Multi-Document Summarization using Multiple-Sequence Alignment

V. Finley Lacatusu, Steven J. Maiorano and Sanda M. Harabagiu

Human Language Technology Research Institute
Department of Computer Science
University of Texas at Dallas
P.O. Box 830688, Richardson, Texas 75083-0688
finley@hlt.utdallas.edu

Abstract

This paper describes a novel clustering-based text summarization system that uses Multiple Sequence Alignment to improve the alignment of sentences within topic clusters. While most current clustering-based summarization systems base their summaries only on the common information contained in a collection of highly-related sentences, our system constructs more informative summaries that incorporate both the redundant and unique contributions of the sentences in the cluster. When evaluated using ROUGE, the summaries produced by our system represent a substantial improvement over the baseline, which is at 63% of the human performance.

1. Introduction

In this paper we present a novel method of producing multi-document summaries based on a technique widely used in bioinformatics: Multiple Sequence Alignment (MSA). MSA is a technique used by biologists to align biological sequences for detecting a common structure, a common function or a common evolutionary source (Gusfield, 1997). Encouraged by the successful implementation of Multiple Sequence Alignment for other Natural Language Processing tasks such as machine translation (Bangalore et al., 2002) and generation (Barzilay and Lee, 2002), we present a system that uses MSA techniques to enhance text summarization by improving the alignment of sentences within topic clusters.

The same topic is often covered concurrently in multiple documents. We usually need to grasp just the main aspects of the topic and for that, summarization techniques are readily available. Such techniques discover the commonalities between documents, eliminate redundancies and most importantly, generate a text of informative content.

Most clustering-based approaches to summarization generate summaries based on the common information expressed by the sentences in a cluster. These systems seek to guarantee the accuracy of their summaries by only incorporating information found in multiple sentences in the cluster; unique or novel contributions are necessarily excluded from the summary as irrelevant or tangential to the documents' topic.

However, eliminating sentences that contribute redundant information must be done carefully. If we do not select the right criterion for eliminating redundant sentences, our summaries run the risk of losing important information that may be contributed only by individual sentences. The approach presented in this paper seeks to improve summarization by using MSA to eliminate redundant information, while preserving the unique information contributions of each sentence in a topic cluster.

Multiple Sequence Alignment can be used to combine together the sentences in topic clusters in order to identify their common content and also the information that each of the original sentences contributes that is unique. Based on such an alignment, we are able to create a new summary sentence that will cover the information in the original sentences.

Our method of producing multi-document summaries consists of two basic steps: (1) the clustering of similar sentences across the set of documents; and (2) the alignment of sentences in each cluster to generate a single sentence per cluster, that is added to the summary. The clustering of sentences is achieved by using two different similarity measures: (a) similarity of each content word in a sentence to a topic description (which typically is associated with a set of related documents); and (b) inter-sentence similarity.

The remainder of this paper is organized as follows: Section 2 describes related work; Section 3 details the clustering method; Section 4 explains how the clusters are consolidated with Multiple Sequence Alignment; Section 5 focuses on the actual summary generation; Section 6 presents the evaluation of the results and Section 7 summarizes the conclusions.

2. Related work

The goal of summarization is to present the most important content of a document (or set of documents) in a condensed way.

Most approaches to single document summarization depend on the potentially monolithic structure of a document. In many single documents, the most important information for a summary is located at the very beginning. Simple single-document summarization systems, then, can extract the first sentences of a document and typically obtain good results.

The same techniques cannot be applied to multi-document summarization, however. Since the input to multi-document summarization is a set of related documents that can feature potentially very different structures, systems cannot depend entirely on generalizations about document structure to produce summaries. Without structural features to rely on, multi-document summarization systems must depend on representing the information content common to many (if not all) of the documents in a collection.

Most multi-document summarization systems identify the important information contained in a set of documents by finding sentences or paragraphs that are closely related. These systems assume a direct relationship between the frequency of these related segments and their relative importance to a multi-document summary. Under this approach, the text snippets that occur the most often in a range of texts are considered to contain the core information that should be included in the summary (Hatzivassiloglou et al., 2001). In order to eliminate redundancy, each cluster of similar text segments is permitted to contribute only one sentence to the summary (Nenkova et al., 2003).

Although Multiple Sequence Alignment has not previously been used for summarization, it has proven successful for other NLP tasks. Work done by Barzilay and Lee (2002) uses MSA for natural language generation, while Bangalore et al. (2002) uses MSA in machine translation for a multilingual instant messaging system.

3. Clustering similar sentences across documents.

Our summarization method is motivated by the DUC competition[1]. Task 2 of DUC 2003 focuses on creating a summary from a given set of related documents and a TDT topic.

