A New Multi-document Summarization System

Yi Guo and George Stylios
RIFLEX
Heriot-Watt University
Scotland, TD1 3HF, U.K.
y.guo@hw.ac.uk, g.stylios@hw.ac.uk

Abstract

A document understanding system has been developed, AMDS_hw, which is based on the synergy of English grammar parsing with sentence clustering and reduction. The system is capable of producing a summary of single or multiple documents, but the present study only focuses on multi-document summary and hence has participated in DUC2003 evaluations under the non-question related task (Task 2). After a thorough and objective evaluation, the system has shown to perform better in Mean Coverage, Mean Length-Adjusted Coverage and Quality Question Score in comparison with other systems.

1. Introduction

This paper describes the structure and algorithms of a newly developed multi-document summarization system, which is a hybrid of a number of related techniques. This system is designed to produce a generic summary, rather than a biased summary towards some topic of special interest or purpose, for a set of multiple documents, no matter whether or not they are closely related with each other. The details of the system structure and algorithms are explained in Section 2. The evaluation results were issued at the end of March 2003 and the analyses, mainly focused on our system, of the evaluation results are discussed in Section 3.

2. System Structure & Algorithms

The new system in question is a multi-document summarization system named as Automatic Multi- Document Summarizer, or AMDS_hw in short. The system structure contains seven key-functional modules, which include Content Reconstruction, Syntactic Parsing, Indices Extraction, Clustering Sentences, Cluster-Filtering, Cluster-Reuction and Size Control (see Figure-1. AMDS_hw System Structure). The arrows in the Figure-1 indicate the data flow directions of sentences, phrases, indices, or clusters which are processed or produced during the whole summarization procedure.

2.1 Content Reconstruction

This module takes the original documents as the input data, uses basic text processing techniques to divide documents into many paragraphs or segments and finally into sentences, by means of sentence-units. Each sentence-unit carries not only the original content of the sentence but also some important initial information, such as in which document and in which paragraph it belongs to, its positions and relative positions in the paragraph and the document. Each sentence-unit is composed by two parts, the content-

section and the information-section. The content-section saves the original content of the sentence and the information-section stores other important information about the sentence. Further more, the content- section of each sentence-unit is fixed, no more changes occur, but the information-section is very flexible and extensive.

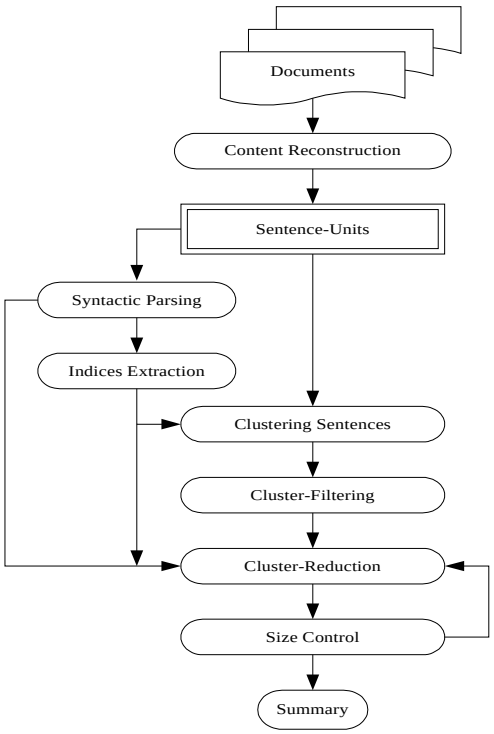

Figure-1. AMDS_hw System Structure

2.2 Syntactic Parsing

In the previous module, the target of summarization is changed from a set of documents to a pool of sentence-units, but no loss of sentence-position information in the documents. In this module, a sentence grammar parser (Link Grammar Parser) [Sleator et al., 1991; Sleator et al., 1993; Grinberg et al., 1995; Lafferty et al., 1992] has been applied, since the collection of sentences, instead of documents, becomes the target of summarization. The Link Grammar Parser is a syntactic parser of English, based on link grammar, an original theory of English syntax [Sleator et al. 2000]. The Link Grammar Parser assigns a syntactic structure, which consists of a set of labeled links connecting pairs of words, and produces a postscript and a constituent tree for the content-section of each sentence-unit in the pool of sentence-units. All these postscripts and consti- tuent trees are added to the information-sections of corresponding sentence-units and managed in XML format files

2.3 Indices Extraction

From the results of syntactic parsing, post-scripts and constituent trees, the indices are being extracted, which include subjects, time, spaces or locations, and actions, for clustering sentences in the next step. Normally, every sentence has at least one subject and one verb phrase (VP), and a complex sentence will have more subjects and verb phrases than a simple sentence. So the ‘Subjects’ and ‘Actions’ indices of a sentence can’t be empty, whether it is a simple or a complex sentence, but if the sentence has more than one subjects or verb phrases, all the subjects are saved in the ‘Subjects’ index and the verb phrases are saved with referential subjects in the ‘Actions’ index.

