Unsupervised Query Segmentation Using Monolingual Word Alignment Method

Dayong Wu (Corresponding author)
School of Computer Science and Technology
Harbin Institute of Technology
PO box 321, Harbin 150001, China
Tel: 86-451-8641-3683 E-mail: dywu@ir.hit.edu.cn

Yu Zhang
School of Computer Science and Technology
Harbin Institute of Technology
PO box 321, Harbin 150001, China
Tel: 86-451-8641-3683 E-mail: zhangyu@ir.hit.edu.cn

Ting Liu
School of Computer Science and Technology
Harbin Institute of Technology
PO box 321, Harbin 150001, China
Tel: 86-451-8641-3683 E-mail: tliu@ir.hit.edu.cn

Received: September 27, 2011 Accepted: November 22, 2011 Published: January 1, 2012
doi:10.5539/cis.v5n1p13 URL: http://dx.doi.org/10.5539/cis.v5n1p13

Abstract

In this paper, we propose a novel unsupervised approach to query segmentation using the word alignment model which is usually adopted in statistical machine translation system. Query segmentation is to obtain complete phrases or concepts in a query by segmenting a sequence of query terms, which is an important query processing procedure for improving information retrieval performance in search engines. In this work, we use a novel monolingual word alignment method to segment queries and automatically obtain the query structure in the form of multilevel segmentation. Our approach is language independent and unsupervised so that it is easy to be applied to various language scenarios. Experimental results on a real-world query dataset show that our approach outperforms the state of the art language model based method, which demonstrates the effectiveness of the proposed approach in query segmentation.

Keywords: Query segmentation, Word alignment, Query processing, Information retrieval

1. Introduction

Query segmentation is a query processing procedure in which a sequence of query terms is segmented in order to obtain complete phrases or concepts. For example, a query "*chanel the fifth avenue new york*" can be segmented the following partitions "*Chanel*" "*the fifth avenue*" "*new york*". Query segmentation is crucial for search engine systems. When using complete phrases and concepts in queries as search units, search engine can provide more relevant retrieval results (Li, Hsu, Zhai, & Wang, 2011). The queries that users issue to search engine usually have special linguistics structures and characteristics which are different with natural languages. The existing natural language processing techniques usually can not be straightly applied to query processing. It is necessary to develop specific approaches to query segmentation for real-world queries.

In this work, we propose an approach to query segmentation by utilizing the word alignment model that is often used in statistical machine translation. The word alignment model, which is normally conducted on parallel

bilingual corpus (Brown, Pietra, Pietra, & Mercer, 1993), is here modified to perform on monolingual corpus which refers to the query data in this work. Our approach can obtain the structural units of queries only by exploiting the query dataset extracted from a real-world query log without other resources. Query logs record all types of user search queries, which really reflect how users construct search queries. The monolingual word alignment based method is unsupervised statistical model that do not need any human-annotation effort. We use the monolingual word alignment based method to align the terms of queries that is related closely to one another. Through iteratively aligning previously obtained segments of the queries, the multilevel segmentation structures of each query are finally parsed. Figure 1 shows an example of the multilevel segmentation of a query. The multilevel segmentation of queries can provide search engines with the information that what terms should appear contiguously in retrieved documents so as to help ranking documents. We evaluate our approach on a query dataset which is manually annotated the multilevel segmentation structures of each query. Experimental results show that our approach achieves better performance in comparison with the state of the art language model based method (Huang *et al.*, 2010), in which our approach achieves improvements of about 5% on the average overall accuracy of query segmentation and about 13% on the average F-value of obtaining query segments.

The main advantages of the proposed query segmentation approach are as follows. (1) It is an unsupervised approach without any manual annotation effort involved in training the segmentation model. (2) Only monolingual corpus instead of bilingual corpus is needed so that it is easy to prepare data resources. (3) Our approach use only the queries extracted from a query log as data resource without any additional resources. Experiments show that the approach can effectively capture the complete phrases and concepts of queries. (4) The approach is language independent and can be easily adapted to other language scenarios. As far as we know, we are the first to use the word alignment model in statistical machine translation to segment queries for information retrieval. The Experimental results demonstrate that our approach achieves more accurate multilevel segmentation structure and segments of queries than language model based method.

