Towards Optimum Query Segmentation: In Doubt Without

(Extended Abstract)*

Matthias Hagen Martin Potthast Anna Beyer Benno Stein

Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

ABSTRACT

Query segmentation is the problem of identifying compound concepts or phrases in a query. We conduct the first large-scale study of human segmentation behavior, introduce robust accuracy measures, and develop a hybrid algorithmic segmentation approach based on the idea that, in cases of doubt, it is often better to (partially) leave queries without any segmentation.

1. INTRODUCTION

Keyword queries are the predominant way of expressing information needs on the web. Search engines nowadays rely on tools that help them to interpret, correct, classify, and reformulate every submitted query in a split second before the actual document retrieval begins. We study one such tool that identifies indivisible sequences of keywords in a query (e.g., `new york times`) that users could have included in double quotes—the task of query segmentation.

Our contributions include the first large-scale analysis of human segmentation behavior (50 000 queries, each segmented by 10 annotators) showing that different segmentation strategies should be applied to different types of queries. In particular, a good strategy often is to refrain from segmenting too many keywords (i.e., in doubt without segmentation).

2. NOTATION AND RELATED WORK

A query q is a sequence $(w_1, \ldots, w_k)$ of k keywords. Every contiguous subsequence of q forms a potential segment. A valid segmentation for q consists of disjunct segments whose concatenation yields q again. The problem of query segmentation is the automatic identification of the "best" valid segmentation, where "best" refers to segmentations that humans would choose or that maximize retrieval performance. Note that a valid segmentation determines for each pair $\langle w, w' \rangle$ of consecutive keywords in q whether or not there should be a segment break between w and w'. Hence, there are 2^{k-1} valid segmentations for a k-keyword query and $k(k-1)/2$ potential segments with at least two keywords.

Risvik et al. [5] were the first to propose an algorithm for query segmentation based on pointwise mutual information. Later on, more sophisticated approaches like the supervised learning method by Bergsma and Wang [1] combined many features (web and query log frequencies, POS tags, etc.). Recently, efficiency issues become more important [3] and evaluation moves away from simple accuracy against human segmentations towards retrieval impact analyses [4].

*Original paper with all the omitted details in CIKM 2012 [2].

DIR 2013, April 26, 2013, Delft, The Netherlands.

3. HOW HUMANS SEGMENT

Our study of human segmentation behavior is based on the Webis-QSeC-10 corpus [3] consisting of 53 437 web queries (3–10 keywords) with at least 10 different annotators per query. One of our intentions is to compare human quoting on noun phrase queries with that on other queries. As automatic POS-tagging in short queries is a difficult task, we restrict our analysis to *strict noun phrases* (SNP) composed of only nouns, numbers, adjectives, and articles. These parts-of-speech can be identified reliably using for instance Qtag.[1] About 47% of the queries are tagged as SNP queries.

Our study of how humans quote queries results in the following major findings. (1) SNP queries are segmented more often than others, (2) in segmented SNP queries more keywords are contained in segments, and (3) annotators agree more on short queries but unanimity is an exception (many queries even do not have a segmentation supported by an absolute majority of annotators). These findings suggest that algorithms aiming at accuracy against human segmentations should take into account the query type. The second implication is to carefully reconsider the traditional accuracy measures (some based on annotator unanimity).

4. ACCURACY MEASURES REVISITED

Segmentation accuracy is typically measured against a corpus of human segmentations on three levels: query accuracy (ratio of correctly segmented queries), segment accuracy (precision and recall of the computed segments), and break accuracy (ratio of correct decisions between pairs of consecutive words). The crucial point is the choice of the reference segmentation from the corpus. Traditionally, the reference is the segmentation that best fits the computed one (i.e., the one with highest break accuracy) without any further considerations. We argue that for corpora with many annotators per query (e.g., the Webis-QSeC-10) this is an oversimplification and scoring references from a set of weighted alternatives should be an integral part of accuracy measuring.

Given a query q, and a list of m reference segmentations $(S_1, \ldots, S_m)$ from m different annotators, we propose the following two strategies to select a reference segmentation. (1) Weighted Best Fit: select the S_i chosen by an absolute majority of annotators if there is one. Otherwise select the S_i as the traditional best fit strategy (i.e., the S_i maximizing break accuracy). But then, the obtained accuracy values are weighted by the ratio of votes allotted to S_i compared to the maximum number of votes on any segmentation in $(S_1, \ldots, S_m)$. (2) Break Fusion: instead of selecting a reference segmentation from $(S_1, \ldots, S_m)$, fuse them into one. For each pair of consecutive words in q: if at least half of the annotators inserted a segment break, so does this strategy. If not, no break is inserted.

[1] http://phrasys.net/uob/om/software

To demonstrate the impact of the new reference schemes, we apply them in a comparison of the segmentation algorithms from the literature (results in the full paper). With our new schemes many of the relative accuracy differences between segmentation algorithms increase and more of these differences become statistically significant. Hence, the new reference selectors provide a more robust means to evaluate segmentation accuracy.

