ARE WEB SEARCH QUERIES AN EVOLVING PROTOLANGUAGE?

RISHIRAJ SAHA ROY
Indian Institute of Technology Kharagpur
Kharagpur, 721302, India
rishiraj@cse.iitkgp.ernet.in

MONOJIT CHOUDHURY and KALIKA BALI
Microsoft Research India
Bangalore, 560025, India
{*monojitc, kalikab*}*@microsoft.com*

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. Web search queries have a unique structure, which is more complex than just a bag-of-words, yet simpler than a natural language. This structure has been evolving over the past decade which is an artefact of the way search engines are evolving and aggressively using feedback from past users to serve current and future users better. In this paper, we argue that queries can be considered as an evolving protolanguage from functional, structural and dynamical points of view. Therefore, Web search logs, a perfectly preserved and rich dataset, can probably reveal several interesting facts about the evolution of protolanguage.

1. Introduction

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 language that has a structure far simpler than NL, but perhaps more complex than the commonly assumed bag-of-words model. In fact, Web search queries define a new and fast evolving language of its own, whose dynamics is governed by the behavior of the search engine towards the user and that of the user towards the engine. With a small number of highly popular terms and a large number of rarer terms, search queries possess properties strikingly similar to NL, yet have several unique features (Spink et al., 2001; Saha Roy et al., 2011) – they define a language on their own.

The objective of this paper is to present and carefully scrutinize the proposition that "Web search queries can be considered as an evolving *protolanguage*." While a couple of past researchers (Guichard, 2002; Dessalles, 2006) have cursorily mentioned this idea, there has been no detailed analysis, at least from the perspective of language evolution, of the structure of queries and the nature of interactions between users and search engines that have led to this structure. Queries are mostly formulated as short and ad hoc strings of words with little or no grammatical structure. They exhibit properties such as a relaxed word order, absence of long range dependencies and relatively few inflections, that have also been suggested as features of a protolanguage. Therefore, studying queries and their evolution can perhaps help us understand the various facets of language and protolanguage evolution. This might be especially useful in light of the fact that historical data and large-scale controlled experiments are major bottlenecks in language evolution research.

At the very outset we want to note that there is a fundamental difference between evolution of human language and that of queries: while human language is used for communication between two human beings (presumably) having very similar cognitive capabilities, queries are used as the means of communication between a human user and a search engine, which are incomparable not only in terms of their cognitive capabilities, but also in their biological and cultural history of language use. This asymmetry between the communicating agents in the context of queries can raise serious doubts about the basic proposition presented in this paper. However, we believe that there is an alternative and perhaps a more realistic interpretation of this communicative behavior, which is to assume that ultimately queries are actions of the users on a shared environment represented by the search engine. While the users might believe that they are communicating their information need to the search engine through their queries (which of course is true), the search engine behaves pretty much like a blackboard and its response is dependent completely on the action-response of other users in the past. This is especially true for modern commercial search engines which extensively rely on user queries, URL clicks and explicit feedbacks on relevance of documents for learning and improvement of the search models. Thus, we can visualize this situation as an indirect communication between two human users mediated through a shared environment (or channel) which is the search engine. It is a well-known fact that the channel (such as the structure of our articulatory and perceptual devices) has a profound effect on the structure and dynamics of the evolving language.

In the next three sections, we present analogies between queries and (proto-)languages from functional, structural and dynamical perspectives. In the last section, we present a synthesis of the salient ideas emerging from these three orthogonal perspectives on queries that point to the strong parallels between Web search queries and existing notions of protolanguage.

2. Functional Aspects of Web Search Queries

Web search queries are small fragments of texts (symbols) that are used to communicate the information need of an individual to a search engine. In this regard, the basic function of queries is similar to that of languages, which is transmission of information. Hockett (1960) proposed thirteen design features of a communication system. NLs possess all these features and in this section we see that a large number of these features are present in queries as well. Some of these features, such as **semanticity**, **arbitrariness**, **discreteness** and **duality of patterning**, are exhibited in queries by virtue of the fact that the building blocks of queries are, after all, words – which are also the basic units of NL. However, with respect to some of the other features, NL and queries are analogous in their structure and function. We discuss the other design features here.

