Unsupervised Query Segmentation Using only Query Logs

Nikita Mishra* Rishiraj Saha Roy* Niloy Ganguly
Indian Institute of Technology Kharagpur
Kharagpur, India 721302
nikitamishra07@gmail.com,
{rishiraj, niloy}@cse.iitkgp.ernet.in

Srivatsan Laxman Monojit Choudhury
Microsoft Research Lab India
Bangalore, India 560080
{slaxman, monojitc}@microsoft.com

ABSTRACT

We introduce an unsupervised query segmentation scheme that uses query logs as the only resource and can effectively capture the structural units in queries. We believe that Web search queries have a unique syntactic structure which is distinct from that of English or a bag-of-words model. The segments discovered by our scheme help understand this underlying grammatical structure. We apply a statistical model based on Hoeffding's Inequality to mine significant word *n*-grams from queries and subsequently use them for segmenting the queries. Evaluation against manually segmented queries shows that this technique can detect rare units that are missed by our Pointwise Mutual Information (PMI) baseline.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Algorithms, Measurement, Experimentation

Keywords

Query Grammar, Query Structure, Unsupervised Query Segmentation, Hoeffding's Inequality

1. INTRODUCTION

Web search queries with length between 3 and 10 words, which constitute approximately 80% of all queries in the query log that we have analyzed here, seem to have a unique structure; they are neither bags-of-words, nor grammatically correct natural language phrases or sentences. For example, the queries ***3g not working nokia n96 telstra australia*** and ***nokia n96 telstra australia 3g not working*** can both be paraphrased in natural language as "3G *is* not working *in a* Nokia N96 *mobile phone bought from the* Telstra *store in* Australia." The queries seem to have been derived from the underlying English sentence by dropping the stop words like *is* and *from*, stripping off the common nouns such as *phone* and *store*, and randomly permuting the left over chunks – ***3g not working***, ***nokia n96***, and ***telstra Australia***. This leads us to the interesting observation that queries are, in fact, *bags-of-units*, as opposed to bags-of-words.

Previous research has expressed and addressed the need for identification of these *units* [1 - 5], a process termed as *query segmentation*. Nevertheless, efforts have been mainly directed towards identification of multiword named entities [1] and natural language phrases [2]. Towards this end, various external resources such as Webpages [3, 5], search result snippets [4] and Wikipedia titles [5] have been used. Although these methods can help in retrieval, query expansion and query suggestion, we strongly believe that they miss out on the unique syntactic properties of queries due to a bias towards projecting natural language structure on queries.

Thus, we think that the linguistic structure of queries is distinct from that of the standard language (i.e., English, in our case); the first step towards understanding this structure is to understand the nature of the constituent word groups. These word groups should be identified solely on the basis of queries, because use of external resources raises the risk of projecting natural language structures onto the queries; and a proper understanding of this structure coupled with automatic techniques for parsing it can lead to significant performance improvements in various IR tasks. In this work, we take the first steps to unravel the structure of queries by proposing an unsupervised method for query segmentation that uses only query logs. As we shall see, the segments identified by our method do not necessarily align with natural language segments, yet it is clear that they are meaningful.

2. METHOD

We are given a large collection of search queries. Consider an *n*-gram $M = (w_1 w_2 \dots w_n)$ where w_j-s denote the words constituting *M*. Let $\{q_1, q_2, \dots, q_k\}$ denote the subset of queries in the log that contain all the words of *M,* though not necessarily occurring together as an *n*-gram. Our premise is that search queries can be viewed as bags of Multi-Word Expressions (MWEs), which is to say that any permutation of the MWEs constituting a particular search query will effectively represent the same query. Thus, to test if an observed *n*-gram is an MWE, we could ask if the constituents of an MWE appear together more frequently than they would under a *bag-of-words* null model. We now formalize this intuition in a new test of significance for detecting MWEs in search queries.

Let us fix our focus on *M,* a candidate MWE. Let X_i be the indicator variable for the event "*M* occurs in the query q_i". Let P_i denote the probability of this event and let $\ell_i$be the length of q_i. There are$(\ell_i - n + 1)$ locations where *M* can be positioned in q_i and for each choice of location there are $(\ell_i - n)!$ ways of permuting the remaining $(\ell_i - n)$ non-MWE words of q_i.

* Part of this work has been done during the authors' internship at Microsoft Research India

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

Thus, we can write the probability of $[X_i = 1]$ under the bag-of-words model (null) as follows:

$$P_i = \frac{(\ell_i - n + 1) \times (\ell_i - n)!}{\ell_i!} = \frac{(\ell_i - n + 1)!}{\ell_i!} \quad \ldots (1)$$

We define $X = \sum_i X_i$ (which models the number of times the words of M appear together in the k queries). We use Hoeffding's Inequality to obtain an upper-bound δ on the probability of $[X \geq N]$, where N denotes the observed value of X in the data (also referred to as the *frequency* of M):

$$Prob[X \geq N] \leq exp\left(-\frac{2(N - E(X))^2}{k}\right) = \delta \quad \ldots (2)$$

where, the expectation $E(X)$ is given by $E(X) = \sum_i P_i$. We obtain δ for each n-gram M and define $(-\log \delta)$ as the *MWE score* for M. If δ is small, then the surprise factor is higher indicating a greater chance of M being an MWE, and vice versa. We note that unigrams have a score of zero, since their observed and expected frequencies are equal.

