A Bottom-up Approach to Sentence Ordering for Multi-document Summarization

Danushka Bollegala [*,1] , Naoaki Okazaki , Mitsuru Ishizuka

*Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1,
Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan*

Abstract

Ordering information is a difficult but important task for applications generating natural-language texts such as multi-document summarization, question answering, and concept-to-text generation. In multi-document summarization, information is selected from a set of source documents. However, improper ordering of information in a summary can confuse the reader and deteriorate the readability of the summary. Therefore, it is vital to properly order the information in multi-document summarization. We present a bottom-up approach to arrange sentences extracted for multi-document summarization. To capture the association and order of two textual segments (e.g. sentences), we define four criteria: *chronology*, *topical-closeness*, *precedence*, and *succession*. These criteria are integrated into a criterion by a supervised learning approach. We repeatedly concatenate two textual segments into one segment based on the criterion, until we obtain the overall segment with all sentences arranged. We evaluate the sentence orderings produced by the proposed method and numerous baselines using subjective gradings as well as automatic evaluation measures. We introduce the average continuity, an automatic evaluation measure of sentence ordering in a summary, and investigate its appropriateness for this task.

Key words: sentence ordering, multi-document summarization, natural language processing

[*] Corresponding author.
Email addresses: danushka@mi.ci.i.u-tokyo.ac.jp (Danushka Bollegala), okazaki@is.s.u-tokyo.ac.jp (Naoaki Okazaki), ishizuka@i.u-tokyo.ac.jp (Mitsuru Ishizuka).

[1] Research Fellow of the Japan Society for the Promotion of Science (JSPS)

1 Introduction

Multi-document summarization (MDS) (Radev and McKeown, 1999; Carbonell and Goldstein, 1998; Elhadad and McKeown, 2001) tackles the information overload problem by providing a condensed and coherent version of a set of documents. Among a number of sub-tasks involved in MDS including sentence extraction, topic detection, sentence ordering, information extraction, and sentence generation most MDS systems have been based on an extraction method, which identifies important textual segments (e.g. sentences or paragraphs) in source documents. It is important for such MDS systems to determine a coherent arrangement for the textual segments extracted from multi-documents, in order to reconstruct the text structure for summarization.

A summary with improperly ordered sentences confuses the reader and degrades the quality/reliability of the summary itself. Barzilay et al. (2002) has provided empirical evidence to show that the proper order of extracted sentences significantly improves their readability. Lapata (2006) experimentally shows that the time taken to read a summary strongly correlates with the arrangement of sentences in the summary.

For example, consider the three sentences shown in Figure 1, selected from a reference summary in Document Understanding Conference (DUC) 2003 dataset. The first and second sentences are extracted from the same source document, whereas the third sentence is extracted from a different document. Although all three sentences are informative and talk about the storm, *Gilbert*, the sentence ordering shown in Figure 1 is inadequate. For example, the phrase, *such storms*, in sentence 1, in fact refers to *Category* 5 *storms*, described in sentence 2. A better arrangement of sentences in this example would be 3-2-1.

In single document summarization, where a summary is created using only one document, it is natural to arrange the extracted information in the same order as in the original document. In contrast, for multi-document summarization, we need to establish a strategy to arrange sentences extracted from different documents . Ordering extracted sentences into a coherent summary is a non-trivial task. For example, identifying rhetorical relations (Mann and Thompson, 1988) in a document has been a difficult task for computers. However, our task of constructing a coherent summary from an unordered set of sentences is even more difficult. Source documents for a summary may have been written by different authors, have different writing styles, or written on different dates, and based on the different background knowledge. We cannot expect a set of extracted sentences from such a diverse set of documents to be coherent on their own.

The problem of information ordering is not limited to automatic text summarization, and concerns natural language generation applications. A typical natural lan-

(1) Such storms have maximum sustained winds greater than 155 mph and can cause catastrophic damage.
(2) Earlier Wednesday, Gilbert was classified as a Category 5 storm, the strongest and deadliest type of hurricanes.
(3) Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.

Fig. 1. Randomly ordered sentences in a summary

guage generation (NLG) (Reiter and Dale, 2000a) system consists of six components: content determination, discourse planning, sentence aggregation, lexicalization, referring expression generation, and orthographic realization. Among those, information ordering is particularly important in discourse planning, and sentence aggregation (Karamanis and Manurung, 2002; Duboue and McKeown, 2002, 2001). In concept-to-text generation (Reiter and Dale, 2000b), given a concept (e.g. a keyword, a topic, or a collection of data), the objective is to produce a natural language text about the given concept. For example, consider the case where generating game summaries, given a database containing statistics of American football. A sentence ordering algorithm can support a natural language generation system by helping to order the sentences in a coherent manner.

In this paper, we propose four criteria to capture the association of sentences in the context of multi-document summarization for newspaper articles. These criteria are then integrated into one criterion by a supervised learning approach. We propose a bottom-up approach in arranging sentences, which repeatedly concatenates textual segments until the overall segment with all sentences is arranged. Moreover, we investigate several automatic evaluation measures for the task of sentence ordering in multi-document summarization because subjective evaluations of sentence orderings is time consuming and difficult to reproduce. The proposed method outperforms existing sentence ordering algorithms, and shows a high correlation (Kendall's τ of 0.612) with manually ordered sentences. Subjective evaluations made by human judges reveal that the majority of summaries (ca. 64%) produced by the proposed method are coherent (i.e. graded as *perfect* or *acceptable*). Among the semi-automatic evaluation measures investigated in our experiments, we found that Kendall's τ coefficient has the best correlation with subjective evaluations.

2 Related Work

Existing methods for sentence ordering are divided into two approaches: making use of chronological information (McKeown et al., 1999; Lin and Hovy, 2001; Barzilay et al., 2002; Okazaki et al., 2004), and learning the natural order of sentences from large corpora (Lapata, 2003; Barzilay and Lee, 2004; Ji and Pulman, 2006). A newspaper usually disseminates descriptions of novel events that have occurred since the last publication. For this reason, the chronological ordering of sentences is an effective heuristic for multi-document summarization (Lin and Hovy, 2001; McKeown et al., 1999). Barzilay et al. (2002) proposed an improved version of chronological ordering by first grouping sentences into sub-topics discussed in the source documents, then arranging the sentences in each group chronologically.