Vocal-auditory channel: All spoken human language is produced using the vocal tract and auditory channel. While the role of vocal-auditory channel is currently irrelevant for queries, they are produced and perceived by writing (typing) and reading of text.

Broadcast transmission and directional reception: Human language can be heard if it is within the range of another person's auditory channel. Additionally, a listener, who shares the same time and space of the speaker, has the ability to determine the source of a sound by binaural direction finding. In the case of Web search, queries issued by a user are recorded in the search engine log files. The engine uses these logs to generate *query completions* for another user. This way, a query can be potentially broadcast to millions of new users. A new user can *choose* to be receptive of these completions (similar to signals) by enabling this specific feature on their search engine.

Rapid fading: Waveforms of spoken language dissipate over time and do not persist. A hearer can only receive specific auditory information at the time it is spoken. This feature is related to the modality of language, and as queries are mainly textual, they are therefore less ephemeral than spoken language.

Interchangeability: A person has the ability to both speak and hear the same signal. Anything that a person is able to hear, s/he has the ability to reproduce through spoken language. Similarly, users have the ability to understand and reformulate somebody else's query. If a person has seen a query, s/he can also use that query.

Total feedback: Speakers have the ability to hear themselves speak. Through this, they are able to monitor their speech production and internalize what they are producing through language. A user also knows what query s/he has issued, monitors it and internalizes its use.

Specialization: Human language sounds are specialized for communication, that is, humans speak mainly to transmit information. Query words too are specialized for specific information needs of the user.

Displacement: NL has the ability to refer to things in space and time and communicate about things that are currently not present. Queries also allow the users to seek information about past and future events or objects.

Productivity: NL allows for the creation of new and unique meanings of utterances from previously existing utterances and sounds. Likewise, a pre-existing set of words (around 1.2 million in our dataset sampled through Bing Australia[a]) can be combined to formulate unseen queries.

Traditional transmission: Human language is not completely innate and acquisition depends in part on the learning of a language. Users can learn how to formulate queries from search experts, engine guidelines, books (Ray et al., 1998), engine feedback and mimicking other users.

Thus, from a purely functional perspective, Web search queries are very similar to NL.

3. Structural Aspects of Web Search Queries

We analyzed around 16.7 million Web search queries collected through Bing Australia in the year 2008–'09 to understand query structure. The average length of a query in this dataset was found to be 4 words, which is much higher than the earlier reported average of 2.21 words per query (Jansen et al., 2000). Short queries (one or two words) constitute 18.96% of the queries and are mostly named entities and dictionary entries. Long queries, which have nine or more words and comprise only 3.23% of all queries in our dataset, are generally grammatically correct natural language sentences, computer generated error messages and song lyrics.

Traditionally, queries were assumed to be an unordered set of keywords, commonly referred to as the bag-of-words model. While this might still be the case for short (< 4 words) and a significant portion of medium length (between 4 and 8 words) queries, the latter often exhibit more complex structure than a bag-of-words model, but not as complex as NL. For example, the query *gprs config nokia n96 telstra australia* is clearly not semantically equivalent to *telstra gprs n96 config nokia australia*. This is largely due to the presence of multiword expressions (MWEs) and conceptual units within such queries. Medium length queries are on the rise today owing to the vast improvements in the performance of search engines on such queries.

Although a random permutation of the words within a query does not necessarily lead us to a query which is semantically equivalent, it turns out that queries can almost always be segmented into contiguous chunks of words such that all permutations of these segments represent semantically equivalent queries. Recent research has established the usefulness of query segmentation and the presence of a structure in queries using a variety of Web resources to accomplish the task (Hagen et al., 2011). Segments can be considered to be broadly of two types – *heads*, representing the core information need, and *modifiers*, indicating specific intent of the user (Saha Roy et al., 2011). The semantics of queries depend largely on the interaction and dependencies between heads and modifiers. For example, in the query *compare internet explorer and mozilla firefox, internet explorer* and *mozilla*

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

firefox represent heads, while *compare* and *and* appear as modifiers. More complex dependency models have also been proposed for queries (Bendersky et al., 2009).

