UNDERSTANDING THE LINGUISTIC STRUCTURE AND EVOLUTION OF WEB SEARCH QUERIES

RISHIRAJ SAHA ROY, M. DASTAGIRI REDDY and NILOY GANGULY

Indian Institute of Technology Kharagpur
Kharagpur - 721302, India.
{*rishiraj, reddy.dastagiri, niloy*}*@cse.iitkgp.ernet.in*

MONOJIT CHOUDHURY

Microsoft Research India
Bangalore - 560001, India.
monojitc@microsoft.com

Web search queries have been observed to exhibit properties of a rudimentary language system, distinct from the mother language from which words of the queries are drawn. It has been hypothesized that the language of search queries is fast growing in complexity, reflected in the steady increase of query lengths over the years. In this research, we make the first attempts to quantify change in the linguistic structure of search queries by examining large query logs spaced four years apart. We adopt a multi-pronged approach and analyze query structure from three different perspectives, namely, language models, complex networks and positional preferences of words. All experimental findings confirm that the linguistic structure of Web search queries is indeed evolving.

1. Introduction

Millions of users around the world interact everyday with search engines to satisy diverse information needs. In our own previous work at Evolang IX (Saha Roy et al., 2012), we argued that the structure of Web search queries is evolving over the years and we established several similarities between the functional, structural and dynamical aspects of queries and natural languages (NL). Based on our own analysis and some of the previous observations (Jansen et al., 2000; Spink et al., 2001; Guichard, 2002; Dessalles, 2006; Huang et al., 2010), we further hypothesized that Web search queries could be an evolving *protolanguage*. Thus, search query logs, which are well-preserved by the various search engine companies, can serve as excellent datasets for studying and understanding language evolution. Nevertheless, neither in our nor in the previous studies that we know of, has there been any attempt to systematically quantify the structural properties of Web search queries and their evolution over time that can convincingly bring out the fact that queries are indeed an evolving linguistic system.

In this study, we try to understand the evolution of the search query language by statistically analyzing the structual properties of two query logs that are four years apart (2006 and 2010). We employ three different and complementary approaches: *(a)* n-gram based modeling of queries, *(b)* complex network analysis of query logs, and *(c)* analysis of word positions in a query. Wherever applicable, we compare the structural properties of queries with NL and draw parallels between the two systems. We find that there is an observable change in the properties of queries across this time span. Interestingly, we also observe that (English) queries are actually deviating from the structural properties of the mother language (i.e., Standard English). Thus, through this paper, we underline our previous claim of using query logs as a potent source of studying language evolution.

This paper is organized as follows. After describing the datasets used in Sec. 2, we study language models and change in pertinent properties in Sec. 3. Complex network modeling and associated statistics are outlined next in Sec. 4. Word positions in queries are analyzed in Sec. 5. Finally, concluding remarks and future work are discussed in Sec. 6.

2. Datasets

We use the 2006 AOL log from USA (Pass et al., 2006) and a 2010 log sampled from Bing Australia[a], which are a good four years apart. Only queries between two and ten word lengths consisting of only alpha-numeric characters were used in this study. The processed 2006 and 2010 logs contained $12.8M$ ($M =$ million) and $11.9M$ queries respectively ($3.6M$ and $4.8M$ distinct queries, but we retained duplicates to preserve frequency distributions). For NL, wherever applicable, we use newswire corpora[b]. All text was case-folded and all punctuation marks were removed appropriately.

3. Perplexity and Entropy of Query Language Models

A statistical language model (SLM) is basically a probability distribution over the various possible strings that can be generated by the language learnt from suitable corpora, and is a well-established tool to understand the randomness or predictability of a symbolic system. The n-gram model is one of the earliest yet very effective SLM techniques (Dunning, 1994). An n-gram model assumes that the probability of the n^{th} word in a sentence depends *only* on the previous $(n-1)$ words (Brown et al., 1992). In spite of their simplicity and memoryless nature, 3-gram models are sufficient for most practical applications (Duan & Hsu, 2011).

Information-theoretic measures. *Perplexity* is defined as $2^{H(X)}$, where $H(X)$ is the entropy (Shannon, 1948) of a probability distribution $p(X)$ and is

[a] http://www.bing.com/?cc=au

[b] http://corpora.uni-leipzig.de/download.html

Table 1. Counts, perplexity and cross-entropy for n-gram models.