We ranked all of the sentences in a collection of documents based on their similarity to a TDT topic description. This ranking is described algorithmically in Step 1 below.

After performing this ranking, we observed that the sentences that were most closely related to the TDT topic were characterized by a high degree of redundancy in terms of their information content. We sought to reduce the amount of redundancy passed on to the summary by clustering the sentences in the collection before generating a summary.

Our clustering technique is produced by a three-step approach that is based on word-to-word similarities that leverages the lexico-semantic information found in WordNet. Given a set of related documents D_1, D_2, ..., D_n, and their topic description TD containing a sequence of content words t_1, t_2, ..., t_m, we generate clusters by the following steps:

Step 1: For each sentence S_a in each document D_i
- For every content word w_b of S_a, compute its similarity to the topic TD, as:
$sim(w_b, TD) = \sum_{t_i \in TD}(s(w_b, t_i))$ *where:*
† *s(w_b, t_i) = 1 if w_b and t_i are identical*
= 0.95 if w_b and t_i belong to the same WordNet synset;
= 0.3 if w_b is in the gloss of t_i, or t_i is in the gloss of w_b
= 1/n if there is an IS-A sequence of length n between the synsets of w_b and t_i
= 1/2n if there is an IS-PART sequence of length n between the synsets of w_b and t_i

Step 2: Rank all sentences by their topic-similarity, produced by:
$ts(S_a, TD) = \sum_{w_b \in S_a}(sim(w_b, TD))$

Step 3: Given S_t, the top-ranked sentence from Step 2, we create a cluster of similar sentences using S_t as a "new topic" descriptor for measuring the similarity of any other high ranked sentence S_u to S_t by using the formula from Step 2.

[1] http://duc.nist.gov/

(1) APW19981017.0507: Former Chilean dictator Gen. Augusto Pinochet has been arrested by British police on a Spanish extradition warrant, despite protests from Chile that he is entitled to diplomatic immunity.
(2) NYT19981017.0177: Gen. Augusto Pinochet, who ruled Chile as a despot for 17 years, has been arrested in London after Spain asked that he be extradited for the presumed murders of hundreds of Chilean and Spanish citizens, the British authorities announced Saturday.
(3) NYT19981018.0098: Gen. Augusto Pinochet, the former Chilean dictator arrested here at the request of a Spanish judge, remained sequestered under police guard Sunday, awaiting a potentially devastating court hearing to weigh his extradition on charges of genocide, terrorism and murder.

Figure 1: The first cluster in the *Pinochet Trial* topic

Topic representations are key to the success of our summarization system: without a topic representation, our system cannot determine which information should appear in a summary.

However, this dependence on a topic representation does not mean that our system has to have a pre-specified (TDT-style) topic representation to construct a summary. If we do not have a topic (or if the topic is under-specified), it can be derived from the set of related documents by using the (Lin and Hovy, 2000) technique for automatically acquiring topic signatures. Although their technique was applied to single documents only, our initial experiments show that the Lin and Hovy method can be extended to handle multiple documents. In short, if we do not have the topic description required to perform efficient clustering over a related document set, we can use an extended version of the Lin and Hovy technique to provide us with a topic signature that works as well in clustering as the TD.

4. Consolidating Clusters with Multiple Sequence Alignments

Word-to-word similarities based on lexico-semantic information indicate possible relatedness between sentences and thus possible commonalities.

This same idea, known as Multiple Sequence Alignment, is used in bioinformatics, where the alignment of multiple sequences can determine commonalities among biological sequences. News articles change as the news story develops, but the story retains a large part of the previously revealed content. A good summary, therefore, should retain the "old" common-denominator content that predominates across the clustered sentences while adding new, less redundant pieces of information. To make this summary even more informative, the information needs to be understood in terms of 1) what is common and pervasive information; 2) what is new and unique; and 3) what is contradictory. MSA enables us to accomplish these three things.

Since MSA is an NP-complete problem (when the number of sequences is a variable), we need approximations to run it in polynomial time. Similar to Barzilay and Lee 2002 (and much other work employing MSA), we used iterative pairwise alignment, an approximation technique which generates good results in distinguishing the common traits of sentences.

The pairwise algorithm employs a distance function between pairs of words from different sentences that computes:

dist(w_i, w_j) = 1.01 if w_i or w_j is a sequence gap,
= 1-s(w_i,w_j) if s(w_i, w_j) > 0.1,
where s(w_i, w_j) is as before (†),
= 1.5 if w_i and w_j are not related.

