University of Lethbridge's Participation in DUC-2007 Main Task

Yllias Chali Shafiq R. Joty
Department of Computer Science
University of Lethbridge
4401 University Drive
Lethbridge, Alberta, Canada, T1K 3M4
E-mail: {chali, jotys}@cs.uleth.ca

Abstract

This paper presents the summarization technique implemented by the University of Lethbridge summarizer in order to generate summaries of maximum 250 words from multiple documents. We describe our system for query-focused summarization based on an enhanced, feature-based framework.

1. Introduction

Document Understanding Conference(s), DUC, organized by NIST[1], provide a framework to evaluate the system-generated summaries. Our system participated in the main task that models the real-world application:

"***Given a topic statement and a set of 25 relevant documents, the task is to synthesize a fluent, well-organized 250-word summary of the documents that answers the question(s) in the topic statement.***"

This paper describes a query-focused multi-document summarizer based on two distinct but complementary concepts: a) how much the sentence is related to the user query and b) how much the sentence is salient to the overall concept. Keeping these in focus we consider 6 important features: (1)Cosine Similarity (2)Lexical chain (3)BE overlaps (4)Question Focus overlap (5) Previous Sentence overlaps and (6)Document overlap. We consider Cosine Similarity measure, for computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences (Erkan and Radev, 2004). Lexical chains efficiently identify the theme of the document. An additional argument for the chain representation to consider as opposed to a simple word frequency model is the case when a single concept is represented by a number of words, each with relatively low frequency. Because the chain combines the number of occurrences of all its members, it can overcome the weight of the single word (Chali and Kolla, 2004). With BE[2] represented as a head-modifier-relation triple, one can quite easily decide whether any two units match (express the same meaning) or not–considerably more easily than with longer units (Hovy et al., 2005). We consider Question Focus Overlap feature to extract the sentences, which are relevant to the topic and narration. We consider the other two features: Previous Sentence overlaps and Document overlap in order to increase the coherence among the sentences in the summary.

---

[1] National Institute of Standards and Technology

[2] Basic Element

2. System Overview

The global architecture of our system is shown in Figure 1. Each of the modules of the system is described below.

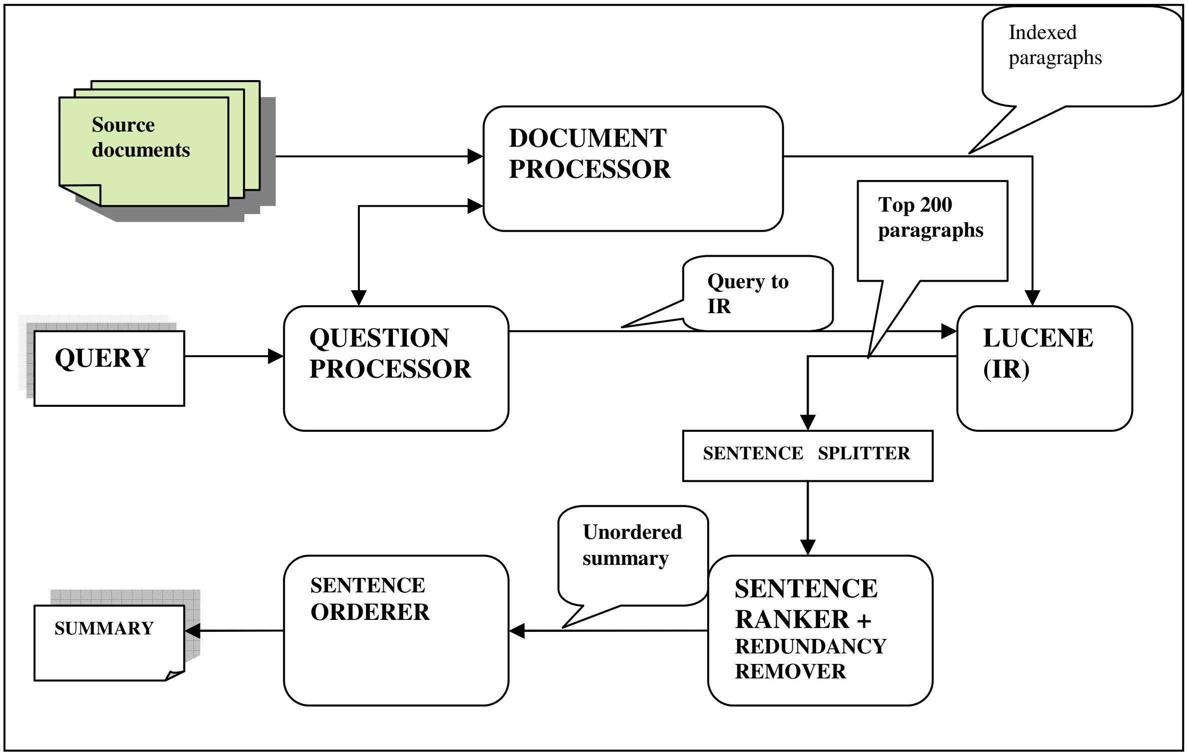

Figure 1: Overview of UofL summarizer

2.1 Document Processor

Document-processing involves preprocessing the documents using several tools. We have used the following tools:

a) Extracting the main text: This module extracts the main text from the source xml document removing the unnecessary tags and makes the documents ready to feed to the co-reference resolution module.

