Web Semantics: Science, Services and Agents on the World Wide Web 30 (2015) 22–38

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents
on the World Wide Web

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

Discovering and understanding word level user intent in Web search queries

Rishiraj Saha Roy[a,*], Rahul Katare[a], Niloy Ganguly[a], Srivatsan Laxman[b,1], Monojit Choudhury[c]

[a] Computer Science and Engineering, Indian Institute of Technology Kharagpur, India
[b] Scibler Technologies Private Limited, India
[c] Microsoft Research India, India

ARTICLE INFO

Article history:
Received 15 July 2013
Received in revised form
27 July 2014
Accepted 28 July 2014
Available online 7 August 2014

Keywords:
Query understanding
Query intent
Intent words
Co-occurrence entropy

ABSTRACT

Identifying and interpreting user intent are fundamental to semantic search. In this paper, we investigate the association of intent with individual words of a search query. We propose that words in queries can be classified as either *content* or *intent*, where content words represent the central topic of the query, while users add intent words to make their requirements more explicit. We argue that intelligent processing of intent words can be vital to improving the result quality, and in this work we focus on intent word discovery and understanding. Our approach towards intent word detection is motivated by the hypotheses that query intent words satisfy certain distributional properties in large query logs similar to function words in natural language corpora. Following this idea, we first prove the effectiveness of our corpus distributional features, namely, word co-occurrence counts and entropies, towards function word detection for five natural languages. Next, we show that reliable detection of intent words in queries is possible using these same features computed from query logs. To make the distinction between content and intent words more tangible, we additionally provide operational definitions of content and intent words as those words that should match, and those that need not match, respectively, in the text of relevant documents. In addition to a standard evaluation against human annotations, we also provide an alternative validation of our ideas using clickthrough data. Concordance of the two orthogonal evaluation approaches provide further support to our original hypothesis of the existence of two distinct word classes in search queries. Finally, we provide a taxonomy of intent words derived through rigorous manual analysis of large query logs.

1. Introduction

Semantic search has attracted a good amount of research in recent years [1–3]. The goal of semantic search is to improve the result relevance by appropriately understanding user intent and using intelligent document retrieval techniques to leverage the knowledge of this intent. Thus, the ability to identify user intent is one of the first steps in semantic search. Most often, the search query is a translation of the user's intent into a short sequence of keywords. This imposes great value on every word in the query from the aspect of a semantic search engine. Past research has mostly focused on inferring the intent of the query as a whole, and the most generic intent classes were found to be informational, navigational and transactional [4–6]. In this research, we take a deeper look at query intent, zooming in on individual words as possible indicators of user intent.

From an information retrieval (IR) perspective, the equivalence of a Web search query with an unordered sequence of words (or a "bag-of-words") has long been challenged, with research on term dependence [7–9] and term proximity models [10–14] showing significant improvements in retrieval performance. Extending this idea of the presence of a *query structure* further, we propose that words or multiword units in queries basically belong to two classes—*content words* that represent the central topics of queries, and *intent words*, which are articulated by users to refine their

[*] Corresponding author. Tel.: +91 9477588851.
E-mail addresses: rishiraj.saharoy@gmail.com, rishiraj@cse.iitkgp.ernet.in (R. Saha Roy), rah.ykg@gmail.com (R. Katare), niloy@cse.iitkgp.ernet.in (N. Ganguly), srivatsan.laxman@gmail.com (S. Laxman), monojitc@microsoft.com (M. Choudhury).

[1] This work was done while the author was at Microsoft Research India.

http://dx.doi.org/10.1016/j.websem.2014.07.010
1570-8268/© 2014 Elsevier B.V. All rights reserved.

information needs concerning the content words. The class of content units include, but are not restricted to named entities (like `brad pitt`, `titanic` and `aurora borealis`)—anything that is capable of being the topic of a query would be the content unit in the context of that query. For example, `blood pressure`, `marriage laws` and `magnum opus` are legitimate examples of content words or units. Intent words or intent units, on the other hand, present vital clues to the search engine regarding the specific information sought by the user about the content units. For instance, intent units like `home page`, `pics` and `meaning`, all specify unique information requests about the content units. The queries `brad pitt website`, `brad pitt news` and `brad pitt videos` all represent very different user intents. It is not hard to see that while content units need to be matched inside document text for relevance, it is possible to leverage the knowledge of intent units to improve user satisfaction in better ways. For example, words like `pics`, `videos` and `map` can all trigger relevant content formats to directly appear on the result page. Words like `near` and `cheap` may be used to sort result objects in the desired order. These ideas motivate us to focus on the discovery and understanding of query intent units in this research.

