Topic-based Summarization at DUC 2005

Horacio Saggion
Department of Computer Science
University of Sheffield
211 Portobello Street - Sheffield, England, UK, S1 4DP
Tel: +44-114-222-1947
Fax: +44-114-222-1810
saggion@dcs.shef.ac.uk

Abstract

We describe a topic-based multidocument sentence extractor developed for the DUC 2005 competition. The system has been designed for the real task of producing summaries given an information need expressed as a set of questions. The implementation of the system takes advantage of an in-house summarization toolkit and available natural language processing technology.

1 Introduction

The National Institute of Standards and Technology (NIST) with support from the Defense Advanced Research Projects Agency (DARPA) is conducting a series of evaluations in the area of text summarization, the Document Understanding Conferences (DUC), providing the appropriate framework for system-independent evaluation of text summarization systems.

This year DUC task was related to a real summarization application: to synthesize from a set of related documents a brief, well-organized, fluent answer to a *need for information* that cannot be met by just stating a name, date, quantity, etc. The *need for information* is expressed through a topic description of the types illustrated in Figure 1. A *user profile* is also identified in the form of a feature "granularity" with possible values "generic" of "specific" expressing the granularity required for the summary. The task requires summarizers to address two problems:

- to find the appropriate pieces of information in different document matching the user requirements; and
- to compile the information in a well organized text.

This year summarizers had to produce summaries for 50 DUC topics developed by NIST assessors. For each topic, the systems received between 25 and 50 document as input.

In this paper we describe the basic components used by our system, how we have used them to create a topic-based summarizer, and the results obtained in the recent DUC 2005 evaluation.

2 The System

We make use of a general purpose summarization system which can be adapted to different summarization tasks (Saggion, 2002): the same underlying components were used to develop a centroid-based summarization system for DUC 2004 (Saggion and Gaizauskas, 2004) and a cross-lingual summarization system for the Multilingual Summarization Evaluation (Saggion, 2005).

<TOPIC ID="d324e" GRANULARITY="specific">
How have relations between Argentina and Great Britain developed since the 1982 war over the Falkland Islands? Have diplomatic, economic, and military relations been restored? Do differences remain over the status of the Falkland Islands?
</TOPIC>

<TOPIC ID="d332h" GRANULARITY="general">
What kinds of non-tax crimes have lead to tax evasion prosecutions (failure to file, inaccurate filing), instead of or in addition to prosecution for the non-tax crimes themselves?
</TOPIC>

Figure 1: Topic description

2.1 Generic Components

The system for document analysis uses tools for text structure identification, tokenization, sentence boundary detection, named entity recognition, coreference resolution, etc. adapted from the GATE library (Cunningham et al., 2002). The summarization system implements a number of scoring functions to assess sentence-summary worthiness including sentence position, similarity of the sentence to the document headline, similarity of the sentence to the leading paragraph of the document, term distribution, named entity distribution, etc. The sentence final score is computed by combining individual feature-values in a linear equation.

We have implemented a vector space model in which texts are represented as vectors of terms $< t_i, w_i >$ where t_i is a word and w_i is *term frequency * inverse document frequency* (idf). A corpus statistics module is in charge of computing term frequencies. Vectors of terms can be produced for different text fragments. In our approach some summarization features are based on the computation of text similarity values. We compute the similarity between textual units by using the *cosine* between their vector representations. The formula we use is as follows:

$$cosine(x, y) = \frac{\sum_{i=1}^{n} w_{i,x} * w_{i,y}}{\sqrt{\sum_{i=1}^{n} (w_{i,x})^2} * \sqrt{\sum_{i=1}^{n} (w_{i,y})^2}}$$

where $w_{i,x}$ is the weight of term i in unit x and n is the numbers of terms.

In a multidocument situation, when the input to the document is a set of n related documents (i.e., cluster), a *centroid* representation of the cluster is constructed. The centroid is a vector of pairs of term-weight, where the weight w_i of term i in the centroid is obtained as follows:

$$w_i = \frac{\sum_{k=1}^{n} w_{i,k}}{n}$$

where $w_{i,k}$ is the weight of term i in document k. This representation is used to compute a centroid based summarization feature as the similarity of each sentence in the cluster to the cluster centroid.

