Real Triplets” column lists the number of triplets that has the presence of a query from the corresponding LM. The “Real Percentage” column lists the percentage of times a query generated by a model was marked as “Real”. The actual number of times that a query from the corresponding LM was marked as “Real” in a triplet is shown in the Judged “Real” column. The triplets

Table 13: AMT experiment details. An individual HIT was defined as the task of rating all the queries in five triplets. A set of five annotations was requested for each HIT.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task description</td>
<td>Identify a real query hidden among model-generated ones. Then give grades to the remaining queries.</td>
</tr>
<tr>
<td>Task keywords</td>
<td>Web search queries, Real and generated queries. Rating queries</td>
</tr>
<tr>
<td>No. of query triplets in 1 HIT</td>
<td>5</td>
</tr>
<tr>
<td>Total no. of triplets</td>
<td>350</td>
</tr>
<tr>
<td>Annotations per triplet</td>
<td>5</td>
</tr>
<tr>
<td>Allowed assignment time for each HIT</td>
<td>20 minutes</td>
</tr>
<tr>
<td>Actual assignment time per HIT (mean)</td>
<td>1 minute 24 seconds</td>
</tr>
<tr>
<td>Turker qualification</td>
<td>Approval rate &gt; 50 tasks</td>
</tr>
<tr>
<td>Turker location</td>
<td>Any</td>
</tr>
<tr>
<td>Reward per HIT</td>
<td>$0.05</td>
</tr>
<tr>
<td>Total completion time allotted</td>
<td>7 days</td>
</tr>
</tbody>
</table>

Table 14: A summary of absolute ratings obtained through AMT. We report the average rating assigned to a query within a triplet by five annotators, averaged over all queries from a Model (reported in the Average Rating column). #Total Triplets count the number of triplets that has the presence of a query from the corresponding LM. The “Real” Percentage column lists the percentage of times a query generated by a model was marked as “Real”. The actual number of times that a query from the corresponding LM was marked as “Real” in a triplet is shown in the Judged “Real” column.

Table 15: Summary of relative ratings from AMT. Results are shown in a tournament-like fashion. Ordering of LMs in rows and columns is such that light and dark cells roughly create upper and lower triangular matrices.

The two highest values in the last two columns marked bold.

5.3. Interpretation

These results provide us with the following interesting insights: (a) If any string was equally acceptable, real queries would get a “Real” rating only 33.3% of the time by random chance. The fact that real queries get the “Real” rating about 60% of the time implies that users already have a notion of queries being well-formed, i.e., the acceptability of queries; (b) The 3-gram model-generated queries can confuse the user about 28% of the time. In contrast, for NL, speakers can easily identify trigram-generated sentences, which are locally readable, but semantically incoherent. This shows that trigrams capture more information than bigrams and probably overfit the data; (c) 2-term and 3-term queries getting lower “Real” percentage scores than 2-gram and 3-gram queries implies that word ordering provides vital clues to the users; (d) Bigrams received a higher “Real” percentage value but a lower average rating than the 2-term and the 3-term models. This is because bigrams generate very realistic queries at times, especially when the query length is small, but meaningless ones on other occasions. This is supported by the observation that the standard deviation of the ratings for bigrams is 1.476, while it is lower for 2-terms and 3-terms (1.334 and 1.302 respectively); (e) 2-term and 3-term models getting almost exactly similar average ratings further emphasizes the importance of word ordering. 3-terms are expected to generate semantically more coherent queries, but an obvious lack of ordering hinders their acceptability to Web users.

6. Discussion

We now discuss some of the inferences that we draw from the research presented in this paper.
Motif analysis is insightful. Biemann et al. [39] observe that box motifs in NL (Figure 3) often occur due to synonymy, and chain motifs due to polysemy. We found that the synonymy and polysemy interpretations from NL hold good for queries as well. This is because synonyms will rarely co-occur in the same query, thus leading to an absence of connecting edges in the WCN. Similarly, the second and the third nodes in 4-chains (or the middle node in a 3-chain) are connected to concepts that are not connected to each other, hence representing multi-senses. Additionally, we observed that box motifs in query logs typically occur with two similar entities (e.g., titanic and spiderman, both movies) forming a pair of disconnected nodes and two attributes (e.g., mp3 and cast) forming the other pair of disconnected nodes. The other common reason for box motifs in queries is spelling mistakes or spelling variations (like pituitary and pituitary) at two opposite ends, and related words like hormone and tumor forming the other two opposite ends of the box motif. We observed star motifs in query logs arise due to a content unit (e.g., titanic) in the centre and three intent units (e.g., cast, mp3, and review) connected to it. Thus, motifs occur due to syntactic and distributional constraints in the linguistic system, and hence, they are relatively harder to capture through a generative model. While bigram models can match the motif signature of the real log to a good extent, the actual motifs are far less intuitive, which is also evident from the low acceptability of bigram-generated queries by Web users.

Trigrams outperform the data. User experiments show that trigram models generate quite realistic queries (Table 14) and have surprisingly low perplexity (Table 2). However, this does not imply that trigrams solve the problem of artificial query generation and thus can be used for creating synthetic logs. In fact, network analysis shows that trigram-generated queries give rise to a higher number of motifs of each kind than even the real WCN, which means that trigrams possibly overfit the training data. Therefore, a log generated with trigrams will contain realistic but frequently seen queries. Such a synthetic log will lack the diversity present in a true query log. This is not surprising because the average length of a query being only four, a large number of queries will be generated with only one or two trigrams. This will effectively generate only queries that have been frequently seen in the training query log.

