Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Topic Concentration in Query Focused Summarization Datasets

Tal Baumel, Raphael Cohen, Michael Elhadad
Ben Gurion University, Dept. of Computer Science
{talbau, cohenrap, elhadad}@cs.bgu.ac.il

Abstract

Query-Focused Summarization (QFS) summarizes a document cluster in response to a specific input query. QFS algorithms must combine query relevance assessment, central content identification, and redundancy avoidance. Frustratingly, state of the art algorithms designed for QFS do not significantly improve upon generic summarization methods, which ignore query relevance, when evaluated on traditional QFS datasets. We hypothesize this lack of success stems from the nature of the dataset. We define a task-based method to quantify *topic concentration* in datasets, *i.e.*, the ratio of sentences within the dataset that are relevant to the query, and observe that the DUC 2005, 2006 and 2007 datasets suffer from very high topic concentration. We introduce TD-QFS, a new QFS dataset with controlled levels of topic concentration. We compare competitive baseline algorithms on TD-QFS and report strong improvement in ROUGE performance for algorithms that properly model query relevance as opposed to generic summarizers. We further present three new and simple QFS algorithms, RelSum, ThresholdSum, and TFIDF-KLSum that outperform state of the art QFS algorithms on the TD-QFS dataset by a large margin.

Introduction

The task of Query-Focused Summarization (QFS) was introduced as a variant of generic multi-document summarization in shared-tasks like DUC 2005 (Dang, 2005). QFS goes beyond factoid extraction and consists of producing “a brief, well-organized, fluent answer to a need for information” (Dang, 2005), which is directly applicable in real world settings.

In multiple DUC datasets (2005, 2006, 2007) (Dang, 2005, 2006), the QFS task asks for an answer to a query as a summary of at most 250 words created from a cluster of 25-50 documents (newspaper articles). As part of the dataset preparation, assessors were instructed to populate the cluster with at least 25 documents that are relevant to the query. The instructions thus encouraged the creation of topically coherent document sets as input to the summarization task. Notably, the extent to which the document clusters are focused on the query is not directly observable: assessors could select between 50% to 100% of the documents as "relevant to the topic”. In this paper, we investigate the level of *topic concentration* (ratio of sentences relevant to the query) in QFS datasets, and the impact of the dataset topic concentration on QFS algorithms

As a research objective, QFS is a useful variant of generic multi-document summarization because it helps articulate the difference between content centrality within the documents in the cluster and query relevance. This distinction is critical when dealing with complex information needs (such as the TREC 2006 Legal Track (Baron *et al*., 2006)) because we expect that the summary should cover multiple aspects of the same general topic. However, the difference between central and topic-relevant content will only be significant when we can observe a clear difference between these two components in the dataset. Interestingly, it has been observed in (Gupta *et al*., 2007) that generic summarization algorithms (which simply ignore the query) perform as well as many proposed QFS algorithms on standard QFS datasets, such as DUC 2005. We hypothesize that this is due to the fact that existing QFS datasets have very high topic concentration in the input (the document cluster). In other words, the datasets used to evaluate QFS are not geared towards distinguishing central and topic content

Topic concentration is an abstract property of the dataset and there is no explicit way to quantify it. A direct method of quantifying this property was introduced before (Gupta *et al*., 2007) and tested on DUC 2005. The method measures similarity between sentences in the documents cluster and an Oracle expansion of the query. As many as 86% of the sentences in the overall document set were found similar to the query. We find however that this direct method has problems that we will discuss later. We introduce an alternative way to assess topic concentration in a dataset: which compares the behavior of summarization algorithms on varying subsets of the document cluster. On the DUC 2005, DUC 2006 and DUC 2007 datasets, our method indicates that these datasets have high topic concentration, which makes it difficult to distinguish content centrality and query relevance.

We aim to define a new QFS dataset that suffers from less topic concentration. In the new dataset we constructed, we explicitly combine documents covering multiple topics in each document cluster. We call this new dataset Topically Diverse QFS (TD-QFS). By construction, TD-QFS is expected to be less topically concentrated than DUC datasets. We confirm that, as expected, our method to measure topic concentration finds TD-QFS less concentrated than earlier DUC datasets and that generic summarization algorithms do not manage to capture query relevance when tested on TD-QFS. We observe that a strong QFS algorithm such as Biased-LexRank (Otterbacher *et al.*, 2009) performs significantly better on TD-QFS than generic summarization baselines whereas it showed relatively little benefit when tested on DUC 2005 (see Fig.1).