MSA through pairwise iterative sequence alignment is performed as follows: first, identify the two closest sentences, next, align them (using pairwise sequence alignment), finally select the closest sentence to one of the already aligned sentences and do the pairwise alignment between the first alignment and the new sentence (this last step is repeated until there are no sentences left in the cluster).

For the cluster in Figure 1, the sentences from documents APW19981017.0507 and NYT19981017.0177 will be aligned first, obtaining the alignment in Table 1(a). Then, this alignment will be aligned with the sentence from document NYT19981018.0098, obtaining the final alignment in Table 1(b).

5. Using Multiple-Sequence Alignments for generating Summaries

The alignment produced by Multiple-Sequence Alignment creates a single summary sentence for each cluster. We claim that this sentence is more informative than any of the cluster sentences because (1) generation prefers new information over background information; and (2) it selects the longest span of content.

The general procedure is as follows:

Step 1: Identify common points across sequences whenever words w_i and w_j are identical.

Step 2: Classify similar vs. complementary sequences between any two successive common points

- sequences are similar if they contain related words
- otherwise, sequences are complementary

Step 3: In the case of similar sequences, select the longest sequence and use it in the summary sentence. When sequences are complementary, both sequences appear in the summary.

For the alignment in Table 1(b), the global common points are: *Gen. Augusto Pinochet* and *arrested*, but between sentences (1) and (2) we have also *has been* and *extradition/extradited* and for sentences (1) and (3) we have also *Spanish* as common point.

The sentence generated for the summary from the alignment of the sentences in Figure 1 is:

"Former Chilean dictator Gen. Augusto Pinochet, who ruled Chile as a despot for 17 years, has been arrested in London after Spain asked that he be extradited for the presumed murders of hundreds of Chilean and Spanish citizens, remained sequestered under police guard Sunday, awaiting a potentially devastating court hearing to weigh his extradition on charges of genocide, terrorism and murder, despite protests from Chile that he is entitled to diplomatic immunity."

Augusto - Augusto	Former - ****** - ******
Pinochet - Pinochet	Chilean - ****** - ******
* - ,	dictator - ****** - ******
** - who	Gen. - Gen. - Gen.
**** - ruled	Augusto - Augusto - Augusto
**** - Chile	Pinochet - Pinochet - Pinochet
** - as	****** - , - ,
* - a	****** - who - ******
***** - despot	****** - ruled - ******
** - for	****** - Chile - ******
** - 17	****** - as - ******
**** - years	****** - a - ******
* - ,	****** - despot - ******
has - has	****** - for - ******
been - been	****** - 17 - ******
arrested - arrested	****** - years - the
** - in	****** - , - former
***** - London	has - has - Chilean
by - after	been - been - dictator
British - Spain	arrested - arrested - arrested
police - asked	****** - in - ******
on - that	****** - London - here
a - he	by - after - at
Spanish - be	British - Spain - the
extradition - extradited	police - asked - request
warrant - for	on - that - of
** - the	a - he - a
***** - presumed	Spanish - be - Spanish
***** - murders	extradition - extradited - judge
* - of	warrant - for - ,
, - hundreds	****** - the - remained
despite - of	*** - presumed - sequestered
protests - Chilean	****** - murders - under
from - and	****** - of - police
Chile - Spanish	, - hundreds - guard
that - citizens	despite - of - Sunday
he - ,	protests - Chilean - ,
is - the	from - and - awaiting
entitled - British	Chile - Spanish - a
to - authorities	that - citizens - potentially
diplomatic - announced	he - , - devastating
immunity - Saturday	is - the - court
. - .	entitled - British - hearing
	to - authorities - to
	****** - ****** - weigh
	****** - ****** - his
	***** - ***** - extradition
	****** - ****** - on
	****** - ****** - charges
	****** - ****** - of
	****** - ****** - genocide
	****** - ****** - terrorism
	diplomatic - announced - and
	immunity - Saturday - murder
(a)	(b)

Table 1: The result of (a) P. S. A. and (b) M. S. A.

In Table 1(a) is presented the result of a Pairwise Sequence Alignment step and in Table 1(b) is presented the result of the whole Multiple Sequence Alignment algorithm applied on a cluster of sentences. The "*****" in the table represent sequence gaps introduced during the alignment process.

6. Results

We have tested our approach on the DUC 2003 Task 2 data. There are 30 document sets and 30 TDT topics, one for each cluster. For each cluster there are 4 manual summaries. For evaluation we have used ROUGE[2], an automatic summarization evaluator based on n-gram co-occurrences between summary pairs (Lin and Hovy, 2003). Lin and Hovy (2003) show that "automatic evaluation using unigram co-occurrences between summary pairs correlates surprising well with human evaluations, based on various statistical metrics".

