Role of Text Structure for Summary Generation: Clues for Sentence Combination

Kazuhiro Takeuchi
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takamaya, Ikoma, Nara 630-0101, JAPAN
kazuh-ta@is.aist-nara.ac.jp

Abstract

In this paper we clarify the role of text structure for summary generation. We investigate two problems. One is to estimate the consistency of human analysis in text structure. The other is to investigate what role the text structure plays to generate a summary. We analyze aligned sentences in a source text and those in its summary that human subject produced. As a result, we find that the relations between adjacent sentences in text structure play special roles for one of the crucial operations in producing a summary. To assess the clues behind the relations and the operation, we apply a machine learning program.

1 Introduction

In the research of automated summarization, some researchers such as Ono et al.(1994) and Marcu(1998) use tree structure models to represent text structure and to select important subparts in texts. By doing this, they exploit relations among the sentences in a text.

Although text structure plays an important role in developing an automated summarization system, there is no concrete model to make a summary through a representation of text structure. Moreover, what representation is suitable for generating a summary is not clear. We investigate mainly two problems in this paper. One is to estimate the consistency of human analysis on text structure. The other is to investigate what kind of role the text structure plays to generate a summary. First of all, we set up an experimental scheme to analyze text structure. We ask human subjects to produce summaries of texts, where the structure of the source texts is analyzed in advance. After those preparations, we examine how the alignment is done between the sentences in the source text and the sentences in its summary. By means of the structure and the alignment analysis, we confirmed that the text structure plays an important role in generating a summary. In particular, pairs of adjacent sentences that have a direct relationship in the text structure exhibit a special role in both of the stages; in the analysis of text structure and in the generation of summaries. We could regard such a relationship as the clues for sentence combination, which is one of the crucial operations for a human to produce a summary. To investigate the characteristics of the relationship further, we apply a machine learning method using some linguistic features and make it sure that the clues are to act as the trigger for sentence combination.

2 Analyzing Text Structures

2.1 Coding Scheme for Text Structures

In general, properties of a text are classified in terms of the linguistic notion of cohesion and coherence. According to Halliday and Hasan(1976), the notion of cohesion is closely related with linguistic cues such as anaphora, ellipses, conjunctions, lexical relations, etc. Those linguistic cues contribute to create semantic connectedness in a text. Compared to cohesion, coherence is related to more abstract semantic structure of a text. Although a number of models for text structure analysis have been proposed, there is not yet a specific model that can provide us with useful information for generating summaries.

In the previous researches we described in the introduction, Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) is used to represent the text structure. RST is one of the well-known models for text structure representation and is mainly used to represent coherence of texts. With RST we can decompose a text into sub-parts forming a hierarchical structure. Every sub-part has a relationship to another sub-part with one of the relation types (rhetorical relations). These relations form an overall coherence structure of the text.

We propose a coding scheme with which a hu-

man subject analyzes text structure in order to investigate how well the representation of text structure works for summarization. In the coding scheme, we modify RST in two ways as described below to help a subject to code the text structure.

First, since the original RST proposes more than 20 types of rhetorical relations, a human subject often finds difficulty in selecting a proper relation between sentences. Furthermore, as Moore and Pollack(1992) point out, a pair of sentences sometimes inherently has a multiple-analysis. To avoid such complexities we introduce two fundamental relations in terms of the degree of importance between a pair of sentences. One is the relation where there is unclear distinction in degree of importance between the pair. The other is the relation where the pair has relative degree of importance. We assume this simplification does not contradict with other's work in automated text summarization using text structure.

Second, we define a sentence as the elementary unit of text structure. However, in the original RST, a more fine-grained fragment (which usually corresponds to a clause) is considered as the elementary unit. Since the coherence between sentences is our main interest in this paper, we currently assume a sentence as the elementary unit.

We developed an annotating tool for text structure analysis as shown in Figure 1. It helps the subject to annotate the texts based on simplified version of RST described above. On the screenshot, boxes correspond to the sentences in the annotating text and arrows stand for coherence relations between sentences. In practice, a subject simply selects a tag through a graphical user interface(GUI) of the tool. For each sentence S_t in the annotating text, a subject acts as follows:
1) selects the most relevant sentence S_t while S_t must be more important than or equal to S_s with regard to the meaning of the whole text. (S_s is the root of the text structure if S_s does not have such a sentence in the text. The root must be only a single sentence in a text.)
2) selects the relation type for the pair of sentences (S_t and S_s).