2.4 Clustering Sentences

In “subject-prominent” languages [Li & Thompson, 1976] like modern English and French, “subject” has been defined as a prominent grammatical relation which is crucially involved in certain syntactic phenomena such as verb agreement, passive, “raising” constructions, etc [Lambrecht, 1996]. Since modern English is a ‘subject-prominent’ language, it is considered that the ‘Subject’ should be used as the first index dimension. The ‘Time’ dimension is appointed as the second one, as most of events happened and

developed in some temporal order. The ‘Spaces/ Locations’ takes the third position of the index dimensions. Finally, the ‘Actions’ is considered as the fourth or supplement index.

After the indices information for each sentence has been established and the index priorities have been set up, all the sentences that have the same or closest ‘Subjects’ index are put in a cluster, and they are sorted out according to the temporal sequence, from the earliest to the latest, then the sentences that have the same ‘Spaces/Locations’ index value in the cluster are marked out.

Practically, the sentences that have different ‘Subjects’ indices but the same ‘Time’ indices can be clustered to see what happened at the same time point; or the sentences that have the same ‘Spaces/Locations’ indices and no matter what their other indices are can be clustered to examine what is going on in the same space or location.

2.5 Cluster-Filtering

From the defined index priorities, many clusters that focused on different ‘Subjects’ indices are established. But how can the outstanding clusters from the others be separated. Since the required multi-document summary should not be more than 100 words, the compression rate, which means the number of words of an original document set compared with the required number of words of the summary, is very high and varies from 16 to 80 or more words. In this situation, it is only arbitrarily to judge whether a cluster is important or not, only by its size. More attentions should be paid on the relation of sizes of these clusters.

A new method has been devised to pick out the most outstanding clusters by computing the dispersion of their sizes. For example, first we rank these clusters by their size from large to small; second, we start from the largest cluster, which has the largest number of words in all clusters, because the largest cluster has to be chosen for its importance. The next question is which cluster should be included as the last, in another words, how many clusters have to be selected. Following the list of ranked clusters, we will find out a cluster, whose size is the largest one among the clusters whose sizes are below 20% of the largest cluster, and we call this cluster as the ‘end-cluster’. Any cluster below the end-cluster in the list of ranked clusters is discarded. But if from the largest cluster to the end-cluster there are more than 10 clusters, only the first 10 clusters will be selected and the 10^{th} cluster will be the new ‘end-cluster’.

2.6 Cluster-Reduction

In this section, WordNet [Miller *et al.* 1990] is applied to process synonym, antonym, hypernymy and hyponymy in the selected clusters. Sentences have been compared on phrase level to get rid of some ‘redundant’ information or sentences/clauses. In order to facilitate the reduction of sizes of the chosen clusters, the positions of sentences has also been taken into consideration. After every sentence in the chosen clusters was touched, the system moves to the next step, size control.

2.7 Size Control

The word count of the output of Cluster-Reduction has been counted. If the word count is over the required size, 100 words, the procedure is replaced by a loop back to Cluster-Reduction and the output is taken as the

new input of Cluster-Reduction until the word count is dropped in the zone of 100± 20 words. Finally the result on which size is closer to 100-word, either above or below 100, is chosen.

3. Evaluations

There were 18 multi-document summarization systems, including two systems (system 2 and 3) used as guidelines, involved in the evaluation for Task 2 in DUC2003. Meanwhile 3 human summarizers and 1 model summarizer were also involved in each document set.

In order to discriminate four different types of summarizers in following analysis, all the 18 multi-document summarization systems will be called as systems or peers; the system 2 and system 3 will be called as guidelines when comparing them with the other participant systems; the 3 human summarizers listed in the column **Peer ID** together with the 18 systems will be called as human-summarizer; the 1 model summarizer will be called as model-summarizer, although it is a human-summarizer itself. With regard to the summaries produced by the above summarizers, the summary means the summarization result from any of the 18 systems, including guidelines; the human-summary means the result come from any of the 3 human-summarizers for each document set; the model-summary means the result from the 1 model-summarizer for each document set.