Property	C (06)	C (10)	C (NL)	P (06)	P (10)	P (NL)	CE (06,10)	CE (10,06)
1-gram	$0.4M$	$0.6M$	$0.3M$	$7,869$	$8,481$	$2,143$	13.550	13.644
2-gram	$3.5M$	$4.7M$	$4.4M$	75	109	188	7.699	7.540
3-gram	$2.1M$	$2.8M$	$11.7M$	5	6	12	4.543	4.561

given by $H(X) = -\sum_{x \in X} p(x) log_2 p(x)$ (Jurafsky & Martin, 2000). The entropy, and therefore the perplexity, of a probability distribution is high when there is a high degree of randomness associated with the distribution, and consequently, low predictability of the data. Given a probability distribution, *cross-entropy* measures the amount of additional information that one would require to estimate another probability distribution (Jurafsky & Martin, 2000). It is computed as $H(X,Y) = -\sum_{x \in X,Y} p(x) log_2 q(x)$, where p and q refer to the two probabilty distributions (an extension of the simple entropy).

Experimental method. From both logs, we estimated 1-gram (*unigram*), 2-gram (*bigram*) and 3-gram (*trigram*) probabilities. Table 1 reports counts (C), perplexities (P) (rounded to nearest integer) and cross-entropies (CE) of the n-gram models.

Results and interpretations. There are a number of interesting insights that we obtain from these results, which we itemize as follows. *(a)* All perplexities increase from 2006 to 2010, which indicate that diversity in search queries is increasing. In this respect, it is approaching the higher perplexity of NL with respect to bi- and trigrams. Supporting evidence is found in the fact that the counts of all n-grams in queries are increasing. We wanted to investigate if words were simply being added to the lexicon. Interestingly, we found that only $\simeq 0.2M$ words were common between 2006 and 2010, while $\simeq 0.5M$ new words were added, and $\simeq 0.2M$ old words were deleted. These reflect the volatility of the query vocabulary. A small part of these figures can be attributed to the geographical difference between the logs, the 2006 and 2010 data being sampled from the USA and Australia respectively. *(b)* It is quite interesting to note that the perplexity of the unigram model for queries is much higher than that of NL. We observed in our data that the rate of encountering a new word in queries is much higher (about one per 20 words) than NL (about one per 58 words). Thus, the unigram model for queries has a much higher perplexity. The lower perplexities for higher order n-gram models for queries are because the frequency of queries is known to follow a power law (Pass et al., 2006). This means that some queries are extremely common and they repeat quite often in the logs. On the other hand, NL sentences are rarely repeated exactly. Therefore, while the vocabulary of Web search queries is extremely rich, regular co-occurrence patterns make queries much more predictable than NL sentences. This much higher predictability is another feature relating queries to a protolanguage, rather than a mature language. *(c)* None of the

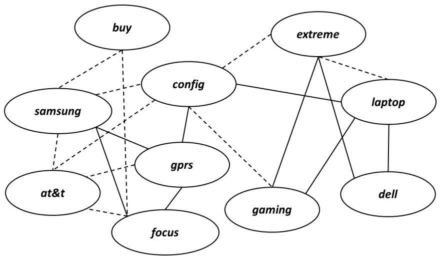

Figure 1. Illustration of a WCN for queries.

cross-entropy and the entropy values (log of perplexity base two) are equal to each other, which means that there are non-trivial shifts in the probability mass distributions of n-grams common between the two logs, and queries are undergoing noticeable internal structural change.

4. Complex Network Modeling of Web Search Queries

Languages, being complex systems, can be modeled as complex networks (Ferrer-i-Cancho & Solé, 2001). Topological analysis of these networks has enabled researchers to analyze the statistical properties of NL texts. The most popular and well-studied representation of a language corpus is the Word Co-occurrence Network (WCN) (Ferrer-i-Cancho & Solé, 2001). 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 labeled by a unique word 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. Co-occurrence can be defined variously; in our research, we consider a local model of co-occurrence where 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. Edges resulting from random collocations are suitably pruned using joint probability measures. Fig. 1 illustrates the concept of a WCN for queries by showing the network generated from a toy query log. Pruned edges are shown using dashed lines.

Experimental method. Our goal was to study the basic statistics of query WCNs from 2006 and 2010 and see how they compare with similar WCNs for Standard English (reported in Cancho and Solé (2001)). We built WCNs from $1M$ queries randomly sampled from the AOL and Bing logs. Words were stemmed using the Porter Stemmer before network construction. Then we measured basic network statistics of the largest connected components (LCC) of these WCNs, namely the number of nodes ($|N|$), edges ($|E|$), average degree (k), clustering coefficient (CC) and average shortest path length (ASPL) (Table 2). CC is a measure of triadic closure in graphs, and a high CC indicates a high proportion of these closures in the network.

Results and interpretations. We make several insightful observations: *(a)*

Table 2. WCN statistics for queries and NL.