2.2 Evaluation of the Analyzed Text Structures

We let three human subjects analyze text structures using our coding scheme described in the previous section. Note that the analysis is done by a different group of subjects from ones that produce summaries. We use 32 Japanese report articles from Nihon-keizai-shinbun (Japanese financial newspaper) in 1995. The total number of sentences in the articles is 500.

To observe the overall tendency among agree-

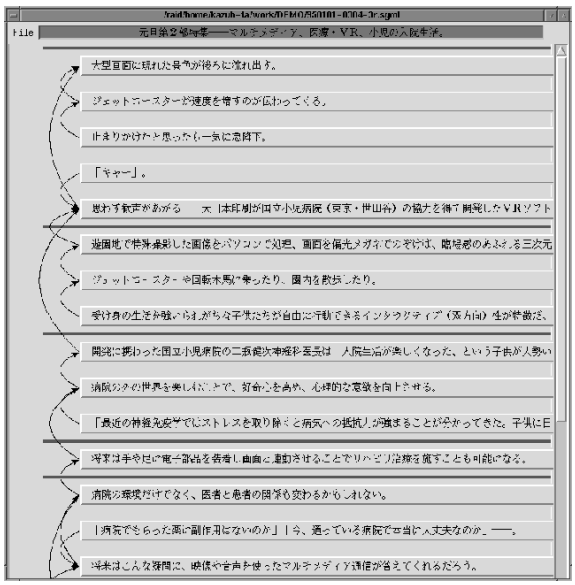

Figure 1: Tool for annotating a text structure

ment of subjects' analysis, we use the Kappa coefficient, which is used to assess agreement in the area of behavioral science. Carletta (1996) introduces and discusses the use of it as an agreement measure in the discourse analysis. The Kappa coefficient measures pairwise agreement among a set of coders making category judgments and is defined as:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of frequency that the subjects agree and $P(E)$ is the proportion of frequency that they agree by chance. Intuitively, the measure shows the degree of agreement adjusted by the agreement by chance.

In our experiment of text structure analysis by human subjects, P(A) stands for the agreement ratio of selected sentences. [1] We obtain Kappa coefficient of 0.58 (P(A) = 0.63, P(E)=0.11). In the evaluation of agreement using Kappa coefficient, there is a guideline that is used to evaluate the degree of reliability of the agreement in a coding scheme; 0 < Kappa < 0.2 is regarded as "slight" agreement, 0.21 to 0.40 as "fair", 0.41 to 0.60 as "moderate", 0.61 to 0.80 as "substantial", and 0.81 to 1.0 as "near perfect" (Carletta and others, 1997). According to the guideline, our

[1] Unfortunately, we have not discovered the role of two relation types that we distinguished in the text structure on the stage of summary generation. Since we need to investigate the problem more in the further work, we concentrate on the selection of sentence pairs on the analysis of the text structure in this paper.

Table 1: Agreement ratio against Distance

n	at least 1 subj. selected	majority selected	ratio(n)
1	352	293	0.832
2	113	48	0.425
3	53	21	0.396
4	36	14	0.389
n≥5	85	29	0.341

coding scheme is evaluated as "moderate" level, close to the "substantial" level with the overall agreement.

Then, we observe the tendency of the assignments of relation types between pairs of sentences. We found that agreement rate of the assignment for the relation between adjacent pairs of sentences is much higher than that of the other pairs. The difference of tendency among subjects' assignment is observed when it is evaluated in terms of the distance between dependent sentences as shown in Table 1. The distance of relation is defined as the relative distance between the dependent sentences. If a sentence relates with the preceding sentence, the distance of relation is 1. We define the ratio as follows:

$$ratio(n) = \frac{\text{the number of pairs that have distance } n \text{ (agreed by the majority of subjects)}}{\text{the number of pairs that have distance } n \text{ (assigned by at least one subject)}}$$

From Table 1, compared with other distances the relations of distance = 1 agree more frequently. This indicates that even a human has difficulty in judging the relation between sentences when the distance is two or more. On the other hand, when two related sentences are adjacent, the judgment by humans is significantly more accurate.

