Information Processing and Management 000 (2016) 1–26

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Syntactic complexity of Web search queries through the lenses of language models, networks and users

Rishiraj Saha Roy[a,1,*], Smith Agarwal[b,2], Niloy Ganguly[c], Monojit Choudhury[d]

[a] Databases and Information Systems Group, Max Planck Institute for Informatics, Saarbrücken, Germany
[b] Experian PLC, Cyberjaya, Malaysia
[c] Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India
[d] Multilingual Systems Research Group, Microsoft Research India, Bangalore, India

ARTICLE INFO

Article history:
Received 17 February 2015
Revised 17 February 2016
Accepted 5 April 2016
Available online xxx

Keywords:
Query complexity
Statistical language models
Word co-occurrence networks
Crowd-sourcing

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.

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 Jansen and Spink (2006). 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 (Ntoulas, Cho, & Olston, 2004; Risvik & Michelsen, 2002; Schwartz, 1998). 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 (Pass, Chowdhury, & Torgeson, 2006; Saha Roy, Choudhury, & Bali, 2012a; Spink, Wolfram, Jansen, & Saracevic, 2001). Search queries represent a unique mode of interaction between humans and artificial systems, and they differ observably in syntax from that of the

* Corresponding author. Tel.: +4917622327841.
E-mail addresses: rishiraj@mpi-inf.mpg.de (R. Saha Roy), smith.agarwal@my.experian.com (S. Agarwal), niloy@cse.iitkgp.ernet.in (N. Ganguly), monojitc@microsoft.com (M. Choudhury).

[1] This work was completed during the author's stay at the Indian Institute of Technology Kharagpur.
[2] This work was done during the author's internship at the Indian Institute of Technology Kharagpur.

http://dx.doi.org/10.1016/j.ipm.2016.04.002
0306-4573/

parent NL. This has led researchers to argue that probably queries are acquiring linguistic properties of their own (Dessalles, 2006; Guichard, 2002; Jansen, Spink, and Saracevic, 2000; Saha Roy et al., 2012a; [74]; Spink et al., 2001). 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* (Saha Roy et al., 2012a). 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 (Brown, Desouza, Mercer, Pietra, & Lai, 1992b). Such a line of research in the past has been extensively used for analyzing ancient languages (like the Indus script hypothesis (Rao et al., 2009)), studying languages with diverse typological properties (Gauvain, Messaoudi, & Schwenk, 2004), and also for understanding non-linguistic systems such as music (Downie, 1999) and the genetic code (Mantegna et al., 1995). Over time, *n*-gram models have been appropriately generalized or restricted using more sophisticated linguistic features capturing various syntactic and semantic properties (see Bellegarda (2004) 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 (Zissman & Singer, 1994), machine translation (Koehn, 2010), spelling correction (Duan & Hsu, 2011) and information retrieval (IR) (Ponte & Croft, 1998).

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 Choudhury and Mukherjee (2009) 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 (Dorogovtsev and Mendes, 2001; Ferrer-i-Cancho & Solé, 2001).

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 (2002): "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 (Paikeday, 1985), 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. (2009) 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 (Sproat, 2010). Subsequent work on Indus scripts used network modeling to further substantiate Rao et al.'s original claim (Sinha, Izhar, Pan, & Wells, 2011). However, as we shall see, such an analysis, by itself, is not sufficient. Therefore, if queries are indeed an 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 (Srikanth & Srihari, 2002; Yan, Guo, Lan, & Cheng, 2013) 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 tri-gram-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.

Please cite this article as: R. Saha Roy et al., Syntactic complexity of Web search queries through the lenses of language models, networks and users, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.04.002