Analyzing Linguistic Structure of Web Search Queries

Rishiraj Saha Roy
IIT Kharagpur, India - 721302.
rishiraj@cse.iitkgp.ernet.in
« Supervised by: Niloy Ganguly (IIT Kharagpur) and »
« Monojit Choudhury (Microsoft Research India) »

ABSTRACT

It is believed that Web search queries are becoming more structurally complex over time. However, there has been no systematic study that quantifies such characteristics. In this thesis, we propose that queries are evolving into a unique linguistic system. We demonstrate proof of this hypothesis by examining the structure of Web queries by applying well-established techniques from natural language understanding. Preliminary results of these experiments show quantitative and qualitative proof that queries are not just some form of text between random sequences of words and natural language – they have distinct properties of their own.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation

General Terms

Measurement, Experimentation, Human Factors

Keywords

Query understanding, Query structure, Query segmentation, Query intent, Word co-occurrence networks

1. PROBLEM

Web users communicate their information need to a search engine through queries. The fact that search engines do not really "understand" or "process" natural languages (NLs) drives average Web users to specify their queries in a form that has a structure far simpler than NL, but perhaps more complex than the commonly assumed bag-of-words model. We hypothesize that Web search queries define a new and fast evolving language of their own, whose dynamics are governed by the behavior of the search engine towards the users and that of the users towards the engine. The objective of this research is to carefully scrutinize this proposition through structural analysis of queries and allied user experiments.

1.1 Motivation

Query understanding [4] is gaining importance as a distinct area of research in query log analysis. It is an umbrella term that includes discovering query intent and modeling term dependencies. The stakes in this problem are high – this is fueled by the idea that a deep understanding of query structure can not only help in better information retrieval (IR) for complex queries and improved assessment of relevance, but also help in several related applications like query completion and sponsored search. Additionally, if it is possible to show, with data spanning across several years, that queries indeed display dynamics similar to NLs, then perfectly preserved search logs can act as a potent dataset for studying the evolution of language. Data scarcity happens to be a perennial problem plaguing researchers in the field of language evolution. Irrespective of these factors, it is interesting to understand how millions od users across the world, without direct interaction, are communicating their information needs to the search system in a similar "language". While some researchers [10] have cursorily mentioned this idea, there has been no systematic investigation.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

1.2 Proposed Approach

Approach. We adopt a multi-pronged approach. First, we apply statistical techniques used in Natural Language Processsing (NLP) to conduct structural analysis of queries from linguistic first principles. Second, we use complex network modeling techniques for understanding aggregate statistical properties of query logs and compare them to those of NLs. Third, we conduct user experiments to understand cognitive processes underlying query formulation and analysis. In the following sections, we apply such techniques from NL understanding to queries in a principled manner.

Organization. We first briefly describe our dataset of Bing query logs in the next section (Sec. 2). Then, we detect the basic structural units of queries using a segmentation technique that relies mainly on query logs (Sec. 3). Next, we devise an unsupervised approach that can predict broad roles of segments in queries with reasonable accuracy (Sec. 4). After these two aspects of query analysis at a micro-level, we explore macro-analysis of query logs using complex network modeling (Sec. 5). Finally, we briefly touch upon an important work-in-progress, query generation using statistical models (Sec. 6). We conclude by briefly summarizing our findings and highlighting the existence of strong parallels between Web search queries and existing notions of language (Sec. 7).

2. DATASET

For all our experiments, we use queries sampled from a Bing Australia log of May 2010. It originally consisted of

16.7M (M = million) queries. 14.5M queries remained after removing those with non-ASCII characters. One word queries do not have any "structure" and were discarded. Similarly, very long queries that consist of more than ten words are typically code excerpts, computer generated messages or NL sentences, and were also omitted. In any case, they form a very small portion (< 0.1%) of the whole log. The final list thus obtained had 11.9M queries, out of which 4.7M were unique. Each query is accompanied by a clicked URL, a click count and several other features.

