Syntactic Analysis of Web Queries with Question Intent

Yuval Pinter¹, Roi Reichart¹,², Idan Szpektor¹, and Avihai Mejer¹

¹Yahoo Labs, Haifa 31905, Israel, {yuvalp, roiri, idan, amejer}@yahoo-inc.com
²Technion IIT, Haifa 32000, Israel

Abstract

Accurate automatic processing of Web queries is important for high quality information retrieval from the Web. While the syntactic structure of a large portion of these queries is trivial, the structure of queries with question intent is much richer. In this paper we therefore extend the standard dependency grammar to describe the syntax of queries with question intent. The extended grammar is driven by the concept of a segment – an independent syntactic unit within a potentially larger syntactic structure. We then develop a general algorithm, based on the idea of query to question mapping, that can adapt any given dependency parser trained on standard edited text to produce syntactic structures that conform to the extended grammar, without requiring training data in the form of queries manually annotated with a dependency structure. On a new dataset of thousands of queries with question intent our algorithm is shown to outperform baselines trained on edited text only and to perform similarly to models trained with as many as several thousand annotated queries.

1 Introduction

As the Web grows in mass, it encompasses ever-increasing amounts of text. A major gate to this invaluable resource is through Web queries which users compose to guide a search engine in retrieving the information they desire to inspect. Automatic processing of Web queries is therefore of crucial importance. Previous research (Bergsma and Wang, 2007; Barr et al., 2008) suggested that a large proportion of Web queries are trivial in structure (usually referring to entity lookup, e.g. “frozen” or “condos in NY”). However, with the increasing popularity of Community Question Answering (CQA) sites, such as Yahoo Answer[1] and StackOverflow[2], as well as other social QA sites such as various forums, more Web queries encompass information needs in the form of questions that can be answered by these sites. We found that this subcategory of queries, which we call CQA queries (Liu et al., 2011), exhibits a wide range of structures, from simple noun phrases to concatenated phrases to full sentences. This suggests that the processing of such queries may benefit from syntactic analysis. Examples for some of the more complex structures are shown in Table 1.

Recent progress in statistical parsing, (Zhang and Nivre, 2011; Choi and McCallum, 2013), has resulted in models that are both fast, parsing several hundred sentences per second, and accurate. These parsers, however, still suffer from the problem of domain adaptation (McClosky et al., 2010), excelling mostly when their training and test domain are similar. This problem is of particular importance in the heterogeneous Web (Petrov and McDonald, 2012) and is expected to worsen when addressing queries due to their non-standard grammatical conventions.

In another line of research, syntactic analysis of User Generated Content (UGC) has become prevalent (Petrov and McDonald, 2012; Kong et al., 2014; Eisenstein, 2013). Yet, these efforts have generally focused on aspects of UGC that pertain to grammatical mistakes made by users (Foster et al., 2008) and to the unique writing conventions of specific Web platforms, such as Twitter (Foster et al., 2011; Kong et al., 2014). Our analysis of thousands of CQA queries, however, reveals that regardless of such issues, CQA queries are generated by a well-defined grammar that sometimes deviates from the one used to generate the standard written language of edited resources such as newspapers.

Consequently, this work has two main contributions. First, we extend the standard dependency grammar to describe the syntactic process which governs the generation of queries with question intent. The extended grammar is driven by the concept of a syntactic segment – an independent syntactic unit within a potentially larger syntactic structure. A query may include several segments, which can be related to each other in a myriad of semantic relations but lacking an explicit syntactic connection. Hence, an analysis of a query
Table 1: Examples of CQA queries of different structural composure.

<table>
<thead>
<tr>
<th>Type</th>
<th>Example queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full sentence</td>
<td>how many bags of food does a horse eat; what does bold mean; my spleen hurts when i walk</td>
</tr>
<tr>
<td>Incomplete or broken sentence</td>
<td>muscle in leg is called; why page takes so long to load; how to find rate</td>
</tr>
<tr>
<td>Complex phrase</td>
<td>bed sheet that goes with pink and white room; inability to make eye contact</td>
</tr>
<tr>
<td>Syntactically disconnected phrase</td>
<td>modem internet off light; missing malaysia airplane psychic; resignation letter unhappy</td>
</tr>
</tbody>
</table>

requires the identification of its segments and of the internal dependency structure of each segment, and may be complemented by finding the inter-segment semantic relationships.

As a second contribution, we develop a general algorithm for parsing CQA queries that can adapt any given dependency parser trained on standard edited text (e.g. the Wall-Street-Journal PTB (Marcus et al., 1993)) to produce syntactic structures that conform to the extended grammar. Specifically, our approach views a CQA query as a reformulation of a grammatical question that expresses the user’s intent. Our algorithm therefore first maps an input CQA query to a grammatical question. Then, it uses an off-the-shelf dependency parser (trained on grammatical text) to parse the question. Finally, the algorithm projects the question parse tree into a syntactic representation of the input query that is grounded in our extended dependency grammar.

Taking a projection-based approach, our algorithm enjoys the abilities of state-of-the-art parsers to accurately parse grammatical sentences. In addition, it does not require annotated queries for training, alleviating the need for a costly and error-prone annotation process. The only supervision it does require, on top of the parser training data, is a set of (query, question) pairs, automatically derived from a query log of a Web search engine, for the training of the query-to-question mapping component.

We constructed a new dataset consisting of thousands of CQA queries from the Yahoo Answers query log, and annotated these queries according to our extended dependency grammar. We evaluated our algorithm on the tasks of syntactic segmentation and root finding at the segment level. Our algorithm outperforms two strong baselines that do not use annotated queries for training and performs similarly to models trained on thousands of manually annotated queries.


References


Jinho D Choi and Andrew McCallum. 2013.