There have not been many studies to understand the unique linguistic structure of queries. Barr et. al. (2008) showed that 70% of all words in queries are nouns, followed by adjectives (7.1%) and prepositions (3.7%). In a recent study, Saha Roy et al. (2011) have shown that word co-occurrence networks constructed from query logs reveal interesting similarities and differences between queries and NL. Like NL, a two-regime degree distribution in word or segment co-occurrence networks of queries reveals the existence of a small kernel and a very large periphery. But unlike NL, where a large fraction of sentences are formed only using the kernel words, most queries consist of units (words/segments) both from the kernel and the periphery. Word co-occurrence networks for NL have small average shortest path (the small world effect) which, it has been argued, facilitates fast word access in the mind. The analysis of query networks thus shows that queries do not exactly behave like languages at a cognitive level, but share sufficient similarities to rule out the possibility of them lacking any linguistic structure. Hence, they might be considered similar to a protolanguage evolving between humans.

There is no concrete empirical evidence for the existence of protolanguage, though “the hypothesis of a protolanguage helps to bridge the otherwise threatening evolutionary gap between a wholly alingual state and the full possession of language” (Bickerton, 1995). There have been several speculations and suggestions on the structural aspects of a protolanguage. Wray (1998) and Smith (2006) suggest that protolanguages have short sequences of symbols strung into ad hoc sentences, a feature characteristic of queries, in that the average query is significantly shorter than the average natural language sentence. Dessalles (2006) suggests that there is no consistent ordering of the symbols in a protolanguage, nor are there inflections or long range dependencies. Queries fit in well with this hypothesis too. Wray (1998) and Dessalles (2006) further suggest that protolanguage has a simple or no grammar, which is true for queries. Wray (1998) also suggests that there are a limited number of referential symbols in a protolanguage, which is not true for queries. This claim, however, has been contested by Smith (2006). We note that for queries there is already an agreed-upon huge vocabulary that consists of words that we have readily borrowed from NL. Words specific to the query language or new interpretations of existing symbols (if any) are few (e.g., *wiki*, which generally indicates that the source of the result pages should be *Wikipedia*).

4. Dynamics of Web Search Queries

Search engines are *complex adaptive systems* that are able to communicate with humans and evolve at two levels – algorithms and models. Search engines have come a long way since the first generation search systems (McBryan, 1994). Even though search engine companies rarely publish parts of their internal algorithms, the huge volume of Web IR literature over the last decade is an indication enough that search algorithms have evolved. Algorithmic evolution for search engines include more sophisticated machine learning algorithms for ranking and use of a higher number of features for retrieval. These changes are analogous to agents undergoing anatomical modifications over several years – like the unique structure of the vocal tract and the descended position of the larynx in humans (Hauser et al., 2002).

Evolution of the model, on the other hand, is information that the search system learns from user interactions and uses to present better results in the future. Models are learnt by the search engine through constant user feedback and preferences gathered during the course of Web searches. More and more Web data is crawled to make better document models. More query logs are used to build better query analysis models. Query log analysis can be used to study individual search behavior, query duplications, user sessions and query correlations (Silverstein et al., 1999). Clickthrough data (Joachims, 2002; Xue et al., 2004) and pseudo-relevance feedback (Yu et al., 2003) are also used by search engines to enrich their models of relevance.

This evolution of models is similar to the cultural transmission of language. Adaptation of the search engine is a population-level phenomenon, where individual users are agnostic to the fact that their interactions with the search engines indeed affect the response of the search engine for other users and vice versa. Cultural transmission for the language of queries can be considered from two aspects: Experts teaching novices how to search, and new users learning search tips and tricks from the collective knowledge of the Web, or relevant books (Ray et al., 1998) – like traditional language transmission. An individual's competence in language is derived from data which is itself a consequence of the linguistic competence of other individuals (Smith et al., 2003). Modern theories of cultural evolution recognize that cultural traditions are socially transmitted from person to person between and within generations (Steele et al., 2010). Individual click data or search engine use affect the engine as a whole. Users unknowingly affect the response of the engine towards other users, effectively transmitting information of some kind through the engine.