Figure 2 shows an example of text structure in the representation of RST. In the figure, a number corresponds to the position of a sentence and an arrow corresponds to a relationship between sentences. To summarize the result using the representation, the relationships between adjacent sentences such as $2 \rightarrow 1$ agrees more often by human subjects compared with the relations like $4 \rightarrow 1$ and $8 \rightarrow 4$.

In the rest of this paper, we regard the coherence structure that are determined by the majority of the subjects as the text structure.

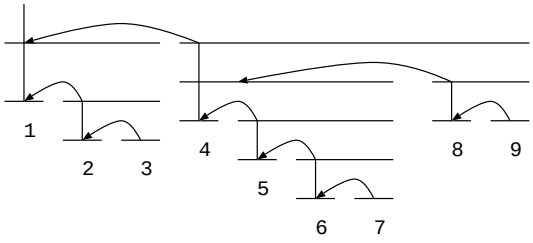

Figure 2: Representation of a text structure in RST

Table 2: Source Texts and Their Summaries

(Average number)	Source Texts	Summaries A	Summaries B
# Characters	882.6	342.0	320.4
# Sentences	16.1	7.2	6.9

3 Analysis on Human-generated Summaries

3.1 Making Summaries

We ask two Japanese native speakers to summarize 20 texts. They are educated in literature, but are not professional summarizers. We ask them to generate summaries up to the length of 40% of the source texts. We pick up the 20 source texts to be summarized from the set of texts of which structure have been analyzed beforehand according to the experimental analyzing scheme described in section 2.1. The group of human subjects who made summaries are different from the group of the subjects who analyzed the text structures. Since we assume that human can generate summaries without explicit information of text structure and word importance, we removed paragraph breaks and the titles from the texts in the experiment.

In the summarization task, we give two instructions:

- The summary should keep the overall story of the source text and author's main opinion.
- If the summarizer uses proper nouns in the summary, the form of the proper nouns should be kept as the originals.

Basic information about the source texts and their summaries are shown in Table 2. In the table and the rest of this paper, we refer to the summaries made by summarizer A as *Summary A* and those of summarizer B as *Summary B*.

Table 3: The number of sentence reduction and sentence combination

Operation Type	Summaries	
	A	B
sentence reduction	94	110
sentence combination	49	27

Table 4: Sentence Position in Source Texts and in Text Structure

Position of pairs	Summary A	Summary B
adjacent	38(34)	19(18)
non-adjacent	11	8
Total	49	27

3.2 Operations for Generating Summary

In order to analyze the human behavior in generating summaries, it is important to know the alignment between the sentences in the human-generated summary and the corresponding source sentences that have been used to generate the sentences. There are some related work of the summarizing task. For example, Jing and McKeown (1999) identified 6 operations in human summary generating process. In this paper, we focus on the operations where a summary sentence is generated from one sentence or more and classify the operations into following two types.

1. sentence reduction: the summarized sentence is generated from exactly one source sentence.
2. sentence combination: the summarized sentence originates from two or more source sentences.

We manually perform the alignment. For each summary sentence, two human subjects select which sentences in the source text are used in summarization. Only case where two subjects agreed with the analysis of the alignment, we accept the human analysis as valid. For other cases, we accept the most similar sentence in the source text to the summary sentence by using word-based cosine similarity.

Table 3 shows the number of two types of operations in summary generation. Comparing the number of sentence reduction with that of sentence combination, the former has more examples than the later in both cases. However, 103 source sentences are used to create 49 combined sentences in summary A, and 58 source sentences are used to create 27 combined sentences in summary B. This shows that the number of the sentences used in sentence combination with respect to the total number of source sentences for summary should not be ignored.

3.3 Coherence Structure and Sentence Combination

From the alignment result, we notice an explicit tendency in the sentence combination operation on the source text. Table 4 shows the number of sentence pairs combine into summary sentences. In both set of summaries, the summary sentences that are generated by sentence combination are mostly the adjacent sentences. However, even if a pair of adjacent sentences in a source text is likely to combine, the pair is not always combined to the summary sentence. We assumed the clues for sentence combination is adjacency relation in the coherence structure. This assumption comes from the fact that the adjacency relationship in coherence structure is comparatively easy for humans to analyze as we described in section 2. The bracketed figures on the upper line in Table 4 shows the number of examples that have coherence relation. This result shows that the text structure can act as the clues for sentence combination.