b) Co-reference Resolution: Our system uses the java based tool Lingpipe[3] for co-reference resolution. For example: "Vladimir Putin is the president of Russia. He decided to boost the pensions by 20 percent". "He" refers to "Vladimir Putin". After resolving the co-reference our system generates: "Vladimir Putin is the president of Russia. Vladimir Putin decided to boost the pensions by 20 percent".

c) Retrieve Paragraphs: This module is responsible to extract the paragraphs from the main text and make the paragraphs ready to be indexed by the Information Retrieval System Lucene[4].

[3] http://www.alias-i.com/lingpipe/
[4] http://lucene.apache.org/

d) Sentence Splitting, Text Stemming and Chunking: This module splits the documents into sentences, then stems out the words and chunks the stemmed words**.** We used OAK systems[5] (Sekine, 2002) for this purpose.

e) Lexical chains: Our system computes the lexical chains (Morris and Hirst, 1991) as an intermediate representation for each document. This intermediate representation is useful in order to rank the sentences during the extraction stage.

f) BE Extraction: At the most basic level, Basic Elements (Hovy et al., 2005) are defined as follows:

- ***The head of a major syntactic constituent (noun, verb, adjective or adverbial phrases), expressed as a single item, or***
- ***A relation between a head-BE and a single dependent, expressed as a triple (head | modifier | relation).***

The BEs (BE-Fs) are generated by BE package 1.0 distributed by ISI[6] . We used the standard BE-F breaker included in the BE package.

2.2 Question Processor

Our Question Processing System is shown in figure 2. The brief description of the modules follows:

a) Title and Narrative Extraction: From the topic description provided in the data this module extracts the title and narrative.

b) Focus and Important Words Extraction: Focus is the list of nouns in the title and Important Words is the list of nouns in narrative.

c) WordNet Synonym Adding: To get the related terms of the topic terms (nouns) as well as the narrative terms we add the first synonym words of those terms in the WordNet.

d) Question Expansion using Topic Signature: Expansion of question is a statistical method that aims to extend the topic identification to the concept level while not relying on the knowledge bases or complex semantic parsers (Biryukov et al., 2005). For each term in the focus and its synonym we find the related terms. Instead of learning the topically related terms from a knowledge base, one can create them from the texts on a given topic using statistical methods like: tf.idf weighting scheme, χ^2 (chi-square), and likelihood ratio for binomial distribution tests. We used the likelihood ratio.

e) Question Expansion using Lexical Chain: In the attempt to take the content analysis beyond the word level, discourse analysis has been suggested. It often relies on the lexical cues or

[5] http://nlp.cs.nyu.edu/oak/

[6] http://haydn.isi.edu

exploits the lexical knowledge bases such as WordNet in order to detect the cohesion of a text as signaled by lexical items, which allows identifying the most salient portions of a text.

