Chapter 24

# Multi-Document Viewpoint Summarization Focused on Facts, Opinion and Knowledge

**Yohei Seki**
*Department of Informatics,*
*The Graduate University for Advanced Studies (Sokendai) / National Institute of Informatics*
*National Institute of Informatics, 2-1-2, Hitotsubashi,*
*Tokyo 101-8430, Japan.*
Email: seki@grad.nii.ac.jp

**Koji Eguchi and Noriko Kando**
*National Institute of Informatics / Department of Informatics,*
*The Graduate University for Advanced Studies (Sokendai)*
Email: 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, situation, or purpose of search can be effective for understanding document content. The purpose of this study is to build a multi-document summarizer, "*Viewpoint Summarizer With Interactive clustering on Multi-documents (v-SWIM)*", which produces summaries according to such viewpoints. We tested its effectiveness on a new test collection, *ViewSumm30*, which contains human-made reference summaries of three different summary types for each of the 30 document sets. Once a set of documents on a topic (e.g., documents retrieved by a search engine) is provided to *v-SWIM,* it returns a list of topics discussed in the given document set, so that the user can select a topic or topics of interest as well as the summary type, such as fact-reporting, opinion-oriented or knowledge-focused, and produces a summary from the viewpoints of the topics and summary type selected by the user. We assume that sentence types and document genres are related to the types of information included in the source documents and are useful for selecting appropriate information for each of the summary types. "Sentence type" defines the type of information in a sentence. "Document genre" defines the type of information in a document. The results of the experiments showed that the proposed system using automatically identified sentence types and document genres of the source documents improved the coverage of the system-produced fact-reporting, opinion-oriented, and knowledge-focused

summaries, 13.14%, 34.23%, and 15.89%, respectively, compared with our baseline system which did not differentiate sentence types or document genres.

**Keywords**: multi-document summarization, viewpoint, opinion, genre classification, sentence type.

## 1. Introduction

Our goal is to summarize multiple documents using specified viewpoints. We implemented "*Viewpoint Summarizer With Interactive clustering on Multi-documents (v-SWIM)*" to achieve this goal. In this system a topical classification methodology with clustering techniques was applied to identify topics discussed in a set of documents, and then identify the most representative topical words for each cluster. For the summary type, we used "fact-reporting", "opinion-oriented", or "knowledge-focused" summaries, where the discrimination is based on the types of information that the user requires.

Text summarization is a reduction process of mostly textual information to its most essential points. Mani (2001) stated that multiple-document summarization (MDS) was the extension of single-document summarization toward collections of related documents, and the goal of MDS was to present the most important content to the user in a condensed form and in a manner sensitive to the user's needs. In this chapter, we define summarization as an interactive process to present information according to each user's information needs.

These needs may be different for each user. Human-made reference summaries tend to differ among summary writers (Rath et al., 1961; Lin and Hovy, 2002; Harman and Over, 2004). This is a result of the differences in viewpoints of users when accessing information, because summary writers assume ideal users would read their summaries. "Viewpoint" is defined as "*a mental position or attitude from which subjects or questions are considered."* (Simpson and Weiner, 1991) Query-biased Summarization (SUMMAC, 1998) has been proposed as a method for generating summaries by focusing on the topics related to a query in the context of information retrieval. This is one aspect of summary viewpoints, because topics related to queries could give a mental position from which document sets are considered. Viewpoints, however, relate not only to the topics that the summary reader is focusing on but can also be extended to include other aspects such as the type of information. In the Document Understanding Conference (DUC) 2003, viewpoint summary was tested as one of the tasks, and viewpoint statements about each topic were given to participants to produce summaries. The viewpoints were not explicitly defined but these viewpoint statements included subjective descriptions such as "authority response" or "causal relation of flood". Angheluta et al. (2003) tried viewpoint summarization with topic segmentation, but its effectiveness was not fully investigated.

The purpose of this study is to build a multi-document summarizer that produces summaries according to viewpoints based on user's information needs. In this paper, "viewpoint" in the summarization is defined as the combination of "topic" and "summary type", such as "fact-reporting", "opinion-oriented", or "knowledge-focused".

The distinction of topics and information types can be found in *question taxonomies* (Pomerantz, 2002), which discriminates between subjects (main topics) of questions and functions of expected

answers, or in *relevance dimensions,* as topical and situational relevance (Borlund, 2003). The above mentioned summary types are also related to Pomerantz (2002, p.70), which surveyed question types for long answers, e.g., definition, example, comparison, and causal antecedent:

1. Fact-reporting summary: event example, causal antecedent or consequence, object or resource
2. Opinion-oriented summary: expectation, judgment, motivation, goal orientation
3. Knowledge-focused summary: definition, comparison, interpretation, assertion

The second summary type, opinion-focused summarization for multi-perspective question-answering (Cardie et al., 2003), has attracted much interest. For the third summary type, an extraction-based approach for definitional questions has been proposed (Xu et al., 2004). This research focused on information extraction techniques based on surface linguistic features and question profiles. In contrast, we focus on information types and explicitly use sentence type and document genre information.

To produce summaries from multiple documents according to the point-of-view, we investigated the advantage of using document genre information. "Document genre" here means document type such as "personal diary" or "report", and is defined as a recognizable form of communization in a social activity (Bazerman, 2004). Document genre is also defined as "an abstraction based on a natural grouping of documents written in a similar style and is orthogonal to topic," as in (Finn et al., 2002). In this chapter, we use "document genre" as a concept that defines the information type described in a document. Researchers in summarization have focused on factual information and topics, but users might require subjective information such as opinions, evaluations, and prospects that are mentioned in the source documents. 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, 2002).

Spärck-Jones (1999) proposed a model of summary factors which are formulated as input factors, purpose factors, and output factors[1]. This chapter focuses on the associations between the purpose factors such as "user's situation in which the summary is used" or "user's intention in information", "the form of the source" as an input factor, and "the expression of the summary" as an output factor. Such association between a user's intention in retrieving information and output summary has not yet been surveyed or proposed clearly.

This chapter consists of six sections. In the next section, our experiment overview comparing several types of multi-document summaries is described. Then, our methods of sentence-type annotation and automatic genre classification are detailed in Sections 3 and 4. The experimental results then follow. Finally, we present our conclusions.

## 2. Experiment Overview: Multi-Document Viewpoint Summarization with Summary Types

To clarify viewpoints that are represented as combinations of topics and summary types, we investigated the effectiveness of using "information type" to discriminate summary types based on information needs for multi-document summarization. In this research, two kinds of information

[1] http://duc.nist.gov/RM0507/ksj/factors

types are defined: sentence type and document genre (text type). They are detailed in Sections 3 and 4. In this section, the experiment overview is described.