From the original evaluation results issued by NIST, the following tables, table 1~3, has been produced to provide the analysis of this system, AMDS_hw, which was marked as system 6 (Peer ID = 6) in the evaluation results and in the tables.

3.1 Evaluations on Mean Coverage and Mean Length-Adjusted Coverage

In the table 1, 'MC Rank' means the rank of the described system, system 6, on the scores of 'Mean Coverage' amongst the 18 systems, excluding the 3 human-summarizers for each document set; 'HM Ahead' means the number of human-summarizers whose ranks are before the described system, system 6; 'Max MC' means the maximum score of Mean Coverage for all 18 systems including the 3 human-summarizers; 'Sys*N* MC', *N*=2,3 or 6, means the Mean Coverage score for the corresponding system, 2,3 or 6, as system 2 and system 3 were taken as guidelines in Task 2. Similarly, in table 2, 'MC' is replaced by 'MlajC', which means the score of Mean Length-Adjusted Coverage, and of the corresponding columns refer to the evaluation results on the Mean Length- Adjusted Coverage as well.

In these tables, **Mean Coverage** (MC) and **Mean Length-Adjusted Coverage** (MlajC) are two key evaluation parameters, which show how much content of model-summary is covered by the summary produced by system 6 for each document set. The difference between the two parameters is that Mean Length-Adjusted Coverage considered the brevity of summaries in calculation.

In column **MC Rank**, it can be seen that, amongst the 18 systems, system 6 is ranked 24 times in the first half, before 9th position, 20 times in the top six and 4 times in the 1st position, within the 30 test-document sets. While in the column **MlajC Rank**, after taking considerations of brevities of the system-produced summaries, system 6 still ranked 22 times in the first half, before 9th position, 17 times in the top six and 4 times in the 1st position, within the 30 test

document sets.

From columns **MC Rank** and **HM Ahead**, it was found that the human-summaries were still much better than the summaries produced by system 6 in most of the 30 document sets. For example, in the 30 document sets, 22 times for all the three human-summaries, 6 times for two of the three human- summaries and twice for only one of the three human-summaries for each document set had higher ranks than system 6. The results were the same as above when the consideration of brevity was given in analysis on columns **MlajC Rank** and **HMlaj Ahead**, as the length of each summary produced by system 6 is around the Target Size, 100-word, which avoided many extra penalties of score zero for brevity.

Some comparisons had to be made between system 6 and the two guidelines, system 2 and 3. From columns **Sys2 MC, Sys3 MC** and **Sys6 MC**, in 20 of the 30 document sets the Mean Coverage score of system 6 was better than both guidelines, while the other 10 sets, either or both of the Mean Coverage scores of the two guidelines was/were better than system 6. The 10 document sets were D30010, D30050, D31001, D31010, D31022, D31027, D31031, D31033, D31041, and D31050. After the data in columns **Sys2 MlajC, Sys3 MlajC** and **Sys6 MlajC** has been compared, D30010 and D31010 are eliminated from above list.

In most conditions, the 3 human-summaries are closer to the model-summary than the summaries produced by the 18 systems for each document set. As **Max MC** is the maximum value of the Mean Coverage values of all 18 systems and human-summarizers for each document set, it can be deduced how much the gap is between system 6 and the best human-summarizer, when compare the data in columns **Max MC** and **Sys6 MC**.

The value of Max MC provided us a possible value to achieve for each document set, but there was a wide gap between system 6 and the best human-summarizer. Even in the document sets, D30012, D30016, D30020 and D31009, amongst which system 6 was ranked as first on Mean Coverage, the gap is still obvious. Similar conclusion could be obtained from comparisons between columns **Max MlajC** and **Sys6 MlajC.**

Table 1. Evaluation Results on Mean_Coverage for System 6

Document Set No.	Peer Size	Peer ID	MC Rank	HM Ahead	Max MC	Sys6 MC	Sys2 MC	Sys3 MC
D30003	98	6	4	3	0.655	0.418	0.145	0.345
D30005	105	6	6	3	0.714	0.257	0.057	0.171
D30010	112	6	4	3	0.718	0.388	0.388	0.388
D30012	98	6	1	3	0.444	0.378	0	0.222
D30016	103	6	1	1	0.545	0.509	0.364	0.127
D30020	101	6	1	2	0.58	0.42	0.02	0.38
D30025	91	6	4	3	0.429	0.243	0.129	0.186
D30028	99	6	3	3	0.74	0.4	0.32	0.3
D30034	97	6	2	3	0.500	0.32	0.14	0.22
D30040	99	6	6	2	0.56	0.3	0.12	0.24
D30042	98	6	3	3	0.82	0.42	0	0.18
D30044	103	6	9	3	0.517	0.3	0.117	0.2