Okazaki et al. (2004) proposed an algorithm to improve the chronological ordering by resolving the presuppositional information of extracted sentences. They assume that each sentence in newspaper articles is written on the basis that presuppositional information should be transferred to the reader before the sentence is interpreted. The proposed algorithm first arranges sentences in a chronological order, and then estimates the presuppositional information for each sentence by using the content of the sentences placed before each sentence in its original article. The evaluation results show that the proposed algorithm improves the chronological ordering significantly.

Lapata (2003) presented a probabilistic model for text structuring and its application in sentence ordering. Her method computes the transition probability from one sentence to the next in two sentences, from a corpus based on the Cartesian product using the following features: verbs (precedent relationships of verbs in the corpus), nouns (entity-based coherence by keeping track of the nouns), and dependencies (structure of sentences). Lapata (2006) also proposed the use of Kendall's rank correlation coefficient (Kendall's τ) for the automatic evaluation that quantifies the differences between orderings produced by an algorithm and by a human. Although she has not compared her method with chronological ordering, it could be applied to generic domains, not relying on the chronological clue specific to newspaper articles.

Barzilay and Lee (2004) proposed *content models* to deal with the topic transition in domain specific text. The content models are implemented by Hidden Markov Models (HMMs), in which the hidden states correspond to topics in the domain of interest (e.g. earthquake magnitude or previous earthquake occurrences), and state transitions capture possible information-presentation orderings. The evaluation results showed that their method outperformed Lapata's approach by a wide margin. They did not compare their method with chronological ordering as an application of multi-document summarization.

Ji and Pulman (2006) proposed a sentence ordering algorithm using a semi-supervised sentence classification and historical ordering strategy. Their algorithm includes three steps: the construction of sentence networks, sentence classification, and sentence ordering. First, they represent a summary as a network of sentences. Nodes in this network represent sentences in a summary, and edges represent transition probabilities between two nodes (sentences). Next, the sentences in the source documents are classified into the nodes in this network. The probability $p(c_k|s_i)$, of a sentence s_i in a source document belonging to a node c_k in the network, is defined as the probability of observing s_k as a sample from a Markov random walk in the sentence network. Finally, the extracted sentences are ordered to the weights of the edges. They compare the sentence ordering produced by their method against manually ordered summaries using Kendall's τ. Unfortunately, they do not compare their results against the chronological ordering of sentences, which has been shown to be an effective sentence ordering strategy in multi-document news summaries.

As described above, several good strategies/heuristics to deal with the sentence ordering problem have been proposed. In order to integrate multiple strategies/heuristics, we have formalized them in a machine learning framework, and have considered an algorithm to arrange the sentences using the integrated strategy.

3 Method

We define notation $a \succ b$ to represent that sentence a precedes sentence b. We use the term *segment*, to describe a sequence of ordered sentences. When segment A consists of sentences a_1, a_2, ..., a_m in this order, we denote it as:

$$A = (a_1 \succ a_2 \succ ... \succ a_m). \tag{1}$$

The two segments A and B can be ordered as either B after A, or A after B. We define the notation $A \succ B$ to show that segment A precedes segment B.

Let us consider a bottom-up approach in arranging the sentences. Starting with a set of segments initialized with a sentence for each, we concatenate two segments, with the strongest association (discussed later) of all possible segment pairs, into one segment. Repeating the concatenating will eventually yield a segment with all sentences arranged. The algorithm is considered as a variation of agglomerative hierarchical clustering, with the ordering information retained at each concatenating process.

The underlying idea of the algorithm, a bottom-up approach to text planning, was proposed by Marcu (1997). Assuming that the semantic units (sentences) and their rhetorical relations (Mann and Thompson, 1988) (e.g., sentence *a* is an *elaboration* of sentence *d*) are given, he modeled the text structuring task as a problem of finding the best discourse tree that satisfies the set of rhetorical relations. He stated that the global coherence could be achieved by satisfying local coherence constraints in ordering and clustering, thereby ensuring that the resultant discourse tree was well-formed.

Unfortunately, identifying the rhetorical relation between two sentences has been a difficult task for computers (Marcu, 2000). However, the bottom-up algorithm for arranging sentences can still be applied only if the direction and strength of

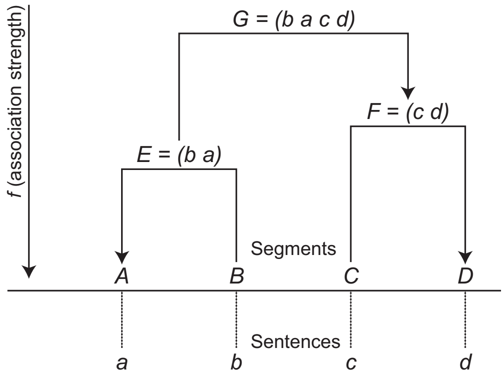


Fig. 2. Arranging four sentences A, B, C, and D with a bottom-up approach.

the association of the two segments (sentences) are defined. Hence, we introduce a function $f(A \succ B)$ to represent the direction and strength of the association of two segments, A and B,

$$f(A \succ B) = \begin{cases} p & \text{(if } A \text{ precedes } B) \\ 0 & \text{(if } B \text{ precedes } A) \end{cases}, \tag{2}$$

where p $(0 \leq p \leq 1)$ denotes the association strength of the segments A and B. The association strengths of the two segments with different directions, e.g., $f(A \succ B)$ and $f(B \succ A)$, are not always identical by our definition,

$$f(A \succ B) \neq f(B \succ A). \tag{3}$$

Figure 2 shows the process of arranging four sentences a, b, c, and d. We first initialize four segments with a sentence for each,

$$A = (a), B = (b), C = (c), D = (d). \tag{4}$$

Supposing that $f(B \succ A)$ has the highest value of all possible pairs, e.g., $f(A \succ B)$, $f(C \succ D)$, etc, we concatenate B and A to obtain a new segment,

$$E = (b \succ a). \tag{5}$$

Then, we search for the segment pair with the strongest association. Supposing that $f(C \succ D)$ has the highest value, we concatenate C and D to obtain a new segment,

$$F = (c \succ d). \tag{6}$$

Finally, comparing $f(E \succ F)$ and $f(F \succ E)$, we obtain the final sentence ordering,

$$G = (b \succ a \succ c \succ d). \tag{7}$$

Algorithm 1 Sentence ordering algorithm.

```
1: $P \leftarrow \{(s_1), (s_2), \ldots\}$
2: while $|P| > 1$ do
3:     $(p_a, p_b) \leftarrow \arg\max_{p_i, p_j \in P,\ p_i \neq p_j} f(p_i \succ p_j)$
4:     for $s \in p_b$ do
5:         $p_a \leftarrow p_a \oplus s$
6:     end for
7:     $P \leftarrow P \setminus \{p_b\}$
8: end while
9: return $P$
```