### 2.1 Experiment: Summary Types for Multi-Document Summarization

We suppose that users recognize information type from their own viewpoint for multi-document summarization. In order to test this hypothesis, we constructed a summary test collection called *ViewSumm30* and tested the effectiveness of the proposed summarization algorithm which differentiated summary types. The human-made reference summaries in *ViewSumm30* were produced with explicit instructions for summary writers to focus on each of the three summary types: fact-reporting, opinion-oriented, and knowledge-focused. This process is detailed in Section 2.2. On this test collection, we tested our baseline multi-document summarization algorithm without differentiating summary types, as well as our proposed algorithm with differentiating summary types using sentence-type annotation and genre classification of the source documents. For the three summary types, we changed the weighting parameters to extract sentences with genre features and sentence-type. Then, coverage and precision for human-made reference summaries was computed. "Coverage" and "precision" are proposed by (Hirao et al., 2004) as metrics to evaluate effectiveness of sentence extraction against reference summaries and were used in the NTCIR-4[2] Text Summarization Challenge (Kando, 2004; Hirao et al., 2004). Finally, the genre dimensions and sentence types effective for MDS of each of the summary types were discussed. The results are detailed in Section 5.

### 2.2 Summary Data with Three Summary types

In this experiment, the authors made a summary test collection, *ViewSumm30.* Like the test collections used in DUC[3] or the NTCIR Text Summarization Challenge (TSC), it consists of a set of document sets with particular topics and a set of human-made reference summaries for each of the document sets. As shown in Table 1, we selected 30 topics and retrieved Mainichi and Yomiuri newspaper articles published in 1998–1999 using an information retrieval system. Then, we manually selected 6–12 documents from the 60 top ranked retrieved documents, and composed 30 document sets. The topics resemble the queries input by the users of the information retrieval systems and the set of documents for each topic can be thought of as a set of retrieved documents for the query. Then human-made reference summaries were created discriminating the three summary types: fact-reporting summaries, opinion-oriented summaries and knowledge-focused summaries. Such differentiation was not included in any existing summary test collections such as those used in DUC or NTCIR. Three different types of reference summaries for a document set were created by the same professional editors. A reference summary was created for each summary type for a document set. In total, three professional editors were used as summary writers. Instructions for the summary writers as to which summary types to produce were given as follows:

1. Fact-reporting summary: Summaries focused on events, which happened in real time or in past times; that is, the summaries for users who want to know facts or to check back for events related to topics.

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

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

2. Opinion-oriented summary: Summaries focused on the authors' opinions or experts' opinions by third parties; that is, the summaries for users who want advice, prospects, or evaluations related to topics.
3. Knowledge-focused summary: Summaries focused on definitional or encyclopedic knowledge; that is, the summaries for users who are interested in descriptive knowledge related to topics.

| ID | Topic | Source Articles | |
|---|---|---|---|
| | | # of Articles | # of Bytes |
| S010 | European monetary union | 10 | 41060 |
| S020 | Annual pension | 10 | 43408 |
| S030 | Accounting fraud | 9 | 42414 |
| S040 | Itoman fraud case | 10 | 41294 |
| S050 | Removal of deposit insurance | 11 | 38502 |
| S060 | Digital cellular phone | 11 | 40706 |
| S070 | Guidelines for Japan-U.S. defense cooperation | 9 | 41374 |
| S080 | Kosovo | 11 | 41166 |
| S090 | Strategic arms reduction | 8 | 30998 |
| S100 | Brain-death diagnosis | 7 | 42104 |
| S110 | Juvenile proceedings | 11 | 41934 |
| S120 | Freedom of Information Act | 8 | 33906 |
| S130 | Donor card | 10 | 31804 |
| S140 | Defined contribution pension plan | 12 | 38262 |
| S150 | Genetically-enginered foods | 12 | 40450 |
| S160 | Organized Crime Control Act | 8 | 42850 |
| S170 | Criticality-caused nuclear accident | 7 | 33870 |
| S180 | Financial Big Bang | 8 | 38822 |
| S190 | Pluthermal | 9 | 38184 |
| S200 | Theater Missile Defenses | 8 | 34646 |
| S210 | Government-owned company in China | 6 | 27058 |
| S220 | Conflict of Nothern Ireland | 10 | 28482 |
| S230 | Russian economic and financial crises | 7 | 31362 |
| S240 | Taepodong missile | 8 | 40260 |
| S250 | International Convenants on Human Rights | 7 | 41904 |
| S260 | Impeachment case | 8 | 38340 |
| S270 | Sunshine Policy | 7 | 33884 |
| S280 | Endocrine-disrupting Chemicals | 10 | 36736 |
| S290 | International Space Station | 8 | 30242 |
| S300 | Convention concering the Protection of the World Cultural and Natural Heritage | 7 | 33624 |
| Max | | 12.0 | 43408.0 |
| Min | | 6.0 | 27058.0 |
| Average | | 8.9 | 37321.5 |
| Standard Deviation | | 1.6 | 4661.7 |

*Table 1. Topics of the document sets in the ViewSumm30 test collection for multi-viewpoint document summarization.*

The maximum length of the reference summaries was 1600 bytes. For some document sets, subtopics on which summaries focused were specified by summary writers.

### 2.3 Baseline Summarization Algorithm

The baseline is a multi-document summarizer using paragraph-based clustering with Ward's Method but without considering summary types. It was tested in the NTCIR-4 TSC and worked well among other participants. The goal of multi-document summarization (MDS) is usually

defined as extracting content from a given set of related documents and presenting the most important content. The baseline system could extract important content sensitive to the user's needs by specifying subtopics in document sets.

Many clustering-based multi-document summarization frameworks (Stein et al., 2000; Hatzivassiloglou et al., 2001; Maña-López et al., 2004; Radev et al., 2004) have been proposed. Their research focused on making the topic structure explicit. By detecting similarities in topic structure, such systems could avoid redundant information in summaries. These methods have four principal aspects: (1) clustering algorithms, (2) cluster units, (3) sentence extraction strategy, and (4) cluster size.

For the clustering algorithm, we used Ward's after testing the clustering using complete link, group average, or Ward's method on the same document collection. For cluster unit, we used paragraphs rather than sentences. It was for the following reasons: (1) it allowed real-time interactivity, and (2) because of the sparseness of vector spaces when using sentence vectors. In addition, we did not cluster source documents by document units because source document counts (from 6 to 12 documents) were too small compared to summary sizes.

An algorithm is detailed below. The evaluation results in NTCIR-4 TSC3 using this algorithm are described in more detail in (Seki et al., 2004a).

