Analysis of Multi-Document Viewpoint Summarization Using Multi-Dimensional Genres

Yohei Seki[1,2], **Koji Eguchi**[2,1], **and Noriko Kando**[2,1]
Department of Informatics, The Graduate University for Advanced Studies (Sokendai)[1]
National Institute of Informatics (NII)[2]
Tokyo 101-8430, Japan
seki@grad.nii.ac.jp eguchi@nii.ac.jp kando@nii.ac.jp

Abstract

An interactive information retrieval system that provides different types of summaries of retrieved documents according to each user's information needs can be effective for understanding the contents. The purpose of this study is to build a multi-document summarizer to produce summaries according to such viewpoints. As an exploratory stage of investigation, we examined the effectiveness of genre for source documents to produce different types of summaries. Once a set of documents on a topic is provided to our summarization system, a list of topics discussed in the given document set is presented, so that the user can select a topic of interest from the list as well as the summary type, such as opinion-oriented, fact-reporting or knowledge-focused, according to their requirements. We assume a relationship between a summary type and human recognition of information types included in the source: a document genre. We also analyzed the results of the multi-document summarization using automatic genre classification to reveal the association between genre dimensions and the summary types.

Introduction

Our goal is to summarize multiple documents using specified viewpoints. In text summarization research, hand-created summaries tend to differ between summary writers.This is caused by the differences in viewpoints of users when accessing information, because summary writers assumed ideal users would read their summaries. Query-biased Summarization (SUMMAC[1], 1998) has been proposed as a method to generate summaries by focusing on the topics related to a query in the context of information retrieval. Viewpoints, however, relate not only to the topics but also to types of information such as opinion, evaluation, commentary or fact-reporting.

In the Document Understanding Conferences (DUC) 2003[2], viewpoint summary was tested as one task. Angheluta et al. (Angheluta, Moens, & De Busser 2003) tried viewpoint summarization with topic segmentation, but its effectiveness was not fully investigated. Viewpoint is a concept that is not only based on the topics which the summary writer focusing on but can also be extended to include other aspects such as type of information. For multi-document summarization, more viewpoints exist in the original documents than for single document summarization. Such multi-document summarization techniques can be positioned as one of the promising applications to concisely present search results, meeting the information needs of users, in information retrieval systems.

In this paper, "viewpoint" in the summarization is defined as the combination of "topic" and "summary type", such as "opinion-oriented", "knowledge-focused", or "fact-reporting". Our goal is to summarize multiple documents by specifying viewpoints. We applied a topical classification methodology with clustering techniques to identify topics discussed in a set of documents, then identified the most representative topical words for each cluster. For the summary type, we used three summary types: "fact-reporting", "opinion-oriented", or "knowledge-focused", where the discrimination is based on the types of information that the user requires. For the second summary type, opinion-oriented summarization for multi-perspective question-answering (Cardie *et al.* 2003) have attracted much interest.

To produce multi-document summaries according to the point-of-view, we investigated the advantage of using genre information. "Genre" here means document type such as "personal diary", "report", etc. Genre is defined as "an abstraction based on a natural grouping of documents written in a similar style and is orthogonal to topic," as in (Finn, Kushmerick, & Smyth 2002). In summarization research, factual information and topics have been focused on, but users might require subjective information such as opinion, evaluation, and prospects.

In this paper, we described the "genre feature" of each document by a combination of four dimensions based on Biber's multi-dimensional register analysis (Biber, Conrad, & Reppen 1998, chap. 6.4).

This paper consists of six sections. In the next section, we detail experiments to compare several types of genre dimensions in our system. Then, we describe our methods of automatic genre classification and multi-document summarization with genre classification. The analysis of results considering summary types follows them. Finally, we present our conclusions.

[1]TIPSTER Text Summarization Evaluation Conference

[2]http://www-nlpir.nist.gov/projects/duc/

Experiments: Effectiveness of Genre Feature for Multi-Document Viewpoint Summarization

As a preliminary study in summarization from viewpoints that are represented as combinations of topics and summary types, we investigated the effectiveness of "Genre" for multi-document viewpoint summarization.

Summary Data

In this experiment, we used the Japanese summary data of NTCIR-3 TSC2 (Oyama, Ishida, & Kando 2003; Kando 2003; Fukusima, Okumura, & Nanba 2003)[3]. There are 30 Japanese document sets in NTCIR-3 TSC2 formal run dataset[4], which were selected from Mainichi newspaper articles in 1998 and 1999. Three reference summaries were created for each document set by three different professional captionists. For TSC2 summaries, the document sets and their topics are given, but instructions which summary types to produce were not given to the summary writers (captionists). Therefore, each captionist might produce different type summaries from the common document set.