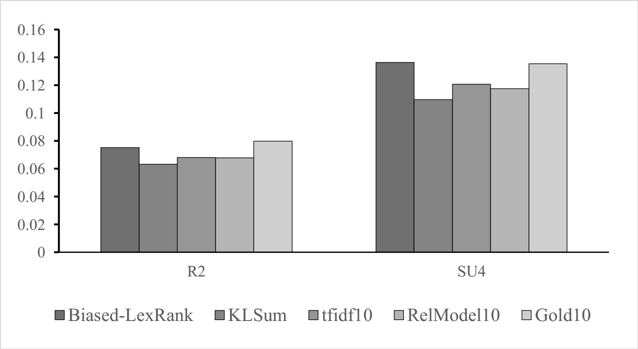

Figure 1. ROUGE-Comparing QFS methods to generic summarization methods: Biased-LexRank is not significantly better than generic algorithms.

To refine our assessment of topic concentration, we analyze a 2-stage model of QFS: (i) first filter the document set to retain only content relevant to the query using various models; (ii) then apply a generic summarization algorithm on the relevant subset. This model allows us to investigate the impact of various relevance models on QFS performance (see Fig.2).

In the rest of the paper, we introduce ways to measure topic concentration in QFS datasets based on this model, and show that existing DUC datasets suffer from very high topic concentration. We then introduce TD-QFS, a dataset constructed to exhibit lower topic concentration. We finally compare the behavior of strong baselines on TD-QFS and introduce three new algorithms that outperform QFS state of the art by a very large margin on the TD-QFS dataset.

Topic Concentrations in Document Clusters

Our objective is to assess the level of "topic concentration" in a QFS document dataset, so that we can determine the extent to which performance of QFS algorithms depends on topic concentration. For example, the DUC 2005 instructions to topic creators when preparing the dataset were to construct clusters of 50 documents for each topic, with 25 documents marked as relevant, so that, we would expect that about 50% of the documents be directly related to the topic expressed by the query.

Gupta *et al*. (2007) proposed to measure topic concentration in a direct manner: a sentence is considered relevant to the query if it contains at least one word from the query. They also measured similarity based on an Oracle query expansion: The Oracle takes the manual summaries as proxies of the relevance model, and assesses that a sentence is "similar to the query" if it shares a content word with one of the manual summaries. With this direct similarity measure, 57% of the sentences in DUC 2005 are found similar to the query; with Oracle similarity, as many as 86% of the sentences are found similar to the query. This is much higher than the expected 50% that was aimed for at construction time.

We have found that this direct measure of similarity predicts levels of topic concentration that are not good predictors of the margin between generic and focused summarization performance. We propose instead a task-based measure of topic concentration with finer granularity. We first describe the method and the new dataset we have constructed, and then show that the direct measure incorrectly predicts high concentration on a topically diverse dataset, while our new topic concentration measure distinguishes between the two datasets.

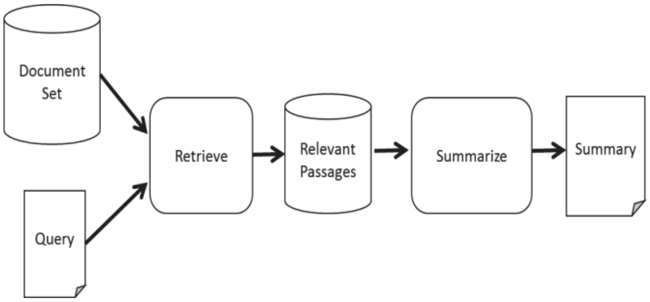

Figure 2. Two-stage QFS Scheme.

We model QFS as a 2-stage process as illustrated in Fig.2: (1) rank passages in the cluster by similarity to the query; (2) filter the document cluster and apply a generic summarization algorithm on the most relevant passages. We can now use various content retrieval methods to assess whether a passage is relevant to the query, and keep the same generic summarization method to organize the set of sentences found relevant into a set of non-redundant central sentences.