[1] Paragraph Clustering Stage

a) Source documents were segmented to paragraphs, and then term frequencies were indexed for each paragraph.
b) Paragraphs were clustered based on Euclidean distance between feature vectors with term-frequency. The clustering algorithm was Ward's method. Cluster sizes varied according to the number of extracted sentences.

[2] Sentence Extraction Stage

a) The feature vectors for each cluster were computed with term frequencies and inverse cluster frequencies: TF * log (Total Clusters / Cluster Frequency).
b) *If* questions or subtopics focusing on a summary were given, clusters were ordered by the similarity between content words in the questions and the cluster feature vectors. Questions were used for expressing information needs for the original documents to produce summaries.
c) *Else* we computed the total term frequencies of all documents and ordered clusters based on similarities between total TF and cluster feature vectors.
d) *End*
e) Sentences in each cluster were weighted based on question words, heading words in the cluster, and TF values in the cluster.
f) One or two sentences were extracted from each cluster in cluster order to reach the maximum allowed number of characters or sentences.

In Chapter 5, this algorithm was compared to the extended algorithm with sentence-type annotation (Chapter 3) and genre classification (Chapter 4).

## 3. Sentence-type Annotation

In this section, sentence types, which represent information type effectively for finer-grained units than documents, are detailed. This information, along with document genres, which are elaborated in Section 4, is used to discriminate summary types in the proposed system.

### 3.1 Sentence Types for News Articles

Sentence types (Seki et al., 2004b; McKnight and Srinivasan, 2003; Teufel and Moens, 2002) were broadly used to discriminate information type with text structure. Kando (1996) has defined five sentence types for newspaper articles: *main description*, *elaboration*, *background*, *opinion*, and *prospect*. The intercoder consistency for these five sentence types was proved by experiments (Kando, 1996). The meanings of the five sentence types are as follows:

1. "Main description": the main contents in a document.
2. "Elaboration": the "main description" is detailed.
3. "Background": history or background is described.
4. "Opinion": author's opinion.
5. "Prospect": likely developments in the future are expressed.

In this experiment, a sixth type, "authority's opinion", was added to the above mentioned five sentence types, and then these six were used to discriminate the summary types.

6. "Authority's opinion": opinion reported by third parties such as experts, authorities, and so on.

### 3.2 Automatic Sentence-type Annotation

#### *3.2.1 Manual annotation for training data*

The training set consisted of 352 articles (5201 sentences total) from the 1994 Nikkei newspaper. All the sentences in the training set were annotated manually with sentence types. The number of sentences for each type was as follows: main description (1052), elaboration (3003), background (585), author's opinion (483), authority's opinion (391), and prospect (506).

#### *3.2.2 Automatic sentence-type annotation with SVM*

An automatic sentence type annotation was implemented using SVM. According to Joachims (2002), SVM is fairly robust to overfitting and can scale up to considerable dimensionalities. The feature set for automatic annotation was as follows:

1. Sentence position in document or paragraph
2. Paragraph position
3. Sentence length
4. The number of heading words in the sentence
5. The number of words with high TF/IDF value in the sentence
6. Voice, tense and modality information judged by auxiliary verb

7. Eight kinds of named entity frequencies extracted with parsers[4]
8. 20–40 kinds of semantic primitives for predicates and subjects extracted using the thesaurus published by The National Institute for Japanese Language (2004)
9. 30–40 kinds of keyword frequencies for background, author's and authority's opinion, and prospect types
10. Sentence type for pre-position sentences and post-position sentences

This classification technique was evaluated for three measures: *precision, recall*, and *accuracy* (Joachims, 2002; Sebastiani, 2002). *Precision* and *recall* are of widespread use in information retrieval. In Table 2, a convenient display of the prediction behavior is provided. We define *precision*, *recall* and *accuracy* based on this table. The diagonal cells count how often the prediction was correct. The off-diagonal cells show the frequency of prediction errors. The sum of all cells equals the total number of predictions.

- *Precision*: $f_{++} / (f_{++} + f_{+-})$
- *Recall*: $f_{++} / (f_{++} + f_{-+})$
- *Accuracy*: $(f_{++} + f_{--}) / (f_{++} + f_{--} + f_{+-} + f_{-+})$

| | label = +1 | label = -1 |
|---|---|---|
| prediction = +1 | $f_{++}$ | $f_{+-}$ |
| prediction = -1 | $f_{-+}$ | $f_{--}$ |

*Table 2. Contingency table for accuracy, precision, and recall (Joachims, 2002).*

In addition, we use *macro-averaging* and *micro-averaging* when averaging the precision, recall and accuracy, respectively, over the four-fold cross-validations that will be described below. *Macro-averaging* corresponds to the standard way of computing an (arithmetic) average, while *micro-averaging* averages each frequencies in Table 2 and computes the precision, recall, and accuracy.

Four-fold cross-validation was applied to 352 newspaper articles. Cross-validation was processed in the following steps. First, the 352 articles were divided into four groups by publishing dates. From the four training sample groups, the first group was removed. The resulting sample groups were used for training, leading to a classification rule. This classification rule was tested on the removed sample group. This process was repeated for all training sample groups. These results are summarized in Tables 3. They showed good accuracy, so we used this set as training data and the feature set for automatic annotation of sentence types in the summary test collections. Of the 11926 sentences in the summary test collections, 797 were annotated as main description, 1871 as elaboration, 189 as background, 1506 as authors' opinion, 1179 as authority's opinion, and 190 as prospects.

[4] http://chasen.org/~taku/software/cabocha

| | Main Description (M) | | | Elaboration (E) | | | Background (B) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | Accuracy | Precision | Recall | Accuracy | Precision | Recall |
| Group A | 97.02 | 89.00 | 96.49 | 85.45 | 83.65 | 91.87 | 92.90 | 75.00 | 60.00 |
| Group B | 97.40 | 90.97 | 98.60 | 86.99 | 89.29 | 88.80 | 93.98 | 78.26 | 57.14 |
| Group C | 95.78 | 89.83 | 90.05 | 84.72 | 86.06 | 89.64 | 94.68 | 64.57 | 58.99 |
| Group D | 96.72 | 86.17 | 98.20 | 85.29 | 82.67 | 91.65 | 90.59 | 81.17 | 60.10 |
| Macro Avg. | 96.73 | 88.99 | 95.84 | 85.62 | 85.42 | 90.49 | 93.04 | 74.75 | 59.06 |
| Micro Avg. | 96.54 | 88.93 | 94.68 | 85.33 | 85.00 | 90.58 | 93.16 | 74.52 | 59.49 |

