Measuring Importance and Query Relevance in Topic-focused Multi-document Summarization

Surabhi Gupta and **Ani Nenkova** and **Dan Jurafsky**
Stanford University
Stanford, CA 94305
surabhi@cs.stanford.edu, {anenkova, jurafsky}@stanford.edu

Abstract

The increasing complexity of summarization systems makes it difficult to analyze exactly which modules make a difference in performance. We carried out a principled comparison between the two most commonly used schemes for assigning importance to words in the context of *query focused multi-document summarization*: raw frequency (word probability) and log-likelihood ratio. We demonstrate that the advantages of log-likelihood ratio come from its known distributional properties which allow for the identification of *a set* of words that in its entirety defines the aboutness of the input. We also find that LLR is more suitable for query-focused summarization since, unlike raw frequency, it is more sensitive to the integration of the information need defined by the user.

1 Introduction

Recently the task of multi-document summarization in response to a complex user query has received considerable attention. In *generic summarization*, the summary is meant to give an overview of the information in the documents. By contrast, when the summary is produced in response to a user query or topic (*query-focused*, *topic-focused*, or generally *focused* summary), the topic/query determines what information is appropriate for inclusion in the summary, making the task potentially more challenging.

In this paper we present an analytical study of two questions regarding aspects of the topic-focused scenario. First, two estimates of importance on words have been used very successfully both in generic and query-focused summarization: *frequency* (Luhn, 1958; Nenkova et al., 2006; Vanderwende et al., 2006) and *loglikelihood ratio* (Lin and Hovy, 2000; Conroy et al., 2006; Lacatusu et al., 2006). While both schemes have proved to be suitable for summarization, with generally better results from log-likelihood ratio, no study has investigated in what respects and by how much they differ. Second, there are many little-understood aspects of the differences between generic and query-focused summarization. For example, we'd like to know if a particular word weighting scheme is more suitable for focused summarization than others. More significantly, previous studies show that generic and focused systems perform very similarly to each other in query-focused summarization (Nenkova, 2005) and it is of interest to find out why.

To address these questions we examine the two weighting schemes: raw frequency (or word probability estimated from the input), and log-likelihood ratio (LLR) and two of its variants. These metrics are used to assign importance to individual content words in the input, as we discuss below.

Word probability $R(w) = \frac{n}{N}$, where n is the number of times the word w appeared in the input and N is the total number of words in the input.

Log-likelihood ratio (LLR) The likelihood ratio λ (Manning and Schutze, 1999) uses a background corpus to estimate the importance of a word and it is proportional to the mutual information between a word w and the input to be summarized; $\lambda(w)$ is defined as the ratio between the probability (under a binomial distribution) of observing w in the input and the background corpus assuming equal probability of occurrence of w in both and the probability of the data assuming different probabilities for w in the input and the background corpus.

LLR with cut-off (LLR(C)) A useful property of the log-likelihood ratio is that the quantity

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 193–196,
Prague, June 2007. ©2007 Association for Computational Linguistics

$-2\log(\lambda)$ is asymptotically well approximated by χ^2 distribution. A word appears in the input significantly more often than in the background corpus when $-2\log(\lambda) > 10$. Such words are called signature terms in Lin and Hovy (2000) who were the first to introduce the log-likelihood weighting scheme for summarization. Each descriptive word is assigned an equal weight and the rest of the words have a weight of zero:

$R(w) = 1$ if $(-2\log(\lambda(w)) > 10)$, 0 otherwise.

This weighting scheme has been adopted in several recent generic and topic-focused summarizers (Conroy et al., 2006; Lacatusu et al., 2006).

LLR(CQ) The above three weighting schemes assign a weight to words regardless of the user query and are most appropriate for generic summarization. When a user query *is* available, it should inform the summarizer to make the summary more focused. In Conroy et al. (2006) such query sensititivity is achieved by augmenting LLR(C) with all content words from the user query, each assigned a weight of 1 equal to the weight of words defined by LLR(C) as topic words from the input to the summarizer.

2 Data

We used the data from the 2005 Document Understanding Conference (DUC) for our experiments. The task is to produce a 250-word summary in response to a topic defined by a user for a total of 50 topics with approximately 25 documents for each marked as relevant by the topic creator. In computing LLR, the remaining 49 topics were used as a background corpus as is often done by DUC participants. A sample topic (d301) shows the complexity of the queries:

Identify and describe types of organized crime that crosses borders or involves more than one country. Name the countries involved. Also identify the perpetrators involved with each type of crime, including both individuals and organizations if possible.