In order to support redundancy detection, the system also computes n-grams for all the input documents. Having computed n-grams sets for each document in the input, the n-gram based similarity metric between two text fragments T_1 and T_2 is computed as follows:

$$similarity(T_1, T_2, n) = \sum_{k=1}^{n} w_k * \frac{|grams(T_1,k) \cap grams(T_2,k)|}{|grams(T_1,k) \cup grams(T_2,k)|}$$

where n means that n-grams 1, 2, ... n are to be considered, $grams(T, k)$ is the set of k-grams of fragment T, and w_k is the weight associated with the k-gram similarity of two sets. In general we use n=4 and the arithmetic series $w_1 = 0.1$, $w_2 = 0.2$, $w_3 = 0.3$, and $w_4 = 0.4$ as weighting scheme. A threshold for identification of similar information in sentences is established in the following way: we assume that in a given document all sentences will report different information, therefore we can use the n-gram similarity values between them

to help estimate a similarity threshold. We compute pairwise n-gram similarity values between sentences in documents estimating a threshold for dissimilarity as the average of the pairwise similarity values. This dissimilarity value is used to decide whether two sentences are different. We consider two units T_1 and T_2 to be different if $similarity(T_1, T_2, n) \leq \alpha$.

2.2 Content Reduction

Content reduction is a process of sentence elimination we have recently implemented. Most sentence extraction algorithms work in a constructive way: given a document and a sentence scoring mechanism, the algorithm ranks sentences by score, and then chooses sentences from the ranked list until a compression rate is reached. We take a different approach which consists in removing sentences from a pool of candidate sentences until the desired compression is achieved. We follow a procedure similar to Marcu's algorithm (Marcu, 1999) for the creation of extracts from pairs of $< document, abstracts >$. In his approach clauses from the document are greedily deleted in order to obtain an extract which is maximally similar to the abstract. We assume that the candidate list of sentences which has been selected is the ideal content to include in the final summary. Given a set of sentences C which match a selection criteria, the algorithm creates an extract which is "close" in content to C but which is reduced in form. The measure of proximity between documents is taken to be the cosine between two term vectors representing the documents. At each step, the algorithm greedily rejects a sentence from the extract. The rejected sentence is one which if removed from the extract produces a pseudo-document which is closer to C among all other possible pseudo-documents. This strategy for sentence rejection has been used in a profile-based summarization system which produced quite good results.

2.3 The System for DUC 2005

The system receives a cluster of documents and a topic description and produces a summary. We have not implemented functionalities to deal with topic granularity, our system produces the same summary regardless of the type of granularity.

- *Topic analysis*: the topic description is tokenised and represented in the vector space model (*topic vector*);
- *Document analysis*: each document in the cluster is tokenised, sentence splitted, and statistics computed for each term. Each sentence is represented in the vector space model using a general idf table where inverted document frequencies are computed from the British National Corpus. Each sentence is annotated with its position and the publication date of the document.
- *Sentence analysis*: for each sentence in the cluster the cosine similarity between the sentence and the *topic vector* is computed (in the vector space model) and the sentence is annotated with the similarity value.
- *Similarity threshold computation*: the average similarity value of the set of sentences in the cluster is computed.
- *Sentence selection*: a sentence is included in a list of candidate sentences if its similarity value is greater than the average similarity and if it is *different* from all other candidate sentences in the list.
- *Sentence ordering*: sentences are ordered by date/time of document publication and position in the document they appear.
- *Content reduction*: the list of candidate sentences is reduced by our greedy-rejection mechanism until the desired compression is reached.

This procedure generates summaries such as the one shown in Figure 2 for the "Relation between Argentina and Great Britain on the Malvinas issue". Our system, SHEF-BSL, is run-ID 32 in the evaluation.