On the other hand, the number of examples of sentence combination of non-adjacent sentences is not enough to draw a conclusion. We think that the structure between non-adjacent sentences involves more complex mechanism than adjacency relation.

4 Clues for Sentence Combination

As we described in Section 3, we discover that the coherence relationship between a pair of adjacent sentences plays an important role in summary generation. In this section, we investigate clues for those relationships. Our investigation consists of two steps. The first step is to see how well a machine predicts whether a pair of adjacent sentences has a coherence relationship or not. The second step is to see whether the clues of the relationships between adjacent sentences work as the clues for the sentence combinations as well.

4.1 Clues for Relationships between Adjacent Sentences

In order to investigate the clues for the relationship between pairs of adjacent sentences, we apply the machine-learning program C4.5 (Quinlan, 1992). C4.5 is a decision tree learning program that acquires general rules from the training examples that consist of features and the target class. In our learning task, the target class is assigned using the relations between the pairs of adjacent sentences in a coherence structure described in section 2.3. Suppose S_{i-1} and S_i are

Table 5: Features for Characterizing a pair between adjacent sentences

Feature Categories	Features	ID
Syntactic Features	cue words of S_i	CUE
	predicate type of S_{i-1}	PRD0
	predicate type of S_i	PRD1
	topic marker type of S_i	TPC
	omission of topic or subject of the S_i	OMIT
Semantic Features	S_i introduce a new proper noun / S_i refers the proper noun in S_{i-1}	NEWT
	Character based similarity between S_i and S_{i-1}	SIM

a pair of adjacent sentences in a text. If S_i has relationship to S_{i-1} in the coherence structure, the pair of adjacent sentences classified as "yes", and "no", otherwise.

We arrange the information of an example into the sets of features as shown in Table 5. Since the relations in coherence structure are abstract and complex, we use not only syntactic information, but also information that influences the meaning. In practice, our features can be divided into two categories; syntactic features and semantic features.

Syntactic features represent the characteristics of a sentence using the result of syntactic structure analysis. We used automated word dependency structure analyzing program developed by Fujio et al.(1998). Japanese dependency structure is usually defined in terms of the relationship between phrases called 'bunsetu.' The relationships reflect the underlying syntactic structure of a sentence. Bunsetu is a segment that consists of one or more words and that includes a head word. Fujio's program outputs not only the syntactic relationship between Bunsetus but also the pos (part-of-speech) tags of the words. Since the reliability of the program is not perfect yet, we manually correct the result of the dependencies when the result has some errors. Value of the syntactic features is assigned using the analyzed dependency structure.

The value of CUE feature is assigned based on the type of conjunctive expressions at the beginning of the corresponding sentence (e.g., *For example, However, Therefore,* etc. in English). Since conjunctive expressions provide cohesive connectivity between adjacent sentences in general, we divide conjunctive expressions into 10 types according to the function of connectivity (e.g. Exemplify, Contrast, Restating, etc.). The features PRD0 and PRD1 represent the type of the predicates of S_{i-1} and S_i. Each predicate type of the sentence is classified according to its modality or its tense. These two features are expected to indicate the writer's attitudes in the sentence. TPC feature represents the type of topic marker in the sentence. In Japanese, the grammatical role such as topic and subject is marked by particles called postpositions. We distinguish the grammatical role using the analyzed dependency structure. OMIT feature shows whether the topic or the subject of the sentence S_i is present or not. The absence will show the stronger cohesion between S_i and S_{i-1}. Some researchers such as Walker et al.(1994) claim that the entity that marked by topic maker and its omission play an important role in the discourse. Note that the values of TPC and OMIT are hard to be assigned only by the pos information. Therefore, we use the analyzed dependency structure to assign the features.