*Table 3 (a). Results of 4-fold cross validation test of automatic sentence-type annotation on main description, elaboration, and background-type.*

| | Authors' Opinion (O1) | | | Authority's Opinion (O2) | | | Prospect (P) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | Accuracy | Precision | Recall | Accuracy | Precision | Recall |
| Group A | 91.61 | 69.51 | 60.64 | 95.13 | 64.20 | 54.74 | 91.00 | 78.95 | 39.89 |
| Group B | 96.91 | 82.61 | 55.88 | 94.96 | 83.33 | 49.02 | 93.98 | 62.50 | 44.44 |
| Group C | 93.01 | 73.29 | 52.97 | 92.70 | 63.64 | 50.84 | 93.90 | 66.82 | 75.00 |
| Group D | 96.39 | 63.79 | 62.71 | 96.89 | 71.01 | 74.24 | 94.37 | 71.43 | 35.29 |
| Macro Avg. | 94.48 | 72.30 | 58.05 | 94.92 | 70.55 | 57.21 | 93.31 | 69.93 | 48.66 |
| Micro Avg. | 93.85 | 70.84 | 57.35 | 94.62 | 67.18 | 55.50 | 93.19 | 70.00 | 52.57 |

*Table 3 (b). Results of 4-fold cross validation test of automatic sentence-type annotation on authors' opinion, authority's opinion, and prospect-type.*

## 4. Genre Classification

In this section, document genres, which represent document-level information types, are detailed. This information and the sentence types described in the previous section were used to determine summary types in the proposed system.

### 4.1 Genre Feature

To begin with, genre taxonomies for news articles were surveyed. International Press Telecommunications Council (here after, IPTC) defined a set of document genres[5] for news delivery. These, however, number more than 40 and are based on several different classification criteria, such as 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". This framework is not appropriate for discriminating among document genres for the summary type because the categorizing criteria are complex and relate to the different attributes of the documents.

Therefore, in this research, document genres were represented by a combination of the values for each of the multiple dimensions representing different genre features. It is based on Douglas Biber's proposal (Biber, 2002). 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 for each of the dimensions.

[5] http://www.iptc.org/download/dliptc.php?fn=topicset/topicset.iptc-genre.xml

The five basic dimensions in Biber's framework were:

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

Of Biber's dimensions, the fifth could not be discriminated using *ViewSumm30* because all documents were categorized as "information production" in this dimension. We used the remaining four dimensions. The definitions are as following:

1. Situation-Dependency (G1): documents marked according to the degree of coincidence between their publishing time and the event time.
2. Argumentation (G2): documents marked according to the degree of persuasion and the author's point of view.
3. Impersonal Styles (G3): documents marked according to criteria such as frequent passive constructions.
4. Fact-Reporting (G4): documents marked that reported facts in the inverse-pyramid discourse structure of newspaper articles.

In this research, the "genre feature" of each document was described by the combination of these four dimensions.

### 4.2 Manual Annotation for Genre Feature

To begin with, we tested the inter-coder consistency of genre feature manual annotation. The corpus consists of 208 newspaper articles which are not included in *ViewSumm30*, but published in the same years as those in *ViewSumm30*. Three coders, a1, a2, and a3, annotated each of the 208 documents independently. The annotation instructions were prepared and updated through pretests with all three coders. As shown in Table 4, the kappa coefficient value showed good agreement between coders.

| Genre Dimension | Pair of Assessors | | | Avg. |
|---|---|---|---|---|
| | (a1,a2) | (a1,a3) | (a2,a3) | |
| Situation-Dependency (G1) | 0.618 | 0.595 | 0.665 | 0.626 |
| Argumentation (G2) | 0.41 | 0.536 | 0.678 | 0.541 |
| Impersonal Styles (G3) | 0.459 | 0.506 | 0.604 | 0.523 |
| Fact-Reporting (G4) | 0.604 | 0.566 | 0.657 | 0.609 |

*Table 4. Kappa coefficients: inter-coder consistency.*

These results suggest that manual annotation can be moderately or substantially consistent (Landis et al., 1977).

### 4.3 Automatic Genre Classification

Similar to the method used for automatic sentence type annotation, we applied SVM to automatic genre classification. In order to examine its effectiveness, we performed 4-fold cross validation

using the 208 annotated documents mentioned in subsection 4.2. For automatic genre classification, we selected about 200 structural features as listed below:

- Five structural features: author signature, section, photo, figure, and news source.
- Nine statistical features ('#' is defined as "numbers"): # of characters, Type-to-Token Ratio, # of sentences, # of opinion sentences, # of prospect sentences, # of background sentences, # of conjunctions, # of quote parentheses, and average sentence length.
- Eight kinds of named entity frequencies extracted with parsers[6].
- 60 function phrases (which relate to opinion, prospect, and background information).
- 93 symbols (which include several punctuation-related symbols).
- 20–40 kinds of semantic primitives for predicates and subjects extracted using the thesaurus published by The National Institute for Japanese Language (2004).

It has been claimed that function phrases and punctuation mark counts (Kessler et al., 1997; Stamatatos et al., 2000) are effective for genre classification. Statistical features (Karlgren and Cutting, 1994) are also often used to classify texts in the corpus linguistics field. Function phrases and symbols were selected from the corpus used for sentence-type annotation clues. For 4-fold cross validation, the 208 documents were divided into four groups, each containing 52 documents. We used one group as a test set and the other three groups as training sets, and evaluated the effectiveness four times. The results are shown in Tables 5 below.

| | Situation-Dependency (G1) | | | Argumentation (G2) | | |
|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | Accuracy | Precision | Recall |
| Group A | 82.69 | 88.57 | 86.11 | 84.62 | 70.00 | 58.33 |
| Group B | 78.85 | 85.19 | 76.67 | 88.46 | 72.73 | 72.73 |
| Group C | 84.62 | 95.83 | 76.67 | 92.31 | 87.50 | 70.00 |
| Group D | 75.00 | 72.22 | 89.66 | 90.38 | 50.00 | 80.00 |
| Macro Avg. | 80.29 | 85.45 | 82.27 | 88.94 | 70.06 | 70.27 |
| Micro Avg. | 80.29 | 84.43 | 82.40 | 88.94 | 70.27 | 68.42 |

*Table 5 (a). Results of 4-fold cross validation test of automatic genre classification for G1 and G2.*

| | Impersonal Styles (G3) | | | Fact-Reporting (G4) | | |
|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | Accuracy | Precision | Recall |
| Group A | 88.46 | 93.33 | 93.33 | 90.38 | 93.18 | 95.35 |
| Group B | 78.85 | 83.33 | 93.02 | 92.31 | 95.35 | 95.35 |
| Group C | 90.38 | 97.67 | 91.30 | 90.38 | 100.00 | 89.36 |
| Group D | 96.15 | 95.83 | 100.00 | 92.31 | 95.65 | 95.65 |
| Macro Avg. | 88.46 | 92.54 | 94.42 | 91.35 | 96.05 | 93.93 |
| Micro Avg. | 88.46 | 92.39 | 94.44 | 91.35 | 96.00 | 93.85 |

*Table 5 (b). Results of 4-fold cross validation test of automatic genre classification for G3 and G4.*

Table 5 shows that G1, G2, G3, and G4 could be classified properly.