In our experiments, we use the KLSum method as the constant summarization method (Haghighi and Vanderwende, 2009). KLSum selects a set of sentences from the source documents such that the distribution of words in the selected sentences is as similar as possible to the overall distribution of words in the entire document cluster. To measure similarity across word distributions, KLSum uses the KL-Divergence (Kullback and Leibler, 1951) measure between the unigram word distributions. KLSum provides a well-motivated way to remove redundancy and select central sentences and obtains near state of the art results for generic summarization. Since we rank passages by similarity to the query, we can control the degree to which the input document cluster is filtered.

We compare three content retrieval methods in our experiments:

- The traditional TF*IDF method (Yu *et al*., 1973)
- Lavrenko and Croft's Relevance Model (2001), which was found effective in a recent survey (Yi and Allan, 2009)
- Oracle gold retrieval model: passages (defined as non-overlapping windows of 5 sentences extracted from each document) are represented as unigram vectors; they are then ranked by comparing the KL-Divergence (Kullback and Leibler, 1951) of the passage vector (interpreted as a word distribution) with the vocabulary distribution in the manual summaries

For each retrieval model, we keep only the top-N sentences before applying the generic KLSum method - so that we obtain variants with the top most-relevant passages containing up to 750, 1,000 … 2,250 words. As a baseline, we also apply KLSum on the whole document set, with no query relevance filtering (thus as a generic summarization method). We report for each configuration the standard ROUGE-2 and ROUGE-SU4 (Lin, 2004) recall metric[1]. Note that these metrics take into account "responsiveness to the query", because they compare the summary generated by the algorithm to human created summaries aimed at answering the query.

In our setting, the retrieval component makes the summary responsive to the query, and the generic summarization component makes the summary non-redundant and focused around the central aspect of the content relevant to the query.

Our hypothesis in this setting is that: if a QFS dataset is not fully saturated by the input topic, the results of the same generic summarization algorithm will improve when the quality of the retrieval component increases. In other words, the ROUGE score of the algorithm will increase when the retrieval improves. In contrast, when the dataset is fully saturated by content that is exclusively relevant to the query, the quality of the retrieval component, and even the level of filtering applied in the retrieval component will not significantly affect the score of the QFS algorithm.

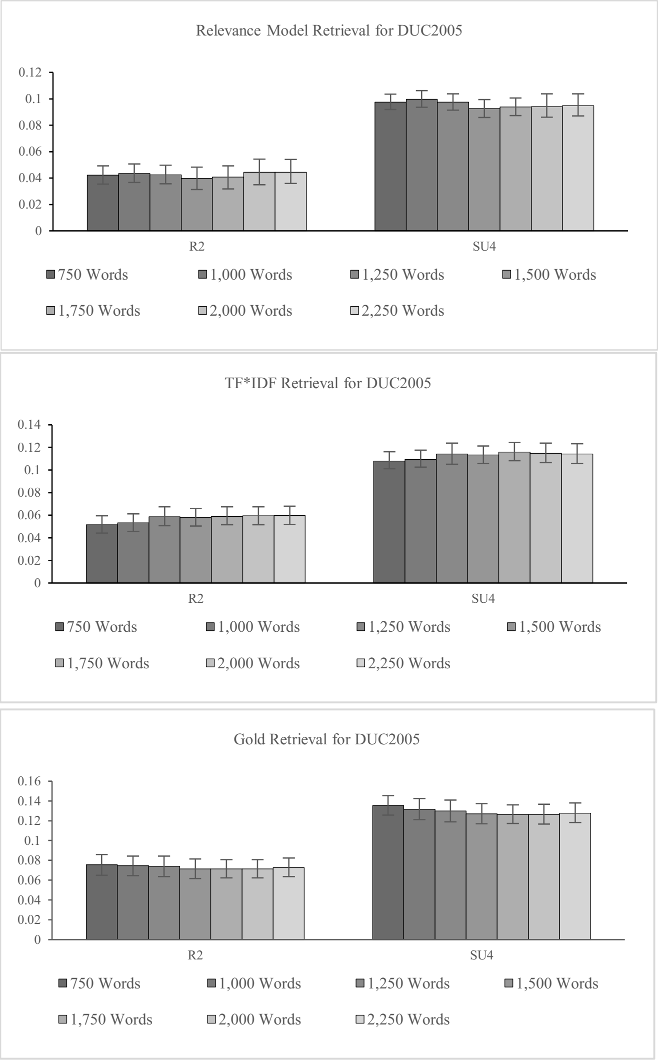

Figure 3. Comparing Retrieval Components on DUC 2005

The results when applied to the DUC-2005 dataset are shown in Fig.3: remarkably, the ROUGE metrics are not significantly different regardless of the level of filtering. The graphs remain flat – generic summarization performs as well on 750 words as on 2,250 words of input (out of about 12,000 total words in each cluster and output summarization length is 250 words).