Incorporation of user feedback has tremendously improved the performance and perception of the popular commercial search engines. While the algorithmic components of a search engine rarely make any attempt to understand NL or complex queries, search engines can intelligently process very complex queries just by learning from past user behavior. This gives an average user the impression that the search engine is indeed getting smarter, and consequently they are motivated to formulate more complex queries. This results in a population level snowball effect leading to increase in the structural complexity of the queries.

5. Discussions and Conclusions

In this research, we have highlighted some similarities that Web search queries are observed to share with an evolving protolanguage, giving evidence from three different aspects. First, the function and some of the basic features of queries are similar to that of NL. Second, the structure of the queries are in between that of a random bag-of-words and a full-fledged NL form. Nevertheless, this structure seems to be evolving in complexity. Third, the evolutionary dynamics of queries is analogous to models of cultural evolution for language. Although this evolution is actually an outcome of the interactions between the users via the search engine, it seems as though the search engine is itself evolving in this process.

In conclusion, we would like to draw the attention of the language evolution research community to Web search queries, which can provide an immensely potent source of data for understanding the structure and dynamics of protolanguage and NL evolution. Nevertheless, one must also keep in mind the fact that after all, queries are issued by agents, and search engines are designed by engineers who already have a very complex linguistic communication system in place to express their ideas. This will surely bias the evolution of queries in certain directions which may or may not reflect the true evolution of a protolanguage.

Acknowledgments

The first author would like to thank Microsoft Corporation and Microsoft Research India for his Microsoft Research India PhD Fellowship Award, which supported this work.

References

Barr, C., Jones, R., & Regelson, M. (2008). The linguistic structure of english web-search queries. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1021–1030). Stroudsburg, PA, USA: Association for Computational Linguistics.

Bendersky, M., Croft, W. B., & Smith, D. A. (2009). Two-stage query segmentation for information retrieval. In *Proceedings of the 32nd international acm sigir conference on research and development in information retrieval* (pp. 810–811). New York, NY, USA: ACM.

Bickerton, D. (1995). *Language and human behavior.* University College London Press.

Dessalles, J.-L. (2006). Du protolangage au langage : modèle d'une transition. *Marges linguistiques*, *11*, 142–152.

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

Hagen, M., Potthast, M., Stein, B., & Bräutigam, C. (2011). Query segmentation revisited. In *Proceedings of the 20th international conference on world wide web* (pp. 97–106). New York, NY, USA: ACM.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569-1579.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*, 88-96.

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. *Information Processing and Management*, *36*(2), 207 - 227.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (pp. 133–142). New York, NY, USA: ACM.

McBryan, O. A. (1994). Genvl and wwww: Tools for taming the web. In *Proceedings of the first international world wide web conference* (pp. 79–90).

Ray, E. J., Ray, D. S., & Seltzer, R. (1998). *Altavista search revolution.* Osborne Publishing.

Saha Roy, R., Ganguly, N., Choudhury, M., & Singh, N. K. (2011). Complex network analysis reveals kernel-periphery structure in web search queries. In *Proceedings of the 2nd international acm sigir workshop on query representation and understanding* (pp. 5–8). New York, NY, USA: ACM.

Silverstein, C., Henzinger, M. R., Marais, H., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, *33*(1), 6-12.

Smith, K. (2006). The protolanguage debate: bridging the gap? In *Proceedings of the 6th international conference on the evolution of language* (p. 315-322).

Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems (ACS)*, *6*(04), 537-558.

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*, 226–234.

Steele, J., Jordan, P., & Cochrane, E. (2010). Evolutionary approaches to cultural and linguistic diversity. *Philos Trans R Soc Lond B Biol Sci*, *365*(1559), 3781-3785.

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication*, *18*(1), 47-67.

Xue, G. R., Zeng, H. J., Chen, Z., Yu, Y., Ma, W. Y., Xi, W., & Fan, W. (2004). Optimizing web search using web click-through data. In *Cikm 2004* (pp. 118–126). New York, NY, USA: ACM.

Yu, S., Cai, D., Wen, J. rong, & Ma, W. ying. (2003). Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Intl. world wide web conf. (www)* (pp. 11–18).