DISCOURSE STRUCTURE ANALYSIS FOR NEWS VIDEO

Yasuhiko Watanabe[†] *Yoshihiro Okada*[†] *Sadao Kurohashi*[‡] *Eiichi Iwanari*[†]

[†] Dept. of Electronics and Informatics, Ryukoku University, Seta, Otsu, Shiga, Japan
[‡]Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo, Kyoto, Japan
watanabe@rins.ryukoku.ac.jp

ABSTRACT

Various kinds of video recordings have discourse structures. Therefore, it is important to determine how video segments are combined and what kind of coherence relations they are connected with. In this paper, we propose a method for estimating the discourse structure of video news reports by analyzing the discourse structure of their transcripts.

1. INTRODUCTION

A large number of studies have been made on video analysis, especially segmentation, feature extraction, indexing, and classification. On the other hand, little attention has been given to the discourse structure (DS) of video data.

Various kinds of video recordings, such as dramas, documentaries, news reports, and sports castings, have discourse structures. In other words, each video segment of these video recordings is related to previous ones by some kind of relation (coherence relation) which determines the role of the video segments in discourse. For this reason, it is important to determine how video segments are combined and what kind of coherence relations they are connected with. In addition, Nagao et.al proposed a method for multimedia data summarization using GDA tags [Nagao 00]. However, the cost of making GDA tagged data is great. Our method will be helpful in reducing the annotation cost.

In this paper, we propose a method for estimating the discourse structure of video news reports. Coherence relations between video segments are estimated in the following way:

1. a video news article is segmented into shots by using DCT components,

2. consecutive shots are merged by using speech information, and

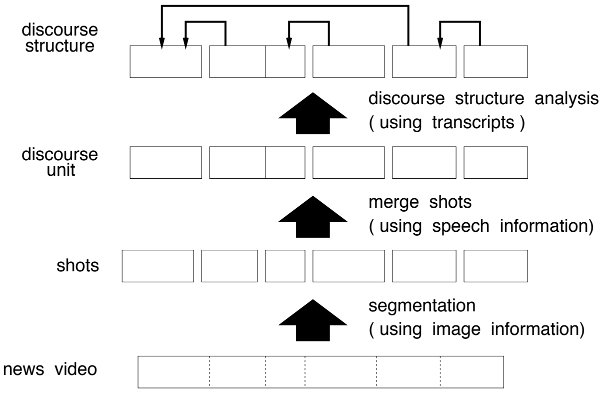


Figure 1: Procedure of discourse structure analysis for news video

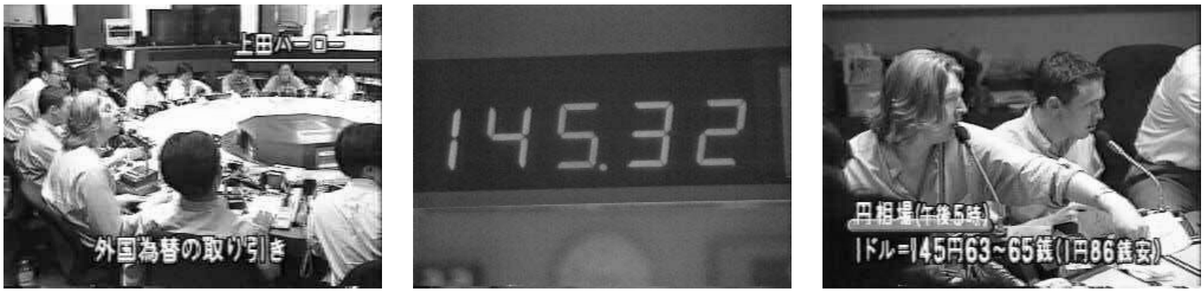
3. coherence relations are estimated by using three kinds of clues in the transcript of the news video:

   - clue expressions indicating a certain relation,
   - occurrence of identical/synonymous words/ phrases in topic chaining or topic-dominant chaining relation, and
   - similarity between two sentences in list or contrast relation.