[6] http://chasen.org/~taku/software/cabocha

## 5. Experiment Results

In the experiment, the *v-SWIM* using sentence types and document genres was tested on the test collection *ViewSumm30*, and its effectiveness over the baseline algorithm was evaluated for the *coverage* and *precision* of the human-made reference summaries. The extended algorithm with the genre classification and the sentence-type annotation algorithm were also evaluated. Hirao et al. (2004) defined *precision* as the ratio of how many sentences in the system output are included in the set of sentences that correspond to sentences in the human-made reference summaries. *Coverage* was defined in (Hirao et al., 2004) as an evaluation metric for measuring how close the system output is to the reference summary, taking into account the redundancy found in the set of sentences in the output.

There are two aspects for evaluation. The first aspect is: 1. Genre classification effect, 2. Sentence-type annotation effect, and 3. Combination of both effects. The second aspect is: A. fact-reporting summary, B. opinion-oriented summary, and C. knowledge-focused summary. These effects are described in this section.

### 5.1 Summarization based on Genre Classification

We first surveyed the *coverage* and *precision* for the extended algorithm with genre feature for baseline sentence extraction, as stated in Section 2.3. Sentences in each article were annotated with genre feature in the article. With this genre information, sentence weights were multiplied by variables ranging from 0 to 4 with 0.1 intervals. For example, when the G1 (G2, G3, or G4) dimension for an article was annotated as positive, sentence weights in the article were multiplied by a variable to extract in the summary. When the dimension was annotated as negative, sentence weights remained unchanged. Then, coverage and precision values for extracting sentences were computed as the variables changed. The results are shown in Figures 1, 2, and 3.

#### *5.1.1 Fact-reporting Summaries*

In Figure 1, the coverage changes for multiplying weighting parameters varied from 0 to 4 with 0.1 intervals for genre feature (G1, G2, G3, and G4) are shown. For "sentence weighting = 1" on the *x*-axis, sentence weights with positive genre features were multiplied by one, so the sentence weights to extract were unchanged. The points at this *x*-axis position were common for G1, G2, G3, and G4. They represent the points evaluated for the baseline system without genre feature. The coverage for the baseline system was 0.175. Compared to this value, the coverage values for G3 were higher than the baseline coverage with variables greater than one. This means that **positive values** for the *impersonal styles* (G3) feature had a positive effect when producing a fact-reporting summary. Overall, the genre feature effect was stable for fact reporting summaries. This result shows that the baseline algorithm is well suited to fact-reporting summary production.

#### *5.1.2 Opinion-oriented summaries*

For opinion-oriented summaries, the coverage changed drastically according to the weighting parameter using the genre feature. This result is shown in Figure 2. This result was totally different from the result in Figure 1 for the change in slope. The coverage for the baseline system was 0.111. Compared to this value, **negative values** for the *impersonal styles* (G3) and *fact* (G4) feature had positive effects on producing opinion-oriented summaries. This effect was shown more explicitly than the effects of the genre feature for fact-reporting summaries. In contrast,

**positive values** for the *argumentation* (G2) feature had a positive effect in producing opinion-oriented summaries.

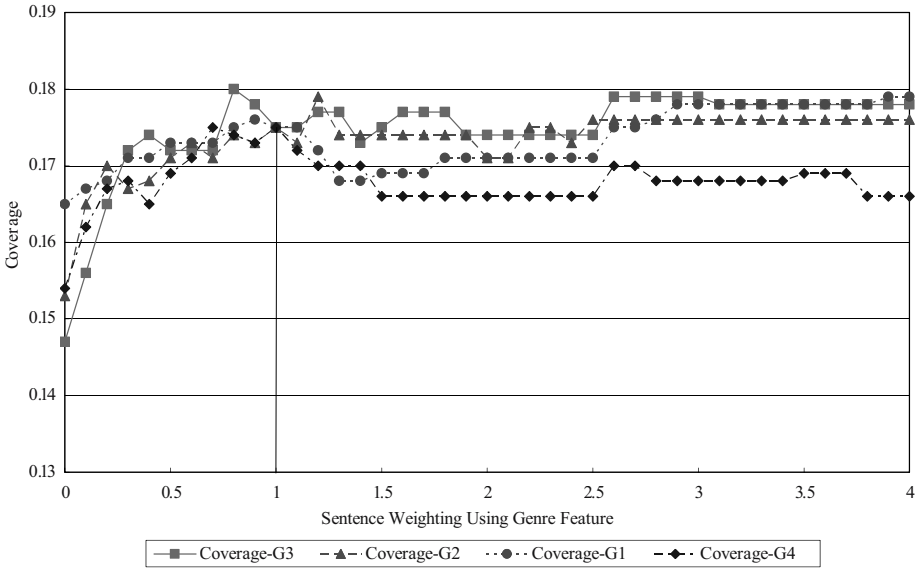


*Figure 1. Coverage for fact-reporting summaries based on sentence weighting using genre feature.*

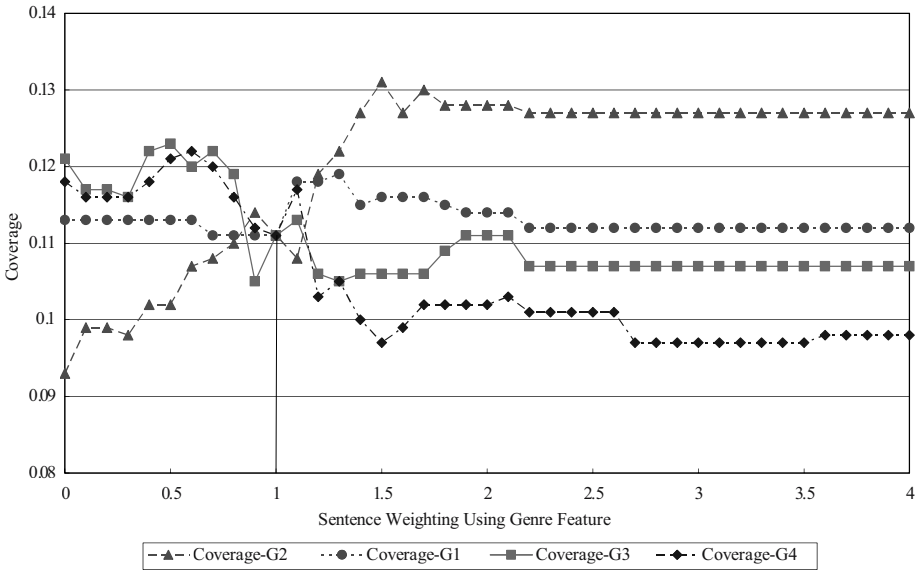


*Figure 2. Coverage for opinion-oriented summary based on sentence weighting using genre feature.*