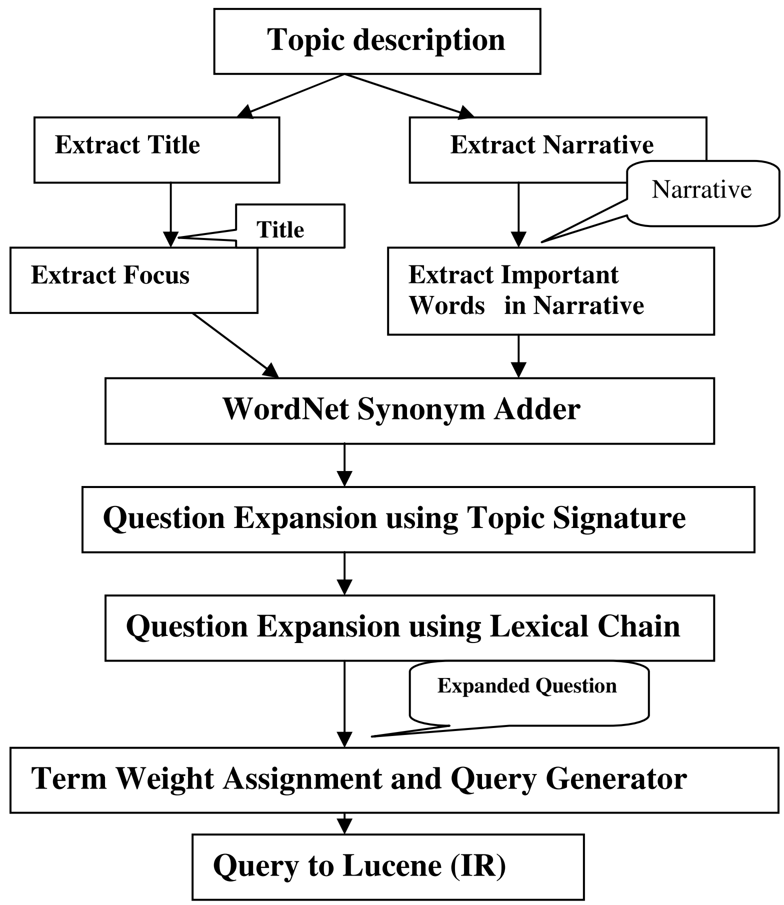

Figure 2: Question Processor

The main gain of the discourse based methods is their ability to capture the dependency relations between the text segments and spot the concepts rather than its word components. Our system finds the top five strong lexical chains then add the words of those lexical chains with the topic terms.

d) Term Weight Assignment and Query Generation: For each word in the question focus and narrative we assign the weight 1. For each term of their WordNet synonyms we assign 0.7 and for each term of their Topic Signature and Lexical Chain expansion we assign 0.5 weights. After getting the terms from the topic, narrative and question expansion we form the query to select the relevant paragraphs from the documents with the help of Lucene.

2.3 Paragraph Extraction from Information Retrieval system (Lucene)

Our system uses java based Lucene as the IR system. We have customized Lucene so that it can index and search at the paragraph level. Lucene ranks the relevant paragraphs. The system takes top 200 paragraphs to be processed further.

2.4 Sentence Splitter

The retrieved paragraphs are split into sentences using the sentence splitter and the sentences are feed to the Sentence Ranking Module.

2.5 Sentence Ranker

In order to rank a sentence the system considers two distinct but complementary concepts: a) how much the sentence is related to the user query and b) how much the sentence is salient to the overall concept. Keeping these in focus we consider six features to rank a sentence:

a) BE overlap
b) Question focus overlap
c) Cosine similarity
d) Lexical Chain
e) Previous Sentence Overlap
f) Document Overlap

a) BE overlap feature

With BE represented as a head-modifier-relation triple, one can quite easily decide whether any two units match (express the same meaning) or not–considerably more easily than with longer units (Hovy et al., 2005). The system scores the BEs and identifies the important BEs.

Identify Important Bes and Rank Sentences: BEs can be used as counting unit. We will compute likelihood ratio (LR) for each BE. The LR score of each BE is an information theoretic measure (Dunning, 1993; Lin and Hovy, 2000) that represents the relative importance in the BE list from the document set that contains all the texts to be summarized. We used the sentence ranking method describe in (Hovy et al., 2005) to find important sentences.

b) Question focus overlap feature

We used the following formula to compute the sentence score, which indicated the relevance of the sentence to the question focus:

$$\boxed{S(s) = \frac{\sum_{t_i \in s} w_{t_i}}{N(s)}}$$

Where s means the sentence to score, S(s) is the score of s, t_i means a topic term in the topic signature, and w_{ti} means weight of the term (i.e. weight that was associated to the term when the topic signature was created) which is present in the sentence. The normalization factor N(s) means the number of words in the sentence.

c) Cosine Similarity Measure

Our system follows the same ranking method described in Lex-Rank (Erkan and Radev, 2004) for ranking the sentences.

d) Lexical Chain feature

Lexical chains efficiently identify the theme of the document. An additional argument for the chain representation to consider as opposed to a simple word frequency model is the case when a single concept is represented by a number of words, each with relatively low frequency (Chali and Kolla 2004).

Scoring Chains: In order to use lexical chains as outlined above, one must first identify the strongest chains among all those that are produced by our algorithm. The strength of the chains is computed according to (Barzilay and Elhadad, 1997) then the sentences are ranked based on these strong chains.

e) Previous Sentence Overlap feature

For a sentence s, we measure the overlap feature by tfidf-weighted word-stem vector cosine distance between s and p where p is the last sentence of the summary under construction.

f) Document Overlap feature

Score is non-zero iff s and p are from the same input document, and p preceded s in original ordering; score is inversely proportional to number of sentences between p and s. where p is the last sentence of the summary under construction.