Algorithm 1 presents the pseudo code of the sentence ordering algorithm. Algorithm 1 takes a set of extracted sentences S as input, and returns a single segment of ordered sentences. First, for each sentence s_i in S, we create a segment p_i that contains s_i only. Subsequently, we find the two segments, p_a and p_b in the set of segments P, that have the maximum strength of association (Line No. 3). The *for* loop in lines 4-6 then appends the sentences in segment p_b to the end of segment p_a. The operator $\oplus$ in Line No. 5 denotes this appending operation. We then remove the segment p_b from P (Line No. 7). This process is repeated until we are left with a single segment in P. In Algorithm 1, we use the notation $|P|$ to denote the number of elements (i.e. segments) in P.

In the above description, we have not defined the association of two segments. We define four criteria to capture the association of two segments: *chronology*, *topical-closeness*, *precedence*, and *succession*. These criteria are integrated into a function $f(A \succ B)$ by using a machine learning approach. The rest of this section explains the four criteria, and an integration method with a Support Vector Machine (SVM) (Vapnik, 1998) classifier.

3.1 Chronology criterion

Chronology criterion reflects the chronological ordering (Lin and Hovy, 2001; McKeown et al., 1999), by which sentences are arranged in the chronological order of publication timestamps. A newspaper usually deals with novel events that have occurred since the last publication. Consequently, the chronological ordering of sentences has shown to be particularly effective in multi-document news summarization. As already discussed in Section 2, previous studies have proposed sentence ordering algorithms using chronological information. Publication timestamps are used to decide the chronological order among sentences extracted from different documents. However, if no timestamp is assigned to documents, or if several documents have the identical timestamp, the chronological ordering does not provide a clue for sentence ordering. Inferring temporal relations among events (Mani et al., 2003; Mani and Wilson, 2000) using implicit time references (such as tense system) (Lapata and Lascarides, 2006), and explicit time references (such as temporal adverbials) (Filatova and Hovy, 2001), might provide an alternative clue for chronological ordering. However, inferring temporal relations across a diverse set of multiple documents is a difficult task. Consequently, by assuming the availability of temporal information in the form of timestamps, we define the strength of association in arranging segments B after A, measured by a chronology criterion $f_{\text{chro}}(A \succ B)$ in the following formula:

$$f_{\text{chro}}(A \succ B) = \begin{cases} 1 & \text{T}(a_m) < \text{T}(b_1) \\ 1 & [\text{D}(a_m) = \text{D}(b_1)] \wedge [\text{N}(a_m) < \text{N}(b_1)] \\ 0.5 & [\text{T}(a_m) = \text{T}(b_1)] \wedge [\text{D}(a_m) \neq \text{D}(b_1)] \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

Here, a_m represents the last sentence in segment A, b_1 represents the first sentence in segment B, $T(s)$ is the publication date of the sentence s, $D(s)$ is the unique identifier of the document to which sentence s belongs, and $N(s)$ denotes the line number of sentence s in the original document. The chronological order of segment B arranging after A is determined by comparing the last sentence in the segment A and the first sentence in the segment B.

The chronology criterion assesses the appropriateness of arranging segment B after A if sentence a_m is published earlier than sentence b_1, or if sentence a_m appears before b_1 in the same article. For sentences extracted from the same source document, preferring the original order in the source document has proven to be effective for single document summarization (Barzilay et al., 2002). The second condition in the chronological criterion defined in formula 8 imposes this constraint. If sentence a_m and b_1 are published on the same day, but appear in different articles, the criterion assumes the order to be undefined. If none of the above conditions are satisfied, the criterion estimates that segment B will precede A. By assigning a score of zero for this condition in formula 8, the chronological criterion guarantees that sentence orderings which contradicts with the definition of chronological ordering are not produced.

In addition to the formulation of chronology criterion defined by Formula 8, in our preliminary experiments we tried alternatives that consider the absolute time difference between the publication dates of articles. For two sentences extracted from different articles (i.e. D(a_m)≠D(b_1)), we defined the chronological distance between them as the difference of publication dates in days. The chronological distances in a summary are normalized to values in range $[0, 1]$ by dividing from the maximum value of chronological distances. However, we did not find any significant improvement in the sentence orderings produced by this alternative approach in our experiments. Therefore, we only consider the simpler version of chronological criterion defined in Formula 8.

(a) The earthquake crushed cars, damaged hundreds of houses and terrified people for hundreds of kilometers around.
(b) A major earthquake measuring 7.7 on the Richter scale rocked north Chile Wednesday.
(c) Authorities said two women, one aged 88 and the other 54, died when they were crushed under collapsing walls.

Fig. 3. Three sentences from a summary about an earthquake.

3.2 Topical-closeness criterion

A set of documents discussing a particular event usually contains information related to multiple topics. For example, a set of newspaper articles related to an earthquake typically contains information about the magnitude of the earthquake, its location, casualties, and rescue efforts. Grouping sentences by topics has shown to improve the readability of a summary (Barzilay et al., 2002; Barzilay and Lee, 2004). For example, consider the three sentences shown in Figure 3, selected from a summary of an earthquake in Chile. Sentences (a) and (c) in Figure 3 present details about the damage by the earthquake, whereas sentence (b) conveys information related to the magnitude and location of the earthquake. In this example, sentences (a) and (c) can be considered as topically related. Consequently, when the three sentences are ordered as show in Figure 3, we observe abrupt shifts of topics from sentence (a) to (b), and from (b) to (c). A better arrangement of the sentences that prevents such disfluencies is (b)-(a)-(c).