3. IDENTIFYING STRUCTURAL UNITS

In order to analyze the language of Web search queries from the basics, the first step would be to identify its fundamental structural units. We know that units like *harry potter* and *blood pressure* should be indivisible during query processing. Hence individual words need not be the basic units for query understanding. This process of dividing a query into its structural units is called query segmentation [5]. For example, the query `australian open 2013 home page` can be segmented as `australian open 2013 | home page`.

3.1 State-of-the-art

The past decade has seen a good amount of work on query segmentation, using diverse supervised and unsupervised algorithms and diverse resources [5]. However, all of these works, in some form, make use of document resources – either Web n-gram frequencies, contents of clicked documents or search result snippets. We believe that such approaches, inherently, project NL structure onto Web queries. Queries, to be understood properly, need to be analyzed based on their own data – just like all inferences on NL structure have been made by studying corpora of text in that same language. Also, supervised methods rely on human annotations to segment unseen data. Supervised approaches, in the Web scenario, suffer from the problem of *coverage*. Manual annotations are noisy, expensive, and not usable for segmenting unseen queries from diverse domains.

3.2 Methodology

To resolve these issues, we propose a novel unsupervised query segmentation algorithm that uses only query logs [6]. The basis of this algorithm is that if an n-gram is to be a meaningful segment, then its words need to appear adjacently significantly more often than with other words in between. We use probabilistic bounds using the *Hoeffding inequality* to judge the significance of observed counts exceeding expected counts and thus derive a *score* for an n-gram. The score for a segmentation is defined as the sum of the scores of its candidate segments. We use a dynamic programming approach to search over all possible segmentations and select the one with the highest score as the optimal segmentation for the query. No manual segmentations are thus involved in the learning process. It is interesting to note that our algorithm is able to detect interesting segments like *how do i* or *spot a fake*, which are generic intent/action phrases in queries but are not standard units in an English document. To detect rare named entities which do not have sufficient statistical evidence in the data, we augment our algorithm using Wikipedia titles [8].

Now, once we are able to segment unseen queries, the next challenge lies in the evaluation phase. Previous approaches have mostly relied on validation against human segmentations. However, the end goal of query understanding (and hence query segmentation) is to improve retrieval performance. It is not clear whether human intuition of query segments is actually the best from an IR perspective. To address this problem, we design an IR-based evaluation framework for Web query segmentation [8]. In this setup, queries segmented by different strategies are evaluated on their *potential* to retrieve the best quality pages from the collection. We note here that the only way current search engines "support" segmentation is to treat each segment as an indivisible unit in documents through the use double quotes. However, we find that treating all segments (c.f. *how do i* versus *lord of the rings*) in the same way – matching exactly in the documents – negatively affects performance. Hence, our evaluation framework is kept flexible and scores a segmentation strategy on its *best possible performance*, by independently considering each detected segment as divisible or indivisible.

Strategy	Unseg.	Our	[5]	Human	Brute
nDCG@5	0.688	**0.767**	0.752	0.759	0.825
nDCG@10	0.701	**0.768**	0.756	0.763	0.832
MAP@5	0.882	**0.945**	0.930	0.936	0.958
MAP@10	0.865	**0.923**	0.910	0.916	0.944
MRR@5	0.538	**0.650**	0.632	0.632	0.711
MRR@10	0.549	**0.658**	0.640	0.640	0.717

Table 1: IR evaluation of query segmentation.

3.3 Results and future work

Some of the results are presented in Table 1. On the basis of experiments conducted on this evaluation framework, we are able to make several interesting conclusions. (1) There is no perfect correlation between rankings of algorithms based on IR evaluation and validation against manual annotations (2) Segmentation can actively improve retrieval performance (3) Human segmentations do not always have the best IR performance (4) Thus, setting human annotations as the gold standard limits the scope of improvement for a segmentation algorithm (5) The metrics that were previously used to compare machine segmentations against human markup also had serious flaws. Incidentally, our segmentation algorithm based on query logs and Wikipedia titles is also shown to have the best retrieval performance, significantly improving over the state-of-the-art. The most important future work in this direction is to explore *nested* query segmentation (`((windows xp) home) ((serial number) format)`), that addresses the problem of granularity inherent in the so-called *flat* segmentation algorithms.