Appropriately understanding the distinction between the two classes of words and concretizing these notions of intent and content required rigorous manual analysis of large volumes of query logs on our part. During this process, we observed that intent units share corpus distributional properties similar to function words of natural language (NL). NLs generally contain two categories of words—*content* and *function* [15]. In English, nouns, verbs, adjectives and most adverbs constitute the class of content words. On the other hand, pronouns, determiners, prepositions, conjunctions, interjections and other particles are classified as function words. While content words express meaning or *semantic content*, function words express important grammatical relationships between various words within a sentence, and themselves have little lexical meaning. The distinction between content and function words, thus, plays an important role in characterizing the syntactic properties of sentences [16–18]. Distributional postulates that are valid for function word detection, like the co-occurrence patterns of function words being more diverse and unbiased than content words, seemed to be valid for query intent units as well. Following these leads, we first segment queries to identify possible multiword units using a state-of-the-art query segmentation algorithm [19], and compute the relevant distributional properties, namely, co-occurrence counts and entropies, for the obtained query units. We found that the units which exhibit high values of these indicators indeed satisfy our notions about the class of intent units. Subsequently, we systematically evaluated our findings against human annotations and clickthrough data (which represent functional evidence of user intent) and substantiate our hypotheses.

In hindsight, we understand that while NL function words have little describable meaning (like `in`, `of` and `what`) and only serve to specify relationships among content words, well-defined semantic interpretations can be attributed to most intent words (like `map`, `pics` and `videos`). Intent words, even though effectively lacking purpose without the presence of a content word(s) in the same query, carry weight of their own within the query. Thus, content and intent units play slightly different roles in the query from the roles of content and function words in NL sentences. It simply turns out that function words in NL and intent words in queries share similar statistical behavior. Function words and intent words are still not fully comparable, and an important difference between the two is the fact that the definition of a function word is not context-dependent, whereas intent words can also behave as content words depending on the context (Section 4).

The objective of this paper is to identify and characterize intent words in Web search queries, words that are explicit indicators of user intent, and it is organized as follows. In Section 2, we begin with a verification of the efficacy of corpus-based distributional statistics towards function word identification and through rigorous experimentation over five languages, discover that *co-occurrence counts and entropies* are the most robust indicators of function words in NL. Having convinced ourselves of the power of co-occurrence statistics in detecting function words across diverse languages, we apply similar techniques to discover intent units in Web search queries (Section 3). This is followed by a simple algorithm to label intent units in the context of individual queries and subsequent evaluations using human annotations and clickthrough data (Section 4). Observing that co-occurrence statistics locate quite a diverse set of intent units, we attempt to provide a taxonomy of such units based on their relationships with content words that we believe can be very useful in semantic search (Section 5). Finally, we present concluding remarks and open directions for future work (Section 6).

2. Distributional properties of NL function words

Function words play a crucial role in many Natural Language Processing (NLP) applications. They are used as features for unsupervised POS induction and also provide vital clues for grammar checking and machine translation. In this section, we first re-examine this popular hypothesis that the most frequent words in a language are the function words. By *function words or units* we refer to all the closed-class lexical items in a language, e.g., pronouns, determiners, prepositions, conjunctions, interjections and other particles (as opposed to open-class items, e.g., nouns, verbs, adjectives and most adverbs). We note that the statistics presented here are applicable for both single-word (`in, about`) as well as multiword (`how to, because of`) function units from corpora, though the latter demands chunking of the NL text. We perform all the NL experiments on unsegmented (or unchunked) sentences and hence report the results for detection of single word function units. Nevertheless, Web search queries, on which we mainly focus, have been suitably segmented by the state-of-the-art algorithm [19].

2.1. Datasets

For the NL experiments, we shall look at five languages from diverse families: English, French, Italian, Hindi and Bangla. English is a *Germanic* language, French and Italian are *Romanic* languages, and Hindi and Bangla belong to the *Indo-Aryan* family. Therefore, any function word characterization strategy that works across these languages is expected to work for a large variety of languages.

The details of the corpora used for these five languages are summarized in Table 1. The sentences were uniformly sampled from larger datasets. M in the value columns denotes million. S, N, V and F denote the *numbers* of all sentences, all words, unique words (vocabulary size) and function words, respectively. We note that the Indian languages have almost twice as many function words as compared to the European ones. This is due to morphological richness and the existence of large numbers of modal and vector verbs.

2.2. Metric

In a distributional property-based function word detection approach, the output is a ranked list of words sorted in descending order of the corresponding indicator value. Here we adopt a popular metric, *Average Precision* (AP) [20,21], used in IR for the evaluation of ranked lists. More specifically, let $w_1, w_2, \ldots, w_n$ be