The topical-closeness criterion deals with the association of two segments, based on their topical similarity. The criterion reflects the ordering strategy proposed by Barzilay et al. (2002), which groups sentences referring to the same topic. To measure the topical closeness of two sentences, we represent each sentence by using a vector. First, we remove *stop words* (i.e. functional words such as *and, or, the*, etc.) from a sentence and lemmatize verbs and nouns. Second, we create a vector in which each element corresponds to the words (or lemmas in the case of verbs and nouns) in the sentence. Values of elements in the vector are either 1 (for words that appear in the sentence) or 0 (for words that do not appear in the sentence). [2]

We define the topical closeness of two segments A and B as follows,

$$f_{\text{topic}}(A \succ B) = \frac{1}{|B|} \sum_{b \in B} \max_{a \in A} \text{sim}(a, b). \tag{9}$$

[2] Using the frequencies of words instead of the binary $(0, 1)$ values as vector elements, did not have a positive impact in our experiments. We think this is because, compared to a document, a sentence typically has a lesser number of words, and a word does not appear many times in a single sentence.

(a) Honduran death estimates grew from 32 to 231 in the first days, to $6,076$ with $4,621$ missing.
(b) Honduras braced as category 5 Hurricane Mitch approached.
(c) The EU approved 6.4 million in aid to Mitch's victims.

Fig. 4. Precedence relations in a summary

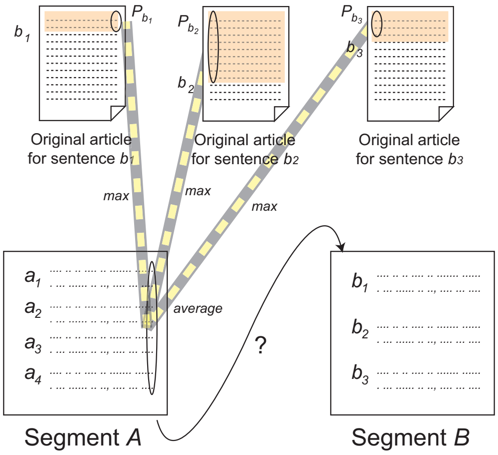


Fig. 5. Precedence criterion

Here, $\mathrm{sim}(a, b)$ denotes the similarity of sentences a and b, calculated by the cosine similarity of two vectors corresponding to the sentences. For sentence $b \in B$, $\max_{a \in A} \mathrm{sim}(a, b)$ yields the similarity between sentences b and $a \in A$, which is the most similar to b. The topical-closeness criterion $f_{\mathrm{topic}}(A \succ B)$ assigns a higher value when the topic referred to by segment B is the same as by segment A.

3.3 Precedence criterion

In extractive multi-document summarization, only the important sentences that convey the main points discussed in source documents are selected to be included in the summary. However, a selected sentence can presuppose information from other sentences that were not selected by the sentence extraction algorithm. For example, consider the three sentences shown in Figure 4, selected from a summary on hurricane Mitch. Sentence (a) describes the after-effects of the hurricane, whereas sentence (b) introduces the hurricane. To understand the reason for the deaths mentioned in sentence (a), one must first read sentence (b). Consequently, it is appropriate to arrange the three sentences in Figure 4 in the order (b)-(a)-(c). In general, it is difficult to perform such an in-depth logical inference on a given set of sentences. Instead, we use source documents to estimate precedence relations. For example, assuming that in the source document where sentence (a) was extracted, there exist a sentence that is similar to sentence (b), we can conclude that sentence (b) should precede sentence (a) in the summary.

To formally define the precedence criterion, let us consider the case illustrated in Figure 5, where we arrange segment A before B. Each sentence in segment B has the presuppositional information such as background information or introductory facts that should be conveyed to a reader in advance. Given sentence $b \in B$, such presuppositional information may be presented by the sentences appearing before the sentence b in the original article. However, we cannot guarantee whether a sentence-extraction method for multi-document summarization chooses any sentences before b for a summary, because the extraction method usually determines a set of sentences within the constraint of summary length that maximizes information coverage and excludes redundant information. The *precedence criterion* measures the substitutability of the presuppositional information of segment B (e.g, the sentences appearing before sentence b) as segment A. This criterion is a formalization of the sentence-ordering algorithm proposed by Okazaki et al. (2004).

We define the precedence criterion in the following formula,

$$f_{\mathrm{pre}}(A \succ B) = \frac{1}{|B|} \sum_{b \in B} \max_{a \in A, p \in P_b} \mathrm{sim}(a, p). \tag{10}$$

Here, $f_{pre}(A \succ B)$ is the strength of association for ordering segment B after A, measured using the precedence criterion. P_b is a set of sentences appearing before sentence b in the original article from which b was extracted. If b is the first sentence in its source document, then P_b is the empty set. For each sentence p in set P_b we compute the cosine similarity $\mathrm{sim}(a, b)$ between p and sentences a in segment A. Cosine similarity between sentences are computed exactly as described in the topical-closeness criterion. We find the maximum similarity between p and any sentence from segment A. Finally, we average the similarity scores by dividing from the number of sentences in segment B. Figure 5 shows an example of calculating the precedence criterion for arranging segment B after A. We approximate the presuppositional information for sentence b by sentences P_b, i.e., sentences appearing before the sentence b in the original article. Calculating the maximum similarity in the possible pairs of sentences in P_b and A, Formula 10 is interpreted as the average similarity of the precedent sentences $\forall P_b (b \in B)$ to the segment A.

3.4 Succession criterion

In extractive multi-document summarization, sentences that describe a particular event are extracted from a set of source articles. Usually, there exist a logical sequence among the information conveyed in the extracted sentences. For example, in Figure 3, sentence (a) describes the results of the earthquake described in sentence (b). It is natural to order a sentence that describes the result or an effect of a certain cause after a sentence that describes the cause. Therefore, in Figure 3, sentence (a) should be ordered after sentence (b) to create a coherent summary. We use the in-

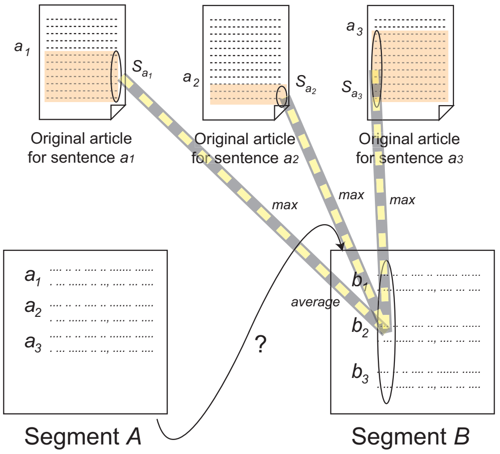

a1
Sa1
Original article for sentence a1
a2
Sa2
Original article for sentence a2
a3
Sa3
Original article for sentence a3
max
max
max
average
?
a1
a2
a3
b1
b2
b3
Segment A
Segment B

Fig. 6. Succession criterion

formation conveyed in source articles to propose *succession criterion* to capture the coverage of information for sentence ordering in multi-document summarization.