4. UNSUPERVISED INDUCTION OF SYNTACTIC CATEGORIES

Once the identification of query segments is possible, the next step in a systematic linguistic analysis is to characterize the roles of different *types* of segments. A similar analysis in NLP, based on distributional features of words, leads to Part-of-Speech (POS) induction – labelling words in sentences as nouns, verbs, or adjectives depending upon the context in which they appear. But around 70% of the words in queries are nouns, and a state-of-the-art POS tagger trained on either NLs or queries, fails to achieve satis-

factory levels of accuracy. Thus, just like a true structural analysis had to based upon query logs themselves, the functions, or the roles that segments play within queries, have to be deduced similarly.

4.1 State-of-the-art

The characterization of segments with respect to their roles is currently a very active area in query understanding. One of the popular lines of research focus on *entities and attributes* in queries. Queries in this model, can either consist only of entities, like `harry potter`, or along with an attribute, like `harry potter cast`. Similarly, with respect to entities, *classes* (like *countries*), *instances* (like *India*) and *relationships* have been defined (like *country-capital*). However, the drawback of this model is that all classes of queries cannot be explained properly. For example, in queries like `how to meditate`, or `steps to reduce blood pressure`, mapping words or segments to the predefined roles is not straightforward. Another philosophy, which involves the concept of *intent words or phrases* [11], is more generic. Under this model, most queries have a core information need – the *content* of the query – (*harry potter*, *analgesic drugs*), which is called the *head/kernel-object* [12], and which could be an *entity* according to the previous notion. Additionally, queries also contain words or phrases that carry user *intent*, which are called modifiers/intent words/intent phrases (*movie* or *side effects*) [11, 12]. These intent segments serve to specify the exact information need of the user with respect to the content segments. However, no generic method exists for mining such content and intent segments, and labelling them in queries. We note that such a task can have a lot of impact on improving retrieval, as different segments need to be processed differently for the best results. For example, the word *titanic* needs to be matched exactly in the document; but *map* should be able to bring up a map of the desired location (the content).

4.2 Methodology

In our attempts at characterizing query segments and being able to label them in queries, we look for inspirations from techniques in NLP. Similar to the notion of content and intent segments in queries, exist concepts of *content and function words* in NL. Content words are those with statable lexical meaning (nouns, verbs, adjectives) and which convey the core concepts in a spoken or written sentence. On the other hand, function words (prepositions, conjunctions) serve to specify important relationships between the content words. Thus, processing both types of words correctly is crucial to the intended interpretation of a sentence.

There are certain important properties of function words that we wish to exploit in order to apply this concept to the "language" of queries. Function words are known to be more frequent and co-occur with more distinct words in a sentence than content words. They have low preference for specific words in their co-occurrence distributions (function words like *and* are equally likely to co-occur with words as diverse as *school, law* or *shark*; however, content words like *shark* are much more likely to co-occur only with specific words like *fin*). Moreover, the general co-occurrence between neighbors of function words is also expected to be low (*shark* and *school* are likely not to co-occur in a sentence). We now formalize these indicators into specific features that are able to quantify this discerning behavior of content and function words.

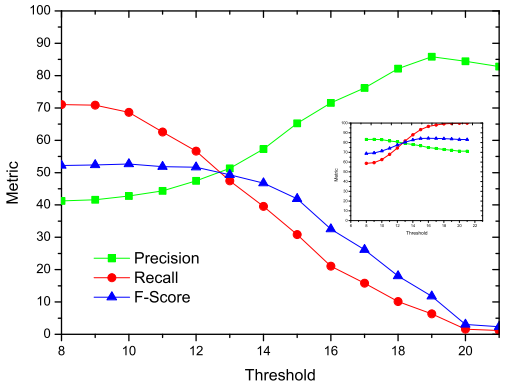

Figure 1: Precision, recall and F-score for intent and content (inset) segment detection.