We have ranked our system using ROUGE and compared it with the scores of three other summaries: We have used as a baseline summary the first sentences from the most recent document in the set of documents. We have also created a summarization system that composes a summary from the top ranked sentences, but eliminating redundant sentences. A sentence is considered redundant if it doesn't bring at least 50% new information to the summary. For comparison we have scored the human performance, as well: one of the 4 manual summaries was considered as *test summary* and the other three as model summaries for each of the document sets.

BL ROUGE-1 Average: 0.21899 (95%-conf.int. +- 0.03339)
BL ROUGE-1 Median: 0.21996 (95%-conf.int. +- 0.03436)
BL ROUGE-1 Maximum: 0.23858 (95%-conf.int. +- 0.05298)
BL ROUGE-1 Minimum: 0.19744 (95%-conf.int. +- 0.05447)

Table 2: ROUGE output for the baseline

TS ROUGE-1 Average: 0.25901 (95%-conf.int. +- 0.02419)
TS ROUGE-1 Median: 0.25989 (95%-conf.int. +- 0.02507)
TS ROUGE-1 Maximum: 0.27539 (95%-conf.int. +- 0.04057)
TS ROUGE-1 Minimum: 0.24088 (95%-conf.int. +- 0.04085)

Table 3: ROUGE output for the *top sentences* approach

MA ROUGE-1 Average: 0.27125 (95%-conf.int. +- 0.02258)
MA ROUGE-1 Median: 0.27239 (95%-conf.int. +- 0.02372)
MA ROUGE-1 Maximum: 0.28875 (95%-conf.int. +- 0.04008)
MA ROUGE-1 Minimum: 0.25148 (95%-conf.int. +- 0.03904)

Table 4: ROUGE output for the MSA approach

As it can be seen from tables 2, 3, 4, and 5 our approach performs better than the *top sentences* approach and definitely better than the baseline. If we take the human performance as point of reference, as in Table 6, MSA obtains a coverage score of 79% while the baseline is at 63% and top sentences at 75%. It is interesting to observe that the human performance is at only 0.34384 which indicates that the degree of agreement between human summarizers is not too high.

7. Conclusions

Our approach was motivated by the DUC competitions, where the sets of documents contained news articles from various sources. The journalistic style makes the most important sentences form documents that refer to the same news to be somewhat similar, even if the documents were not issued by the same agency. This fact enabled the multiple-sequence alignment approach to give good results.

The MSA approach presented in this paper can be extended for cases where the topic description is not given, and thus indexing and retrieval of documents can be considered based on summaries of sets of related documents. MSA can provide a different way of organizing texts, providing an alternative to the inverted lists employed by vector space retrieval models. Most importantly, our multi-document summarization technique may be used for organizing the lexico-semantic information derived from texts crucial for such tasks as Information Extraction or Question Answering.

MS ROUGE-1 Average: 0.34384 (95%-conf.int. +- 0.02620)
MS ROUGE-1 Median: 0.34545 (95%-conf.int. +- 0.02636)
MS ROUGE-1 Maximum: 0.36909 (95%-conf.int. +- 0.05000)
MS ROUGE-1 Minimum: 0.31697 (95%-conf.int. +- 0.05307)

Table 5: ROUGE output for the manual summaries

Baseline	Top Sentences	MSA	Manual
0.219	0.259	0.271	0.344
63%	75%	79%	100%

Table 6: Comparison of the results

8. References

Bangalore, Srinivas, Vanessa Murdock, and Giuseppe Riccardi, 2002. Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system. In *Proceedings of the International Conference on Computational Linguistics (COLING 2002)*.

Barzilay, Regina and Lillian Lee, 2002. Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gusfield, Dan, 1997. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press.

Hatzivassiloglou, V., J. Klavans, M. Holcombe, R. Barzilay, M. Kan, and K. McKeown, 2001. Simfinder: A flexible clustering tool for summarization. In *NAACL Workshop on Automatic Summarization, 2001*.

Lin, Chin-Yew and Eduard Hovy, 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the COLING 2000 Conference*.

Lin, Chin-Yew and Eduard Hovy, 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*.

Nenkova, A., B. Schiffman, A. Schlaiker, S. Blair-Goldensohn, R. Barzilay, S. Sigelman, V. Hatzivassiloglou, and K. McKeown, 2003. Columbia at the document understanding conference 2003. In *HLT Workshop on Text Summarization (DUC 2003)*.

[2]http://www.isi.edu/ cyl/ROUGE/