But the planned trip by an aircraft which is normally used by the Argentine navy to transport military equipment would have almost certainly fuelled stiff opposition from the Falkland islanders and from MPs opposed to any suggestion of a 'sell-out' over the islands. The Falklands may have become a peripheral issue in Britain, but Argentina remains obsessed with the islands and clearly hopes that Mr Hurd's five-day visit will signify another step on the tortuous road to 'recovering' the islands, this time through diplomatic means. A first round of Anglo-Argentine talks on seismic exploration around the disputed Falkland Islands will take place in Buenos Aires later this month. In the war over the islands, called the Malvinas by Argentina, 712 Argentines, 255 Britons and three islanders died. The decision removed a major stumbling block toward re-establishing diplomatic relations between the nations, broken during the conflict over the islands. Argentina and Britain announced an agreement Thursday to restore full diplomatic ties, nearly eight years after they fought a 74-day war over the Falkland Islands, a sparsely populated archipelago off Argentina's coast in the South Atlantic Ocean. If oil is found in commercial quantities around the islands, co-operation with Argentina would be essential to develop the reserves. Britain and Argentina re-established diplomatic relations in 1990, a year after President Carlos Menem took office.

Figure 2: Topic-based summary

Quality Test	(1)	(2)	(3)	(4)	(5)
System Rank	7	6	12	12	21

Table 1: Linguistic quality ranks for our system. Rank number 1 is best.

3 Evaluation and Results

DUC 2005 has evaluated summaries in two ways: human evaluation using the pyramid method, summary responsiveness to the topic and linguistic quality; and automatic evaluation using the ROUGE statistic.

3.1 Responsiveness

NIST assesors assigned a responsiveness score between 1 (least response) and 5 (most responsive) to each of the automatic and human summaries. Responsiveness is the amount of information in the summary that helps to satisfy the information in the topic. As responsiveness scores can not be used to compare summarizers, then scaled responsiveness scores were computed and systems ranked using that scaled score. Our system was placed at rank 18 (out of 33) on responsiveness.

3.2 Linguistic Quality

Summaries were judged according to the following five criteria:

- (1) Grammaticality: the summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences that make the text difficult to read.
- (2) Non-redundancy: there should be no unnecessary repetition in the summary.
- (3) Referential clarity: it should be easy to identify who or what the pronouns and noun phrases in the summary are referring to.
- (4) Focus: the summary should have a focus; sentences should only contain information that is related to the rest of the summary.
- (5) Structure and Coherence: the summary should be well-structured and well-organized.

All linguistic quality questions require a certain readability property to be assessed on a five-point scale from "A" to "E", where "A" indicates that the summary is good with the respect to the quality under question, "E" indicates that the summary is bad with respect to the quality stated in the question, and "B" to "D" show the gradation in between. Our system ranked as shown in Table 1 out of 33 peer systems.

3.3 Rouge

The Document Understanding Conferences have adopted ROUGE (Lin.C.-Y., 2004) a statistic for automatic evaluation of summaries. ROUGE allows the computation of recall-based metrics using n-gram matching between a *candidate summary* and a *reference set of summaries*. ROUGE-n is n-gram recall, ROUGE-L is a recall metric based on the longest common subsequence match and ROUGE-W is a weighted longest common subsequence that takes into account distances when applying the longest common subsequence.

When multiple references are available in an evaluation, the ROUGE statistic is defined as the best score obtained by the summary when compared to each reference. However, the Jackknifing procedure can also be used when M reference summaries are present in the evaluation, this procedure will estimate ROUGE scores by averaging over M sets of $M-1$ possible references. Recent experiments have shown that some ROUGE scores correlate with rankings produced by humans (Lin.C.-Y., 2004).

In ROUGE, word overlap is the basis for similarity computation, the problem with this approach is that multi-word units (e.g. "Prime Minister") are not treated as units of meaning while unimportant function words are as rewarding as meaningful content. In order to address this issue, *Basic Elements* (BE) have recently been proposed for automatic evaluation of summaries. These BEs are defined as: the head of a major syntactic constituent; or a relation between a head-BE and a single dependent and can be automatically produced using syntactic analysis together with a set of rules to extract valid BEs from parsing trees. When BEs are extracted using MINIPAR (Lin, 1998), they are called *BE-Fs*. When BEs have been produced for automatic summaries and references, then the BEs of a peer summary can be compared with the BEs of a reference summary to evaluate content in the same way n-grams are used in the conventional ROUGE.