The features adopted are frequency, co-occurrence count, co-occurrence entropy (i.e. the entropy of the co-occurrence distribution for a word) and the clustering coefficient. Entropy is an information-theoretic measure of randomness; a low entropy implies more bias towards specific items in the co-occurrence list of a word. Clustering coefficient of a node is a graph-based metric that assumes low values when the number of interconnections between the neighbors of a node in a graph is low. Thus, intent segments can be expected to have high frequency, high co-occurrence count, high co-occurrence entropy and low clustering coefficient. To begin with, we compute a simple log-linear combination (to make individual feature *values* comparable) of these features for each distinct segment appearing in the query log. Subsequently, all the segments are sorted in a descending order of the final combination score. Intent segments can be expected to have higher values of the final score than content segments.

4.3 Results and future work

First, it was not clear as to what constitutes the correlates for function words in queries. We observed that the words that appear in the upper part of the list include, along with English function words like *and, in* and *of*, segments like *pics, maps, how to, cast* and *song lyrics*. Thus, there is a high correspondence between these words ($\simeq$ top 500 positions) and existing notions of intent words in queries. Content segments like *roger federer* and *barack obama* appear much lower down (beyond rank 2000). Hence, this simple and lightweight method can effectively discriminate between content and intent segments.

We understand that such a method is not useful unless we are able to tell apart content and function segments inside queries. So we also implemented a naïve labeling algorithm for two-segment queries (forming $\simeq$ 44% of the queries in the log) as an initial step. In this method, since a query must have at least one content segment, the segment with a lower score is labelled as content. The other segment is labelled as intent if its score exceeds a user-defined threshold tuned on a development set. We asked three experienced Web users to label content and intent segments in 1000 two-segment queries. We found that even such a simple score-based technique is able to achieve $\simeq$ 80% precision for content units and $\simeq$ 65% precision for intent units (Fig. 1). The recall, for

both classes, though, is low at $\simeq 50\%$. The inter-annotator agreement is high, being close to 80%. We have, subsequently, also developed a taxonomy of such intent segments which aligns well with the notion of query facets [3].

There are several avenues of future work that will be addressed along this line: (1) Finding a better way of combining the indicator features using machine learning techniques (2) Developing a more sophisticated labelling algorithm that extends to multi-segment queries and can account better for the specific context in the query before assigning labels (3) Formalizing the intent class taxonomy coupled with a proper evaluation methodology (4) Formulating intent-class specific detection and labelling algorithms.

Content and intent segments can be said to be the broad *lexical categories* for the language of Web queries. Query semantics are actually governed by the interactions between these segments *within queries*. Thus, being able to formulate supervised and unsupervised dependency grammars for queries based on these lexical categories is our long-term goal, which could have a huge impact if coupled with intelligent retrieval mechanisms.

5. UNDERSTANDING CORPUS-LEVEL STRUCTURE

In the last two sections, we have looked at two aspects of structure that are important to understanding individual queries. But how can we efficiently model queries at a corpus-level? Fortunately, in the last fifteen years, complex network theory has been shown to be provide an extremely useful representation of a body of text, that can reveal deeper insights about the holistic properties of language [2]. They also act as a good visual representation tool.

5.1 State-of-the-art

Complex network theory provides a powerful mathematical framework to study various complex systems, and its success is primarily due to the fact that a network can simultaneously capture both the 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. A language corpus, which is a running body of text in a language, collected from various sources, can be modeled as a complex network. Very recently, topological analysis of these networks has enabled researchers to summarize the statistical properties of real NL texts that sets it apart from artificially generated corpora. The most popular and well-studied representation of a language corpus is the Word Co-occurrence Network (WCN) [2], which has also been recently applied to term weighting for IR. These facts motivate us to choose a network model, *viz.* WCN for query logs. As far as we know, we are the first to study such network modeling for Web search queries [9].