Types of Document Sets

In the NTCIR-3 TSC2, document sets are categorized as either: "news stories only" or "news stories and editorials". Of the 30 sets, 21 are categorized as the former, and nine are the latter.

Genre Feature

IPTC (International Press Telecommunications Council)[5] proposed a set of genres for news delivery. These, however, include more than 40 categories from "opinion" and "background" down to resource-type information, such as "music" and "raw sound", or type of news source, such as "press release". The categories were based on several different classification criteria, but they were allocated in only one dimension. In general, criteria for categorizing genre are complex and hard to annotate. This framework was not appropriate to test the effectiveness of summarization.

Therefore, in this research, genre categories were formalized with a combination of dimensions. The multiple dimensions were based on Douglas Biber's proposal (Biber, Conrad, & Reppen 1998, pp. 135–155). The merits of using this idea are as follows.

- The effectiveness of each dimension is explicit.
- New genre dimensions can be added easily without changing the entire framework.
- Annotation rules were expected to be simple.

The five basic dimensions in Biber's framework were:

1. Elaborated/Situation-Dependent Reference
2. Overt Expression of Argumentation
3. Impersonal/Non-Impersonal Style.
4. Narrative/Non-Narrative Discourse
5. Involved/Information Production

[3] http://research.nii.ac.jp/ntcir/

[4] These "document sets" were called "clusters" in DUC 2003.

[5] http://www.iptc.org/site/subject-codes/genre.html

Table 1: Kappa Coefficients: Intercoder Consistency

Genre Dimension	Pair of Assessors			
	(a1,a2)	(a1,a3)	(a2,a3)	Avg.
Situation-Depend	0.597	0.607	0.587	0.627
Argumentation	0.685	0.440	0.394	0.506
Impersonal	0.729	0.585	0.447	0.587
Fact-Reporting	0.606	0.724	0.484	0.605

Of Biber's dimensions, the fifth could not be discriminated using the test data because all test documents were categorized to "information production" in this dimension. We used the remaining four dimensions.

1. Situation-Dependency: marked documents according to the degree of coincidence between their publishing time and event time.
2. Argumentation: marked documents according to the degree of persuasion and the author's point of view.
3. Impersonal Styles: marked documents according to frequent passive constructions, and so on.
4. Fact-Reporting: marked documents that report facts in an inverse-pyramid discourse structure in newspaper articles.

In this paper, **the "genre" of each document was defined with a combination of these four dimensions.**

Genre Classification

In this section, we explain the consistency of genre dimensions brought by manual annotation and by automatic genre classification.

Manual Annotation for Genre Feature

To assess the four dimensions of the genre feature, we took the following steps:

i) The first author annotated values (1/0) of the four genre dimensions each in 156 documents (TSC2, 19 topic sets) manually and also produced a first version of instructions for annotation.

ii) Three assessors (not the first author) annotated several documents with the instructions for genre coding. We interviewed them about the instructions. The instructions were modified to reduce the ambiguity of annotation.

iii) Two other assessors annotated 156 documents with the modified instructions.

After these steps, the κ coefficient value showed good agreement between judges, as shown in Table 1. These results suggest that human judgment can be reliable.

Automatic Genre Classification

The TSC2 test collection for multi-document summarization consisted of 30 topic sets. To estimate the effectiveness of the genre dimensions and thus improve the selection of relevant sentences to be included in a summary for each summary type, we investigated the improvement of the coefficient of determination in multiple linear regression analysis based on the features described in the section of "Genre Feature". We could not estimate the effectiveness of the four genre dimensions in topic sets that included less than four documents. As a result, TSC2 contained 10 topic sets which could not be estimated for effectiveness. We found a further three topics that could not be discriminated based on the four genre dimensions. Therefore, we divided the 30 topic sets into two groups: 17 topics (148 documents) and 13 topics (69 documents).

We applied one of the major machine learning techniques that is often used in text categorization: support vector machines (SVM) (Joachims 2002). We used the 17 topics as test sets and the 13 topics as training sets. Because the number of training documents was limited, the first author annotated an additional 107 documents for the four genre dimensions each and added these to training sets. For genre classification, we defined a set of features, as follows, hereinafter referred to as "small features".

- Five structural features: author signature, section, photo, figure, and news source.
- Nine statistical features ('#' is defined as "numbers"): # of characters, Type/Token Ratio, # of sentences, # of opinion sentences, # of prospect sentences, # of background sentences, # of conjunctions, # of quote parentheses, and average sentence length.
- 60 function phrases (which relate to opinion, prospect, and background information).
- 93 symbols (which include several punctuation related symbols).