Our configuration did not obtain very good rouge scores: a Rouge-2 score of 0.0534 or rank 24 out of 33 and a Rouge-SU4 (which uses BEs as units of meaning) of 0.1041 or rank 24 out of 33.

3.4 Pyramid

Human evaluation of content was performed following the Pyramid method (Nenkova and Passonneau, 2004). The method seeks to match content units in peer summaries (e.g., produced automatically) with similar content units found in a pool of human summaries. In this evaluation, a good peer summary is one where its contents units are observed across many human summaries.

In a nutshell, the method is based on: (i) the construction of a pool of human summaries for each document cluster; (ii) the identification of summarization content units (SCU) in each summary, where content units are proposition-like, atomic representations; (iii) the association of weights to the different SCU based on its frequency of occurrence in the pool of summaries (if a SCU occurs in k summaries, then its weight is k and in an evaluation with s summaries, the maximum possible weight of a SCU is s); (iv) the matching of SCU in the human summaries with SCU in peer (e.g. automatic) summaries; (v) the calculation of a pyramid formula.

Steps (i)-(iii) give rise to a *pyramid*. A pyramid of order n has n tiers T_i, where each T_i contains SCUs with weight i. Given a pyramid of order n the best possible summary with X units, where content is concerned, is one that contains all SCUs of weight n, all SCUs of weight $n-1$, etc. This is called an *optimal summary*. In a pyramid of order n, a peer summary will have D_i SCUs appearing in T_i (with $D_i \leq |T_i|$). To evaluate the content of the peer, the following formula is used:

$$\mathtt{D} = \sum_{\mathtt{i}=1}^{\mathtt{n}} \mathtt{i} * \mathtt{D_i}$$

The content value associated with an optimal summary with X SCUs is given by:

$$\mathtt{Max} = \sum_{\mathtt{i}=\mathtt{j}+1}^{\mathtt{n}} \mathtt{i} * |\mathtt{T_i}| + \mathtt{j} * (\mathtt{X} - \sum_{\mathtt{i}=\mathtt{j}+1}^{\mathtt{n}} |\mathtt{T_i}|)$$

where:

$$j = max_i(\sum_{t=i}^{n} |T_t| \geq X)$$

and the final score given to the peer summary is given by the ratio of its score to the maximum possible score $\frac{\texttt{D}}{\texttt{Max}}$.

Our system performed relatively well on pyramid scores obtaining an "original" pyramid score of 0.22 (9 out of 25) and a "modified" pyramid score of 0.16 (13 out of 25). The "original score" uses as X the same number as units appearing in the peer. The "modified score" uses as X the average number of units found in the human (model) summaries.

4 Conclusion

Thanks to our adaptable summarization technology, the topic-based system was developed in a very short period of time. We are very glad that the system obtained a reasonable performance, however we note that much work is needed to obtain better system performance.

In our future work we will address the issue of granularity which we have not explored in the current system. We will also apply question answering techniques which seem appropriate for this task and complementary to our summarization technology.

References

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *ACL 2002*.

Lin, D. (1998). Dependency-based Evaluation of MINIPAR. In *Proceeding of the Workshop on the Evaluation of Parsing Systems*.

Lin.C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization*, Barcelona. ACL.

Marcu, D. (1999). The automatic construction of large-scale corpora for summarization research. In Hearst, M., F., G., and Tong, R., editors, *Proceedings of SIGIR'99. 22nd International Conference on Research and Development in Information Retrieval*, pages 137–144, University of California, Beekely.

Nenkova, A. and Passonneau, R. (2004). Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of NAACL-HLT 2004*.

Saggion, H. (2002). Shallow-based Robust Summarization. In *Automatic Summarization: Solutions and Perspectives*, ATALA.

Saggion, H. (2005). Experiments in Arabic/English Summarization. In *Multilingual Summarization Evaluation Workshop/ACL 2005*.

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