Property	Queries (2006)	Queries (2010)	NL
$\|N\|$	$83,525$	$136,555$	$460,902$
$\|E\|$	$1.1M$	$1.4M$	$16.1M$
k	25.404	20.660	69.863
CC	0.592	0.630	0.437
ASPL	3.193	3.305	2.670

Even though the dataset size is comparable for both ($\simeq 1M sentences$), the number of nodes (unique words) are much less for queries. This is because of two reasons. First, queries are smaller and often repeated; second, these are figures for the LCC, and connectivity is poorer for queries. *(b)* Interestingly, the cumulative degree distributions of the WCNs for queries (both 2006 and 2010) are observed to be two-regime power laws. This is a strikingly similar behavior between WCNs built from NL sentences and queries. It is known that such degree distributions correspond to two types of words in the vocabulary – the *kernel* and the *peripheral lexicon* (Ferrer-i-Cancho & Solé, 2001). Hence, such a division is applicable for query words as well. *(c)* While the kernel in Standard English, the mother language for our queries, has about $5,000$ words, the corresponding number is only $1,000$ for queries. The periphery-to-kernel size ratio is much larger for queries than NL. The kernel in queries is much less tightly coupled than NL and kernel-periphery edges dominate the network, while intra-kernel edges form the majority in NL. *(d)* We observe that the CC of the network increases from 2006 to 2010, while the ASPL also increases. One would expect that an increase in CC leads to more triadic closures, which would eventually lead to decreased ASPL. This apparent paradox is explained by the kernel-periphery structure – the new words added mostly extend the periphery, and owing to their low connectivity (in turn due to low frequency and co-occurrence) increase the ASPL. On the other hand, most of the new edges are added between existing words, which increase the CC. The effect of new incoming nodes with few edges seems to outweigh the impact of new edges among existing nodes. *(e)* The values of average degree, CC and ASPL all indicate that queries are moving further away from NL. Interestingly, this is in contrast to the trends indicated by 2- and 3-gram entropy, and in support of the 1-gram trend. All these observations provide evidence in support of the hypothesis that queries, though having a structure comparable to NL in many respects, are really not moving towards the mother language – Standard English; rather, they are evolving as a new and unique linguistic system, perhaps a protolanguage.

5. Word Position Analysis for Queries

One of the important aspects of NL is the significance of the position and ordering of the words in a sentence in determining the meaning. Queries have traditionally

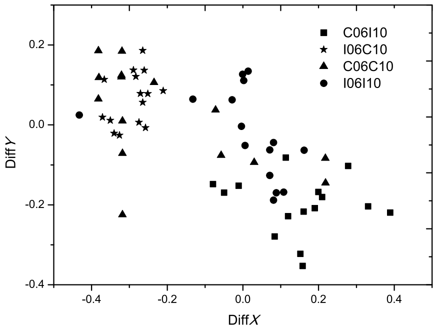

Figure 2. Difference vectors represented on a 2-D plane.

been assumed to be *bags-of-words*, where ordering is considered to have no relevance to meaning. Nevertheless, recent research has challenged this notion (Manshadi & Li, 2009). To understand the importance of word positions in a query, we did the following experiments.

Content and intent segments. We identified segments or multiword phrases in queries using our segmentation algorithm (Saha Roy et al., 2012). Then, we classified all the (single or multiword) *segments* of a query as *intent* (I) or *content* (C). Content segments refer to the topic or the main keywords present in the query (`barack obama`), whereas intent segments represent the type of information the user is interested in (`images`). Segments that *co-occur* with a large number of distinct segments over the entire query log are more likely to indicate user intent (Yin & Shah, 2010). Here we consider the top 2000 segments, when sorted in descending order of co-occurrence counts, as intent, and the rest as content.

Experimental method. Absolute positions are dependent on several factors such as query length. Therefore, we introduce the notion of relative positions – *beginning (b)*, *middle (m)* and *end (e)* for a segment s in a query. From the query logs, we compute three probabilities – $P_b(s)$, $P_m(s)$ and $P_e(s)$, which are respectively the probabilities of observing s in the beginning, middle and end of a query, for the 2010 Bing log and the 2006 AOL log. Then, we computed the change in these probabilities over the four years for each segment, which constitute the difference vector. Since $P_b(s) + P_m(s) + P_e(s) = 1$ for both logs, the difference vectors are co-planar. In Fig. 2, we represent this plane on the 2-D $X - Y$ plane by appropriate rotation of the vectors. The four classes of common segments are represented by different symbols.