The rest of this paper is structured as follows. In Section 2, related work is reviewed. Section 3 describes the proposed approach in detail. The experiment results and analysis are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Related Work

In recent years, many research efforts are devoted to query segmentation due to the significance of this task in information retrieval. The pointwise mutual information (MI) based methods is initially employed to the task of query segmentation (Risvik, Mikolajewski, & Boros, 2003; Jones, Rey, Madani, & Greiner, 2006). The MI methods have a shortage that they can not represent dependencies of phrases or concepts in a query. Bergsma and Wang (2007) proposed a supervised query segmentation approach, in which they trained a segmentation boundary decision model with the syntactic and semantic features extracted from a manually annotated query segmentation data. In their approach, to label training data obviously is time consuming and cost intensive so as to limit the scale of training data. Some researchers investigated unsupervised method to query segmentation, such as the literature (Tan & Peng, 2008; Hagen, Potthast, Stein, & Bräutigam, 2010; Li *et al.*, 2011). Tan and Peng (2008) proposed a query segmentation approach adopting n-gram frequency obtained a large scale text corpus, in which they use expectation maximization (EM) algorithm to estimate segmentation scores. In addition, they exploited the Wikipedia titles to be a complementary concept resource. Hagen *et al.* (2010) designed a heuristic algorithm with the Google n-gram frequency (Brants & Franz, 2006) counts from a large Web corpus, which also attained a comparable performance of query segmentation. Li *et al.* (2011) also used the n-gram statistic to construct a query segmentation model and applied EM to estimate the model parameters. Moreover, they adopted the user-clicked document titles stored in query logs to enhance performance of query segmentation.

Note that the aforementioned works do not perform multilevel segmentation for queries but only produce flat query segments. Huang *et al.* (2010) initially proposed to segment long queries in form of the parse tree structure in their work, in which they use web-scale language models to segment queries. They consider that the multilevel query segmentation is more useful both for human interpretation and for retrieved document ranking, which inspire us to carry out the research on multilevel query segmentation in this work.

The word alignment methods are usually used to align the words in parallel bilingual sentences in statistical machine translation (Brown *et al.*, 1993). Liu, Wang, Wu and Li (2009) originally proposed a modified word alignment method on monolingual corpus, and used it to extract word collocation. Brody (2010) utilized the monolingual word alignment method to conduct unsupervised dependency parsing. In this work, we adopted the

monolingual word alignment method to automatically segment queries for information retrieval.

3. Query Segmentation with Monolingual Word Alignment Method

In this section, we firstly introduce the monolingual word alignment (MWA for short) method, and then describe the algorithm of multilevel query segmentation based on the MWA method.

3.1 Monolingual Word Alignment Method

The word alignment method has been well studied in the context of statistical machine translation. Given a pair of bilingual sentences that are translation each other, the word alignment method aims to align a source language word to its corresponding target language word. Figure 2 shows an example of Chinese-to-English word alignment. In this work, we adapt the word alignment method to the monolingual scenario, which are heavily influenced by the work (Liu *et al.*, 2009). The monolingual queries are replicated to generate the parallel query dataset in which each query pair consist of two identical queries. When aligning the terms of each query pair, the MWA method is constrained with the condition that a term is not allowed to align to itself. We illustrate the aligned terms in a Chinese query pair in Figure 3. As can be see in Figures 2 and 3, in a bilingual word alignment, the words in one language are aligned to their corresponding translation words, while for a monolingual query, each term are aligned with its most related other terms. Our approach utilizes the functionality of monolingual word alignment to find the mutually related term in queries.

The MWA method is a statistical method which models the co-occurrence frequency and position information. Given a query with l terms Q = $\{t_1,t_2,\ldots t_l\}$, the word alignments can be present as $A = \{(i, a_i)|i\in[1,l]\}$ which can be obtained by maximizing the word alignment probability of the query, as shown in Equation (1).

$$A = \arg\max_{\forall A'} Pr(A' \mid Q) \tag{1}$$

Where A' refers to all possible alignments of query terms, (i, a_i) means that the term t_i is aligned with the term t_{ai}. Due to the constraint that the term is not allow to align with itself, the alignment set is represented as $A = \{(i, a_i)|i\in[1,l]\ \&\ a_i\neq i\}$. In this work, the IBM Model 3 (Brown *et al.*, 1993) is modified and adopted to the monolingual word alignment method. The alignment probability $Pr(A'|Q)$ can be computed according to Equation (2).

$$\Pr(A \mid Q) \propto \prod_{i=1}^{l} N(\phi_i \mid t_i) \prod_{j}^{l} T\left(t_j \mid t_{a_j}\right) D\left(j \mid a_j, l\right) \tag{2}$$

