User-focused Multi-document Summarization with Paragraph Clustering and Sentence-type Filtering

Yohei Seki‡, Koji Eguchi†,‡, and Noriko Kando†,‡

† National Institute of Informatics, Japan

‡ Department of Informatics, The Graduate University for Advanced Studies, Japan

User-focused Multi-document Summarization

OBJECTIVES

The objective of our participatin in NTCIR-4 TSC3 has two goals. The first goal is ``User-focused interactive summarization for topical requirements.'' We took paragraph clustering-based summarization approach. The second goal is to produce knowledge-focused summaries. We took sentence-type filtering approach. These approaches was evaluated in extract evaluation in Table 1 and responsiveness evaluation to questions in Table 2. Our system ID was F0301. In F0301(a), we took the ``group-average" clustering algorithm. In F0301(b), the`` Ward's method" was used.

Topical/Situational Information Requirements

The goal of Multi-Document Summarization is defined as ``*to extract content from a collection of related documents and present the most important content sensitive to the user's needs*" [Mani, 2001]. With queries, the user's requirements requirements can be expressed as subtopics.

For topical requirements, we took the approach for multi-document summarization with document clustering techniques. This idea was based on the concept to apply cluster hypothes to produce non-redundant summaries.

In the multi-genre document summarization case, we also focused on the situational relevance, as shown in Figure 1. We defined three summary types: fact-reporting, opinion-oriented, and knowledge-focused.

In NTCIR-4 TSC3, we surveyed the effectiveness for knowledge-focused summarization for responsiveness to questions.

Table 1. Extract Evaluation

SYSTEM－ID	Short		Long	
	Cov.	Prec.	Cov.	Prec.
F0301(a)	0.315	0.494	0.355	0.554
F0301(b)	0.372	0.591	0.363	0.587
F0303(a)	0.222	0.314	0.313	0.432
F0303(b)	0.293	0.378	0.295	0.416
F0304	0.328	0.496	0.327	0.535
F0306	0.283	0.406	0.341	0.528
F0307	0.329	0.567	0.391	0.68
F0308	–	–	–	–
F0309	0.308	0.505	0.339	0.585
F0310	0.181	0.275	0.218	0.421
F0311	0.251	0.476	0.247	0.547
LEAD	0.212	0.426	0.259	0.539

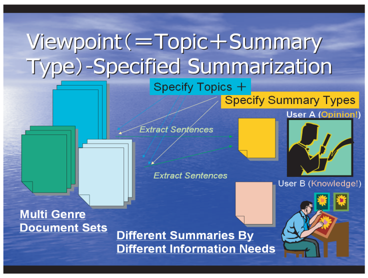

Figure 1. User-focused Multi-document Summarization

Table 2. Responsivenss Evaluation to Questions

SYSTEM－ID	Short		Long	
	exact	edit	exact	edit
F0301	0.394	0.677	0.399	0.706
F0303	0.257	0.556	0.266	0.602
F0304	0.367	0.653	0.356	0.677
F0306	0.342	0.614	0.327	0.63
F0307	0.439	0.71	0.442	0.751
F0308	0.321	0.601	0.313	0.611
F0309	0.39	0.684	0.356	0.633
F0310	0.133	0.427	0.201	0.549
F0311	0.304	0.579	0.308	0.628
LEAD	0.3	0.589	0.275	0.602
HUMAN	0.461	0.716	0.426	0.721

Interactive Summarization with Paragraph Clustering

In order to treat the two aspects of information requirements, we implemented an interactive summarization system, as shown in Figure 2.

Subtopic-focused Summarization

We implemented a subtopic-focused summarization with a document clustering technique. We segmented the source documents into paragraph units and clustered them using the output summary size.

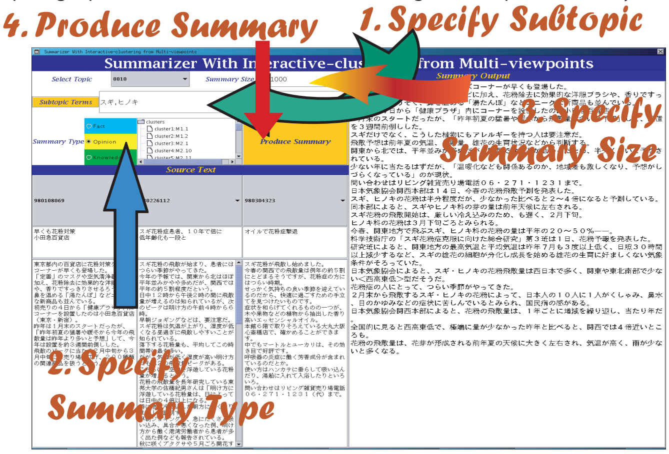

Figure 2. Summarizer With Interactive-clustering from Multi-Viewpoints (SWIM)

We compared the effectiveness our clustering-based summarization techniques from six cluster options: 1: cluster units, 2: features and cluster similarities, 3: clustering algorithm, 4: cluster size, 5: sentence extraction clues, and 6: queries.
We found that the method with the following options best peformed:
1: paragraph cluster units, 2: unnormalized TF features and euclidean distance, 3: Ward's method, 4: # of clusers by ×1 - ×1.5 according to # of senteces extracted, 5: term frequencies and title weighting, and 6: with queries.
Without queries, the responsiveness to questions decreased 0.02-0.03 points.

Sentence-type Filtering

For extracting sentences from each cluster, we proposed new approach called ``sentence-type filtering.'' This approach tries to extract at most two sentences from each cluster with checking redundant sentence types.

Test Collection for Multi-Genre Document Summarization

We evaluated the multi document summaries in terms of four genre features. We made test collections, there were 22 topics for three type multi-document summaries. Topics were shown in Table 1.

Table 3. Test Collection for Multi-Viewpoint Summarization

Task		Source Articles	
ID	Topic	# of Characters	# of Articles
S010	European monetary union	20530	10
S020	Annual pension	21704	10
S030	Accounting fraud	21207	9
S040	Itoman fraud case	20647	10
S050	Removal of deposit insurance	19251	11
S060	Digital cellular phone	20353	11
S070	Guidelines for Japan–U.S. defense cooperation	20687	9
S080	Kosovo	20583	11
S090	Strategic arms reduction	15499	8
S100	Brain–death diagnosis	21052	7
S110	Juvenile proceedings	20967	11
S120	Freedom of Information Act	16953	8
S130	Donor card	15902	10
S140	Defined contribution pension plan	19131	12
S150	Genetically–engineered foods	20225	12
S160	Organized Crime Control Act	21425	8
S170	Criticality–caused nuclear accident	16935	7
S180	Financial Big Bang	19411	8
S190	Pluthermal	19092	9
S200	Theater Missile Defenses	17323	8
S210	Government–owned company in China	13529	6
S220	Conflict of Nothern Ireland	14241	10

Conclusions

For NTCIR-4 TSC3, we focused on multi-document summarization from two different aspects.
1. Compare parameters for paragraph clustering techniques against topical information requirements:
Ward's Methods, unnormalized TF, Euclidean distance, # of clusters by ×1 – ×1.5 according to # of sentences extracted performed best.
2. Sentence-type filtering to improve the responsiveness to questions:
To extract the most important sentence and ``prospective"-type sentence from each cluster improved responsiveness for several topics.

NTCIR Wrokshop 4 Meeting
June 2-June 4, 2004, National Institute of Informatics