D30048	99	6	3	2	0.475	0.275	0.1	0.025
D30050	94	6	6	3	0.371	0.143	0.286	0.057
D30051	107	6	4	1	0.74	0.58	0.4	0.48
D30056	116	6	7	3	0.4	0.222	0.178	0.089
D31001	100	6	13	3	0.667	0.2	0.3	0.133
D31002	103	6	4	2	0.68	0.28	0.14	0.2
D31009	104	6	1	2	0.6	0.4	0.311	0.133
D31010	85	6	6	3	0.527	0.164	0.218	0.127
D31011	108	6	9	3	0.82	0.38	0	0.32
D31013	91	6	10	3	0.54	0.2	0.18	0.04
D31022	98	6	16	3	0.75	0.117	0.25	0.2
D31027	99	6	11	3	0.7	0.175	0.275	0.075
D31028	100	6	2	3	0.617	0.367	0	0.217
D31031	99	6	9	3	0.422	0.022	0.067	0.022
D31033	104	6	11	3	0.444	0.133	0.044	0.178
D31038	98	6	5	3	0.667	0.133	0.089	0.111
D31041	97	6	6	2	0.3	0.1	0.025	0.2
D31050	102	6	11	3	0.283	0.067	0.05	0.15

Table 2. Evaluation Results on Mean_LengthAdjusted_Coverage for System 6

Document Set No.	Peer Size	Peer ID	MlajC Rank	HMlaj Ahead	Max MlajC	Sys6 MlajC	Sys2 MlajC	Sys3 MlajC
D30003	98	6	4	3	0.439	0.283	0.08	0.197
D30005	105	6	7	3	0.476	0.163	0.035	0.133
D30010	112	6	6	3	0.478	0.231	0.21	0.209
D30012	98	6	1	3	0.301	0.255	0	0.128
D30016	103	6	1	1	0.364	0.33	0.215	0.094
D30020	101	6	1	2	0.387	0.277	0.011	0.23
D30025	91	6	4	3	0.288	0.173	0.076	0.102
D30028	99	6	4	3	0.493	0.268	0.165	0.185
D30034	97	6	2	3	0.333	0.218	0.088	0.13
D30040	99	6	8	2	0.374	0.201	0.119	0.14
D30042	98	6	5	3	0.547	0.284	0	0.106
D30044	103	6	11	3	0.344	0.194	0.07	0.117
D30048	99	6	4	2	0.317	0.186	0.056	0.014
D30050	94	6	7	3	0.248	0.101	0.172	0.031
D30051	107	6	4	1	0.555	0.361	0.22	0.324
D30056	116	6	8	3	0.274	0.128	0.106	0.053
D31001	100	6	11	3	0.444	0.133	0.165	0.08
D31002	103	6	5	2	0.453	0.181	0.075	0.112
D31009	104	6	1	2	0.4	0.256	0.174	0.079
D31010	85	6	5	3	0.352	0.123	0.112	0.087
D31011	108	6	11	3	0.547	0.235	0	0.223
D31013	91	6	9	3	0.36	0.148	0.1	0.025

D31022	98	6	17	3	0.5	0.079	0.159	0.12
D31027	99	6	12	3	0.467	0.117	0.148	0.05
D31028	100	6	2	3	0.411	0.244	0	0.13
D31031	99	6	11	3	0.281	0.015	0.038	0.012
D31033	104	6	12	3	0.296	0.085	0.024	0.103
D31038	98	6	7	3	0.444	0.091	0.055	0.085
D31041	97	6	6	2	0.2	0.069	0.015	0.116
D31050	102	6	12	3	0.193	0.044	0.029	0.088

3.2 Evaluations on Counts of Quality Questions with Non-zero answers (CQQN) and Mean of the Quality Question Scores (MQQS)

12 quality questions were asked for the counting of ERRORS in each system/peer-produced summary, while *Qn* [n = 1, 2, … 12] in the evaluation were calculated as below:

Qn = 0, if *NoE* = 0;

Qn = 1, if 1<*NoE*<5;

Qn = 2, if 6<*NoE*<10;

Qn = 3, if *NoE*>11.

