A Term Frequency Distribution Approach for the DUC-2007 Update Task

Lawrence H. Reeve, Hyoil Han

College of Information Science and Technology

Drexel University

Philadelphia, PA

lhr24@drexel.edu, hyoil.han@acm.org

Abstract

We present our system used in the DUC 2007 update task, which is our first entry in any of the DUC evaluations. We make use of ideas within our existing FreqDistSumm text summarizer, which has been shown to perform well in biomedical text summarization. Our system submitted to the DUC Update Task, called FreqDistUpdate, uses a context sensitive approach to scoring sentences based on a frequency distribution model. FreqDistUpdate performed in the middle of all systems in three out of the four evaluations. We believe the frequency distribution method is a promising approach for the update task, and that improvements in implementation and approach will lead to better performance in the future DUC evaluations.

1 Introduction

In the DUC 2007 update task, text summarization systems produce short summaries of newswire articles, assuming a user has read a set of previous, related article texts. The idea is to present new information that the user has not already read from the set of preceding article texts. This task is an appropriate place to test our existing summarizer work, called FreqDistSumm, in a new way and in a new domain. The basic idea behind the FreqDist algorithm on which FreqDistSumm is built is to create a summary which has approximately the same frequency distribution of unit items (i.e., terms or concept) as the source text. In this way, the summary captures the expressions of a text in the same degree they are expressed in the source text. This approach has worked well in our biomedical text summarization work (L. Reeve et al., 2006), (L. H. Reeve, Han, & Brooks, to appear).

For the update task, three summaries were generated for each topic. The summaries were based on three document sets labeled A, B, and C. To generate a summary, the system gathers the sentences from all documents in a document set, determines the frequency distribution of all terms within the document set, and then builds a summary so that the summary term frequency distribution is as close as possible to the current document set's term frequency distribution. To account for information accumulated from a previous summary, the summary term frequency distribution is initialized to the previous summary's term frequency distribution. Sentences from the current document set are then scored based on how well they presented new information (terms) as compared to the previous summary.

Our system performed in the middle of all systems in three out of the four evaluations. In the BE evaluation, we placed 14 out of 22, while in the Responsiveness evaluation we placed 15 out of 22. The Pyramid evaluation assigned FreqDistUpdate an average score of 2.23 out of a possible five. FreqDistUpdate had scores ranging from 1 to 4 for

each of the document set summaries. The ROUGE evaluation was inconsistent with the evaluations, placing us 21 out of 22 systems based on the ROUGE-2 and ROUGE-SU4 scores.

The rest of this paper is organized as follows: Section 2 presents background on the frequency distribution algorithm while Section 3 describes how we adapted the existing biomedical FreqDistSumm summarizer for the DUC update task. Section 4 discusses the official results, and Section 5 concludes.

2 Frequency Distribution

2.1 Term Frequency in Summarization

In this section, we provide some background on the use of term frequency in text summarization. The FreqDistSumm summarizer, described in the next section, uses unit item frequency as its unit of measurement. In our DUC system, the chosen unit item is a term. This is different from our biomedical text summarization system, which uses domain-specific concepts as unit items. Term frequency was first used in extractive text summarization in the late 1950's (Luhn, 1958). A follow-up study of an analysis of several term frequency methods showed high agreement in sentence selection among the methods (Rath, Resnick, & Savage, 1961). Subsequent research using frequency methods focused on the use of frequency as one feature among many for identifying important sentences, such as cue phrases (Pollock & Zamora, 1975) (Edmundson, 1999).

Summarization using larger units of text has also been researched. The LAKE system uses keyphrases for summarization (D'Avanzo, Magnini, & Vallin, 2004). The SUMMARIST system (E. Hovy & Lin, 1999) uses WordNet (Fellbaum, 1998) concept counting not for identifying salient sentences, but for topic interpretation. In topic interpretation, concept frequency counting is used to find a node in the concept hierarchy which sufficiently generalizes more specific concepts (e.g., {pear, apple} → fruit). The SUMMARIST authors cite the lack of domain-specific resources as a serious drawback to this approach.

Most recently, the SumBasic algorithm uses term frequency as part of a context-sensitive approach to identifying important sentences while reducing information redundancy (A. Nenkova & Vanderwende, 2005). The use of frequency as a feature in locating important areas of a text has been proven useful in the literature (Luhn, 1958) (Rath et al., 1961) (Pollock & Zamora, 1975) (Edmundson, 1999). This is most likely due to reiteration, where authors state important information in several different ways, in order to reinforce main points (Sparck Jones, 1999).