3 The Experiment

In the summarizers we compare here, the various weighting methods we describe above are used to assign importance to individual content words in the input. The weight or importance of a sentence S in the input is defined as

$$Weight_{R(S)} = \sum_{w \in S} R(w) \quad (1)$$

where $R(w)$ assigns a weight for each word w.

For GENERIC summarization, the top scoring sentences in the input are taken to form a generic extractive summary. In the computation of sentence importance, only nouns, verbs, adjectives and adverbs are considered and a short list of light verbs are excluded: "has, was, have, are, will, were, do, been, say, said, says". For FOCUSED summarization, we modify this algorithm merely by running the sentence selection algorithm on only those sentences in the input that are relevent to the user query. In some previous DUC evaluations, relevant sentences are explicitly marked by annotators and given to systems. In our version here, a sentence in the input is considered relevant if it contains at least one word from the user query.

For evaluation we use ROUGE (Lin, 2004) SU4 recall metric[1], which was among the official automatic evaluation metrics for DUC.

	GENERIC	FOCUSED
Frequency	0.11972 (0.11168–0.12735)	**0.11795** (0.11010–0.12521)
LLR	*0.11223* (0.10627–0.11873)	0.11600 (0.10915–0.12281)
LLR(C)	*0.11949* (0.11249–0.12724)	0.12201 (0.11507–0.12950)
LLR(CQ)	*not app*	**0.12546** (.11884–.13247)

Table 1: SU4 ROUGE recall (and 95% confidence intervals) for runs on the entire input (GENERIC) and on relevant sentences (FOCUSED).

4 Results

The results are shown in Table 1. The focused summarizer using LLR(CQ) is the best, and it *significantly* outperforms the focused summarizer based on frequency. Also, LLR (using log-likelihood ratio to assign weights to *all* words) perfroms significantly worse than LLR(C). We can observe some trends even from the results for which there is no significance. Both LLR and LLR(C) are sensitive to the introduction of topic relevance, producing somewhat better summaries in the FOCUSED scenario

[1] -n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -d

compared to the GENERIC scenario. This is not the case for the frequency summarizer, where using only the relevant sentences has a negative impact.

4.1 Focused summarization: do we need query expansion?

In the FOCUSED condition there was little (for LLR weighting) or no (for frequency) improvement over GENERIC. One possible explanation for the lack of clear improvement in the FOCUSED setting is that there are not enough relevant sentences, making it impossible to get stable estimates of word importance. Alternatively, it could be the case that many of the sentences are relevant, so estimates from the relevant portion of the input are about the same as those from the entire input.

To distinguish between these two hypotheses, we conducted an oracle experiment. We modified the FOCUSED condition by expanding the topic words from the user query with *all* content words from any of the human-written summaries for the topic. This increases the number of relevant sentences for each topic. No automatic method for query expansion can be expected to give more accurate results, since the content of the human summaries is a direct indication of what information in the input was important and relevant and, moreover, the ROUGE evaluation metric is based on direct n-gram comparison with these human summaries.

Even under these conditions there was no significant improvement for the summarizers, each getting better by 0.002: the frequency summarizer gets R-SU4 of 0.12048 and the LLR(CQ) summarizer achieves R-SU4 of 0.12717.

These results seem to suggest that considering the content words in the user topic results in enough relevant sentences. Indeed, Table 2 shows the minimum, maximum and average percentage of relevant sentences in the input (containing at least one content words from the user the query), both as defined by the original query and by the oracle query expansion. It is clear from the table that, on average, over half of the input comprises sentences that are relevant to the user topic. Oracle query expansion makes the number of relevant sentences almost equivalent to the input size and it is thus not surprising that the corresponding results for content selection are nearly identical to the query independent runs of generic summaries for the entire input.

	Original query	Oracle query expansion
Min	13%	52%
Average	57%	86%
Max	82%	98%

Table 2: Percentage of relevant sentences (containing words from the user query) in the input. The oracle query expansion considers all content words form human summaries of the input as query words.

These numbers indictate that rather than finding ways for query expansion, it might instead be more important to find techniques for *constraining* the query, determining which parts of the input are directly related to the user questions. Such techniques have been described in the recent multi-strategy approach of Lacatusu et al. (2006) for example, where one of the strategies breaks down the user topic into smaller questions that are answered using robust question-answering techniques.