This experiment shows that the specific DUC 2005 dataset does not exercise the content retrieval component of QFS. The dataset behaves as if all sentences were relevant, and the QFS algorithms must focus their energy on selecting the most central sentences among these relevant sentences. This task-based evaluation indicates that DUC-2005 suffers from excessive topic concentration. We ob-

[1] ROUGE parameters used: -n 4 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a -x

serve exactly the same pattern on DUC 2006 and DUC 2007.

The TD-QFS Dataset

We introduce and make available a new dataset that we call the Topically Diverse QFS (TD-QFS) dataset to try to create a QFS benchmark with less topic concentration. The TD-QFS re-uses queries and document-sets from the Query Chain Focused Summarization (QCFS) (Baumel *et al*., 2014) but adds new manual summaries that are suitable for the traditional QFS task

QCFS defined a variant summarization task combining aspects of update and query-focused summarization. In QCFS, a chain of related queries is submitted on the same document cluster (up to three queries in a chain). A new summary is produced for each query in the chain, that takes into account the current query qi and the previous summaries produced to answer the previous queries in the same chain.

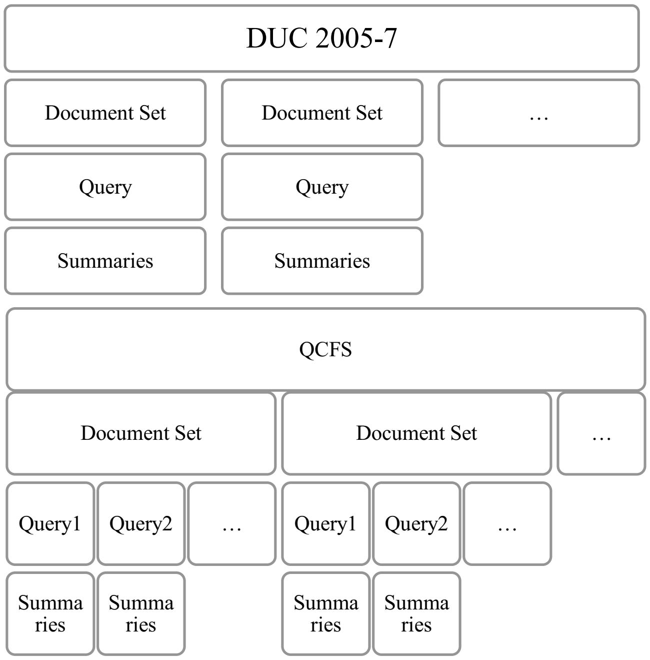

Figure 4. DUC 2005-7 vs. QCFS Dataset Structure

Multiple queries are associated to each document cluster (as seen in Fig.4). All the queries were extracted from PubMed query logs. These query formulations are much shorter than the topic descriptions used in DUC datasets, but the context provided by the chain helps elucidate the information need. To construct the document clusters, medical experts were asked to collect documents from reliable consumer health web-sites relating to the general topic covered by the query chains (Wikipedia, WebMD, and the NHS).

In this paper, we compare the TD-QFS dataset with traditional QFS datasets. We expect that TD-QFS, by construction will be less topic-concentrated than traditional QFS datasets because each document cluster is collected to answer multiple queries.

When constructing the TD-QFS dataset, we first observe that producing a summary for the first query of each chain in QCFS is identical to a QFS task, since there is no prior context involved. To compare different queries on the same document cluster, we asked multiple annotators to generate manual summaries for the second query in each query chain out of context (that is, without reading the first query in the chain). The statistics of the expanded dataset, TD-QFS[2] appear in Table 1.