2.1 Frequency Distribution Summarizer

Extractive approaches to text summarization usually follow a model of scoring sentences based on a set of features. The highest scoring sentences are then extracted to form a summary. When using frequency as the only feature, unit items are counted and then each sentence is given a score based on the frequency count of each unit item in the sentence. A key problem in generating summaries is reducing redundancy. Each new sentence in the summary should add new information rather than repeating already included information. Using the highest frequency terms will likely result in the same information repeatedly being selected, with the chance that some additional information is included. In the SumBasic (A. Nenkova & Vanderwende, 2005) frequency approach, a probability distribution model is first generated, and as each term is used to select sentences, the term probabilities are reduced so that lower probability terms have a better chance of selecting sentences with new information content. This approach is called context sensitive since the summarizer considers sentences already in the summary before selecting a new sentence to add to the summary. This is also related to the idea of finding Maximal Marginal Relevance (MMR), where marginal relevance is defined as finding relevant sentences which contain minimal similarity to previously selected sentences (Carbonell & Goldstein, 1998).

Our frequency distribution algorithm, FreqDist, uses a context sensitive approach to scoring sentences based on a frequency distribution model rather than a probability distribution model (L. Reeve et al., 2006). The rationale of the frequency distribution approach is that the frequency distribution of terms or concepts in the source text ought to appear in the generated summary as

closely as possible to the source text. That is, the frequency distribution models of the source text and its corresponding summary should be as similar as possible.

It is well known that terms in a text follows a Zipf distribution (Zipf, 1949). In previous work, we showed that across 24 paper abstracts and full-text sources, the distributions can be characterized as Zipfian distributions (L. Reeve et al., 2006). Using the observation that both a version of an ideal summary and its corresponding full-text have the same frequency distribution, the frequency distribution approach was conceived to generate a summary based on the frequency distribution of the unit items (i.e., terms or concepts) within a full-text.

Figure 1 shows an outline of our algorithm ("FreqDist") to generate a summary given the full-text of some source text using a frequency distribution approach (L. Reeve et al., 2006). There are two stages in the algorithm: Initialization and Summary Generation. In the initialization stage, the unit items (terms or concepts) of the source text are counted to form a frequency distribution model of the text, and a pool of sentences from the source text is created, called the sentence pool. A summary frequency distribution model is created from the unit items found in the source text. The summary frequency distribution model frequency counts are initially set to zero to indicate an empty summary. In the Summary Generation stage, new sentences are evaluated and then selected for inclusion in the summary. Identifying the next sentence to be added to the summary is accomplished by finding the sentence which most closely aligns the frequency distribution of the summary generated so far to the frequency distribution of the original source text. A candidate summary is first initialized to the summary generated so far. For each sentence in the sentence pool, the sentence is added to the candidate summary to see how much it contributes to the candidate summary. To determine the sentence's contribution, the candidate summary frequency distribution is compared for similarity to the source text's frequency distribution. The comparison generates a similarity score. This similarity score is assigned to the sentence as the sentence's score. After all sentences from the sentence pool have been evaluated for their contribution to the candidate summary, the highest scoring sentence is added to the summary. The sentence added to the summary is then removed from the sentence pool. The sentence selection process is iterative, and repeats until the desired length of the summary is reached.

3 DUC Update Task System Description

The update summarization task required the generation of three 100-word multi-document summaries for each of ten topics. Within each topic, there are three document clusters labeled A, B, and C. Each document cluster is chronologically ordered and contains approximately ten documents related to a topic. The task is to generate three summaries from the contents of each document set given a topic statement (information need). Summary A summarizes the texts in document cluster A. Summary B summarizes the texts in document cluster B assuming the reader already has the information from the documents in document cluster A. Summary C proceeds the same way, assuming the reader has already read the documents in document clusters B and C. There are approximately 10 topics in the test data, with 25 documents per topic.

Our DUC Update Task system, FreqDistUpdate, incorporates the base FreqDist algorithm as well as some scoring heuristics adapted for the DUC Update task. FreqDistUpdate starts by first constructing a list of important words from the topic statement provided by DUC. The topic statement words are generated using a simple method of first replacing a known set of delimiters, defined as {(,), ;, :}, with spaces. The topic sentence is then split into words based on a space character as the delimiter. Semantically unimportant words, such as 'a' and 'the', were removed from the list. The words remaining in the important word list boost the scores of these words if they are found in the texts within a document cluster.

Initialize:

- Determine frequency distribution model of the source text
- Create a summary frequency distribution model, using same term set as source text model, but with frequency values set to zero.

Generate Summary:

- Iteratively select a sentence from the source text and add it to the summary
- Compare summary frequency distribution model with the source text frequency distribution model
- Add the source text sentence which results in a summary best matching the source text frequency distribution model
- Remove the selected sentence from the source text sentence pool
- Repeat the generation process until the desired summary size is reached.