The succession criterion assesses the coverage of the succeeding information for segment A by arranging segment B after A:

$$f_{\mathrm{succ}}(A \succ B) = \frac{1}{|A|} \sum_{a \in A} \max_{s \in S_a, b \in B} \mathrm{sim}(s, b). \tag{11}$$

Here, for each sentence a in segment A, S_a denotes the set of sentences appearing after sentence a in the original article (i.e. article from which a was extracted). For each sentence s in set S_a, we compute the cosine similarity $\mathrm{sim}(s, b)$, between sentences s and sentences b in segment B. Cosine similarity is computed exactly as described in the topical-closeness criterion. Figure 6 shows an example of calculating the succession criterion to arrange segments B after A. We approximate the information that should follow segment A by the sentences in segments S_a. We then compare each segment S_a with segment B to measure how much segment B covers this information. The succession criterion measures the substitutability of the succeeding information, (e.g., the sentences appearing after the sentence $a \in A$) such as segment B.

3.5 SVM classifier to assess the integrated criterion

We described four criteria for measuring the strength and direction of association between two segments of texts. However, it is still unclear which criteria and conditions increase the performance. A human may use a combination of criteria to produce a summary. Thus, we use summaries created by humans as training data to find the optimum combination of the proposed criteria so that the combined function fits to the human-made summary. We integrate the four criteria: chronology,

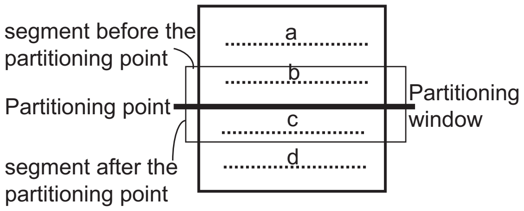

segment before the partitioning point
Partitioning point
segment after the partitioning point
a
b
c
d
Partitioning window

Fig. 7. Partitioning a human-ordered extract into pairs of segments

$$\begin{aligned}+1 &: [f_{\text{chro}}(A \succ B), f_{\text{topic}}(A \succ B), f_{\text{pre}}(A \succ B), f_{\text{succ}}(A \succ B)]\\ -1 &: [f_{\text{chro}}(B \succ A), f_{\text{topic}}(B \succ A), f_{\text{pre}}(B \succ A), f_{\text{succ}}(B \succ A)]\end{aligned}$$

Fig. 8. Two vectors in a training data generated from two ordered segments $A \succ B$

topical-closeness, precedence, and succession, to define the function $f(A \succ B)$ to represent the association direction and strength of the two segments A and B (Formula 2). More specifically, given the two segments A and B, function $f(A \succ B)$ yields the integrated association strength based on four values, $f_{\text{chro}}(A \succ B)$, $f_{\text{topic}}(A \succ B)$, $f_{\text{pre}}(A \succ B)$, and $f_{\text{succ}}(A \succ B)$. Formalizing the integration task as a binary classification problem, we employ a Support Vector Machine (SVM) to model the function.

We partition a human-ordered extract into pairs each of which consists of two non-overlapping segments. Let us explain the partitioning process taking four human-ordered sentences, $a \succ b \succ c \succ d$ shown in Figure 7. Firstly, we place the partitioning point just after the first sentence a. Focusing on sentences a and b at the boundary of the partition point, we extract the pair $\{(a),(b)\}$ of two segments (a) and (b). Enumerating all possible pairs of two segments appearing just before/after the partitioning point, we obtain the following pairs, $\{(a),(b)\}$, $\{(a),(b \succ c)\}$ and $\{(a),(b \succ c \succ d)\}$. Similarly, segment pairs, $\{(b),(c)\}$, $\{(a \succ b),(c)\}$, $\{(b),(c \succ d)\}$, $\{(a \succ b),(c \succ d)\}$, are obtained from the partitioning point between sentences b and c. Collecting the segment pairs from the partitioning point between sentences c and d (i.e. $\{(c),(d)\}$, $\{(b \succ c),(d)\}$ and $\{(a \succ b \succ c),(d)\}$), in total, ten pairs were extracted from the four sentences shown in Figure 7. In general, this process yields $n(n^2-1)/6$ pairs from ordered n sentences. From each pair of segments, we generate one positive and one negative training instance as follows.

Given a pair of two segments A and B, arranged in an order $A \succ B$, we obtain a positive training instance (labeled as $+1$) by computing a four dimensional vector (Figure 8) with the following elements: $f_{\text{chro}}(A \succ B)$, $f_{\text{topic}}(A \succ B)$, $f_{\text{pre}}(A \succ B)$, and $f_{\text{succ}}(A \succ B)$. Similarly, we obtain a negative training instance (labeled as -1) corresponding to $B \succ A$. We use a manually ordered set of summaries and assume an ordering $A \succ B$ as a positive sentence ordering (for training purposes) if in a manually ordered summary the sentence A precedes the sentence B. For such two sentences A and B, we consider the ordering $B \succ A$ as a negative sentence ordering (for training purposes). Accumulating these instances as training data, we construct a binary classifier modeled by a Support Vector Machine. The SVM classifier yields the association direction of two segments (e.g. $A \succ B$ or $B \succ A$) with the class information (i.e., $+1$ or -1).

We assign the association strength of two segments by using the posterior probability that the instance belongs to a positive $(+1)$ class. When an instance is classified into a negative (-1) class, we set the association strength as zero (see the definition of Formula 2). Because SVM is a large-margin classifier, the output of an SVM is the distance from the decision hyperplane. However, the distance from a hyperplane is not a valid posterior probability. We use sigmoid functions to convert the distance into a posterior probability (see Platt (2000) for a detailed discussion on this topic).