Figure 1 shows the procedure of discourse structure analysis for news video. We applied our method to NHK[1] News [2]. This method is aimed to make the process of retrieval, summarization, and information extraction more efficient.

[1] Nippon Hoso Kyokai (Japan Broadcasting Corporation)

[2] NHK news reports do not have closed captions. Instead of closed captions, we used scripts which were read out by newscasters as transcripts.



shot (a) shot (b) shot (c)

(I) The US dollar inched up against the yen as the stock market continued the selling trend in Tokyo today.

(II) The US currency traded at 145.63–65 yen at 5 p.m. Tokyo.

Figure 2: An example of shots and their transcript in a news video (NHK evening news, August/3/1998)

2. DISCOURSE STRUCTURE AND VIDEO

Little attention has been given to discourse structure of video data in image processing. This is because it is difficult to determine it only by analyzing image data. In contrast to this, discourse structure is the subject of a large number of studies in natural language processing. So several methods for estimating the discourse structure of a text have been explored[Sumita 92] [Kurohashi 94]. Therefore, these methods can be applied to language data of video data in order to determine discourse structure in video data.

In addition, some researchers in natural language processing showed that the information of discourse structure is useful for extracting significant sentences and summarizing a text [Miike 94] [Marcu 97]. It suggests that information of video discourse structure is utilized for extracting significant video segments and skimming. It may be useful to look at video skimming and extraction of significant segments before we discuss some points about discourse structure analysis because they are closely related to the discourse structure estimation.

One of the simple ways to skim a video is by using the pair of the first frame/image of the first shot and the first sentence in the transcript. However, this representative pair of image and language is often a poor topic explanation. To solve this problem, Zhang, et.al, proposed a method for key-frame selection by using several image features such as colors, textures, and temporal features including camera operations [Zhang 95]. Also, Smith and Kanade proposed video skimming by selecting video segments based on TFIDF, camera motion, human face, captions on video, and so on [Smith 97]. These techniques are broadly applicable, however, still have problems. One is the semantic classification of each segment. To solve this problem, Nakamura and Kanade proposed the spotting by association method which detects relevant video segments by associating image data and language data [Nakamura 97]. Also, Watanabe, et.al, proposed a method for analyzing telops (captions) in video news reports by using layout and language information [Watanabe 96]. However, these studies did not deal with coherence relations between video segments.

In contrast to this, several works on discourse structure have been made by researchers in natural language processing. Pursuing these studies, we are confronted with two points of discourse structure analysis:

- available knowledge for estimating discourse structure, and
- definition for discourse units and coherence relations.

First, we shall discuss the available knowledge for estimating discourse structure. Most studies on discourse structure have focused on such questions as what kind of knowledge should be employed, and how inference may be performed based on such knowledge (e.g., [Grosz 86], [Hobbs 85], [Zadrozny 91]). In contrast to this, Kurohashi and Nagao pointed out that a detailed knowledge base with broad coverage is unlikely to be constructed in the near future, and that we should analyze discourses using presently available knowledge. For these reasons, they proposed a method for estimating discourse structure by using surface information in sentences [Kurohashi 94]. In video analysis, the same problems occurred. Therefore, we propose here a method for estimating the discourse structure in a news report by using surface information in the transcript.

Next, we shall discuss the definition for discourse unit and coherence relation. As mentioned, discourses are composed of segments (discourse units), and these are connected to previous ones by coherence relations. However, there has been a variety of definitions for discourse unit and coherence relation. For example, a discourse unit can be a frame, a shot, or a group of several consecutive shots. In this study, we consider as a discourse unit, one or more shots which are associated with one part of announcer's speech. For example, shot (a), (b), and (c) in Figure 2 represent consecutive shots in the news video "Both yen and stock were dropped"(Aug/3/1998), while sentence (I) and (II) are parts of the transcript. Sentence (I) was spoken in shot (a) and (b), and correspondingly, sentence (II) was spoken in shot (c). As a result, shot (a) and (b) were merged and the result was considered as one discourse unit. On the other hand, shot (c) alone constituted one discourse unit. We will explain how to extract the discourse units in Section 3.1.