5.2 Methodology

A WCN for any given text corpus is defined as a network $\mathcal{N} : \langle N, E \rangle$, where N is the set of nodes each labelled by a unique word (or segment) and E is the set of edges. Two nodes $\{i, j\} \in N$ are connected by an edge $(i, j) \in E$ if and only if i and j "co-occur" in a sentence [2]. Co-occurrence can be defined variously; in our research, we consider local and global models of co-occurrence as follows. According to the local co-occurrence model of WCN, immediate word neighborhood is considered important and an edge is added between two words if they occur within a distance of two (i.e. separated by zero or one word) in a query (one query is considered as a single sentence in this context). On the other hand, in global co-occurrence, an edge is added between two words if they occur within the same query, irrespective of the position. Thus, in general, a global co-occurrence network has more edges than a local co-occurrence network. For both local and global networks, edges resulting from random collocations are suitably pruned using joint probability measures. Fig. 2 illustrates the concept of WCN by showing the network generated from the toy query log below. Pruned edges are shown using dashed lines.

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

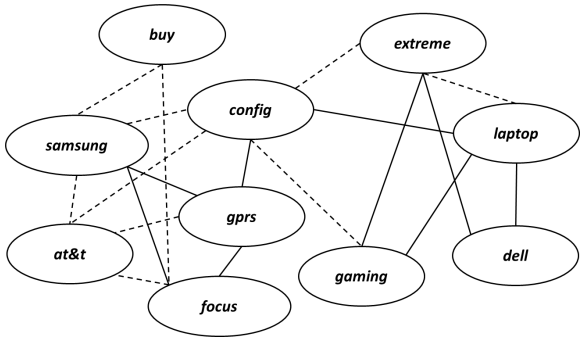

Figure 2: Illustration of a WCN for queries.

5.3 Results and future work

First, our goal is to study the basic statistics of query WCNs and see how they compare with similar WCNs for Standard English. We build WCNs from $1M$ query samples from our large query log. Then, we measure basic network statistics like the cumulative degree distribution, clustering coefficient and average shortest path length. We find a number of insightful results from our experiments. (1) The cumulative degree distributions of the WCNs for queries are

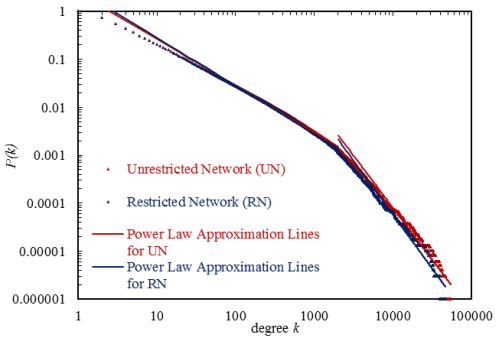

Figure 3: Degree distributions for query WCNs.

observed to be two-regime power laws (Fig. 3, local models), i.e. the plots can be approximated by two piecewise linear segments when the plot is doubly logarithmic. This is a strikingly similar behavior between WCNs built from NL sentences. It is known that such degree distributions correspond to two types of words in the vocabulary – the *kernel* and the *peripheral lexicon* [2]. Hence, such a division is applicable for query words as well. However this is where the differences begin: (2) While the kernel in Standard English, the mother language for our queries, has about 5000 words (common day-to-day nouns, verbs), the corresponding number is only 1000 for queries (mostly intent words and popular names). (3) Small-world property is observed in NL, but not for queries (4) The periphery to kernel size ratio is much larger for queries (5) The kernel is much less tightly coupled than NL and kernel-periphery edges dominate the network, while intra-kernel edges form the majority in NL. All these observations point towards the fact that large query logs have properties that are distinct from NL – providing evidence of a unique underlying linguistic system. We believe that such an analysis is only the first step; the scope of WCN analysis goes much beyond a comparative tool for static NL and query corpora. Currently, we are exploring the use of WCNs as a tool to evaluate generative models for queries.