### *5.1.3 Knowledge-focused summaries*

The coverage and precision for knowledge-focused summaries did not change as drastically as for the opinion-oriented summaries. The coverage for the baseline system was 0.151. **Positive values** for the *impersonal styles* (G3) and *fact* (G4) features had positive effects in producing knowledge-focused summaries. **Negative values** for the *argumentation* (G2) feature had a positive effect in producing knowledge-focused summaries.

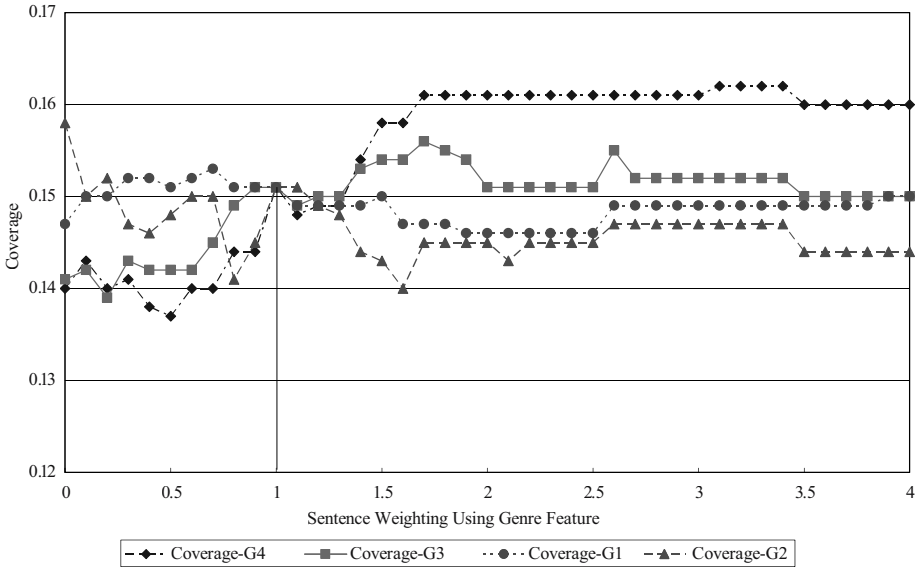


*Figure 3. Coverage for knowledge-focused summary based on sentence weighting using genre feature.*

## 5.2 Summarization based on Sentence-type Annotation

The extension of the baseline algorithm with sentence-type annotation across all source documents did not show much improvement. The results for fact-reporting, opinion-oriented, and knowledge-focused type summaries are shown in Figures 4, 5, and 6. In Figure 4, positive values for the *authority's opinion*-type (O2) had positive effects in producing fact-reporting summaries. In Figure 5, the *author's opinion* (O1) and *authority's opinion*-types (O2) had positive effects in producing opinion-oriented summaries. Finally, in Figure 6, the *elaboration*-type (E) had positive effects in producing knowledge-focused summaries.

We thought some ineffective results might be caused by the different distributions of sentence types among the different document genres. Another possibility was the low quality of automatic sentence-type annotation. In the future, we hope to compare this result with the manual annotation result for sentence types. In the next subsection, the experiment results for the sentence-annotation effects peculiar to each genre dimension are detailed.

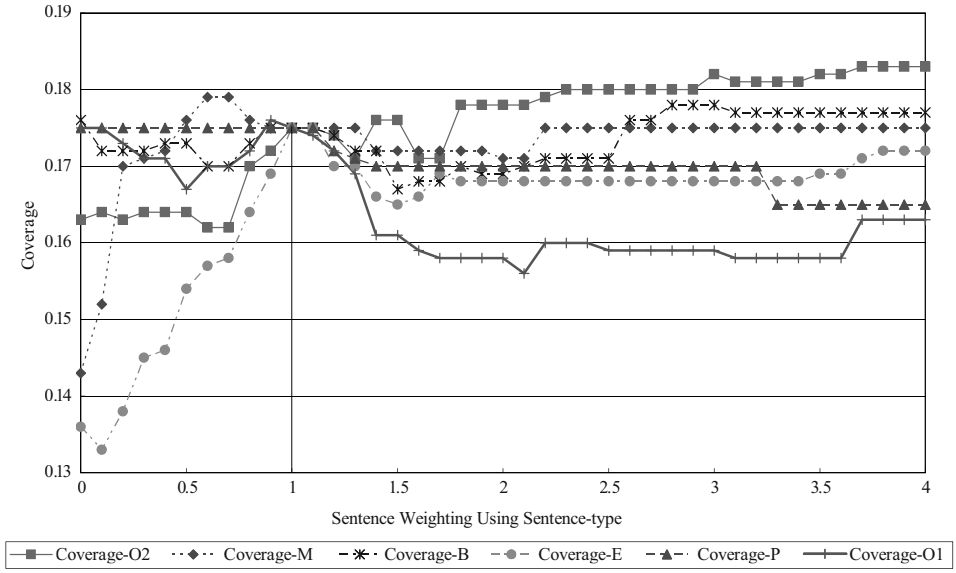


*Figure 4. Coverage for fact-reporting summary based on sentence weighting using sentence type.*