Document clusters	**# Docs**	**# Sentences**	**#Tokens/ Unique**
Asthma	125	1,924	19,662 / 2,284
Lung-Cancer	135	1,450	17,842 / 2,228
Obesity	289	1,615	21,561 / 2,907
Alzheimer’s Disease	191	1,163	14,813 / 2,508
Queries	**# Queries**		**#Tokens/ Unique**
Asthma	9		21 / 14
Lung-Cancer	11		47 / 23
Obesity	12		36 / 24
Alzheimer’s Disease	8		19 / 18
Manual Summaries	**# Docs**		**#Tokens/ Unique**
Asthma	27		3,415 / 643
Lung-Cancer	33		3,905 / 660
Obesity	36		3,912 / 899
Alzheimer’s Disease	24		2,866 / 680

Table 1. TD-QFS Dataset Statistics.

We first verify that, as hypothesized, the TD-QFS dataset has lower topic concentration than DUC 2005. The document clusters have been constructed so that they contain answers to multiple queries (about 15 short queries for each of the four topics). To confirm this, we measure the KL-Divergence of the unigram distribution of the manual summaries obtained for each query with that of the overall document cluster. While in DUC 2005, this KL-Divergence was 2.3; in the QCFS dataset, we obtain 6.7 - indicating that the manual summaries in TD-QFS exhibit higher diversity.

We then reproduce the task-based experiment de-scribed above on the TD-QFS dataset and compare it to the DUC dataset. The results are now markedly different: Fig.5 reports the ROUGE-recall metrics when performing TF*IDF ranking of the documents, selecting the top N passages (750, 1,000 … 2,250 words) and then applying the generic summarization KLSum method to eliminate redundancy and meet the summary length constraint. As expected, we find that filtering out irrelevant content produces better results: instead of the flat curves observed on DUC da-

[2] TD-QFS is freely available at http://www.cs.bgu.ac.il/~talbau/TD-QFS/dataset.html

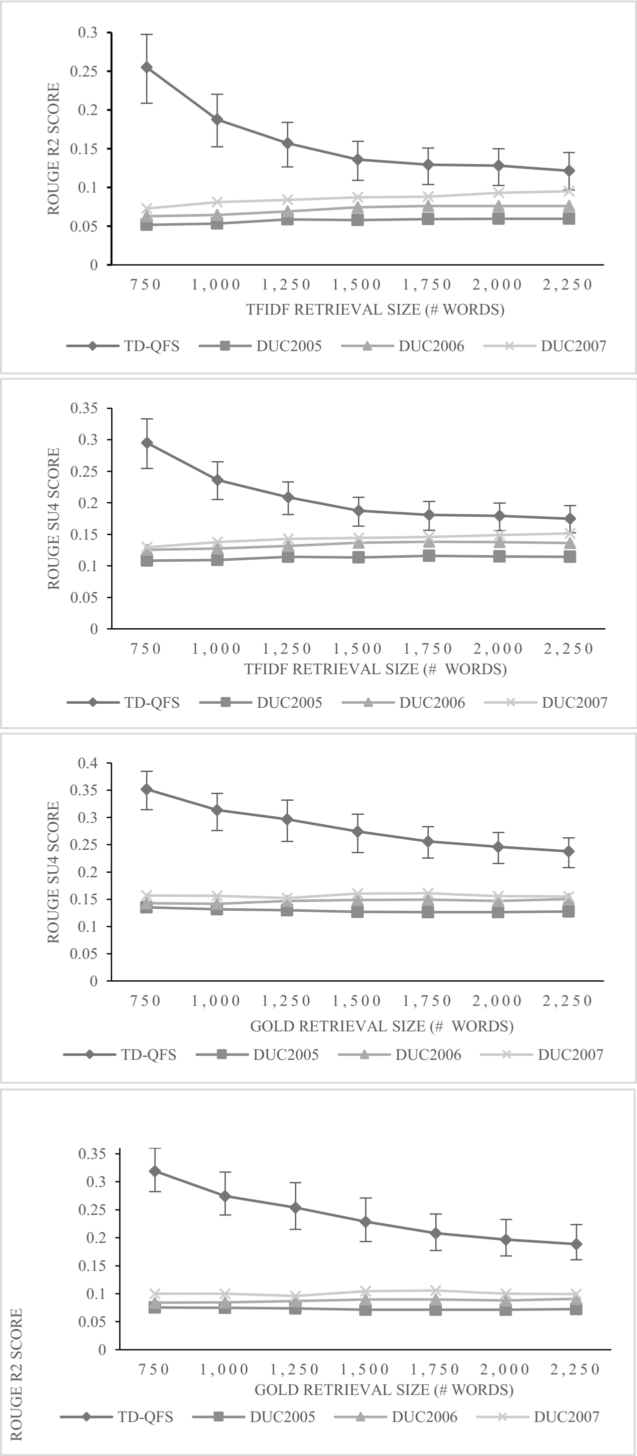

Figure 5. ROUGE-Recall results of KLSum on relevance-filtered subsets of the TD-QFS dataset compared to DUC datasets.

tasets, the quality of the retrieval clearly influences ROUGE results on the TD-QFS dataset, with curves decreasing sharply as less relevant content is added.