Here, *NoE* meant the Number of Errors for this quality question in the summary.

Count of Quality Questions with Non-zero answers (**CQQN**) indicated the total number of non-zero answers of the 12 quality questions. **Mean of the Quality Question Scores** (**MQQS**) was the mean value of Count of Quality Questions with Non-zero answers and calculated with

$$MQQS = \frac{TQQS}{CQQN},$$

Here, *TQQS* meant the Total Quality Questions Score with Non-zero answers,

$$TQQS = \sum_{n=1}^{12} Qn \cdot$$

In order to examine the overall performances of all 18 systems, three new parameters have been calculated in the following table 3, **Total_CQQN**, **Mean_CQQN** and **Mean_MQQS**. **Total_ CQQN** is the sum-up of values of **CQQN** across 30 document sets for each of all 18 systems; **Mean_CQQN** is the mean of values of **CQQN** across 30 document sets for each of all 18 systems; **Mean_MQQS** is the mean of values of **MQQS** across 30 document sets for each of all 18 systems and have been calculated in the following formula:

$$Mean\_MQQS = \frac{\sum_{30}(CQQN * MQQS)}{30} \cdot$$

Table 3. Evaluations on Mean of the Quality Question Scores (MQQS)

System ID	Total CQQN	Mean CQQN	Mean CQQS	Mean MQQS Rank
2	40	1.333333333	1.333333333	4
3	87	2.9	2.966666667	16
6	**42**	**1.4**	**1.433333333**	**5**
10	55	1.833333333	1.866666667	8
11	29	0.966666667	0.966666667	1
12	48	1.6	1.6	6
13	96	3.2	3.299966667	17
14	76	2.533333333	2.6334	13
15	236	7.866666667	10.53323333	18

16	38	1.266666667	1.266666667	3
17	54	1.8	1.833333333	7
18	77	2.566666667	2.566666667	14
19	75	2.5	2.566666667	12
20	70	2.333333333	2.566666667	10
21	80	2.666666667	2.699966667	15
22	36	1.2	1.2	2
23	70	2.333333333	2.433366667	11
26	60	2	2.033333333	9

4. Conclusions

From the above analysis, the newly devised system, system 6, showed a good performance in Mean Coverage, Mean Length-Adjusted Coverage and Quality Question Score compared with other participating systems, which proves that the new algorithm of our multi-document summarization system is working well. But the human-summarizers and model-summarizers are still better than system 6 in the evaluation results. The newly proposed system is still under development and this exercise was extremely useful for revealing the need of improvement in the phrase-level comparison and Cluster-Reduction module.

5. Future Research

The exercise has helped in identifying several areas for improving the performance of system 6:
(1) further analysis on why the performances of system 6 are so different among the given document sets, are the reasons related with the content or styles of the texts in each document set?
(2) how to increase the number of units, reduce the content redundancy and increase the coverage of each unit in every summary.

References

[Grinberg et al. 1995] D. Grinberg, J. Lafferty and D. Sleator, 1995, *A robust parsing algorithm for link grammars*, Proceedings of the 4th International Workshop on Parsing Technologies, Prague, September, 1995.

[Lafferty et al. 1992] J. Lafferty, D. Sleator, and D. Temperley (1992), *Grammatical Trigrams: A Probabilistic Model of Link Grammar*, Proceedings of the AAAI Conference on Probabilistic Approaches to Natural Language, October, 1992.

[Lambrecht, 1996] Lambrecht, K., 1996, *Information structure and sentence form: topic, focus, and the mental representations of discourse referents.* Cambridge University Press. ISBN: 0-521-58704-2(pbk).

[Li & Thompson, 1976] Li, Charles and Thompson, Sandra A., 1976. '*Subject and topic: a new typology of language.*' In Li (ed.) 1976, P457-490.

[Miller *et al.* 1990] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, **3**(4):235–312, 1990.

[Sleator et al. 1991] D. Sleator and D. Temperley (1991), *Parsing English with a Link Grammar*, Carnegie Mellon University Computer Science technical report CMU-CS-91-196, October 1991.

[Sleator et al. 1993] Daniel Sleator and Davy Temperley (1993), *Parsing English with a Link Grammar*, Third International Workshop on Parsing Technologies, August 1993.

[Sleator et al. 2000] D. Sleator, D. Temperley & J. Lafferty, http://www.link.cs.cmu.edu/link/, *Link Grammar Parser.*