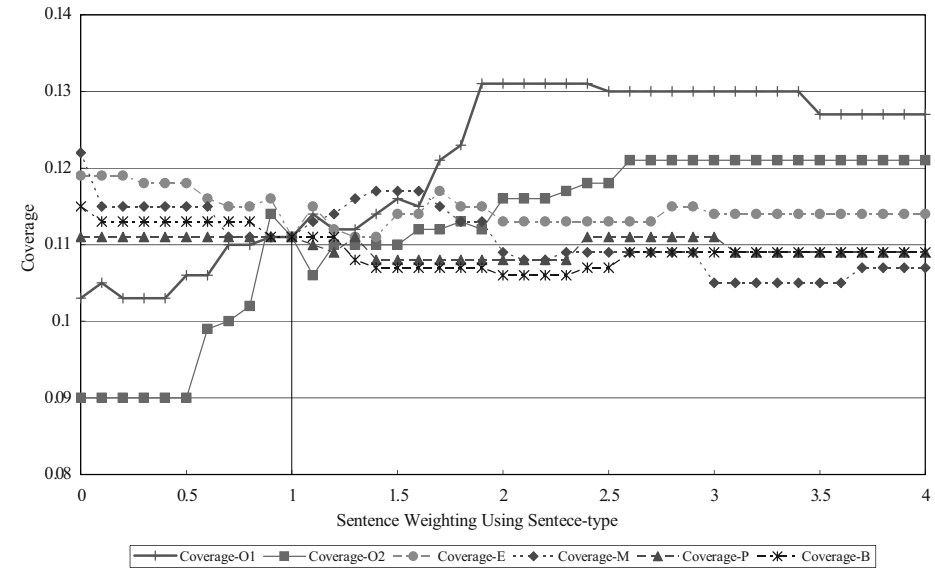


*Figure 5. Coverage for opinion-oriented summary based on sentence weighting using sentence type.*

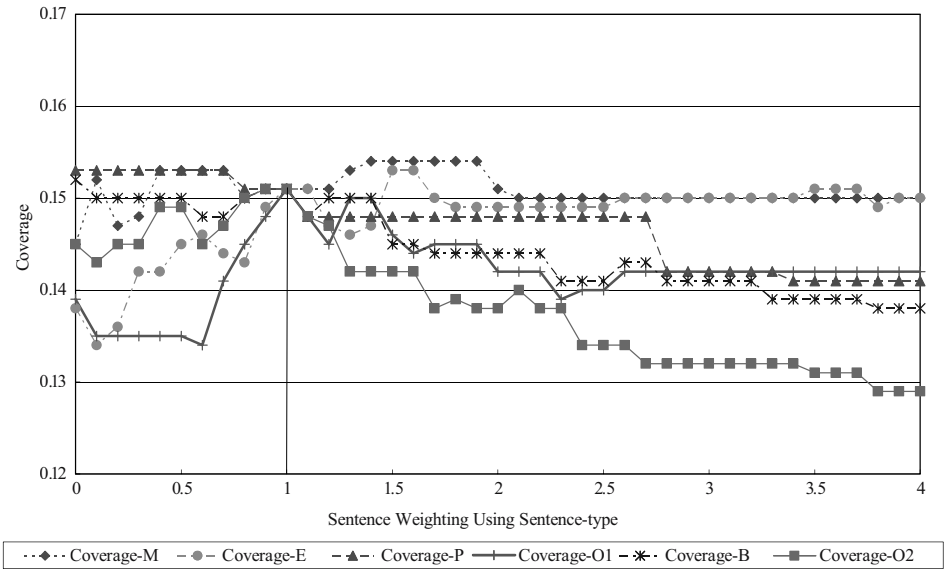


*Figure 6. Coverage for knowledge-focused summary based on sentence weighting using sentence type.*

### 5.3 Summarization based on Combining Genre and Sentence-type

In this subsection, the experiments for combining genre classification and sentence-type annotation effect are described. The improvement results for the three type summaries are shown in Table 6.

#### *5.3.1 Fact-reporting summaries*

For fact-reporting summaries, we found the improvement from combining only information on sentence types was 0.198 for coverage and 0.189 for precision. Compared with the result from the baseline system without sentence-type annotation and genre classification, this result was a significant improvement according to *Wilcoxon* tests for 30 topics.

#### *5.3.2 Opinion-oriented summaries*

For opinion-oriented summaries, the sentence type effect in specific document genres for improving coverage and precision was surveyed by weighting specific sentence-types in specific document genres. We found that the coverage was 0.149 and the precision was 0.136 with weighting 2.3 for *author's opinion*-type (O1) in the **positive values** for the *argumentation* (G2) feature, in the **negative values** for the *fact-reporting* (G4) feature, or in the **negative values** for the *impersonal styles* (G3) when combined with the other sentence-type weightings. This result was a significant improvement according to *Wilcoxon* tests for 30 topics, compared with the result from the baseline system.

| Summary Type | Baseline | | v-SWIM | | Improvement | Weighting parameter | | |
|---|---|---|---|---|---|---|---|---|
| | coverage | precision | coverage | precision | (percent) | Sentence-type | Genre | Weighting |
| Fact-reporting | 0.175 | 0.166 | **0.198*** | **0.189** | **13.14** | Main Description | G4 positive | 2 |
| | | | | | | Elaboration & Background & not Reported Opinion | G3 positive | 5 |
| | | | | | | Background | G1 positive | 4 |
| | | | | | | | G3 positive | 4 |
| | | | | | | Reported opinion & not Prospective | All | 5 |
| | | | | | | Author's opinion | All | 0.9 |
| | | | | | | Prospective | All | 0 |
| Opinion-oriented | 0.111 | 0.104 | **0.149*** | **0.136** | **34.23** | All | G2 positive | 1.5 |
| | | | | | | Author's opinion | G2 positive | 2.3 |
| | | | | | | | G3 negative | |
| | | | | | | | G4 negative | |
| | | | | | | Reported opinion | G1 negative | 3 |
| | | | | | | not Author's Opinon & not Reported Opinion | G1 positive | 0.4 |
| | | | | | | Elaboration & not Author's Opinon & not Reported Opinion | G3 positive | 0.5 |
| | | | | | | Main Description & not Author's Opinion & not Reported Opinion | All | 0 |
| | | | | | | Background & not Author's Opinion & not Reported Opinion | All | 0 |
| Knowledge-focused | 0.151 | 0.156 | **0.175** | **0.170** | **15.89** | Elaboration | G4 positive | 3.2 |
| | | | | | | Author's opinion | | 3.2 |
| | | | | | | Elaboration & not Author's Opinion | G1 negative & G4 positive | 2.3 |
| | | | | | | Main Description | All | 1.7 |
| | | | | | | Reported opinion && not Elaboration | G2 negative | 0 |
| | | | | | | Background | All | 0 |
| | | | | | | Prospective | All | 0 |

*: statistically significant with *Wilcoxon* tests: $p < 0.05$

*Table 6. Coverage and precision improvement effect for three type summaries based on sentence weighting in the specific document genre.*

#### *5.3.3 Knowledge-focused summaries*

For knowledge-focused summaries, the improvement effect of *elaboration*-type sentences was not significant but the improvement can be observed in Figure 6. We surveyed the association between *elaboration*-type (E) and document genre and found the detailed weighting rule that combined with **negative values** for the *author's opinion*-type (O1). This weighting was applied only to **negative values** for the *situation-dependency* (G1) and **positive values** for the *fact-reporting* (G4) features. This result was not a significant improvement according to *Wicoxon* tests for 30 topics, compared with the result from the baseline system.