4.2 Why is log-likelihood ratio better than frequency?

Frequency and log-likelihood ratio weighting for content words produce similar results when applied to rank all words in the input, while the cut-off for topicality in LLR(C) does have a positive impact on content selection. A closer look at the two weighting schemes confirms that when cut-off is not used, similar weighting of content words is produced. The Spearman correlation coefficient between the weights for words assigned by the two schemes is on average 0.64. At the same time, it is likely that the weights of sentences are dominated by only the top most highly weighted words. In order to see to what extent the two schemes identify the same or different words as the most important ones, we computed the overlap between the 250 most highly weighted words according to LLR and frequency. The average overlap across the 50 sets was quite large, 70%.

To illustrate the degree of overlap, we list below are the most highly weighted words according to each weighting scheme for our sample topic concerning crimes across borders.

LLR *drug*, *cocaine*, *traffickers*, *cartel*, *police*, *crime*, *enforcement*, *u.s.*, *smuggling*, *trafficking*, *arrested*, *government*, *seized*, *year*, *drugs*, *organised*, *heroin*, *criminal*, *cartels*, *last*, *official*, *country*, *law*, *border*, kilos, arrest, *more*, *mexican*, laundering, *officials*, *money*, *accounts*, *charges*, *authorities*, corruption, anti-drug, international, banks, operations, seizures, federal, italian, smugglers, dealers, narcotics, criminals, tons, most, planes, customs

Frequency *drug*, *cocaine*, *officials*, *police*, *more*, *last*, *government*, *year*, *cartel*, *traffickers*, *u.s.*, other, *drugs*, *enforcement*, *crime*, *money*, *country*, *arrested*, federal, most, now, *trafficking*, *seized*, *law*, years, new, *charges*, *smuggling*, being, *official*, *organised*, international, former, *authorities*, only, *criminal*, *border*, people, countries, state, world, trade, first, *mexican*, many, *accounts*, according, bank, *heroin*, *cartels*

It becomes clear that the advantage of likelihood ratio as a weighting scheme does not come from major differences in overall weights it assigns to words compared to frequency. It is the significance cut-off for the likelihood ratio that leads to noticeable improvement (see Table 1). When this weighting scheme is augmented by adding a score of 1 for content words that appear in the user topic, the summaries improve even further (LLR(CQ)). Half of the improvement can be attributed to the cut-off (LLR(C)), and the other half to focusing the summary using the information from the user query (LLR(CQ)). The advantage of likelihood ratio comes from its providing a principled criterion for deciding which words are truly descriptive of the input and which are not. Raw frequency provides no such cut-off.

5 Conclusions

In this paper we examined two weighting schemes for estimating word importance that have been successfully used in current systems but have not to-date been directly compared. Our analysis confirmed that log-likelihood ratio leads to better results, but not because it defines a more accurate assignment of importance than raw frequency. Rather, its power comes from the use of a known distribution that makes it possible to determine which words are truly descriptive of the input. Only when such words are viewed as equally important in defining the topic does this weighting scheme show improved performance. Using the significance cut-off and considering all words above it equally important is key.

Log-likelihood ratio summarizer is more sensitive to topicality or relevance and produces summaries that are better when it take the user request into account than when it does not. This is not the case for a summarizer based on frequency.

At the same time it is noteworthy that the generic summarizers perform about as well as their focused counterparts. This may be related to our discovery that on average 57% of the sentences in the document are relevant and that ideal query expansion leads to a situation in which almost all sentences in the input become relevant. These facts could be an unplanned side-effect from the way the test topics were produced: annotators might have been influenced by information in the input to be summarizied when defining their topic. Such observations also suggest that a competitive generic summarizer would be an appropriate baseline for the topic-focused task in future DUCs. In addition, including some irrelavant documents in the input might make the task more challenging and allow more room for advances in query expansion and other summary focusing techniques.

References

J. Conroy, J. Schlesinger, and D. O'Leary. 2006. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the COLING/ACL'06 (Poster Session)*.

F. Lacatusu, A. Hickl, K. Roberts, Y. Shi, J. Bensley, B. Rink, P. Wang, and L. Taylor. 2006. Lcc's gistexter at duc 2006: Multi-strategy multi-document summarization. In *Proceedings of DUC'06*.

C. Lin and E. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING'00*.

C. Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

C. Manning and H. Schutze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

A. Nenkova, L. Vanderwende, and K. McKeown. 2006. A compositional context sensitive multi-document summarizer: Exploring the factors that influence summarization. In *Proceedings of ACM SIGIR'06*.

A. Nenkova. 2005. Automatic text summarization of newswire: lessons learned from the document understanding conference. In *Proceedings of AAAI'05*.

L. Vanderwende, H. Suzuki, and C. Brockett. 2006. Microsoft research at duc 2006: Task-focused summarization with sentence simplification and lexical expansion. In *Proceedings of DUC'06*.