6. QUERY GENERATIVE MODELS

A vital aspect of NL complexity is the measure of how easily one can artificially generate sentences of that language. Very recently, Biemann et al. [1] have shown that 4-node motifs, a property of WCNs, can precisely quantify the distance that a generated corpus is from real language text. They also show that corpora created using traditional n-gram language models, even with $n = 4$, still have significant gap with real data. A similar analysis on queries reveals an interesting phenomenon: corpora generated using bigrams have very high proximity to real logs, while those generated using trigram models begin to move farther away, as observed using *motif signatures* [1]. However, when individual queries generated using trigrams were shown to a large swathe of Web users in crowdsourcing experiments, they received an average rating of 3.2 on a 6-point scale and were considered to be as good as real $\simeq 28\%$ of the time, while bigram queries obtained a mean rating of 2.9 and were thought realistic only $\simeq 22\%$ of the time. This brings to light an important conclusion: the ideal generative model for queries (which are much smaller than the average NL sentence) is somewhere *in between* 2-grams and 3-grams, and thus an orthogonal modeling strategy that takes into account query semantics like content-intent dependencies is necessary to explain query structure better. Another important fallout of studying query generation complexity is to understand, and finally solve, problems in synthetic query generation, an extremely potent application.

7. CONCLUSIONS AND FUTURE WORK

In this thesis, we try to examine the hypothesis of queries evolving into a linguistic system of their own. Initial contributions of this work are as follows (1) Developing a novel query segmentation approach [6] that reuses query logs and outperforms the state-of-the-art [5] when compared on an IR-based framework [8], (2) Providing simple distributional features as reliable indicators of intent (and content) phrases in queries, and (3) Applying WCNs to Web search query logs and quantifying its distinctness from NL [9]. Future works along each of the examined lines have been identified in the respective sections. Our preliminary results underline the necessity of using multiple independent perspectives and adopting a holistic view. We believe that the unique properties exhibited by Web search queries can indeed be considered positive cues in favour of acceptance of our original hypothesis [7]. However, this research still has a long way to go. These are only the first steps towards the final goal – when queries, communicating information needs of millions of users every day, can be established to be an independent language system through a seamless convergence of all the considered structural aspects.

8. ACKNOWLEDGMENTS

The author is supported by Microsoft Corporation and Microsoft Research India under the Microsoft Research India PhD Fellowship Award.

9. REFERENCES

[1] C. Biemann, S. Roos, and K. Weihe. Quantifying semantics using complex network analysis. In *COLING '12*, 2012.

[2] R. Ferrer-i-Cancho and R. V. Solé. The small world of human language. *Proceedings of the Royal Society of London B*, 268(1482):2261–2265, 2001.

[3] C. González-Caro and R. Baeza-Yates. A multi-faceted approach to query intent classification. In *SPIRE '11*, pages 368–379, 2011.

[4] H. Li, G. Xu, B. Croft, et al. Query representation and understanding. In *Proceedings of the 2nd Workshop on query Representation and Understanding*, 2011.

[5] Y. Li, B.-J. P. Hsu, C. Zhai, and K. Wang. Unsupervised query segmentation using clickthrough for information retrieval. In *SIGIR '11*, pages 285–294. ACM, 2011.

[6] N. Mishra, R. Saha Roy, N. Ganguly, S. Laxman, and M. Choudhury. Unsupervised query segmentation using only query logs. In *WWW '11*, pages 91–92, 2011.

[7] R. Saha Roy, M. Choudhury, and K. Bali. Are web search queries an evolving protolanguage? In *Evolang 9*, pages 304–311, 2012.

[8] R. Saha Roy, N. Ganguly, M. Choudhury, and S. Laxman. An IR-based Evaluation Framework for Web Search Query Segmentation. In *SIGIR '12*, pages 881–890. ACM, 2012.

[9] R. Saha Roy, N. Ganguly, M. Choudhury, and N. K. Singh. Complex network analysis reveals kernel-periphery structure in web search queries. In *QRU '11*, pages 5–8, 2011.

[10] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the web: the public and their queries. *J. Am. Soc. Inf. Sci. Technol.*, 52:226–234, February 2001.

[11] X. Yin and S. Shah. Building taxonomy of web search intents for name entity queries. In *WWW '10*, pages 1001–1010, 2010.

[12] H. Yu and F. Ren. Role-explicit query identification and intent role annotation. In *CIKM*, pages 1163–1172. ACM, 2012.