On the other hand, semantic features represent the relevancy between the contents described in the two sentences. In order to represent the meaning of the sentences, we use two features as an approximate relevancy between the sentences: NEWT and SIM features. NEWT feature represents whether S_i introduces new proper nouns as a topic or not. We assign four different values to this feature based on four cases of S_i described bellow [2] :

1) the topic/subject includes the proper nouns that S_{i-1} does not include.
2) the topic/subject includes referring expressions that refer to the proper nouns of S_{i-1}.
3) a part of S_i (except for the topic/subject) refers to the proper noun of S_{i-1}.
4) S_i satisfies none of the above cases.

Proper nouns and expressions referring the proper nouns in the text are manually annotated. In this experiment, we annotated only person names, place names and organization names as proper nouns and referring expressions are restricted to pronouns and nouns that refer to the proper nouns without any inference. SIM feature uses similarity

[2] The notation of 'topic/subject' in this paper stands for the subpart of a sentence. The subpart contains not only the entity marked as topic or subject, but also the phrases modifying the entity.

between the subpart that expresses topic/subject in S_i and the whole S_{i-1}. The similarity is calculated by character (Kanji) based cosine similarity. If topic/subject is absent in S_i, the value of the SIM is assigned 1.0 (i.e. SIM includes a piece of information of OMIT feature). We expect that the SIM feature gives the simple approximation for NEWT features. We, however do not claim that our features set is an exhaustive one.

We evaluate the learned decision tree using the "leave-one-out" cross validation as follows. For every example x_i in the training example set, a decision tree learns from all the examples except the x_i and the learned tree is evaluated by x_i. Table 6 shows the results in terms of precision, recall and accuracy that are defined by following equations.

$$Precision = \frac{\#\text{ examples decision-tree classified correctly}}{\#\text{ examples decision-tree classified}}$$

$$Recall = \frac{\#\text{ examples decision-tree classified correctly}}{\#\text{ examples human classified}}$$

$$Accuracy = \frac{\#\text{ examples decision-tree classified correctly}}{\#\text{ examples}}$$

We assume a system which always classify examples as "yes" as a baseline, whose accuracy is 0.626. Comparing with the baseline, from the obtained results, we can conclude that the features are able to predict whether a sentence has coherence relationship to the preceding sentence.

Moreover, we also evaluate the impact of each individual feature for the learned decision tree to classify the example into the target class correctly. We remove each feature in turn from the set of features listed in Figure 5 and apply the same process to construct the decision tree with the reduced member of the features. The accuracy of each learned decision tree in shown in Figure 3. The figure shows when CUE, PRD1 or SIM feature is removed from the original features, the accuracy decreases significantly. The result shows that these three features have the ability to capture the characteristics of the relationship between the adjacent sentences. Although we carefully designed NEWT in the semantic features, the feature does not show good effect on the accuracy.

4.2 Discussion on the clues for sentence combination

In this section, we want to verify that the clues for relations between adjacent sentences are also the clues for sentence combination. We conduct an experiment by using the reliability of prediction for coherence relationship between pairs of

Table 6: Evaluation for Learned Decision Tree

	target class	
	yes	no
Precision	0.715	0.638
Recall	0.850	0.434
Overall Accuracy	0.697	

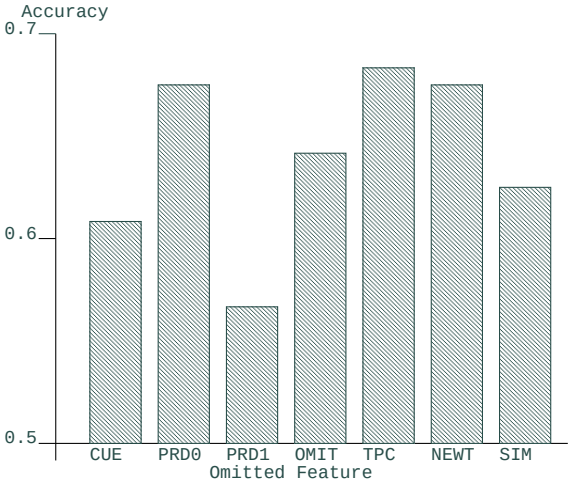

Figure 3: Accuracy of the decision tree for classifying the relations between adjacent sentences without each feature

adjacent sentences. The reliability is provided by the learned decision tree and takes value from 0 to 1. The higher the reliability is, the more likely the sentence relates to the preceding adjacent sentence. In practice, we calculate the reliability as follows: A leaf node of the learned decision tree is constructed to classify the training examples. Note that the leaf node does not always classify the "real" training examples when the decision tree is pruned. We regard the confidence-ratio of classifying the training examples with the leaf node of the learned decision tree as the reliability. For example, if the all of the corresponding training examples to a particular leaf node are classified into class "yes", the reliability of the leaf node is 1.0.