For computational reasons, we compute the *MWE scores* only for n-grams whose constituent words have each appeared in at least α queries in the database (where α is a user-defined threshold). We add an n-gram to the list of *significant n-grams* if its MWE score exceeds β (a second user-defined threshold). In our experiments we used $\alpha = 10$ and $\beta = 0.6k$ (where k is the number of queries in which all the words of the n-gram occur, though not necessarily together).

We now have a list of significant n-grams and their associated MWE scores. We use this list to perform unsupervised query segmentation as follows: First, we compute a final score for each possible segmentation by adding the MWE scores of individual segments. Then we pick the segmentation that yields the highest segmentation score. Here we use a dynamic programming approach to search over all possible segmentations.

3. EVALUATION

All our experiments have been performed on a subset of one million queries (from a total of 342 million) collected through Bing Australia (http://www.bing.com.au). The segmentation accuracy was evaluated using four standard metrics discussed in [5] against a manually segmented set of one thousand six-word queries (handling upto 5-grams). The PMI threshold for MWE significance is 8.2. Results are shown in Table 1.

Table 1. Segmentation Accuracies (in %)

Method	Seg-Acc	Precision	Recall	F-score
PMI	70.69	49.23	54.59	51.77
Proposed Scheme	**75.20**	**54.95**	**60.09**	**57.41**

The results show that our scheme performs better than a baseline method that uses PMI. On close examination of the segmentation results, we found that many segments discovered by our scheme did not match with human annotations because human segmentation is largely influenced by natural language grammar. For example, the query (***how to spot***) (***a fake bill***), where parentheses mark the segmentation boundaries by manual annotators, is segmented as (***how to) (spot a fake***) (***bill***) by our method. While ***a fake bill*** is a noun phrase, and therefore, a valid segment according to the Standard English grammar, one cannot deny the fact that ***how to*** expresses a class of intent in queries and is found to be associated with diverse concepts such as ***save money***, ***play guitar*** or ***make tea***. Interestingly, ***spot a fake***, which makes very little sense as an MWE, is in fact quite commonly seen in queries expressing a generic *action phrase* applicable to diverse objects such as ***video***, ***gucci bag*** or ***mona lisa painting***. Some other examples of generic query intents discovered by this method are ***information about***, ***difference between*** and ***history of the***.

The proposed solution is also capable of detecting named entities such as ***windows media player*** and ***nikon d5000***, including rare ones like ***very hungry caterpillar***. The disagreements between the segmentation by this method and manually annotated data are partly due to influence of English grammar on annotators and inherent ambiguities in some queries, and partly due to lack of domain knowledge which makes it hard to judge the statistical significance of rarer named entities and multiword expressions. The latter can be suitably addressed by using external resources such as Wikipedia, though adequate care has to be taken so that the generic intent phrases are not lost in the process. The accuracy figures reported here are lower than the state-of-the-art, but it should be emphasized that since we do not use any external resources or manually segmented data to learn the models, our results are not comparable to those reported earlier. Moreover, the motivation and goals of our work are fundamentally different.

4. CONCLUSIONS AND FUTURE WORK

In this work, we have proposed an unsupervised method of query segmentation that uses Web queries as the only resource. The method unravels structural units of queries that are distinct from natural language phrases and outperforms the PMI baseline in every metric. Currently we are enriching the segmentation scheme by using lists of named entities obtained from other sources and conducting linguistic and statistical analysis of the segmented queries to discover deeper structural patterns.

5. ACKNOWLEDGMENTS

We would like to thank Bhaskar Mitra, Anjana Das and Victor Das from Bing, India for providing us with the data.

6. REFERENCES

[1] Guo, J., Xu, G., Cheng, X. and Li, H. 2009. Named Entity Recognition in Query. In *Proc. of SIGIR '09*. pp. 267-274

[2] Bergsma, S. and Wang, Q. I. 2007. Learning Noun Phrase Query Segmentation. In *Proc. of EMNLP '07*. pp. 819-826

[3] Risvik, K. M., Mikolajewski, T. and Boros, P. 2003. Query Segmentation for Web Search. In *Proc. of WWW '03 (Poster session)*.

[4] Brenes, D. J., Gayo-Avello, D. and Garcia, R. 2010. On the Fly Query Segmentation Using Snippets. In *Proc. of CERI '10*. pp. 259-266

[5] Tan, B. and Peng, F. 2008. Unsupervised Query Segmentation Using Generative Language Models and Wikipedia. In *Proc. of WWW '08*. pp. 347-356