Figure 1: Base FreqDist summarization algorithm ((L. Reeve et al., 2006))

For document Clusters A, B, and C, FreqDistUpdate reads into memory all documents within each cluster and parses them into sentences using the LingPipe sentence chunker (Carpenter & Baldwin, 2007). The sentence chunker is initialized to use the Indo-European sentence model, which is provided as part of the LingPipe toolkit. The sentences from all documents in the cluster are combined to form a single list representing all sentences within the cluster. The result of the reading and parsing is three lists of sentences, one for each cluster.

For Cluster A, the FreqDistUpdate summarizer is then passed the list of sentences and the list of important words. The first step in the FreqDistUpdate summarizer is to initialize all of the sentences with a score of zero. A hash table containing all words in the sentence list and their frequency counts is generated. FreqDistUpdate incorporates several modifications to the base FreqDist algorithm shown in Figure 1. These modifications account for important words from the topic statement and also the 100-word maximum summary length requirement. Important words within each sentence are counted. If a sentence does not contain one or more important words, it is penalized so that its chance of being selected is very low. The idea is to select sentences which have words in common with the topic statement. For summary length, a sentence is not selected unless its length plus the length of the summary generated so far is less than 100 words. The result is that a lower-scoring sentence will be selected if a higher-scoring sentence causes the summary length to exceed 100 words. Once all sentences are selected, they are sorted into their original order of appearance and a summary is generated.

In Cluster B, the same basic approach is applied, but with Cluster A's summary sentences passed as a parameter to the FreqDistUpdate summarizer, in addition to the set of Cluster B sentences and important topic words. The words from Cluster A's sentences are used to prime the frequency distribution of the summary to be generated for Cluster B. The idea is to account for frequencies of words already seen and selected in Cluster A so that the likelihood of words from the Cluster A summary being selected again in the Cluster B summary will decrease.

Finally the summary for Cluster C is generated identically to the summary for Cluster B, except for using sentences from Cluster C as the source text input. There was a mistake made in the summary generation for Cluster C which we realized after submission. In Cluster C's summary generation, Cluster A and Cluster B summary sentences are passed to the summarizer. Instead, Cluster A was passed in. The result is that FreqDistUpdate summary for Cluster C considered only the summary generated for Cluster A, instead of considering the summaries generated for Clusters A and B.

Results

NIST provided four different evaluations of each of the 22 systems submitted: ROUGE, Basic Elements (BE), Pyramid, and Responsiveness. The ROUGE, Basic Element, and Pyramid evaluations use increasingly larger units of text to measure overlap between a system-generated summary and a set of model summaries. ROUGE (Lin, 2005) measures the n-gram overlap between a system summary and the model summaries. Basic Elements moves beyond simple n-gram matching to find minimal semantic units, which are defined to be heads of syntactic units, such as noun phrases, and also relationship triples (E. Hovy, Lin, & Zhou, 2005), (E. Hovy, Lin, Zhou, & Fukumoto, 2006). Pyramid uses a set of manually annotated semantic units derived from the system summary and the model summaries (A. Nenkova & Passonneau, 2004). The Responsiveness score is a human assessment on a scale of 1 (low) to 5 (high) of how much information the summary provides in order to address the information need defined in the topic statement.

Model summaries were written using ten different human summarizers. There were four update summaries written for each document cluster. Two baseline summarizers were also provided by NIST. The Baseline 1 summarizer returns the first 100 words from the most recent document in a cluster. The Baseline 2 summarizer is an HMM-based summarizer which performed well in DUC-2004.

The official results placed FreqDistUpdate in the middle of all systems in three out of the four evaluations. FreqDistUpdate placed 14 out of 22 in the BE evaluation, ahead of the Baseline 1 Summarizer but below the Baseline 2 Summarizer. The Pyramid evaluation assigned FreqDistUpdate an average score of 2.23 out of a possible five. FreqDist update had scores ranging from 1 to 4 for each of the document set summaries. The Baseline 1 Summarizer produced an average Pyramid score of 1.69, while the Baseline 2 Summarizer had an average Pyramid score of 2.70. In the Responsiveness evaluation FreqDistUpdate placed 15 out of 22. The ROUGE scores placed FreqDist at 21 out of 22 systems. We found it interesting that the ROUGE scores placed us much lower than both BE and Responsiveness evaluations. It appears the ROUGE scores reflect that FreqDistUpdate did not select the same terms as the model summaries, while the BE evaluation, which focus more on semantic than syntactic units, and the Responsiveness evaluation, which is performed by humans, reflected that we did select information which was considered important to the human assessors. We found this encouraging, and credit the use of different evaluation approaches for providing insight into different aspects of a summarization system.