4 Evaluation

4.1 Outline of Experiments

This section outlines the numerous experiments that we conduct to evaluate the proposed sentence ordering algorithm. Further details of each experiment will be given in the sections to follow. It is noteworthy that the proposed method is only a *sentence ordering* algorithm and not a complete text summarization system. For example, a typical extractive multi-document summarization system would first select sentences from a given set of documents, and then order the selected sentences to produce a coherent summary. However, the proposed method does not extract any sentences from a set of documents, but assumes that sentence extraction has already completed, and only focuses on ordering those extracted sentences. Therefore, it cannot be directly compared with text summarization algorithms which perform sentence extraction such as maximum marginal relevance (MMR) (Carbonell and Goldstein, 1998). Consequently, we compare the proposed sentence ordering algorithm with previously proposed sentence ordering algorithms for multi-document summarization. In our experiments, all sentence ordering algorithms are given the same set of extracted sentences (a set of sentences are extracted for each topic in advance), and only the ordering of sentences is different in the summaries produced by different algorithms.

Our experiment dataset is described in Section 4.2. First, we subjectively evaluate the summaries produced by the proposed method and numerous other sentence ordering algorithms in Section 4.3. Specifically, we compare the proposed agglomerative clustering-based sentence ordering algorithm (**AGL**) with six other sentence ordering algorithms: random ordering (**RND**), human-made ordering (**HUM**), chronological ordering (**CHR**), topical-closeness ordering (**TOP**), precedence ordering

(**PRE**), and succedence ordering (**SUC**). Here, **RND** and **HUM** respectively correspond to the lower and upper baselines of sentence ordering. Details of these sentence ordering algorithms are presented later in Section 4.2. We asked three human judges to independently rate each sentence ordering produced by the different sentence ordering algorithms using four grades: *perfect*, *acceptable*, *poor*, and *unacceptable*. The guidelines for grading and the results of the subjective evaluation are detailed in Section 4.3.

Section 4.4 introduces three semi-automatic evaluation measures for sentence ordering. Specifically, we define Kendall's rank correlation coefficient (i.e. Kendall's τ), Spearman's rank correlation coefficient, and the average continuity measure. Kendall's τ and Spearman coefficient are used in previous work on evaluating sentence orderings. Average continuity is a novel metric that we propose in this paper. All three semi-automatic evaluation measures compare a sentence ordering produced by a system under evaluation against a human-made reference sentence ordering. Moreover, in Section 4.5 we extend the semi-automatic evaluation measures to incorporate more than one reference orderings. We present the experimental results of our semi-automatic evaluation in Section 4.6. A good semi-automatic evaluation measure must have a high degree of correlation with subjective gradings provided by humans. In Section 4.7, we compare the correlation between semi-automatic evaluation measures defined in the paper with subjective gradings. Finally, in Section 4.8 we experimentally evaluate the contribution of the four sentence ordering criteria (i.e. chronology, topical-relatedness, precedence, and succession) used by the proposed method. Specifically, we train and test the proposed sentence ordering algorithm by holding out each criterion at a time and measure the difference in performance using semi-automatic evaluation measures.

4.2 Experiment Dataset

We evaluated the proposed method using the 3rd Text Summarization Challenge (TSC-3) corpus [3] . Text Summarization Challenge is a multiple document summarization task organized by the "National Institute of Informatics Test Collection for IR Systems" (NTCIR) project [4] . TSC-3 dataset was introduced in the 4th NTCIR workshop held in June 2-4, 2004. The TSC-3 dataset contains multi-document summaries for 30 news events. The events are selected by the organizers of the TSC task. For each topic, a set of Japanese newspaper articles are selected using some query words. Newspaper articles are selected from Mainichi Shinbun and Yomiuri Shinbun, two popular Japanese newspapers. All newspaper articles in the dataset have their date of publication annotated. Moreover, once an article is published, it is not revised or modified. Therefore, all sentences in an article bares the time

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

[4] http://research.nii.ac.jp/ntcir/index-en.html

Table 1
Correlation between two sets of human-ordered extracts

Metric	Mean	Std. Dev	Min	Max
Spearman	0.739	0.304	-0.2	1
Kendall	0.694	0.290	0	1
Average Continuity	0.401	0.404	0.001	1

stamp of the article.

Although we use Japanese text summaries for experiments, it is noteworthy that there are no fundamental differences between Japanese and English text summarization. In fact, popular summarization algorithms originally designed for English text summarization, such as the maximum marginal relevance (MMR) (Carbonell and Goldstein, 1998), have been successfully employed to summarize Japanese texts (Mori and Sasaki, 2002).

For each topic, the organizers of the TSC task provide a manually extracted set of sentences. On average, a manually extracted set of sentences for a topic contains 15 sentences. The participants of the workshop are required to run their multi-document summarization systems on newspaper articles selected for each of the 30 topics and submit the results to the workshop organizers. The output of each participating system is compared against the manually extracted set of sentences for each of the topics using precision, recall and F-measure. Essentially, the task evaluated in TSC is sentence extraction for multi-document summarization.

In order to construct the training data applicable to the proposed method, we asked two human subjects to arrange the extracts. The two human subjects worked independently and arranged sentences extracted for each topic. They were provided with the source documents from which the sentences were extracted. They read the source documents before ordering sentences in order to gain background knowledge on the topic. From this manual ordering process, we obtained $30(\mathrm{topics}) \times 2(\mathrm{humans}) = 60$ sets of ordered extracts. Table 1 shows the agreement of the ordered extracts between the two subjects. The correlation is measured by three metrics: Spearman's rank correlation, Kendall's rank correlation, and average continuity. Definitions of these automatic evaluation measures are described later in section 4.4. The mean correlation values (0.74 for Spearman's rank correlation and 0.69 for Kendall's rank correlation) indicate a strong agreement in sentence orderings made by the two subjects. In 8 out of the 30 extracts, sentence orderings created by the two human subjects were identical.