In contrast to this, coherence relations strongly depend on the genre of video data: dramas, documentaries, news reports, sports castings, and so on. From the number of coherence relations suggested so far, we selected the following relations for our target, news reports:

List: S_i and S_j involve the same or similar events or states, or the same or similar important constituents

Contrast: S_i and S_j have distinct events or states, or contrasting important constituents

Topic chaining: S_i and S_j have distinct predications about the same topic

Topic-dominant chaining: A dominant constituent apart from a given topic in S_i becomes a topic in S_j

Elaboration: S_j gives details about a constituent introduced in S_i

Reason: S_j is the reason for S_i

Cause: S_j occurs as a result of S_i

Example: S_j is an example of S_i

where S_i denotes the former segment and S_j the latter.

3. ESTIMATION OF DISCOURSE STRUCTURE

Our determination of how video segments are combined and what kind of coherence relations are involved is made in the next way:

1. extract discourse units from a news report,
2. extract three kinds of clue information from transcripts, and then, transform them into reliable scores for some relations, and
3. choose the connected sentence and the relation having the maximum reliable score. If two or more connected sentences have the same maximum score, the chronological nearest segment is selected.

3.1. Extraction of Discourse Units

A shot is generally regarded as a basic unit in video analysis. In this study, however, not only a shot but also more consecutive ones are considered a basic unit (discourse unit). This is because there are some cases where several consecutive shots correspond with one sentences in a transcript. In this case, these consecutive shots should be regarded as a discourse unit. In contrast to this, one shot should be regarded as a discourse unit when it correspond with one or more sentences in a transcript. In both cases, the start/end point of a discourse unit often lies in the pause because the announcer needs to take breath at the end of a sentence. As a result, discourse units are extracted in the next way:

1. detect scene cuts in a video by using DCT components [Iwanari 94],
2. detect speech pauses in the video, and
3. extract the start/end points of discourse units by detecting the cuts in the pause.

For evaluating this method, we used 105 news reports of NHK News. The recall and precision of discourse unit detection were 71% and 97%, respectively, while those of scene change detection were 80% and 90%. We modified the extracted discourse units by hand and used them in the discourse structure analysis described in Section 3.2. In addition, each discourse unit was associated with the corresponding sentences in a transcript by hands. This is because NHK news reports do not have closed captions.

3.2. Detection of Coherence Relations

In order to extract discourse structure, we use three kinds of clue information in transcripts:

- clue expressions indicating some relations,
- occurrence of identical/synonymous words/phrases in topic chaining or topic-dominant chaining relation, and
- similarity between two sentences in list or contrast relation.

Then they are transformed into reliable scores for some relations. In other words, as a new sentence (NS) comes in, reliable scores for all possible connected sentences and relations are calculated by using above three types of clues. As a final result, we choose the connected sentence (CS) and the relation having the maximum reliable score.

3.2.1. Detection of Clue Expressions

In this study, we use 41 heuristic rules for finding clue expressions by pattern matching and relating them to proper relations with reliable scores. A rule consists of two parts: (1) conditions for rule application and (2) corresponding relation and reliable score. Conditions for rule application consist of four parts:

- rule applicable range,
- relation of CS to its previous DS,
- dependency structure pattern for CS, and
- dependency structure pattern for NS.

Pattern for CS and NS are matched not for word sequences but for dependency structures of both sentences. We apply each rule for the pairs of a CS and NS. If the condition of the rule is satisfied, the specified reliable score is given to the corresponding relation between the CS and the NS.

