UofL at TAC 2011 Guided Summarization Task

Yllias Chali, Sadid A. Hasan, Kaisar Imam, and Siddharth Subramanian
University of Lethbridge
4401 University Drive West
Lethbridge, AB, Canada T1K 3M4

Abstract

In this paper, we describe our guided summarization system that participated in the TAC 2011 competition. We submitted two runs for the guided summarization task by following a random walk paradigm. Two different approaches were applied for the *update component* to create two runs of our guided summarization system: 1) using ROUGE (Recall-Oriented Understudy for Gisting Evaluation), and 2) using Latent Semantic Analysis (LSA). Evaluation results are shown to compare the performance of the two runs.

1 Introduction

As the size of the world-wide-web has vastly increased, the demand for access to different types of information have led researchers to a renewed interest in a broad range of Information Retrieval (IR) related areas. In recent years, a great amount of attention has grown in Multi-Document Summarization (MDS) communities to deal with the topic-focused summarization research and it has been one of the main tasks in recent Text Analysis Conferences (TAC[1]). The main goal of the TAC summarization track is to promote research on systems that produce summaries of documents. The systems have to produce well-organized, fluent, query-focused summaries of text.

Two major problems in automatic text summarization are: 1) the absence of a single "gold standard" that automatic systems can model, and 2) the use of solely extractive methods that ignore contextual information while selecting the candidate sentences to form a summary. In TAC 2010, a new direction in focused summarization research was presented with a novel task termed– *guided summarization*[2] to address these issues. TAC 2011 continued to focus on this task[3]. The objective of guided summarization is to encourage a deeper linguistic (semantic) analysis of the source documents instead of relying only on document word frequencies to select important concepts. The guided summarization task is to write a 100-word straightforward query-focused summary of a set of 10 newswire articles for a given topic, where the topic falls into a predefined template-like category. Participants are given a list of aspects for each category, and a summary must include all aspects found for its category.

In addition to this, an "update component" of the guided summarization task is proposed to write a 100-word "update" summary of a subsequent 10 newswire articles for the topic, under the assumption that the user has already read the earlier articles. Users looking for information about a series of related events, often face an intimidating task of filtering out redundant information. To help combating this problem, update summarization task is piloted in DUC[4] 2007 with the hope to deliver focused distilled information to a user who has already read a set of older documents covering the same topic. Update summarization is similar to query-focused summarization in that the system is presented with a topic statement (consisting of one or more questions) and a cluster of on-topic documents; however,

[1] http://www.nist.gov/tac/

[2] http://www.nist.gov/tac/2010/Summarization/GuidedSumm.2010.guidelines.html

[3] http://www.nist.gov/tac/2011/Summarization/GuidedSumm.2011.guidelines.html

[4] http://duc.nist.gov/duc2007/tasks.html#pilot

in this scenario, it is assumed that the user is already familiar with some aspects of the topic (represented by a set of earlier documents). In TAC 2011, update summaries were judged against information extracted from both initial and update model summaries that eventually identified relevance and redundancy at the same time.

Our guided summarization framework operates on a Markov chain model and follows a random walk paradigm in order to generate possible summary sentences. Two different approaches are applied for the *update component*: 1) using ROUGE (Recall-Oriented Understudy for Gisting Evaluation), and 2) using Latent Semantic Analysis (LSA). Rest of the paper is organized as follows: Section 2 gives a detailed description of the approaches used. Section 3 presents the evaluation while Section 4 concludes the paper.

2 Our Approach

2.1 Query-focused Summarization

To generate the query-focused summaries, we exploit the predefined list of important aspects to find the most relevant sentences from the document collection. For each question (i.e. aspect) of a topic, we perform keyword expansion using WordNet[5] (Fellbaum, 1998). For example, the word "happen" being a keyword in the given *aspect*: "What happened?" returns the words: *occur, pass, fall out, come about, take place* from WordNet. On the other hand, for each document sentence in the collection we perform Named Entity (NE) tagging using the *OAK* system (Sekine, 2002). Named Entities (NE) are defined as terms that refer to a certain entity. For instance, *USA* refers to a certain *country*, and *$200* refers to a certain quantity of money. Each sentence is weighted based on the following two criteria:

1. Similarity of each sentence with the expanded aspect (in terms of word matching), and

2. weight assigned to each sentence by the NE tagging procedure[6].

[5]For simplicity, we consider the synsets up to level 1 in this research.

[6]For example, for an aspect like *"When did the accident happened?"*, we search for $< Time >$ tag in the NE tagged sentences and give them higher weights if found.