We applied the leave-one-out method to the proposed method, to produce a set of sentence orderings. In this experiment, the leave-out-out method arranges an extract by using an SVM model trained from the rest of the 29 extracts. Repeating this process 30 times with a different topic for each iteration, we generated a set of

30 orderings for evaluation. In addition to the proposed method, we prepared six sets of sentence orderings produced by different baseline algorithms. We outline the seven algorithms (including the proposed method):

Agglomerative ordering (AGL) is an ordering arranged by the proposed method.

Random ordering (RND) is the lowest anchor, in which sentences are arranged randomly.

Human-made ordering (HUM) is the highest anchor, in which sentences are arranged by a human subject.

Chronological ordering (CHR) arranges sentences with the chronology criterion defined in Formula 8. Sentences are arranged in chronological order of their publication date (i.e. sentences belonging to articles published earlier are ordered ahead of sentences belonging to articles published later. Among sentences belonging to the same source article, we order them according to the order in which they appear in the article. Chronological ordering cannot define an order for sentences belonging to articles with identical publication dates/times. Ordering among such sentences are decided randomly.

Topical-closeness ordering (TOP) arranges sentences with the topical-closeness criterion defined in Formula 9. Ties are resolved randomly.

Precedence ordering (PRE) arranges sentences with the precedence criterion defined in Formula 10. Ties are resolved randomly.

Succedence ordering (SUC) arranges sentences with the succession criterion defined in Formula 11. Ties are resolved randomly.

The last four algorithms (CHR, TOP, PRE, and SUC) arrange sentences by the corresponding criterion alone, each of which uses the association strength directly without integrating other criteria. These orderings are expected to show the performance of each criterion, and their contribution to the sentence ordering problem.

4.3 Subjective grading

Evaluating sentence orderings is a challenging task. Intrinsic evaluation, which involves human judges to rank a set of sentence orderings, is a necessary approach to this task (Barzilay et al., 2002; Okazaki et al., 2004; Lapata, 2006; Karamanis and Mellish, 2005; Madnani et al., 2007). This section describes an intrinsic evaluation with subjective grading, followed by semi-automatic evaluation measures described in Section 4.4.

We asked three human judges to rate sentence orderings according to the following criteria. [5]

[5] The human judges that participated in this evaluation are different from the two annota-

Perfect A *perfect* summary is a text that we cannot improve any further by re-ordering.

Acceptable An *acceptable* summary is one that makes sense, and is unnecessary to revise even though there is some room for improvement in terms of its readability.

Poor A *poor* summary is one that loses the thread of the story at some places, and requires minor amendments to bring it up to an acceptable level.

Unacceptable An *unacceptable* summary is one that leaves much to be improved and requires overall restructuring rather than partial revision.

To avoid any disturbance in rating, we inform the judges that the summaries were made from a same set of extracted sentences, and that only the ordering of sentences is different. Furthermore, the judges were given access to the source documents for each summary. Figure 9 shows a summary that obtained a *perfect* grade. The ordering $1-4-5-6-7-8-2-3-9-10$ was assigned an *acceptable* grade, whereas $4-5-6-7-1-2-3-8-9-10$ was given a *poor* grade. A random ordering of the ten sentences $4-7-2-10-8-3-1-5-6-9$ received an *unacceptable* grade.

We conduct subjective grading on random ordering (RND), chronological ordering (CHR), topical-closeness ordering (TOP), precedence ordering (PRE), succession ordering (SUC), the proposed sentence ordering algorithm (AGL), and the human-made orderings (HUM). It is noteworthy that the TOP, PRE, and SUC criteria cannot be used to produce a total ordering of sentences on their own, because they cannot decide the first sentence in a summary. To produce a sentence ordering using these criteria on their own, we used them as the strength of association in Algorithm 1. For example, to produce a sentence ordering using only precedence criterion, we use Formula 10 as the strength of association function f in Algorithm 1. Kendall's coefficient of concordance (Kendall's W), which assesses the inter-judge agreement of overall ratings, reported a higher agreement between the two judges ($W = 0.873$). Figure 10 shows the distribution of the subjective grading made by the three judges. Each set of orderings has 30(topics) $\times$ 3(judges) $= 90$ ratings. Most RND orderings are rated as *unacceptable*. Out of the four criteria introduced in section 3, CHR has the largest number of *perfect* orderings. Although CHR and AGL orderings have roughly the same number of *perfect* orderings (ca. 24% for CHR and 27% for AGL), the AGL algorithm gained more *acceptable* orderings (37%) than the CHR algorithm (29%). When we counted both *perfect* and *acceptable* summaries, the ratio for the proposed AGL algorithm was 64%, while the CHR reached only 53%. This result suggests that the proposed method successfully incorporated the chronological ordering with other criteria. However, a huge gap between AGL and HUM orderings was also found. In particular, the judges rated 27% AGL orderings as *perfect*, while the figure rose as high as 77% for HUM

tors that created the two sets of reference summaries. All three judges are native Japanese speakers and graduate school students, majoring in information engineering.

(1) Hurricane Gilbert, one of the strongest storms ever, slammed into the Yucatan Peninsula Wednesday and leveled thatched homes, tore off roofs, uprooted trees and cut off the Caribbean resorts of Cancun and Cozumel.
(2) Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
(3) Gilbert reached Jamaica after skirting southern Puerto Rico, Haiti and the Dominican Republic.
(4) The Mexican National Weather Service reported winds gusting as high as 218 mph earlier Wednesday with sustained winds of 179 mph.
(5) More than 120,000 people on the northeast Yucatan coast were evacuated, the Yucatan state government said.
(6) Shelters had little or no food, water or blankets and power was out.
(7) The storm killed 19 people in Jamaica and five in the Dominican Republic before moving west to Mexico.
(8) Prime Minister Edward Seaga of Jamaica said Wednesday the storm destroyed an estimated 100,000 of Jamaica's 500,000 homes when it throttled the island Monday.
(9) The National Hurricane Center said a hurricane watch was in effect on the Texas coast from Brownsville to Port Arthur and along the coast of northeast Mexico from Tampico north.
(10) The National Hurricane Center said Gilbert was the most intense storm on record in terms of barometric pressure.

Fig. 9. An example of a *perfect* grade summary.