## 6. Conclusion

In this chapter, we considered multi-viewpoint document summarization that is focused on topics and summary types to suit users' information needs. We found significant improvements in summary coverage by combining sentence type and genre classification information to discriminate among fact-reporting, opinion-oriented, and knowledge-focused summaries in experiments with our new test collection.

## 7. Acknowledgement

This work was partially supported by the Grants-in-Aid for Exploratory Research (#16650053) and Scientific Research in 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.

## 8. Bibliography

Angheluta, R., Moens, M. F., and De Busser, R. (2003) K. U. Leuven Summarization System - DUC 2003. In *Proceedings of the Workshop on Text Summarization (DUC 2003) at the 2003 Human Language Technology Conference (HLT/NAACL 2003)*, Edmonton, Canada.

Bazerman, C. (2004) Speech Acts, Genres, and Activity Systems: How Texts Organize Activity and People. In Bazerman, C. and Prior, P. (Eds.) *What Writing Does and How It Does It – An Introduction to Analyzing Texts and Textual Practices*. 309-339. Lawrence Erlbaum Associates, Mahwah, NJ.

Biber, D., Conrad, S., and Reppen, R. (1998) Corpus Linguistics - Investigating Language Structure and Use (Reprinted, 2002). Cambridge Approaches to Linguistics. Cambridge University Press.

Borlund, P. (2003) The Concept of Relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10), 913-925.

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.) *Proceedings of ECIR 2002 Advances in Information Retrieval, 24th BCS-IRSG European Colloquium on IR Research*, Glasgow, UK, 353-362. Published in *Lecture Notes in Computer Science 2291*, Springer-Verlag, Heidelberg, Germany.

Harman, D. and Over, P. (2004) The Effects of Human Variation in DUC Summarization Evaluation. In *Proceedings of Text Summarization Branches Out, Workshop at the 42nd ACL 2004*, Barcelona, Spain, 10-17.

Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M. Y., and McKeown, K. R. (2001) Simfinder: A Flexible Clustering Tool for Summarization. In *Proceedings of the Workshop on Automatic Summarization at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, Pittsburgh, PA, 41-49.

Hirao, T., Okumura, M., Fukushima, 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 Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Summarization*. National Institute of Informatics, Japan. Available from: <http://research.nii.ac.jp/ntcir>.

Joachims, T. (2002) Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. Kluwer Academic Publishers.

Kando, N. (2004) Overview of the Fourth NTCIR Workshop. In *Proceedings of the Fourth NTCIR Workshop on Research in Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Summarization*. National Institute of Informatics, Japan. Available from: <http://research.nii.ac.jp/ntcir>.

Kando, N. (1996) Text Structure Analysis Based on Human Recognition: Cases of Japanese Newspaper Articles and English Newspaper Articles (in Japanese). In *Research Bulletin of National Center for Science Information Systems*, 8, 107-126.

Karlgren, J. and Cutting, D. (1994) Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*, Kyoto, Japan, 1071-1075.

Kessler, B., Nunberg, G., Schuetze, H. (1997) Automatic Detection of Text Genre. In *Proceedings of the 35th ACL/8th EACL 1997*, Madrid, Spain, 32-38.

Landis, J. R. and Koch, G. G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33, 159-74.

Lin, C-Y., and Hovy, E. (2002) Manual and Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Automatic Summarization at the 40th ACL 2002*, University of Pennsylvania, PA.

Maña-López, M. J., Buenaga, M. D., and Gómez-Hidalgo, J. M. (2004) Multidocument Summarization: An Added Value to Clustering in Interactive Retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2), 215-241.

Mani, I. (2001) Automatic Summarization. Volume 3 of Natural Language Processing, John Benjamins Pub, Amsterdam, Netherlands.

McKnight, L. and Srinivasan, P. (2003) Categorization of Sentence Types in Medical Abstracts. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium,* Ottawa, Canada, 440-444.

Pomerantz, J. (2002) Question Taxonomies for Digital Reference. Ph. D. thesis, Syracuse University.

Radev, D. R., Jing, H., Styś, M., and Tam, D. (2004) Centroid-based Summarization of Multiple Documents. *Information Processing and Management*, 40(6), 919-938.

Rath, G. J., Resnick, A., and Savage, T. R. (1961) The Formation of Abstracts by the Selection of Sentences. *American Documentation*, 2(12), 139-208.

Sebastiani, F. (2002) Machine Learning in Automated Text Categorization. *ACM Computing Surveys,* 34(1), 1-47.

Seki, Y., Eguchi, K., and Kando, N. (2004a) User-focused Multi-Document Summarization with Paragraph Clustering and Sentence-type Filtering. In *Proceedings of the Fourth NTCIR Workshop on Research in Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Summarization*. National Institute of Informatics, Japan. Available from: <http://research.nii.ac.jp/ntcir>.

Seki, Y., Eguchi, K., and Kando, N. (2004b) Compact Summarization for Mobile Phones. In Crestani, F., Dunlop, M. and Mizzaro, S. (Eds.) *Mobile and Ubiquitous Information Access*. 172-186. Published in *Lecture Notes in Computer Science 2954,* Springer-Verlag, Heidelberg, Germany.

Simpson, J. A. and Weiner, E. S. C. (1991) The Oxford English Dictionary (second edition). Clarendon Press, New York.

Spärck-Jones, K. (1999) Automatic Summarizing: Factors and Directions. In Mani, I., and Maybury, M. T. (Eds.) *Advances in Automatic Text Summarization*. 1-12. MIT Press, Cambridge, MA.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000) Text Genre Detection Using Common Word Frequencies. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING2000),* Saarbrücken, Germany, 808-814.

Stein G. C., Strzalkowski, T., and Wise, G. B. (2000) Evaluating Summaries for Multiple Documents in an Interactive Environment. In *Proceedings of the Second International Conference on Language Resources & Evaluation (LREC2000*), Athens, Greece, 1651-1657.

Teufel, S. and Moens, M. (2002) Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4), 409-445.

The National Institute for Japanese Language (2004) Bunruigoihyo – enlarged and revised edition. Dainippon-Tosho.

Xu, J. Weischedel, R., and Licuanan, A. (2004) Evaluation of an Extraction-Based Approach to Answering Definitional Questions. In *Proceedings of the 27th ACM SIGIR 2004*, Sheffield, UK, 418-424.