Automatic Generation of Query Sessions using Text Segmentation

Debasis Ganguly, Johannes Leveling, and Gareth J.F. Jones

CNGL, School of Computing, Dublin City University, Dublin-9, Ireland
{dganguly, jleveling, gjones}@computing.dcu.ie

Abstract. We propose a generative model for automatic query reformulations from an initial query using the underlying subtopic structure of top ranked retrieved documents. We address two types of query reformulations a) *specification* where the reformulated query expresses a more particular information need compared to the previous query; and b) *generalization* where the query is reformulated to retrieve more general information. To test our model we generate the two reformulation variants starting with topic titles from the TREC-8 ad hoc track as the initial queries. We use the average clarity score as a specificity measure to show that the specific and the generic query variants have a higher and lower average clarity score respectively. We also use manual judgements from multiple assessors to calculate the accuracy of the specificity and generality of the variants, and show that there exists a correlation between the relative change in the clarity scores and the manual judgements of specificity.

1 Introduction

Traditional Information Retrieval (IR) models assume that the information seeking process is static and involves successively refining a query to retrieve documents relevant to the original information need. However, observational studies of information seeking find that searchers' information needs change as they interact with a search system. Searchers learn about the topic as they scan retrieval results and term suggestions, and formulate revised information needs as previously posed questions are fully or partially answered [1]. The topic collections of standard ad hoc evaluation tracks fail to model this user behaviour.

The Session Track organized for the first time at TREC 2010 [2] is an effort to evaluate retrieval systems over an entire session of user queries rather than on separate independent topics. The topic creation phase involved starting with Web-track diversity topics sampled from the query logs of a commercial search engine. Specific variants of the initial topic were created by manually extracting keywords from the different subtopics. The general variants were formed in two ways: a) by constructing an over-specified query from one of the subtopics and removing words manually, and b) by adding manually selected related words from a different subtopic. Related work on automatic query reformulation includes that of Dang and Croft [3] which uses anchor text to reformulate a query by

substituting some of the original terms, assuming that the information need in the reformulated query is identical to that of the initial one. Our work is different in the sense that we seek to move the query towards a more specific subtopic or a broader topic which is associated with a change of information need.

This paper tries to answer the research questions of how to build test collections to study the "Session IR" task by modeling user interactions over a session. We aim to develop topic variants on a large scale for ad hoc retrieval collections which do not possess such meta-information as query logs or anchor texts. The novelty of the paper is that we use the underlying semantic structure of top ranked retrieved documents by applying text segmentation aiming to design a generative model for query reformulation.

2 Automatic generation of topic variants

Motivation A document retrieved in response to a query may comprise of multiple subtopics which are related to the more specific aspects of the information need expressed in the query. For example, the document *FBIS3-20090* fetched in response to the TREC query title *Foreign minorities Germany* contains a segment on *Synagogue attack*. If the user is interested in a more specific reformulation, he is likely to chooses terms which occur frequently in one or a few subtopics. Whereas if he is interested in a more general formulation, it is more likely that he would choose terms which are not concentrated in one of the subtopics but occur abundantly throughout the entire document. Applying a text segmentation based approach in our method for simulated query generation is an attempt to model this behaviour.

Jansen et. al. [4] view generalization as removal of query terms and specialization as addition of terms. This is particularly true when the implicit connectives of the query terms are strictly conjunctive in nature i.e. the relation between query terms (subordination of concepts) plays an important role in determining the type of reformulation. For example the query "osteoporosis" (TREC topic 403) could be considered as a more specific reformulation of "osteoporosis bone disorder" but only if the underlying information need involves an implicit conjunction of all the terms in the later i.e. the searcher being not interested in other bone diseases. An alternative interpretation is that the user is interested in bone disorders in general with a reference to *osteoporosis* in particular. Thus, addition of terms can also contribute to the generalization of a query if the added terms have semantic relations such as *has-a* or *is-a* with respect to the original terms i.e. there is an implied focus on one particular aspect of a query. While a possible approach to generalize a query could involve starting with a longer initial query such as the *description* part of the TREC topics followed by a removal or substitution of specific terms with more general ones, in this paper we concentrate only on the additive model of query reformulation.