5. HYBRID QUERY SEGMENTATION

The decision whether or not to introduce segments into a query is a risky one: a bad segmentation leads to bad search results or none at all, whereas a good one improves them. Since keeping users safe from algorithm error is a core principle at most search engines, and since even a small error probability yields millions of failed searches given billions of searches per day, a risk-averse strategy is the way to go. In doubt, it is always safer to do without any query segmentation. This observation suggests to use a hybrid strategy that treats different types of queries in different ways. One of the main findings on human segmentation behavior is to distinguish SNP queries from others. As potential strategies for either type, we consider algorithms from the literature and two newly developed baselines that only segment Wikipedia titles (WT) or only Wikipedia titles and SNPs (WT+SNP) following our dictionary based scheme [3].

6. EVALUATION

In our evaluation, we compare instances of hybrid query segmentation to traditional approaches with respect to three performance measures. (1) We measure segmentation accuracy using the Webis-QSeC-10. (2) We measure retrieval performance in a TREC setting using the commercial search engine Bing and the Indri ClueWeb09 search engine hosted at Carnegie Mellon University.[2] (3) We measure runtime performance and memory footprint.

We have systematically combined traditional segmentation algorithms (including the option "none" of not segmenting) to form instances of hybrid segmentation. As expected, there is no one-fits-all combination which maximizes performance with respect to all of the above measures. The following table shows the best performing combinations.

Query type	Hybrid segmentation instance		
	HYB-A (accuracy)	HYB-B (Bing)	HYB-I (Indri)
SNP	[3] (= WT+SNP)	None	None
other	WT	WT	[3]

In what follows, we give brief descriptions of the experimental results (more details in the full paper). An explanation for the variant HYB-A can be found in our analysis of human quoting behavior. There, it is shown that accuracy-oriented algorithms should segment SNP queries more aggressively (more keywords in segments) than other queries, which in turn should be segmented conservatively (less keywords in segments). This is exactly the strategy of HYB-A. On SNP queries, the algorithm [3] aggressively segments all phrases that appear at least 40 times on the web, whereas the WT baseline on the other queries conservatively segments only Wikipedia titles.

With respect to retrieval performance we evaluate on the TREC topics in the Web tracks 2009–2011 and the Million Query track 2009 with at least one document being judged as relevant and at least 3 keywords (61 topics from the Web tracks, 294 from the Million query track). Our results suggest that different search engines (i.e., retrieval models) each require specifically tailored hybrid segmentation algorithms. Otherwise, query segmentation may not improve significantly over not segmenting at all.

The main findings of evaluating accuracy and retrieval performance are the following: (1) better accuracy not necessarily improves retrieval performance, (2) SNP queries can often be left unsegmented in terms of retrieval performance. However, there is a grain of salt: our TREC experiments are small-scale compared to the number of queries that went into measuring accuracy. The retrieval performance experiments should be scaled up significantly in order to draw more reliable conclusions. In any case, our experiments have shown that the decision of when to segment at all is an important one.

Besides accuracy and retrieval performance, also runtime and memory consumption are crucial criteria to judge the applicability of a segmentation algorithm in a real-world setting. Runtime is typically measured as throughput of queries per second while memory consumption concerns the data needed for operation. Regarding throughput, a pointwise mutual information baseline is by far the fastest approach (with bad accuracy and retrieval performance). The WT and WT+SNP baselines are faster than [3] since they sum up fewer weights of potential segments. The hybrid approaches are slowest due to the POS tagging step. With respect to memory consumption the WT baseline needs an order of magnitude less data than mutual information or WT+SNP which in turn need much less than [3]. Taking into account the rumored monthly throughput of major search engines of about 100 billion queries (i.e., about 40 000 queries per second), all segmentation approaches can easily handle such a load when run on a small cluster of standard PCs.

7. CONCLUSION AND OUTLOOK

Our study of human query segmentation behavior inspired a new hybrid framework that treats SNP queries different than other queries and that can be tailored to mimic human query quoting better than the state-of-the-art algorithms. However, an important and somewhat unexpected outcome of complementary TREC style evaluation is that maximizing segmentation accuracy not necessarily maximizes retrieval performance as well. Nevertheless, we show the flexibility of the hybrid framework and optimize it for two retrieval models. There, not segmenting SNP queries at all is best, opposing our finding that humans quote SNP queries more aggressively.

We hypothesize that query segmentation is especially beneficial on long non-SNP queries, which currently are underrepresented in the TREC corpora. Hence, scaling up retrieval performance evaluation with a broad range of retrieval models is an important future direction. This could shed light on the question of why SNP queries apparently are better off without any segmentation. One starting point could be an analysis of the best segmentations for different retrieval models in order to better understand what differentiates a "perfect" retrieval-oriented segmentation from those of the algorithms developed so far.

[2] http://boston.lti.cs.cmu.edu/Services/batchquery

8. REFERENCES

[1] S. Bergsma and Q. Wang. Learning noun phrase query segmentation. In *EMNLP-CoNLL 2007*, pp. 819–826.

[2] M. Hagen, M. Potthast, A. Beyer, and B. Stein. Towards optimum query segmentation: in doubt without. In *CIKM 2012*, pp. 1015–1024.

[3] M. Hagen, M. Potthast, B. Stein, and C. Bräutigam. Query segmentation revisited. In *WWW 2011*, pp. 97–106.

[4] Y. Li, B.-J. P. Hsu, C. Zhai, and K. Wang. Unsupervised query segmentation using clickthrough for information retrieval. In *SIGIR 2011*, pp. 285–294.

[5] K. Risvik, T. Mikolajewski, and P. Boros. Query segmentation for web search. In *WWW 2003 (Posters)*.