We next compare different retrieval models: Fig.6 shows the respective ROUGE results when applying KLSum as a generic summarization method, Biased-LexRank as a state of the art QFS algorithm and the Gold Retrieval model - where the most relevant passages are passed to KLSum up to a number of words limit and relevance is measured as KL-Divergence to the manual summaries. The Gold Retrieval model performance indicates the theoretical higher bound we can achieve by improving the retrieval model.

The results demonstrate the critical importance of the relevance model on ROUGE performance for QFS when the dataset contains sufficient variability: ROUGE-SU4 scores vary from 0.155 to 0.351– while the whole range of scores observed on DUC 2005 was limited to [0.119 – 0.136].

	Original query	Oracle query expansion
Min	1.4%	67.8%
Average	28.5%	83.7%
Max	57.0%	92.0%

Table 2. Topic Concentration as predicted by the Direct Method on the TD-QFS Dataset.

Note that, in contrast to what our task-based evaluation demonstrates, the direct method described above to measure topic-concentration using the binary relevance model of Gupta *et a*l. would have predicted that TD-QFS is also highly concentrated – see Table 2. This could be explained by the fact that Gupta's Oracle Expansion test measures lexical overlap be-tween the manual summary and the document cluster; key terms found in the document cluster are bound to appear in both manual summaries and most sentences from the cluster. For example, it is unlikely that all of these sentences match a given query just because both of them contain the term "asthma".

Relevance-based QFS Models

We introduce three new QFS algorithms that account for query relevance in different ways. Those methods attempt to eliminate the need of determining a specific threshold size that was used in the experiments above. We compare the methods to the three baselines presented above: KLSum as generic summarization, Biased-LexRank, and Gold Retrieval as a theoretical upper bound.

In the *RelSum* method, instead of using N-gram distribution to represent the document set we construct a hierarchical model that increases the probability of words taken from relevant documents. In pure KLSum, the probability of each word in the document cluster is modelled as: $P(w) = \sum_{d\in C} freq(w,d)$⟧ . In contrast, RelSum introduces the document relevance in the formula as: $P(w) = \sum_{d\in C} rel(d) * freq(w,d)$, where $rel(d)$[3]is the normalized relevance score of document *d*.

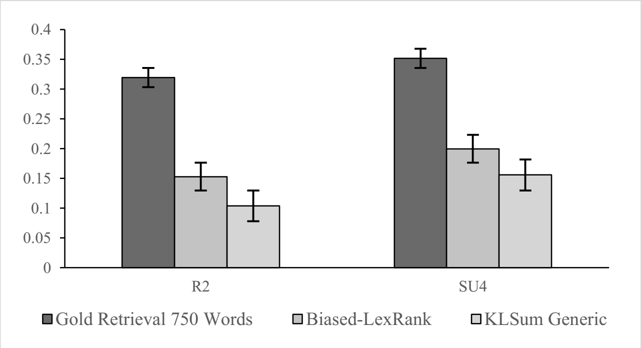

Figure 6. Comprastion of QFS to non-QFS Algorithms Performance on the TD-QFS Dataset

In the *KLThreshold* method, we attempt to identify the threshold that separates relevant from non-relevant documents as part of the retrieval component. We rank documents by their relevance score (measured as the cosine similarity over the TF*IDF vectors of the documents). We then decide the threshold at which document candidates for the summarization are cut: to this end, we compare the KL-divergence of the unigram model of each document to the unigram model of all the documents ranked higher. We repeat this process until the KL-divergence is lower than a fixed threshold (*i.e.*, the next document candidate is too similar to documents ranked higher).