Assigning the weights of the features

Our system assigns the following weights to the features:

Features	Weights
BE overlap	0.7
Question focus overlap	0.4
Cosine similarity	0.7
Lexical Chain	0.6
Previous Sentence Overlap	0.4
Document Overlap	0.4

2.6 Removing Redundancy

a) Redundancy removing by date: We consider this technique with the assumption that similar dates often describe the same event. But considering only date we do not remove a sentence we will just raise its deletion score by 0.3 and the sentence is deleted only if it satisfies the score to be deleted (BE overlap score + date overlap score > 0.7).

b) Redundancy removing by BE overlap: We consider BE overlap between an intermediate summary and a to-be-added candidate summary sentence. We call this overlap ratio R, where R is between 0 and 1 inclusively. For example, R = 0.8 means that a candidate summary sentence, *s*, can be added to an intermediate summary if the sentence has a BE overlap ratio (+ date overlap score) less than or equal to 0.7.

2.7 Ordering Summary

Our system uses the algorithm that is based on the directed backward graph where the edges of the graph are oriented from a sentence to previous sentences in the text (Mihalcea, 2004).

3. Evaluation

In DUC2007 manual evaluation our system scored 1.933 in average content responsiveness. In automatic evaluation our system scored 0.08453 in rouge2, 0.14100 in rougeSU4 and 0.04252 in BE.

4. Future Work

We developed an improved framework for using all the six features to do the task of Summarization. We have the plan to train the system separately on DUC07 (main task), DUC06, DUC05 and DUC04 topics using in each case a randomly selected 80 percent of the data for training, and 20 percent for testing. Our intuition for examining the data sets separately was that these sets had somewhat different properties; DUC04 summaries used biographically focused topics and 100-word models, whereas DUC05 summaries (like those of DUC06 and DUC07) are on broader topics with 250-word models.

5. Conclusion

In this paper, we presented briefly our query based summarization technique to generate summaries of multiple documents in context of DUC 2007. We used several features in order to rank the sentences in the document collection. In one hand, BEs, Cosine Similarity and Lexical Chain features are intended to rank a sentence with respect to its importance in the document collection, on the other hand, Question Focus Overlap, BE overlap with Query are intended to rank the sentences based on its relevancy to the query. Other two features: Previous Sentence and Document Overlap features are intended to increase the coherence among the sentences in the summary.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) research grant and the University of Lethbridge.

References

1. Barzilay, R. and Elhadad, M. (1997) Using Lexical Chains for Text Summarization. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th European Chapter Meeting of the Association for Computational Linguistics*, Workshop on Intelligent Scalable Test Summarization, pages 10-17, Madrid.

2. Blair-Goldensohn, S. and McKeown, K. (2006) Integrating Rhetorical-Semantic Relation Models for Query-Focused Summarization. In *Proceedings of 6th Document Understanding Conference (DUC2006).* June.

3. Chali, Y. and Kolla, M. (2004). Summarization techniques at DUC 2004. In *Proceedings of the Document Understanding Conference,* pages 105 -111, Boston. NIST.

4. Chieu, H. L. and Lee, Y. K. (2004). Query Based Event Extraction along a Timeline. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004).*

5. Erkan, G. and Radev, D. (2004) "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization", *Journal of Artificial Intelligence Research (JAIR),* 22:457-479.

6. Galley, M. and McKeown, K. (2003) Improving Word Sense Disambiguation in Lexical Chaining. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, pages 1486-1488, Acapulco, Mexico.

7. Hovy, E., Lin C.-Y. and Zhou L. (2005). A BE-based Multi-document Summarizer with Sentence Compression. In *Proceedings of the Multilingual Summarization Evaluation Workshop at the ACL 2005 conference*. Ann Arbor, MI.

8. Maria, B., Angheluta, R. and Moens, M. (2005). Multidocument question answering text summarization using topic signatures, *5th Dutch-Belgian Information Retrieval Workshop (DIR'05)*, appeared in a special issue of Journal of Digital Information Management, March.

9. Mihalcea, R. (2004) Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization, in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, *companion volume (ACL 2004)*, Barcelona, Spain, July.

10. Morris, J. and Hirst, G. (1991), Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text .*Computational Linguistics*, 17(1):21-48.

11. Saggion, H., Bontcheva, K. and Cunningham, H. (2003) Robust Generic and Query-based Summarization.. *10th Conference of the European Chapter of the Association for Computational Linguistics. EACL-2003*. Hungary, April 12-17.

12. Saggion, H., and Gaizauskas, R. (2004) Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the Document Understanding Conference,* pages 105 -111, Boston. NIST.

13. Silber, H. G. and McCoy, K. F. (2002). Efficiently Computed Lexical Chains As an Intermediate Representation for Automatic Text Summarization. *Computational Linguistics,* 28(4):487-496.