For example, Rule-1 in Figure 3 gives a score (20 points) to the reason relation between two adjoining sentences if the NS starts with the expression "*nazenara* (because)". Rule-2 in Figure 3 is applied not only for the neighboring CS but also for farther CSs, by specifying the occurrence of identical words "X" in the condition.

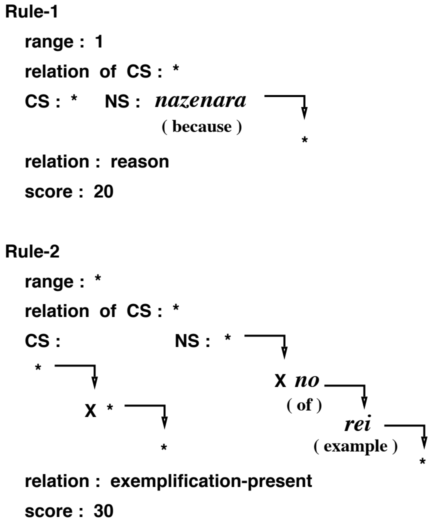


Figure 3: Examples of heuristic rules for clue expressions

3.2.2. Detection of Word/Phase Chain

In general, a sentence can be divided into two parts: a topic part and a non-topic part. When two sentences are in a topic chaining relation, the same topic is maintained through them. Therefore, the occurrence of identical/synonymous word/phrase (the word/phrase chain) in topic parts of two sentences supports this relation. On the other hand, in the case of topic-dominant chaining relation, a dominant constituent introduced in a non-topic part of a prior sentence becomes a topic in a succeeding sentence. As shown, the word/phrase chain from a non-topic part of a prior sentence to a topic part of a succeeding sentence supports this relation.

For these reasons, we detect word/phrase chains and calculate reliable scores in the next way:

1. give scores to words/phrases in topic and non-topic parts according to the degree of their importance in sentences,
2. give scores to the matching of identical/synonymous words/phrases according to the degree of their agreement, and
3. give these relations the sum of the scores of two chained words/phrases and the score of their matching.

For example, by Rule-a and Rule-b in Figure 4, words in a phrase whose head word is followed by a topic marking postposition "*wa*" are given some scores as topic parts. Also, a word in a non-topic part in the sentential style, "*ga aru* (there is ...)" is given a large score (11 points) by Rule-c in Figure 4 because this word is an important new information in this sentence and topic-dominant chaining relation involving it often occur. Matching of phrases

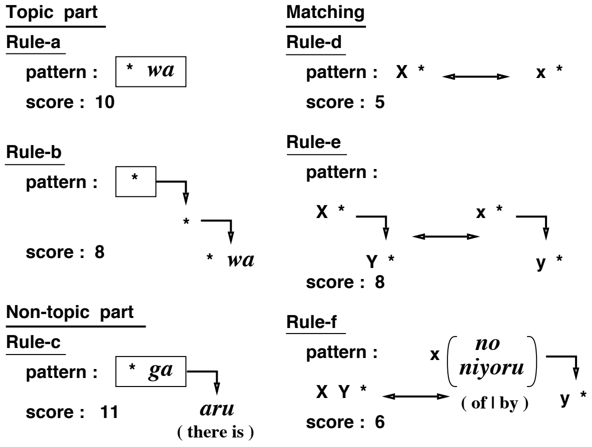


Figure 4: Examples of rules for topic/non-topic parts

like "A of B" is given a larger score (8 points) by Rule-e than that of word like "A" alone by Rule-d (5 points) in Figure 4.

3.2.3. Calculation of Similarity between Sentences in a Transcript

When two sentences have list or contrast relation, they have a certain similarity. As a result, we measure such a similarity for finding list or contrast relation in the next way. First, the similarity value between two words are calculated according to exact matching, matching of their parts of speech, and their closeness in a thesaurus dictionary. Second, the similarity value between two word-strings is calculated roughly by combining the similarity values between words in the two word-strings with the dynamic programming method for analyzing conjunctive structures [Kurohashi 94]. Then, we give the normalized similarity score between a CS and an NS to their list and contrast relations as a reliable score.