Results and interpretations. We can clearly see squares (C to I) and stars form clusters (I to C), thus showing that there are noticeable statistical differences between the coordinates of such segments. The triangles (remain I) and circles (remain C) are overlapping, even though most of the triangles are on the left (negative X) and most of the circles are towards the centre. These observations are further analyzed next. Words such as `wiki` were labeled as content in 2006, but as intent in 2010. These segments (643 in number out of a total of $76,229$ common segments) have emerged as intents through popular usage patterns. $1,260$ segments that were labeled as *intent* in both 2006 and 2010. We see no significant changes in their positional statistics, indicating their stabilization in the query structure. Such words prefer to be at the ends (like `reviews`) or the beginnings (like `how to`) of queries (except for items like `and`), and show observable rise in the associated probabilities $P_b(s)$ or $P_e(s)$. However, a majority of intent segments prefer to be at the end of the query. For example, `sony stocks latest updates` is preferred over `latest updates on sony stocks`. This can be explained by a user model of *query formulation* where the content part is conceived first, followed by the specification of the associated requirements. Several intent segments show significant gains in occurrence probabilities, reflecting the relative abundance with which users add qualifiers to their queries now. While `titanic` would be a more commonly expected query in 2006, it is not surprising to come across `titanic movie review imdb` today. This *stacking* of intent segments as well as a general increase in the number of intent segments and their abundance are the dominant factors towards increased query length.

695 segments were identified as content in 2010 that were previously labeled as intent (like `white pages`). These segments were mostly popular intent identifiers in 2006 which have gradually fallen out of favour over the next few years. In 2010, these units were mostly issued as standalone segments, possibly as esoteric interests of specific users. Earlier, they appeared mostly to the right of content segments. Thus, they show significant drops in $P_e(s)$, and associated rise in $P_b(s)$. Segments that have remained as content ($73,635$) are mostly entities or classes of some kind. While content segments were popularly issued as standalone queries or with a single qualifier in 2006, increased specificity of user needs have added intent words to the left (`what are`) or right (`bio`) of the former class of segments. So they show noticeable drops in their $P_b(s)$ or $P_e(s)$.

6. Conclusions and Future Work

In this work, we analyzed the evolution of the structural properties of Web search queries over four years through three different approaches. The three studies together point to the fact that even though Web search queries are structurally simpler than NL, they are much more complex than the usually assumed bags-of-words model. Further, while some SLM based measures show that queries are approaching NL-like properties, CNT based analysis shows the reverse trend of divergence from NL. These observations underline the uniqueness of the linguistic structure of queries and show that they are undergoing a complex phase transition.

Web search queries provide a very interesting case of a self-organizing communication system which has its unique characteristics, but also has several similarities with NL that make this system interesting to study from a language evolution perspective. We believe that such a line of research can significantly enrich our knowledge of NL evolution, utilizing large volumes of well-preserved query logs over more than a decade of search engine existence.

References

Brown, P., Desouza, P., Mercer, R., Pietra, V., & Lai, J. (1992). Class-based n-gram models of natural language. *Computational linguistics*, *18*(4).

Dessalles, J.-L. (2006). Du protolangage au langage : modèle d'une transition. *Marges linguistiques*, *11*.

Duan, H., & Hsu, B.-J. P. (2011). Online spelling correction for query completion. In *WWW '11* (pp. 117–126). New York, NY, USA: ACM.

Dunning, T. (1994). *Statistical identification of language.* Computing Research Laboratory, New Mexico State University.

Ferrer-i-Cancho, R., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London B*, *268*(1482).

Guichard, E. (2002). *L'internet : mesures des appropriations d'une technique intellectuelle.* These, Ecole des hautes études en sciences sociales.

Huang, J., Gao, J., Miao, J., Li, X., Wang, K., Behr, F., & Giles, C. L. (2010). Exploring web scale language models for search query processing. In *WWW '10.* New York, NY, USA: ACM.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.*, *36*(2).

Jurafsky, D., & Martin, J. (2000). *Speech & language processing.* Pearson Education.

Manshadi, M., & Li, X. (2009). Semantic tagging of web search queries. In *ACL '09.* Stroudsburg, PA, USA: Association for Computational Linguistics.

Pass, G., Chowdhury, A., & Torgeson, C. (2006). A picture of search. In *Infoscale '06.* New York, NY, USA: ACM.

Saha Roy, R., Choudhury, M., & Bali, K. (2012). Are Web search queries an evolving protolanguage? In *Evolang IX* (pp. 304–311). Singapore: World Scientific Publishing Co.

Saha Roy, R., Ganguly, N., Choudhury, M., & Laxman, S. (2012). An ir-based evaluation framework for web search query segmentation. In *SIGIR '12.* New York, NY, USA: ACM.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Spink, A., Wolfram, D., Jansen, M. B. J., & Saracevic, T. (2001). Searching the web: the public and their queries. *J. Am. Soc. Inf. Sci. Technol.*, *52*.

Yin, X., & Shah, S. (2010). Building taxonomy of Web search intents for name entity queries. In *WWW '10.* New York, NY, USA: ACM.