Specific reformulation Our generative model tries to utilize the fact that a term indicative of a more specific aspect of an initial information need, typically

Algorithm 1 Reformulation(Q, R, n_s, n_g)

Q : The original query, R : Number of top ranked documents to use, n_s : Max. # of specific terms to add from each document, n_g : Max. # of general terms to add from each document,
$SpExpQry \leftarrow \emptyset$; $GnExpQry \leftarrow \emptyset$
for $i = 1$ to R **do**
    $d \leftarrow i^{th}$ document
    Segment d into segments $\{s_1, s_2, \ldots s_n\}$ by C99 algorithm
    $s_{max} \leftarrow$ segment with maximum number of matching query terms
    Score each term t in s_{max} by $\phi(t, s_{max})$ and add the top n_s terms to $SpExpQry$ if the term is already not in Q.
    Score each term t of d by $\psi(t)$ and add the top n_g terms to $GnExpQry$ if the term is already not in Q.
end for
return ($SpExpQry$, $GnExpQry$)

is densely distributed in a small part of the text [5]. We segment a document by applying the state-of-the-art segmentation algorithm C99 [6]. To characterize specific reformulation terms we assign scores to terms considering the following two factors: a) how frequently a term t occurs in a segment s, denoted by $\texttt{tf}(t, s)$, and how exclusive the occurrence of t in s is as compared to other segments of the same document, denoted by $\frac{|S|}{\texttt{sf}(t)}$, where $|S|$ is the number of segments in that document and $\texttt{sf}(t)$ is the number of segments in which t occurs; b) how rare the term is in the entire collection, measured by the document frequency (df), the assumption being rare terms are more likely to be specific terms. We use a linear combination to calculate term scores, as shown in Equation 1.

$$\phi(t, s) = a \cdot \texttt{tf}(t, s)\frac{|S|}{\texttt{sf}(t)} + (1 - a) \cdot \log \frac{|D|}{\texttt{df}(t)} \tag{1}$$

$$\psi(t) = a \cdot \texttt{tf}(t, d)\frac{\texttt{sf}(t)}{|S|} + (1 - a) \cdot \log \frac{|D|}{\texttt{df}(t)} \tag{2}$$

Equation 1 assigns higher values to terms which occur frequently in a segment, occur only in a few segments, and occur infrequently in the collection.

General reformulation In contrast to a more specific term, a more general term is distributed uniformly throughout the entire document text [5]. So an obvious choice is to score a term based on the combination of term frequency in the whole document (instead of frequency in individual segments) and segment frequency (instead of inverse segment frequency) where $\texttt{tf}(t, d)$ is the number of occurrences of t in d (see Equation 2). Algorithm 1 is used to create the two types of reformulations of an initial query Q. Another possible approach to generalization can involve removal or substitution of terms of higher $\phi(t, s)$ in the initial query with those having lower ones, thus making general reformulation an inverse to specialization.

Evaluation The clarity score [7] of a query is the KL divergence between the estimated distribution of generating the query from the top ranked pseudo-relevant documents and the probability of generating the query from the collection model. Since the specific version of an initial query aims at a narrower information need, we hypothesize that the clarity score of the specific version should increase. Also, for a more general information need we add terms which are expected to occur in more number of documents potentially making the query more ambiguous hence potentially resulting in a decrease of the clarity score.

3 Experiments

We start with the titles of the TREC-8 topics as initial queries and use Algorithm 1 to form the variants. The parameters were set to $(R, n_s, n_g) = (5, 3, 2)$ after a set of initial experiments with an aim to increase the average clarity of the specific variants and decrease that of the general ones. Figure 1 shows that for specific queries we obtain the maximum clarity at $a = 0.8$ and for general queries we get the minimum clarity at $a = 0.1$. We generated 100 simulated queries (two variants for each TREC-8 topic) with the above settings of parameter a. Five assessors manually judged the quality of the reformulated queries with *yes*/*no* answers. In order to seek a possible correlation between the relative changes in clarity scores and the average score of the manual judgements, for every query variant $\texttt{QV}_\texttt{i}$, obtained from $\texttt{Q}_\texttt{i}$ $(i = 1 \ldots n)$, we compute the following:

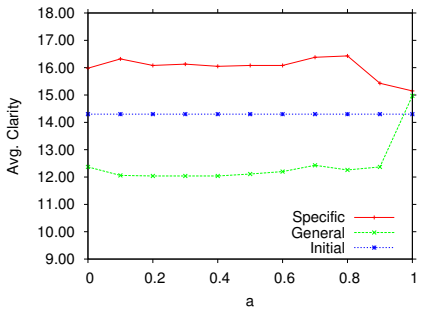

Fig. 1: Average clarity versus a.

$$\delta_i = \frac{\texttt{clarity}(\texttt{QV}_\texttt{i}) - \texttt{clarity}(\texttt{Q}_\texttt{i})}{\texttt{clarity}(\texttt{Q}_\texttt{i})}, \quad m_i = \frac{1}{N_a} \sum_{j=1}^{N_a} d_j \tag{3}$$

In Equation 3, δ_i is the relative change in the clarity score and m_i is the average decision score ($d_j = 1$ for *yes*, $d_j = -1$ for *no*), N_a being the number of assessors for the i^{th} topic. We scale the m_i values by the magnitude of clarity change to avoid ties in the m_i scores across topics, and then compute the Spearman

Table 1: Accuracy of the generated query variants

Variant	Accuracy	Spearman Coeff.	Fleiss' κ	Deduction
Specific	0.83	0.30	0.68	High accuracy, Medium correlation
Generic	0.63	-0.17	0.62	Fair accuracy, Small correlation

correlation between δ_is and $m_i|\delta_i|$s. Accuracy of the generated queries for both variants is measured by a majority decision (1 if majority agree and 0 otherwise) for each topic and averaging it out over all topics. The results are shown in Table 1. The Fleiss' κ for the assessments of both the variants show a substantial inter-assessor agreement. For the specific variant, a good example output is "poaching, wildlife preserves (bear african tiger)", where parenthesized words indicate the new words added, whereas "killer bee attacks (agricultural experts)" is indicative of an imprecise specialization. For the generalization variant "carbon monoxide poisoning (hyperbaric chamber)" is an instance of good generalization but "cosmic events (religion)" is an instance of inaccurate generalization.

4 Conclusions

We have shown that the proposed model of simulated query generation can be used to produce query reformulations with 83% and 63% accuracies for the specific and general cases respectively. We find that there are positive and negative correlations between the changes in clarity scores and the manual judgements of specificity and generality respectively which means that most of the assessors agree on a specific reformulation when the clarity score increases and most of them agree on a generalization if it decreases. A correlation of manual judgement with an automatic measure like the clarity score suggests that clarity scores alone, without the need of manual assessments, can be good indicators of the nature of information need change, thus suggesting that automatic development of user sessions on a large scale could be possible with a little or no manual post-processing effort.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project.

References

1. Bates, M.J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface. Online Review **13**(5) (1989) 407–424
2. Evangelos Kanoulas, Paul Clough, B.C.M.S.: Session track at TREC 2010. In: SIMINT workshop SIGIR '10, New York, NY, USA, ACM (2010)
3. Dang, V., Croft, B.W.: Query reformulation using anchor text. In: Proceedings of WSDM '10, New York, NY, USA, ACM (2010) 41–50
4. Jansen, B.J., Booth, D.L., Spink, A.: Patterns of query reformulation during web searching. J. Am. Soc. Inf. Sci. Technol. **60** (July 2009) 1358–1371
5. Hearst, M.: TextTiling: Segmenting text into multi-paragraph subtopic passages. CL **23**(1) (1997) 33–64
6. Choi, F.Y.Y.: Advances in domain independent linear text segmentation. In: Proceedings of the NAACL. (2000) 26–33
7. Cronen-Townsend, S., Croft, W.B.: Quantifying query ambiguity. In: Proceedings of HLT '02, San Francisco, USA (2002) 104–109