ABSTRACT

The co-evolution of the Web and commercial search engines, and the inability of such search engines to process natural language (NL) questions, have resulted in search queries being formulated in a syntax which is more complex than a bag-of-words model, but more flexibly structured than sentences conforming to NL grammar. In this thesis, we take the first steps to understand this unique syntactic structure of Web search queries in an unsupervised framework, and apply the acquired knowledge to make important contributions to Information Retrieval (IR). First, we develop a query segmentation algorithm that uses query logs to discover syntactic units in queries. We find that our algorithm detects several syntactic constructs that differ from NL phrases. We proceed to augment our method with Wikipedia titles for identifying long named entities. Next, we develop an IR-based evaluation framework for query segmentation which is superior to previously employed evaluation schemes against human annotations. Here, we show that substantial IR improvements are possible due to query segmentation. We then develop an algorithm that uses only query logs to generate a nested (or hierarchical) query segmentation, where segments can be embedded inside bigger segments. Importantly, we also devise a technique for directly applying nested segmentation to improve document ranking. Subsequently, we use segment co-occurrence statistics computed from query logs to find that query segments broadly fall into two classes – content and intent. While content units must match exactly in the documents, intent units can be used in more intelligent ways to improve the quality of search results. More generally, the relationship between content and intent segments within the query is vital to query understanding. Finally, we generate large volumes of artificial query logs constrained by n-gram model probabilities estimated from real query logs. We perform corpus-level and query-level comparisons of model-generated logs with the real query log based on complex network statistics and (crowdsourced) user intuition of real query syntax, respectively. The two approaches together provide us with a holistic view of the syntactic complexity of Web search queries which is more complex than what n-grams can capture, but yet more predictable than NL.

Keywords: Query understanding, Query syntax, Query segmentation, Query intent, Query complexity