Then, inspired from (Harabagiu et al., 2006), we select the most relevant sentences by following a random walk on a graph where each node is a document sentence and the edges represent similarity between sentences. The whole procedure operates on a Markov chain (MC) (Lafferty and Zhai, 2001). A Markov chain is a process that consists of a finite number of states and some known probabilities p_{ij}, where p_{ij} is the probability of moving from state j to state i. For each node (i.e. sentence) and each edge in the graph, we calculate *"node weight"* and *"edge weight"*, respectively. Once we find all the node weights and edge weights, we perform a random walk on the graph following a Markov chain model in order to select the most important sentences. Node (sentence) weights are calculated using the following formula:

$$InitialSentence = \arg\max_{i=1}^{N} (weight\,(S_i)) \quad (1)$$

where N is the total number of nodes in the graph. After finding the initial best sentence, in each step of the random walk we calculate the probability (transition probability) of choosing the next relevant sentence based on the following equation:

$$P(S_j|S_i) = \frac{1}{\alpha} \arg\max_{j=1}^{Z} (weight\,(S_j) * similarity\,(S_i, S_j)) \quad (2)$$

where S_i is the sentence chosen early, S_j is the next sentence to be chosen, Z is the set of sentence indexes that does not contain i, the $similarity(S_i, S_j)$ function returns a similarity score between the already selected sentence and a new sentence under consideration, and α is the normalization factor that is determined as follows:

$$\alpha = \sum_{j=1}^{Z} (weight\,(S_j) * similarity\,(S_i, S_j)) \quad (3)$$

We associate each node (sentence) in the graph a weight that indicates the importance of the node with respect to the document collection. Node weights are calculated based on a Topic Signature (TS) model (Lin and Hovy, 2000), and then combined with the weights obtained from the list of aspects' information (described above). We normalize it to get the final weights of the sentences/nodes.

Inspired by the idea presented in (Lin and Hovy, 2000), for each topic present in the data set, we calculate its topic signature defined as below:

$$\begin{aligned} TS &= \{topic, signature\} \\ &= \{topic, \langle (t_1, w_1), \cdots, (t_n, w_n) \rangle\} \quad (4) \end{aligned}$$

where $topic$ is the target concept and signature is a vector of related terms. Each t_i is a term highly correlated to the $topic$ with association weight, w_i. We use the following log-likelihood ratio to calculate the weights associated with each term (i.e. word) of a sentence:

$$w_i = log \frac{occurrences\, of\, t_i\, in\, topic\, j\, sentences}{occurrences\, of\, t_i\, in\, all\, topics'\, sentences} \quad (5)$$

To calculate the topic signature weight for each sentence, we sum up the weights of the words in that sentence and then, normalized the weights. Thus, a sentence gets a high score if it has a set of terms that are highly correlated with a target concept (topic). Our second run considered title matching (i.e. similarity between the given topic title and a sentence) and cue word matching[7] as additional features (Edmundson, 1969; Chali et al., 2009) to contribute to the node weights.

On the other hand, edge weight is determined by measuring similarity between the sentences. Initially, we remove the stopwords from the sentences using a stopword list. Then, we use the *OAK* system (Sekine, 2002) to get the stemmed words of a sentence. We expand the remaining keywords of the sentence using WordNet. Finally, we find the similar words between each pair of sentences that denotes the edge weight between the two sentences. We build a similarity matrix by populating into it the edge weights between sentences.

2.2 Update Summarization

For update summarization, the random walk model is applied at first to select a list of top query-focused sentences from the second set of given documents. Then, we follow two approaches to generate update summaries for the submitted two runs.

[7]The probable relevance of a sentence is affected by the presence of pragmatic words such as "significant", "impossible", "in conclusion", "finally" etc. We use a cue word list of 228 words (see a sample cue word list in Appendix).