We compare the average reliability for the pairs of adjacent sentences that have coherence relationship with that of the pairs of sentences in which sentence combination takes place. As shown in Table 7, both the average reliability and its standard deviation(stds) give similar tendencies between the pair of sentences that have coherence relationship and the pair of sentences where sentence combination occurs. Thus, the clues for sentence combination characterized by our features are quite similar to the ones for pairs of adjacent

Table 7: Averaged Reliability for Pair of Adjacent Sentences

type of the pair of sentences	Reliability	
	Ave.	stds
coherence relation	0.631	0.301
sentence combination	0.615	0.295

sentences with coherence relation.

In this section, we show that the strong relationship between adjacent sentences give us some hints to understand the operation of sentence combination. The relationships between adjacent sentences must represent the relevancy between the contexts or meaning of the corresponding sentences. Since the sentences produced by the operation of sentence combination are more constrained in terms of the relevancy, the investigation of the sentence combination will help us understand how to represent the coherence structure. Practically, the strength of coherence relation triggers the sentence combination operation in generating summary.

5 Conclusion

In this paper, we investigate human-generated summaries from the point of view of text structure. Even if human generates a summary without explicit discourse clues (such as paragraph breaks), the coherence structure analyzed in this paper can represent implicit clues for automated summarization. Our conclusion includes the followings.

- In analysis of coherence structure of text, even a human has difficulty in judging the relation between sentences when the related sentences stay apart from each other. On the other hand, when two related sentences are adjacent, the human judgment is far more accurate.
- Human summary generation is based on two types of operation; sentence reduction and sentence combination. An existence of relationship between the pair of adjacency sentences is a trigger to determine whether the operation of sentence combination is activated or not.
- Coherence relationships between adjacent sentences are automatically identified using features. We confirm that the features characterizing the relation can also characterize the occurrence of sentence combination.

Our work provides new perspective for automated summary generation. The perspective has three steps. At the first step, we analyze the strength of coherence relation between every adjacent pair in its source sentences. Second, we combine strongly related adjacent pairs to a new summary sentence. Finally, we reduce redundant clauses and words from the combined summary sentences. Based on this experimental results, we are motivated to develop a full-fledged summary generating system in the future.

References

Jean Carletta et al. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics, Vol.23, No.1*, pages 13–31.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, Vol.22, No.2*, pages 249–254.

Masakazu Fujio and Yuji Matsumoto. 1998. Japanese dependency structure analysis based on lexicalized statistics. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing*, pages 88–96.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English.* Longman.

Hongyan Jing and Kathleen R.McKeown. 1999. The decomposition of human-written summary sentences. In *SIGIR '99 *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, University of California, Berkeley*, pages 129–136.

William C. Mann and Sandra A. Thompson. 1987. *Rhetorical Structure Theory: A Theory of Text Organization.* Tech. Report ISI/RS-87-190.

Daniel Marcu. 1998. To build text summaries of high quality, nuclearity is not sufficient. In *Intelligent Text Summarization *Papers from the 1998 AAAI Spring Symposium Technical Report SS-98-06*, pages 1–8.

Johanna D. Moore and Martha E. Pollack. 1992. A problem for rst: The need for multi-level discourse analysis. *Computational Linguistics, Vol.18, No.4*, pages 537–544.

Kenji Ono et al. 1994. Abstract generation based on rhetorical structure extraction. In *COLING-94, Vol.1*, pages 344–348.

J. Ross Quinlan. 1992. *C4.5: Programs for Machine Learning.* The Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann.

Matilyn Walker et al. 1994. Japanese discourse and the process of centering. *Computational Linguistics, Vol.20, No.2*, pages 193–232.