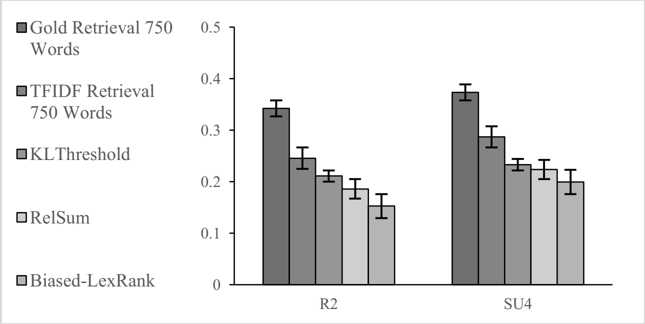

Figure 7. Comprastion of Retrieval Based Algorithms Performance on the TD-QFS Dataset

Finally, we assess the threshold in the list of ranked candidate documents for summarization by learning the average number of documents actually used in the manual summaries. This is a weakly supervised method - which learns the cutoff parameter from the manual document dataset. We find that five documents are used as sources for manual summaries on average. We define the TFIDF-KLSum method as the method that consists of ranking all documents by similarity to the query and passing the top five documents to the KLSum generic summarizer.

We observe (Fig.7) that the TFIDF-KLSum method outperforms RelSum and KLThreshold and closes the gap between Biased-LexRank and the theoretical upper bound represented by the Gold Retrieval method. All three methods based on the $retrieve + generic\_summarize$ meth ods show impressive ROUGE improvements compared to QFS state of the art.

Conclusion

We have investigated the topic concentration level of the DUC datasets for query-focused summarization. We found that the very high topic concentration of those datasets removes the challenge of identifying relevant material from the QFS task. We have introduced the new TD-QFS dataset for the QFS task, and have showed that it has much lower topic concentration through a task-based analysis. The low topic concentration setting allows us to articulate the difference between passage retrieval (a typical Information Retrieval task) and QFS. We discovered that given perfect IR, the gold retrieval model, a standard sum summarization algorithm achieves an order of magnitude improvement in rouge score.

We introduce three algorithms that combine an explicit relevance model to select documents based on the input query, and then apply a generic summarization algorithm on the relevant documents. While these three algorithms significantly outperform state of the art QFS methods on the TD-QFS dataset, the gap with the theoretical upper bound identified by the Gold Retrieval method remains high (from ROUGE 0.25 to 0.34). We make the TD-QFS dataset available to the community. We intend to continue analyzing IR models that can help us further bridge that gap. We also attempt to develop joint models that combine relevance, centrality and redundancy avoidance in a single model.

References

Baron, Jason R., David D. Lewis, and Douglas W. Oard. "TREC 2006 Legal Track Overview." TREC. 2006.

Baumel, Cohen and Elhadad. 2014. “Query-Chain Focused Summarization”. Proceedings of the 52nd Annual Meeting of the

[3] For this paper we tested TF*IDF relevance as $rel(d)$

Association for Computational Linguistics (Volume 1: Long Papers). 913—922.

Dang H.T., 2005. Overview of DUC 2005. National Institute of Standards and Technology (NIST), http://duc.nist.gov/pubs/2005papers/OVERVIEW05.pdf.

Dang H.T., 2006. Overview of DUC 2006. National Institute of Standards and Technology (NIST), http://duc.nist.gov/pubs/2006papers/duc2006.pdf.

Gupta, Nenkova and Jurafsky. 2007. Measuring Importance and Query Relevance in Topic-focused Multi-document Summarization, ACL 2007.

Haghighi and Vanderwende. 2009. Exploring content models for multi-document summarization. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. ACL 2009.

Kullback, S., & Leibler, R. A. 1951. On information and sufficiency. The annals of mathematical statistics, 79-86.

Lavrenko, Victor, and W. Bruce Croft. 2001. Relevance based language models. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001.

Lin, Chin-Yew. 2004. "Rouge: A package for automatic evaluation of summaries." In Text summarization branches out: Proceedings of the ACL-04 workshop, vol. 8.

Otterbacher J., Erkan G., and Radev D., 2009. Biased LexRank: Passage retrieval using random walks with question-based priors. Information Processing & Management 45.1 (2009): 42-54.

Yi X. and Allan J. 2009. A comparative study of utilizing topic models for information retrieval. Advances in Information Retrieval. Springer Berlin Heidelberg, 2009. 29-41.

Yu, C. T., Salton, G., & Yang, C. S. (1973). Contribution to the Theory of Indexing. Cornell University.