These selected features totaled only 167. However, they were extremely effective for genre classification. The results for genre classification are shown in Table 2 below. We compared our method with conventional text categorization methods (based on frequency of 13,825 content words and content words, hereinafter referred to as "TC features").

Table 2: Accuracy of Genre Classification

Features	Dimension			
	Sitn.	Argn.	Impersonal	Facts
Small Features	79.1	86.5	90.5	88.5
TC Features	63.5	85.8	93.2	89.2

Sitn. = Situation Dependency
Argn. = Argumentation

Readers may think TC (Text Categorization) features are effective for impersonal or fact-reporting genre dimensions, however, we do not think so for the reason that SVM techniques with TC features did not discriminate impersonal or fact-reporting genre dimensions at all, as shown in Table 3.

Multi-Document Summarization with Genre Classification

We had already developed a multi-document summarization system and now enhanced this system with genre classification. With the TSC2 multi-document summary set, "coverage" and "precision" can be computed automatically using tools and correct sentences (the definition of "coverage" and "precision" were given in (Hirao *et al.* 2004); the correct sentences data were the same as used in (Hirao *et al.* 2003); the tools were provided by NTCIR-4 TSC3 committee [6]).

[6] http://www.lr.pi.titech.ac.jp/tsc/tsc3-en.html

Table 3: Genre Classification Results

Genre Dimension	Features	System/Manual			
		p/p	p/n	n/p	n/n
Sitn.	Small Features	105	28	3	12
Sitn.	TC Features	65	11	43	29
Argn.	Small Features	6	5	15	122
Argn.	TC Features	0	0	21	127
Impersonal	Small Features	129	5	9	5
Impersonal	TC Features	138	10	0	0
Facts	Small Features	127	12	5	4
Facts	TC Features	132	16	0	0

p/p : System and human classified as positive.
p/n : System classified as positive, human as negative.
n/p : System classified as negative, human as positive.
n/p : System and human classified as negative.

Sitn. and Argn. have the same meaning as in Table 2.

Table 4: Coverage and Precision for Multi-Document Summarization System

	Reference Summaries					
	S1		S2		S3	
	cov.	pre.	cov.	pre.	cov.	pre.
All topics	0.346	0.388	0.228	0.283	0.246	0.336
Usable	0.257	0.317	0.209	0.265	0.178	0.257
Unusable	0.463	0.481	0.253	0.307	0.335	0.440

cov. = coverage
pre. = precision
Usable = 17 Genre Distinction Usable Topics
Unusable = 13 Genre Distinction Unusable Topics

Without genre classification, the coverage and precision results for summaries are shown in Table 4. The TSC2 test set contains three reference summaries; henceforth, three results were shown. Of 30 topics, 17 topics each contained more than five documents with different genre dimensions. Of the remaining 13 topics, 10 contained less than four documents and three contained no different genre dimensions.

These results show that multi-document summarization (without genre classification) was effective for the 13 topics, but not effective for the 17 topics. For the 17 topics, we applied genre classification and constructed summaries from documents with positive or negative features for four genre dimensions. The result is shown in Table 5.

Table 5 shows that the genre dimensions brought improve-

Table 5: Improvement of Coverage and Precision with Genre Classification

Genre	Reference Summaries					
	S1		S2		S3	
	cov.	pre.	cov.	pre.	cov.	pre.
base	0.257	0.317	0.209	0.265	0.178	0.257
G1-p	0.236	0.292	0.195	0.248	**0.184**	0.269
G1-n	0.066	0.205	0.055	0.185	0.049	0.173
G2-p	0.063	0.197	0.067	0.181	0.064	0.211
G2-n	0.243	0.311	0.207	0.262	**0.190**	0.277
G3-p	**0.261**	0.323	**0.212**	0.269	**0.186**	0.254
G3-n	0.016	0.148	0.037	0.153	0.049	0.194
G4-p	0.243	0.307	**0.216**	0.270	0.173	0.246
G4-n	0.024	0.062	0.019	0.044	0.017	0.052

G1 = Situation-Depend G2 = Argumentation
G3 = Impersonal G4 = Fact-Reporting

Table 6: The Number of Summary Types for Short Summaries

Summary Type	Captionists		
	C1	C2	C3
Fact	6	9	11
Opinion	1	3	5
Knowledge	10	5	1

C1-C3: Captionists who produced reference summaries S1-S3.

Table 7: Effective Genre Dimensions for Improvement of Coverage

Summary Type	Genre	Captionists		
		C1	C2	C3
Fact	Base	0.246	0.202	0.074
	Fact	0.199	**0.225**	0.074
Opinion	Base	0.066	0.147	0.073
	Argn.	0	0.037	**0.089**
Knowledge	Base	0.209	0.135	0.036
	Fact	0.202	**0.138**	0.036

ments in coverage and precision. However, on the average, these improvements were hardly significant. Therefore, we discriminated summary types and investigated this result, which will be discussed in the next section.