Where ϕ_i refers to the number of words that is aligned with t_i. There are the three probabilities involved in Equation 2. Term alignment probability $T(t_j \mid t_{aj})$ describes the probability of the term t_j is aligned with the term t_{aj}. Position alignment probability $D(j \mid a_j, l)$ describes the probability of a term in position j is aligned with another term in position a_j. Fertility probability $N(\phi_i \mid t_i)$ describes the probability of the number of terms that is aligned with the term t_i. For example, in Figure 3, $T(t_1|t_3)$ and $T(t_2|t_3)$ describe the probabilities that both terms of "历年 (over years)", "英语 (English)" are aligned with the term "四级(Band four)". $D(1|3,5)$ and $D(2|3,5)$ describe the probabilities that the terms of position 1 and 2 are aligned with term in position 3 in the query of 5 term length. In probability $N(\phi_3 \mid t_3)$, ϕ_3 is 2, which means that the terms "英语(English)" is aligned with two terms in the query. Training method of MWA model is same as training bilingual word alignment with only a constraint that a term is not allowed to be aligned with itself.

3.2 Multilevel Query Segmentation

We apply the MWA model to find the mutually related terms of each query in a query dataset. In order to perform multilevel segmentation, we repeatedly perform term alignment on each query segmentation result in turn until the each query is merged to one term. The Figure 4 shows the algorithm of multilevel query segmentation based on the MWA method. It can be seen that the algorithm performs the process of the multilevel query segmentation in a bottom up fashion. The input of the algorithm is a query dataset, in which each query consists of its original word sequence. For example, in a language English scenario, the input queries are exactly themselves, and in Chinese scenario, each input query is a word sequence obtained by a certain process of Chinese word segmentation because no explicit delimiter such as space character is used between consecutive Chinese words. The MWA method actually can also be used to segment Chinese characters to words in queries, the discussion about which is out of the scope of this paper. When query terms have been aligned, we merge the mutually aligned terms to a new term as Step 5 in the algorithm. For example, in Figure 3 both terms "英语

(English)" and "四级(Band four)" are merged since they are mutually aligned, but terms"历年(over years)" and "四级(Band four)" could not be merged. We impose the limit that the distance of the merged word is not more than two, which means that when a term is among both mutually aligned terms, the three terms are merged to one. The algorithm repeatedly performs aligning and merging until each query is merged to one term.

4. Experiments

4.1 Data Preparation

In this work, the query dataset comes from a query log of a commercial Chinese search engine Sogou (Note 1). We randomly extracted 5 million queries without replacement from the query log for our experiments. The extracted query dataset is processed by a Chinese word segmentation tool implemented on a MWA method, in which each query is presented as a Chinese word sequence. We employ the algorithm described in section 3 to train query segmentation model on the query dataset. For evaluating the query segmentation algorithm, we randomly sampled 700 unique queries from the query dataset, where each query is constrained to be more than two words. We removed the unsuitable (consist of some bizarre symbols) and natural language sentence (NLS) queries from the query. The reason for removing NLS queries is that 1) The NLS queries are only small scale in queries (the number of them is 46 in the query dataset); 2) the NLS queries can be processed using some mature natural language processing technologies. We got a set of 604 queries after the preprocessing and then manually labeled them in a multilevel segmentation form. Three annotators were employed for the annotation task. We compared the three annotation results on observed agreement. As shown in Table 1, the three annotation results reach a relatively high agreement level.

4.2 Baseline Method based on Language Model

We implemented the long query segmentation method based on language model described in the literature (Huang *et al.*, 2010) as the baseline in our experiments. Because the Chinese queries are adopted in the experiments, we employed the Google Chinese n-gram (Brants & Franz, 2006) data to build the language model in the baseline method. The language model based method performs query segmentation in a top down fashion, which is briefly introduced here. The method defined the segment-based point mutual information (SPMI), as shown in Equation (4), where query $q = <w_1, w_2, \ldots w_k>$is split into two parts by boundary t, left part $q_l = <w_1, w_2, \ldots w_{t-1}>$and right part $q_r = <w_t, w_{t+1}, \ldots w_k>$.

$$SPMI(q,t) = \log \frac{P(q_l q_r)}{P(q_l)P(q_r)} \tag{4}$$

The SPMI can be computed as the Equation (5) based on Markov property and the order r language model. In our work, we use the order 3 language model, namely trigram model in the baseline method. Hence we can segment a given query according to Equation (6), where $q(n)$ denotes the query segment of node n. n^* denotes the node is chosen and then is segmented into no-empty left and right parts, C denotes all possible nodes considered for segmenting in each segmentation step. More details can be found in the abovementioned literature.