4. EXPERIMENTS AND DISCUSSION

For evaluating this method, we used 22 news reports of NHK News. Each report was a few minutes in length. The experimental results are shown in Table 1. As mentioned, news reports of NHK News do not have closed captions. For this reason, each video segment (discourse unit) was associated with the corresponding sentences in a transcript by hands.

Table 1: Analysis results

Relation	Success	Failure
List	2	1
Contrast	1	0
Topic chaining	38	11
Topic-dominant chaining	20	2
Elaboration	0	3
Reason	0	0
Cause	4	0
Example	0	0
Total	65	17

Figure 5 shows the video news report we used in our experiment. As shown, shot (b) and (c) were merged together because there was no pause at the cut point between them. Sentence (I), (II), (III), (IV), and (V) were associated with shot (a), (b)(c), (d), (e), and (f), respectively. Coherence relations between video segments were estimated in the following way: a topic-dominant chaining relation was estimated between shot (a) and (b)(c) because "Prime Minister Obuchi" was found in the topic part of sentence (II) and in the non-topic part of sentence (I). The same relation was also estimated between shot (b)(c) and (d) because "the Fiscal Structural Reform Law" was found in the topic part of sentence (III) and in the non-topic part of sentence (II). On the contrary, topic chaining relation was estimated between shot (a) and (e) because "the Ministry of Finance" was found in the topic parts of sentence (I) and (IV). In this case, the system detected also another relation: a topic-dominant chaining relation between shot (b) and (e). However, the system selected the former one because the former exceeded the latter in score. The system also determined a topic chaining relation between shot (e) and (f). In this case, the system additionally detected two other relations: topic chaining relation between shot (a) and (f), and topic-dominant chaining relation between shot (b) and (f). But the relation between (a) and (f) was chosen, because its reliable score was greater than the score between (b) and (f) and equal to the score between (e) and (f), but there the distance between the shots was greater. Figure 6 shows the result of this analysis.

As shown in Table 1, 11 topic chaining and 2 topic-dominant chaining relations could not be extracted. The reasons were (1) the topic words of the following sentences were omitted [3] and (2)

[3] There are many ellipses in Japanese sentences.

	shot	transcript
(a)		(I) In accordance with instructions of Prime Minister Obuchi, the Ministry of Finance will decide about new guidelines for budget requests which is free from the restrictions of the Fiscal Structural Reform Law.
(b)		(II) Prime Minister Obuchi called Vice Minister Tanami, the Ministry of Finance, into the Official Residence, and instructed him to make new budget request guidelines in line with the freeze policy of the Fiscal Structural Reform Law.
(c)		
(d)		(III) The Fiscal Structural Reform Law sets upper limits for expenditures in all categories but social security.
(e)	大蔵省 公共事業・中小企業関係費 ➡上限超え認める方向で新基準策定へ	(IV) In accordance with prime minister's instruction, the Ministry of Finance establishes new guidelines in which key government expenditures, for example, public works projects, are permitted to go beyond the limits of the Law.
(f)		(V) the Ministry of Finance will decide about the new guidelines by the middle of the next week, and government ministries and agencies will submit their budget requests in accordance with this guideline by the end of the month.

Figure 5: An example of news video ("New guidelines for budget requests", NHK evening news, August/3/1998)

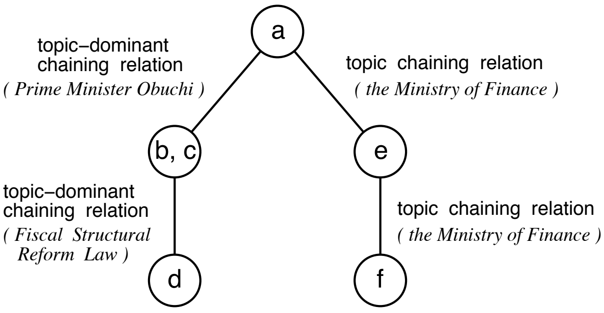

a
topic–dominant chaining relation
(Prime Minister Obuchi)
topic chaining relation
(the Ministry of Finance)
b, c
e
topic–dominant chaining relation
(Fiscal Structural Reform Law)
topic chaining relation
(the Ministry of Finance)
d
f