2.2.1 Using ROUGE

In the first run, we use ROUGE similarity measures (Lin, 2004) to reduce the inter-cluster redundancy (between the to-be-generated summary of cluster B and the cluster A sentences) while producing the update summaries. *ROUGE* stands for "Recall-Oriented Understudy for Gisting Evaluation". It is a collection of measures that determines the quality of a summary by comparing it to reference summaries created by humans. The measures count the number of overlapping units such as n-gram, word-sequences, and word-pairs between the system-generated summary to be evaluated and the ideal summaries created by humans. The available ROUGE measures are: ROUGE-N (N=1,2,3,4), ROUGE-L, ROUGE-W and ROUGE-SU. ROUGE-N is n-gram recall between a candidate summary and a set of reference summaries. ROUGE-L measures the longest common subsequence (LCS) which takes into account sentence level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically. ROUGE-W measures the weighted longest common subsequence (WLCS) providing an improvement to the basic LCS method of computation to credit the sentences having the consecutive matches of words. ROUGE-SU is the overlap of skip-bigrams between a candidate summary and a set of reference summaries where skip-bigram is any pair of words in their sentence order allowing for arbitrary gaps. Most of these ROUGE measures have been applied in automatic evaluation of summarization systems and achieved very promising results (Lin, 2004). We apply all the ROUGE measures and calculate the average ROUGE similarity score between the candidate summary sentences of the cluster B and the sentences of cluster A. In the end, the less similar candidate sentences were selected to be included in the final update summaries of the cluster B.

2.2.2 Using LSA

In the second run, we use LSA (Latent Semantic Analysis) (Landauer et al., 1998) to reduce the inter-cluster redundancy for producing the update summaries. LSA uses a sophisticated approach to decode the inherent relationships between contexts (typically a sentence, a paragraph or a document)

and the words that they contain. The main idea behind LSA technology is to extract the close relationship between the meaning of a text and the words that are present in that text. The main ability of LSA is to identify the similarity between two texts even they do not have any words in common, thus providing at least a similarity score by taking synonymy and polysemy into consideration. In the first phase of LSA, a word-by-context (WCM) matrix is constructed that represents the number of times each distinct word appears in each context. Next, weighting may be applied to the values contained in this matrix in relation to their frequency not only in the individual contexts, but through the corpus collection overall; raw term frequency has the drawback that all the terms in the contexts are considered equal and this may not be the case. The next phase is called the dimensionality reduction step. In this phase, the dimension of the WCM is shortened by applying Singular Value Decomposition (SVD) and then reducing the number of singular values in SVD. This is done in order to access the ability of LSA in determining similarity scores (other than zero) in case where two documents have nothing in common between them. By reducing the dimensions, LSA can enhance the score of two similar documents whilst decreasing the score of non similar documents. Thus the process makes the context and the words more dependent to each other by reducing the inherent noise of the data set. We use LSA to measure the similarity of a candidate summary sentence of cluster B with all the sentences of cluster A. The sentences that have the lowest similarity scores are selected to be included in the final update summary. We use a publicly available implementation of LSA[8] for this task. We did not apply dimension reduction in LSA since this setting gave us the most accurate scores.

3 Evaluation

3.1 Task Overview

The Guided Summarization task at TAC-2010 aims to encourage summarization systems to make a deeper linguistic (semantic) analysis of the source documents instead of relying only on document word frequencies to select important concepts. The task is to write a 100-word summary of a set of 10 newswire articles for a given topic, where the topic falls into a predefined category. There are five topic categories:

1. Accidents and Natural Disasters
2. Attacks
3. Health and Safety
4. Endangered Resources
5. Investigations and Trials

Participants are given a list of important aspects for each category, and a summary must cover all these aspects (if the information can be found in the documents) including any other information relevant to the topic. Additionally, an "update" component of the guided summarization task is to write a 100-word "update" summary of a subsequent 10 newswire articles for the topic assuming that the user has already read the earlier articles.

3.2 Test Data

The TAC 2011 test dataset is composed of 44 topics divided into five categories. Each topic has 20 relevant documents which have been divided into 2 sets: Document Set A and Document Set B. Each document set has 10 documents, and all the documents in Set A chronologically precede the documents in Set B. The system task is to write 2 summaries (one for Document Set A and one for Document Set B) that describe the event indicated in the topic title, according to the list of aspects given for the topic category where the first summary is a straightforward query-focused summary whereas the second summary should be written under the assumption that the user of the summary has already read the documents in Document Set A.

3.3 Corpus

The documents for summarization come from the newswire portion of the TAC 2010 KBP Source Data (LDC Catalog Number: LDC2010E12). The collection spans the years 2007-2008 and consists of documents taken from the New York Times, the Associated Press, and the Xinhua News Agency newswires. Along with the raw data, a clean version of it was

[8] http://code.google.com/p/lsa-lda/

also available from LDC that included sentence segmentation and marking of certain non-sentence segments. We used clean data for all our experiments.