$$SPMI(q,t) = \sum_{i=0}^{\min\{r-2, k-(r-2)\}} \log \frac{P(w_{t+i} \mid w_{t+i-r+1} \ldots w_{t+i-1})}{P(w_{t+i} \mid w_t \ldots w_{t+i-1})} \tag{5}$$

$$n^* = \arg\min_{n \in C} \min_{2 \le t \le |q(n)|} SPMI(q(n), t) \tag{6}$$

We estimate the probabilities in Equation (5) using Equation (7) because only the count of each n-gram token is provided in the Google Chinese n-gram data.

$$p(w_t \mid w_{t-1}, \ldots w_{t-i}) = \frac{C(w_t, w_{t-1}, \ldots w_{t-i}) + 1}{C(w_{t-1}, \ldots w_{t-i}) + |V|} \tag{7}$$

Where $C(w_t, w_{t-1}, \ldots w_{t-i})$ is the count of the word sequence of $w_t, w_{t-1}, \ldots w_{t-i}$, $|V|$ is the number of unique words occurring in the n-gram data.

4.3 Evaluation of Multilevel Query Segmentation

We evaluate the performance of our approach with the overall accuracy of whole query segmentation and precision, recall, F-value based on all query segments. The overall accuracy Ac is defined as the Equation (8).

$$Ac = \frac{\text{the number of correctly segmented queries}}{\text{the number of queries in query datset}} \tag{8}$$

We extract the all segments excluding the single word for each query in the manually labeled query datasets and automatically segmented query datasets respectively. The precision (*P*), recall (*R*), F-value (*F*) of our approach is measured based on the query segments. The precision is the percentage of correct segments in the automatically achieved segment set. The recall is the percentage of correct segments in the manually labeled segment set. The F-value is defined as follow:

$$F = \frac{2 \times P \times R}{P + R} \tag{9}$$

The evaluation results of the overall accuracy for query segmentation and *PRF* of query segments are reported in Table 2 and Table 3 respectively, which shows the performance of both MWA method and the baseline method on three labeled query datasets. As can be seen by comparing the experiments results of both methods, the MWA method provides better performance than the language model method. The MWA method achieves the improvements of about 5% on the average overall accuracy of query segmentation, and about 13% on the average F-value of obtaining query segments, which shows that the MWA method is more effective for query segmentation. We believe the reason for better performance is that the MWA method takes into account the relationship of terms in a long span and the alignment of multiple terms.

5. Conclusion

In this paper, we put forward a novel unsupervised approach to multilevel query segmentation based on the monolingual word alignment method. The word alignment model, which is well studied in statistical machine translation, is adapted to the monolingual scenario to perform the query segmentation task. Our approach is unsupervised without any manual efforts, which make the approach possible to adopt a large scale query dataset for training. The approach is designed to be language independent so that it can be easily applied to other languages. The real-world queries can be straightly used as data resource to train the query segmentation model in our approach, in which the inherent phrases and concept of queries can be latently achieved and using them to segment queries. We experimented the proposed approach with a query dataset extracted from a real-world query log and compared to the state of the art method based on language model. The experimental results show that our approach achieved better performance on multilevel query segmentation.

In future, we plan to explore more features and combine them to the proposed model in this paper to further enhance the performance of multilevel query segmentation. In addition, we also plan to integrate the multilevel query segmentation into the information retrieval model to improve the effectiveness of search systems.

References

Bergsma, S., & Wang, Q. I. (2007). Learning Noun Phrase Query Segmentation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. (EMNLP-CoNLL), 819–826.

Brants, T., & Franz, A. (2006). Web 1T 5-gram Version 1. *Linguistic Data Consortium LDC2006T13*. Philadelphia.

Brody, S. (2010). It Depends on the Translation: Unsupervised Dependency Parsing via Word Alignment. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1214-1222.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., & Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2), 263-311.