Summary Types

In the last section, we discussed the improvement effect for overall summaries. In this research, we defined three summary types, which are orthogonal to topical elements, as a basis for abstracting information needs:

1. Fact-reporting: Summaries focused on events, which happened in real time or in past times.
2. Opinion-oriented: Summaries focused on the authors' opinions or experts' opinions by third parties.
3. Knowledge-focused: Summaries focused on definitional or encyclopedic knowledge.

In TSC2, captionists were not given instructions for summary types. Therefore, different summary types were produced for the same document sets. With the manual analysis of reference summaries, we classified 17 short summaries into three categories.

Personal Preference

Reference summaries were produced by three captionists. The first captionist tended to produce "knowledge-focused" Summaries, while the third captionist tended to produce "opinion-oriented" summaries. The number of summary types produced by each captionist is shown in Table 6.

Analysis of Each Summary Type

Coverage for each dimension and summary types is shown in Table 7. We observed the results for the effect of summary types as follows:

- For fact-reporting summaries, the fact-reporting genre was effective for summaries by captionists 2.
- For opinion-oriented summaries, the argument dimension was effective for summaries by captionist 3.
- For knowledge-focused summaries, the fact-reporting genre dimension was also effective for captionist 2.

We can conclude that captionists 2 and 3 selected documents by genre dimension for summarization.

Conclusion

In this paper we tested the effectiveness of genre, which gives us specified viewpoints. Four genre dimensions were annotated by three assessors and the κ coefficient proved the reliability of these dimensions. The genre dimensions were automatically categorized using functional phrases, and the accuracies of genre classification were over 80 %. Although we tested on Japanese, we think this approach is applicable to other languages, i.e. English.

We also applied genre classification results for multi-document summarization application. We take the direct approach to construct summaries only from positive sets of single genre dimension. Because the important sentences in the articles negative for the dimension were all abandoned, the improvement of coverage was a hard task. In spite of this problem, we found several improvement results which were different according to summary types and captionists.

Acknowledgments

This work was partially supported by the Grants-in-Aid for Scientific Research on Priority Areas of "Informatics" (#13224087) from the Ministry of Education, Culture, Sports, Science and Technology, Japan. This research was also partially supported by the Kayamori Foundation of Informational Science Advancement.

References

Angheluta, R.; Moens, M. F.; and De Busser, R. 2003. K. u. leuven summarization system - duc 2003. In *Workshop on Text Summarization (DUC 2003) in conj. with the 2003 Human Language Technology Conference (HLT/NAACL 2003)*.

Biber, D.; Conrad, S.; and Reppen, R. 1998. *Corpus Linguistics - Investigating Language Structure and Use -*. Cambridge Approaches to Linguistics. Cambridge University Press.

Cardie, C.; Wiebe, J.; Wilson, T.; and Litman, D. 2003. Combining low-level and summary representations of opinions for multi-perspective question answering. In *AAAI Spring Symposium on New Directions in Question Answering*, 20–27.

Finn, A.; Kushmerick, N.; and Smyth, B. 2002. Genre classification and domain transfer for information filtering. In Crestani, F.; Girolami, M.; and van Rijsbergen, C. J., eds., *Proc. of ECIR 2002 Advances in Information Retrieval, 24th BCS-IRSG European Colloquium on IR Research, Lecture Notes in Computer Science*, volume 2291. Glasgow, UK: Springer-Verlag. 353–362.

Fukusima, T.; Okumura, M.; and Nanba, H. 2003. Text Summarization Challenge 2: Text summarization evaluation at NTCIR Workshop 3. In *(Oyama, Ishida, & Kando 2003)*.

Hirao, T.; Takeuchi, K.; Isozaki, H.; Sasaki, Y.; and Maeda, E. 2003. SVM-Based Multi-Document Summarization Integrating Sentence Extraction with Bunsetsu Elimination. *IEICE Transactions on Information and Systems* E86-D(9):1702–1709.

Hirao, T.; Okumura, M.; Fukusima, T.; and Nanba, H. 2004. Text Summarization Challenge 3: Text summarization evaluation at NTCIR Workshop 4. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering, and Summarization*.

Joachims, T. 2002. *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers.

Kando, N. 2003. Overview of the Third NTCIR Workshop. In *(Oyama, Ishida, & Kando 2003)*.

Oyama, K.; Ishida, E.; and Kando, N., eds. 2003. *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*. National Institute of Informatics.