3.4 Results and Analysis

Eight NIST assessors selected and wrote summaries for the 44 topics in the TAC 2011 guided summarization task. Each topic had 2 document sets (A,B), and NIST assessors wrote 4 model summaries for each document set. For each document set, the assessors created a 100-word model summary covering all the aspects listed for the topic category (if such information was found in the documents). The assessors could also include other information relevant to the topic. NIST conducted a manual evaluation of summary content based on the Pyramid Method[9] using the multiple model summaries created by the assessors. The assessor also gave a readability/fluency score and an overall responsiveness score to each peer summary. In addition to the Pyramid evaluation, NIST used automatic evalution tools ROUGE and BE (Basic Elements) to measure the performance of the systems.

Table 1 to Table 3 show the manual, ROUGE and BE evaluation results of our systems, respectively. Each column (except the first) of the tables stands for the run id of our systems along with the NIST assigned peer id. From Table 1, we see that according to the pyramid evaluation, for the query-focused summaries of document set A, run 1 performs better than run 2 whereas linguistic quality and responsiveness scores show that run 2 is better. This shows the significance of using the title matching and cue word matching features in run 2. On the other hand, for the update summaries of document set B, we find that, run 2 (that uses LSA) performs significantly better than run 1 (that uses ROUGE) in all kinds of manual evaluation. In Table 2 we see that, most of the time run 2 is having a better ROUGE score than run 1 proving the effectiveness of LSA. Similar results are found from the BE evaluation also. Analyzing all these results reveals the fact that the LSA technology can be very effective for the update summarization systems.

[9] http://duc.nist.gov/pubs/2005papers/columbiau.passonneau2.pdf

Score	UofL1:45	UofL2:23
Modified Pyramid Score-A	0.322	0.305
Number of SCUs-A	4.273	4.114
Linguistic Quality-A	2.273	2.500
Overall Responsiveness-A	2.273	2.341
Modified Pyramid Score-B	0.115	0.158
Number of SCUs-B	1.477	1.886
Linguistic Quality-B	1.932	2.205
Overall Responsiveness-B	1.614	1.727

Table 1: Manual Evaluation

Score	UofL1:45	UofL2:23
ROUGE-2 Recall-A	0.07120	0.07274
ROUGE-2 Recall-B	0.03857	0.04411
ROUGE-SU4 Recall-A	0.11580	0.11478
ROUGE-SU4 Recall-B	0.08137	0.08755

Table 2: ROUGE Evaluation

4 Conclusion

In this paper, we described our participation in TAC 2011 guided summarization task. The evaluation results showed that Latent Semantic Analysis (LSA) can be very effectively applied for the update summarization task.

Acknowledgments

The research reported in this paper was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada – discovery grant and the University of Lethbridge.

References

Yllias Chali, Shafiq R. Joty, and Sadid A. Hasan. 2009. Complex Question Answering: Unsupervised Learning Approaches and Experiments. *Journal of Artificial Intelligence Research*, 35:1–47.

Harold P. Edmundson. 1969. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery (ACM)*, 16(2):264–285.

Christiane Fellbaum. 1998. WordNet - An Electronic Lexical Database. Cambridge, MA. MIT Press.

Score	UofL1:45	UofL2:23
BE Recall-A	0.03414	0.03511
BE Recall-B	0.01278	0.01711

Table 3: BE Evaluation

Sanda Harabagiu, Finley Lacatusu, and Andrew Hickl. 2006. Answering Complex Questions with Random Walk Models. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 220 – 227. ACM.

John Lafferty and Chengxiang Zhai. 2001. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, (25):259–284.

Chin-Yew Lin and Edward H. Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of Association for Computational Linguistics*, pages 74–81, Barcelona, Spain.

Satoshi Sekine. 2002. Proteus Project OAK System (English Sentence Analyzer), http://nlp.nyu.edu/oak.

Appendix: Sample Cue Words List

indeed	further	as well
as this	either	neither
not only	but also	the reason is
as well as	also	moreover
what is more	as a matter of fact	furthermore
in addition	besides	to tell you the truth
in fact	actually	amazingly
to say nothing of	too	let alone
much less	additionally	nor
alternatively	on the other hand	not to mention
such as	this time	at this time
this also	several years ago	long ago
during	eventually	meanwhile
essentially	enormously	majority of the
absolutely	necessary	especially
specially	after	before
at least	at most	most
therefore	this is	that is
reasonable	according to	throughout
at this point	along with	previously
as	particularly	including
as an illustration	for example	like
in particular	for one thing	to illustrate
for instance	notably	by way of example
speaking about	considering	regarding
with regards to	as for	concerning
on the subject of	the fact that	similarly
in the same way	by the same token	in a like manner