There are several areas where we think the FreqDistUpdate system can be improved. 1) The use important words from the topic statement is a heuristic. We did not empirically evaluate whether its inclusion is helpful or harmful. Also, it may be that our penalization scheme for sentences not including important words is too harsh. 2) We need to evaluate whether or not for Document Clusters A and B it is more appropriate to include prior summaries when selecting content, or the entire prior source text. FreqDistUpdate currently uses the summary to see what information has already been provided to a reader, but it may be more valuable to consider the entire text when selecting new information for the reader. 3) The documents in each cluster were treated as one large source text, but it may be more valuable to generate update summaries of each document in the cluster, and then generate a final cluster summary from the individual document summaries in the cluster. 4) We had a bug which for generating the update summary for Cluster C which considered only Cluster A's information content, when it really should have considered information content from Clusters A and B.

4 Conclusion

The 2007 DUC was our first chance to participate in a DUC event. We adapted our existing frequency-distribution text summarizer, which had good results in the biomedical domain, to the DUC Update Task. The summarizer creates a summary of a source text which has approximately the same frequency distribution of terms as the source text. We made several modifications to the base summarizer for the DUC Update Task to account for a statement expressing an information need, as well as summary length

limitations. While we did not do as well as we had hoped, we have gathered some ideas to improve our performance.

We believe the DUC Update Task, while a pilot, represents an important real-world task, and we look forward to contributing to it in future years.

References

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval,* Melbourne, Australia. 335-336.

Carpenter, R. L., & Baldwin, F. B. (2007). *LingPipe - A suite of Java libraries for the linguistic analysis of human language.* Retrieved January 26, 2007, from http://www.alias-i.com

D'Avanzo, E., Magnini, B., & Vallin, A. (2004). Keyphrase Extraction for Summarization Purposes: The LAKE System at DUC-2004. *Proceedings of the 2004 Document Understanding Conference,* Boston, USA.

Edmundson, H. P. (1999). New Methods in Automatic Extracting. In I. Mani, & M. T. Maybury (Eds.), (pp. 23-42). Cambridge, MA: MIT Press.

Fellbaum, C. (1998). *WORDNET: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Hovy, E., & Lin, C. (1999). Automated Text Summarization in SUMMARIST. In I. Mani, & M. T. Maybury (Eds.), *Advances in Automatic Text Summarization* (pp. 81-94). Cambridge, MA: MIT Press.

Hovy, E., Lin, C., Zhou, L., & Fukumoto, J. (2006). Automated Summarization Evaluation with Basic Elements. *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006),* Genoa, Italy.

Hovy, E., Lin, C., & Zhou, L. (2005). Evaluating DUC 2005 using Basic Elements. *Proceedings of Document Understanding Conference (DUC-2005),* Vancouver, B.C. Canada.

Lin, C. (2005). *Recall-Oriented Understudy for Gisting Evaluation (ROUGE).* Retrieved August 20, 2005, from http://www.isi.edu/~cyl/ROUGE/

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development, 2*(2), 159-165.

Nenkova, A., & Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. Paper presented at the *Proceedings of HLT/NAACL 2004*, Boston, MA.

Nenkova, A., & Vanderwende, L. (2005). *The Impact of Frequency on Summarization* No. MSR-TR-2005-101). Redmond, Washington: Microsoft Research.

Pollock, J. J., & Zamora, A. (1975). Automatic Abstracting Research at Chemical Abstracts Service. *Journal of Chemical Information and Computer Sciences, 15*(4), 226-232.

Rath, G. J., Resnick, A., & Savage, R. (1961). The Formation of Abstracts by the Selection of Sentences.[Electronic version]. *American Documentation, 2*(12), 139-208.

Reeve, L. H., Han, H., & Brooks, A. D. The Use of Domain-Specific Concepts in Biomedical Text Summarization. To appear in *Journal of Information Processing and Management, Special Issue on Summarization.* Elsevier. 2007.

Reeve, L., Han, H., Nagori, S. V., Yang, J., Schwimmer, T., & Brooks, A. D. (2006).

Concept Frequency Distribution in Biomedical Text Summarization. *Proceedings of the ACM Fifteenth Conference on Information and Knowledge Management (CIKM'06)*, Arlington, VA, USA. 604-611.

Sparck Jones, K. (1999). Automatic Summarizing: Factors and Directions. In I. Mani, & M. T. Maybury (Eds.), *Advances in Automatic Text Summarization* (pp. 2-12). Cambridge, MA: MIT Press.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.