There is scope for better generative models. Since trigram models overfit and bigram models fall short of generating good individual queries, realistic queries can only be generated using more sophisticated models that can capture the structural constraints of queries both at syntactic and semantic levels. Since motif analysis indicates the importance of content-intent relationships [72] in queries, we believe that a better quantification of the distribution of these relationships can lead to improved generative models for Web search queries.

Relative word ordering is important for the user. Researchers in the past have criticized the bag-of-words model for queries [73, 74, 75]. Our analysis strengthens earlier findings by showing the importance of word ordering constraints in queries, as the bag-of-words model-based query generation (using the n-term models) is shown to be inadequate in both network and user-based setups.

There is a cognitive model for queries. Finally, it is interesting to note that users, or as we can say, the native speakers of the language of queries, are indeed able to differentiate real queries from artificially generated ones. This shows that an average user has already internalized a cognitive model for queries, or for language in general. It is indeed believed to be true that there is a common faculty of language in all humans that helps in building the cognitive model for a language. Interestingly, this faculty mediating human communication appears very different from that of other living creatures; the human faculty of language bears resemblance to the genetic code—hierarchical, generative, recursive, and virtually limitless with respect to its scope of expression [76]. However, cognitive models built for the different languages will have different parameters that have to be internalized by its native “speakers”. In this respect, we believe that the parameters for the language of queries (say, average length in words, relatively free word order, and omission of function words) are different from that of the parent language from which the query words are borrowed. Moreover, these parameters may depend on the parent language itself, i.e., “English” queries may have different parameters from “German” queries. Our experimental result that users are able to recognize real queries among synthesized ones about 60% of the time, shows that the average Web searcher has acquired the parameters of the English query language with a satisfactory level of proficiency. Further probing of this cognitive model through psycholinguistic experiments would be an interesting exercise that can provide interesting insights into the organization of query syntax, and also how a new language might evolve and automatically acquire a syntax of its own. It is worth noting here that since queries are not used for direct human communication, they cannot be collectively compared to the concept of a register in linguistics [77], which is a style of language used for a specific reason or in a specific social scenario. Finally, we do not know of prior work that takes a holistic approach towards the analysis of the syntactic complexity of Web search queries from the first principles. Previous works related to statistical and network analysis of queries have been reported in the respective sections. Nevertheless, there are two lines of research pertaining to the syntactic analysis of queries.

Linguistic analysis of queries. First, linguistic analysis and annotation of queries at the level of segmentations or chunking [43] and parts-of-speech tagging [78, 79, 80, 81] have been important directions of research since the early 2000s [8]. While these studies reveal interesting syntactic properties and trends, such as more than 70% of the query terms are nouns [78], and NL question queries are on the rise [82], they are based on the fundamental assumption that queries issued
using the words of a certain parent language, say Standard English, will borrow grammatical artefacts of that parent language (i.e., nouns, verbs, noun and verb phrases, etc.). This assumption is biased because a noun in English is called a noun because it follows a particular syntactic distribution; it is quite unlikely that the same word will behave as a noun in a query either from the point of statistical distribution or its cognitive interpretation by the users. Thus, if queries are to be understood linguistically, they should be analyzed from the first principles rather than superimposing the grammatical syntax of NLs [33], which masks their true syntactic properties. We note that techniques based on the superimposition of syntax from the parent NL can still be useful for practical applications, but they cannot tell us much about the true syntax of Web search queries.

**Entities and intents.** The second line of research, which we believe is more promising, is the analysis of queries in terms of user intents. Such studies have looked into queries from various perspectives and have come up with various concepts such as entities and attributes [84-86, 87, 88], kernel-objects and modifiers [89] and query facets [90, 91], to factor the parts of a query and place it within a taxonomy of semantic or syntactic patterns. While it is not possible to review all these studies here, a closer look at the actual network motifs of the WCN for real queries reveals interesting synergy between the concepts of *intent words* [92-93, 72] (also called modifiers or attributes) and *content words* (or entities or kernel-objects), which is worth mentioning here.

7. Conclusions

In this paper, we have tried to understand the syntactic complexity of Web search queries, a distinct mode of interaction between man and man-made systems. We have adopted a three-pronged approach: applying statistical language models (using n-grams and n-terms), using complex network modeling (with WCNs), and asking native speakers (Web search users). Our results underline the necessity of using multiple independent perspectives. Having entropy or perplexity similar to or lower than NL need not, by itself, be indicative of an underlying language system. Network analysis shows bigrams to be within striking distance of replicating real log syntax at a corpus-level. However, when native speakers are consulted, individual queries generated by trigrams are found to be much more acceptable than those by bigrams. Only a combined approach is successful in bringing out the complete picture of the WCN for real queries, communicating information needs of millions of users, can be established to be an independent language system from all the three aspects—structure, function and dynamics.

8. Acknowledgments

The first author was supported by Microsoft Corporation and Microsoft Research India under the Microsoft Research India PhD Fellowship Award (Grant No. 8901901/IITKGP-Fship-2011). We would like to thank Rohan Ramanathan from R. V. College of Engineering (RVCE), Bengaluru, India, for helping us in posting tasks on AMT, Parantapa Bhattacharya from Indian Institute of Technology (IIT) Kharagpur, India, and Pavan Kumar from National Institute of Technology (NIT) Durgapur, India, for helping us with the experiments on network stability and motif analysis.

References


