Syntactic Analysis of Web Queries with Question Intent

Yuval Pinter[1], Roi Reichart[1,2], Idan Szpektor[1], and Avihai Mejer[1]

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

[1]`answers.yahoo.com`
[2]`stackoverflow.com`

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

Table 1: Examples of CQA queries of different structural composure.

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

Cory Barr, Rosie Jones, and Moira Regelson. 2008. The linguistic structure of english web-search queries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shane Bergsma and Qin Iris Wang. 2007. Learning noun phrase query segmentation. In *EMNLP-CoNLL*. Citeseer.

Jinho D Choi and Andrew McCallum. 2013. Transition-based dependency parsing with selectional branching. In *ACL (1)*, pages 1052–1062.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*.

Jennifer Foster, Joachim Wagner, and Josef van Genabith. 2008. Adapting a wsj-trained parser to grammatically noisy text. In *Proceedings of ACL-HLT: Short Papers*.

Jennifer Foster, Özlem Çetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, Josef Van Genabith, et al. 2011. # hardtoparse: Pos tagging and parsing the twitterverse. In *proceedings of the Workshop On Analyzing Microtext (AAAI 2011)*, pages 20–25.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Qiaoling Liu, Eugene Agichtein, Gideon Dror, Evgeniy Gabrilovich, Yoelle Maarek, Dan Pelleg, and Idan Szpektor. 2011. Predicting web searcher satisfaction with existing community-based answers. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 415–424.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, volume 59. Citeseer.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 188–193. Association for Computational Linguistics.