Hagen, M., Potthast, M., Stein, B., & Bräutigam, C. (2010) The Power of Naive Query Segmentation. In Proceeding of the 33rd international ACM SIGIR conference, (SIGIR'10), 797–798. http://dx.doi.org/10.1145/1835449.1835621

Huang, J., Gao, J. F., Miao, J. B., Li, X. L., Wang, K. S., & Behr, F. (2010). Exploring Web Scale Language Models for Search Query Processing. In Proceeding of the 19th international conference on World Wide Web (WWW'10). http://dx.doi.org/10.1145/1772690.1772737

Jones, R., Rey, B., Madani, O., & Greiner, W. (2006). Generating Query Substitutions. In Proceedings of the 15th international conference on World Wide Web (WWW'06), 387–396. http://dx.doi.org/10.1145/1135777.1135835

Li, Y. E., Hsu, B J, Zhai, C. X., & Wang, K. S. (2011). Unsupervised Query Segmentation Using Clickthrough for Information Retrieval. In Proceeding of the 34th Annual ACM SIGIR Conference, 285-294. http://dx.doi.org/10.1145/2009916.2009957

Liu, Z. Y., Wang, H. F., Wu, H. & Li, S. (2009). Collocation Extraction Using Monolingual Word Alignment Method. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 487-495.

Risvik, K. M., Mikolajewski, T., & Boros, P. (2003). Query Segmentation for Web Search. In Proceeding of the 12th international conference on World Wide Web (WWW'03).

Tan, B., & Peng, F. C. (2008). Unsupervised Query Segmentation Using Generative Language Models and Wikipedia. In Proceeding of the 17th international conference on World Wide Web (WWW'08), 347–356. http://dx.doi.org/10.1145/1367497.1367545

Notes

Note 1. http://www.sogou.com

Table 1. Agreement of Three Annotation Results

Annotators	1, 2	2, 3	3, 1	1, 2, 3
Observed agreement	71.5%	71.2%	72.4%	60.6%

Table 2. Overall Accuracy of Query Segmentation

Query Segmentation Method	Annotators			Average
	1	2	3	
Language model method (Baseline)	44.4%	43.5%	44.5%	44.1%
MWA method	47.5%	51.8%	48.8%	49.4%

Table 3. Precision, Recall, F-value of Query Segmentation

Query Segmentation Method	Annotators									Average		
	1			2			3					
	P	R	F	P	R	F	P	R	F	P	R	F
Language model method (Baseline)	0.473	0.471	0.472	0.488	0.472	0.479	0.492	0.489	0.490	0.484	0.477	0.480
MWA method	0.636	0.583	0.608	0.664	0.593	0.626	0.631	0.578	0.604	0.644	0.585	0.613

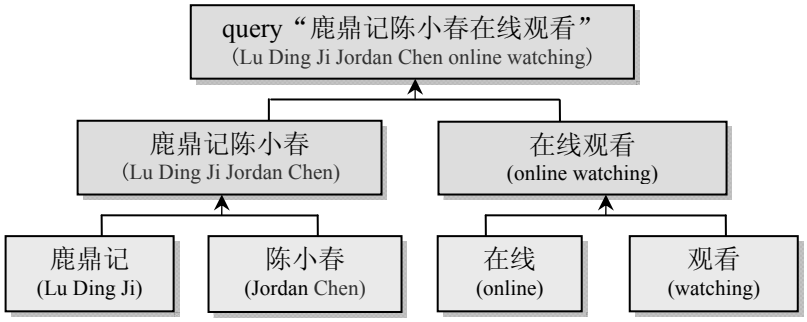

Figure 1. Example of multilevel segmentation of a query

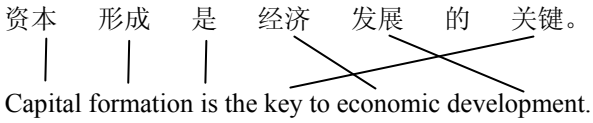

Figure 2. Example of bilingual word alignment in a pair of Chinese-to-English Sentences

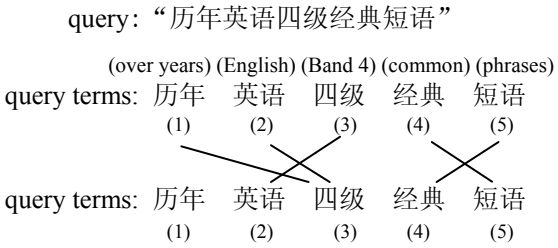

Figure 3. Example of monolingual term alignment for a Chinese query

Algorithm: Multilevel Query Segmentation

Input: Q, a query dataset

Output: S, the query dataset parsed in a multilevel segmentation form.

Procedure:

1: represent each query in Q as a sequence of terms $q_i=<t_1,t_2...t_n>$.

2: **Repeat**

3: replicate each q_i to generate query pair $q_{i\text{-}pair}$.

4: align terms in $q_{i\text{-}pair}$ with MWA method.

5: merge the mutually aligned terms in q_i to one new term and hence obtain a new sequence q_i.

6: **Until** each query in Q are merged to one term.

Figure 4. Algorithm of multilevel query segmentation