Figure 6: The result of discourse structural analysis for the news video shown in Figure 5

the topic word was changed (e.g., driver → man who drove the car) or abbreviated. Also, 3 elaboration relations could not be extracted. This was because there were no clue expressions for the elaboration relation in the sentences. However, the system could mostly detect clue expressions and occurrence of identical/synonymous words/ phrases. In some cases (e.g., a compound sentence), there were many clues for an NS supporting various relations to several CSs. The system could detect them, however, extracted only one CS and relation. In this study, we introduce a reliable score for determining the most plausible CS and relation. As shown in Table 1, this method is useful, however, we should investigate a method for extracting more CSs and relations than one when several CSs and relations exist.

In this study, we assumed that image and language data correspond to the same portion of a news report. For this reason, it is likely that the relation between images slightly differs from the analysis result when image and language are taken form different portions (correspondence problem between image and language).

At the end of this section, we discuss video summarization using discourse structure information. First, we consider summarization of the news video shown in Figure 5 with the summarization topic concerning the Ministry of Finance. The summarization system traces topic chaining relations and generates video summarization which consists of shots (a), (e), and (f). Next, we consider summarization of the same news video with the summarization topic concerning the Prime Minister. The system traces a topic-dominant chaining relation and generates video summarization which consists of shot (a), (b), and (c).

References

[Grosz 86] Grosz and Sidner: Attention, Intentions, and the Structures of Discourse, Computational Linguistics, 12-3, (1986).

[Hobbs 85] Hobbs: On the Coherence and Structure of Discourse, Technical Report No. CSLI-85-37, (1985).

[Iwanari 94] Iwanari and Ariki: Scene Clustering and Cut Detection in Moving Images by DCT components, (in Japanese), technical report of IEICE, PRU-93-119, (1994).

[Kurohashi 92] Kurohashi and Nagao: Dynamic Programming Method for Analyzing Conjunctive Structures in Japanese, COLING-92, (1992).

[Kurohashi 94] Kurohashi and Nagao: Automatic Detection of Discourse Structure by Checking Surface Information in Sentences, COLING-94, (1994).

[Marcu 97] Marcu: From Discourse Structures to Text Summaries, ACL workshop on Intelligent Scalable Text Summarization, (1997).

[Miike 94] Miike, Itoh, Ono, and Sumita: A Full-Text Retrieval System with a Dynamic Abstract Generation Function, SIGIR-94, (1994).

[Nagao 00] Nagao, Shirai, and Hashida: Multimedia Data Summarization Based on the Global Document Annotation, (in Japanese), 6th Annual Meeting of The Association for Natural Language Processing, (2000).

[Nakamura 97] Nakamura and Kanade: Semantic Analysis for Video Contents Extraction – Spotting by Association in News Video, ACM Multimedia 97, (1997).

[Smith 97] Smith and Kanade: Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques, IEEE CVPR, (1997).

[Sumita 92] Sumita, Ono, Chino, Ukita, and Amano: A Discourse Structure Analyzer for Japanese Text, International Conference of Fifth Generation Computer Systems, (1992).

[Watanabe 96] Watanabe, Okada, and Nagao: Semantic Analysis of Telops in TV Newscasts, (in Japanese), technical report of Information Processing Society of Japan, NL-116–16, (1996).

[Zadrozny 91] Zadrozny and Jensen: Semantics of Paragraphs, Computational Linguistics, 17-2, (1991).

[Zhang 95] Zhang, Low, Smoliar, and Wu: Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution, ACM Multimedia 95, (1995).