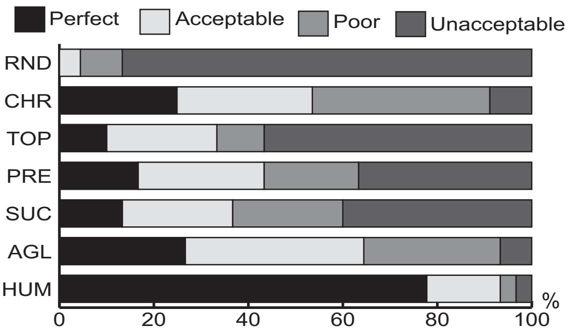


Fig. 10. Subjective grading

orderings.

4.4 Methods for semi-automatic evaluation

Even though subjective grading consumes much time and effort, we cannot reproduce the evaluation afterwards. Automatic evaluation measures are particularly useful when evaluations must be performed quickly and repeatedly to tune an algorithm . Previous studies (Barzilay and Lee, 2004; Lapata, 2003, 2006) employed

$$T_{eval} = (e \succ a \succ b \succ c \succ d)$$
$$T_{ref} = (a \succ b \succ c \succ d \succ e)$$

Fig. 11. An example of an ordering under evaluation T_{eval} and its reference T_{ref}.

rank correlation coefficients including Spearman's rank correlation coefficient, and Kendall's rank correlation coefficient (Kendall's τ), to compare a sentence ordering produced by a system against a manual ordering. In this section, we briefly survey the existing semi-automatic evaluation measures (specifically, Kendall τ and Spearman rank correlation coefficient), and introduce *average continuity* as an alternative evaluation measure for sentence orderings.

Let $S = s_1 \ldots s_N$ be a set of N items to be ranked. Let π and σ denote two distinct orderings of S. Then, Kendall's rank correlation coefficient (Kendall, 1938) (also known as Kendall's τ) is defined as follows,

$$\tau = \frac{4C(\pi, \sigma)}{N(N-1)} - 1. \tag{12}$$

Here, $\mathrm{C}(\pi, \sigma)$ is the number of concordant pairs between π and σ (i.e. the number of sentence pairs that have the same relative positions in both π and σ). For example, in Figure 11 between T_{eval} and T_{ref}, there are six concordant sentence pairs: (a, b), (a, c), (a, d), (b, c), (b, d), and (c, d). These six concordant pairs yield a Kendall's τ of 0.2. Kendall's τ is in the range $[-1, 1]$. It takes the value 1 if the two sets of orderings are identical, and -1 if one is the reverse of the other.

Likewise, Spearman's rank correlation coefficient (r_s) between orderings π and σ is defined as follows,

$$r_s = 1 - \frac{6}{N(N+1)(N-1)} \sum_{i=1}^{N} (\pi(i) - \sigma(i))^2. \tag{13}$$

Here, $\pi(i)$ and $\sigma(i)$ respectively denote the ith ranked item in π and σ. Spearman's rank correlation coefficient for the example shown in Figure 11 is 0. Spearman's rank correlation, r_s, ranges from $[-1, 1]$. Similarly to Kendall's τ, the r_s value of 1 is obtained for two identical orderings, and the r_s computed between an ordering and its revers is -1.

In addition to Spearman's and Kendall's rank correlation coefficients, we propose an *average continuity* metric, which extends the idea of the continuity metric (Okazaki et al., 2004), to continuous k sentences.

A text with sentences arranged in proper order does not interrupt the process of a human reading from one sentence to the next. Consequently, the quality of a sentence ordering produced by a system can be estimated by the number of continuous sentence segments that it shares with the reference sentence ordering. For example, in Figure 11 the sentence ordering produced by the system under evaluation (T_{eval}) has a segment of four sentences ($a \succ b \succ c \succ d$), which appears exactly in that order in the reference ordering (T_{ref}). Therefore, a human can read this segment without any disfluencies and will find to be coherent.

This is equivalent to measuring a precision of continuous sentences in an ordering against the reference ordering. We define P_n to measure the precision of n continuous sentences in an ordering to be evaluated as,

$$P_n = \frac{m}{N - n + 1}. \tag{14}$$

Here, N is the number of sentences in the reference ordering, n is the length of continuous sentences on which we are evaluating, and m is the number of continuous sentences that appear in both the evaluation and reference orderings. In Figure 11, we have two sequences of three continuous sentences (i.e., $(a \succ b \succ c)$ and $(b \succ c \succ d)$). Consequently, the precision of 3 continuous sentences P_3 is calculated as:

$$P_3 = \frac{2}{5 - 3 + 1} = 0.67. \tag{15}$$

The Average Continuity (AC) is defined as the logarithmic average of P_n over 2 to k:

$$\text{AC} = \exp\left(\frac{1}{k-1}\sum_{n=2}^{k} \log(P_n + \alpha)\right). \tag{16}$$

Here, k is a parameter to control the range of the logarithmic average, and α is a fixed small value. It prevents the term inside the logarithm from becoming zero in case P_n is zero. We set $k = 4$ (i.e. more than five continuous sentences are not included for evaluation), and $\alpha = 0.001$. The average continuity is in the range $[0, 1]$. It becomes 0 when the evaluation and reference orderings share no continuous sentences and 1 when the two orderings are identical.

Let us compute the average continuity between the two orderings T_{eval} and T_{ref} as shown in Figure 11. First, we observe that there are three segments of two consecutive sentences (i.e. in Formula 14, for $n = 2$, $m = 3$) between T_{eval} and T_{ref}. Namely, ($a \succ b$), ($b \succ c$), and ($c \succ$ d). Therefore, P_2 computed using Formula 14 is 0.75 (i.e. $3/(5-2+1)$). Similarly, we observe there are two segments of three consecutive sentences between T_{eval} and T_{ref}. They are ($a \succ b \succ c$) and ($b \succ c \succ d$). Therefore, P_3 is 0.67 (i.e. $2/(5-3+1)$). Finally, we observe that there exist a single segment of four consecutive sentences (i.e. $a \succ b \succ c \succ d$) between T_{eval} and T_{ref}.

$$\begin{aligned} T_{eval} &= (e \succ a \succ b \succ c \succ d) \\ T_A &= (a \succ b \succ c \succ d \succ e) \\ T_B &= (b \succ c \succ d \succ e \succ a) \end{aligned}$$

Fig. 12. Comparing multiple reference orderings with a system ordering.

Therefore, P_4 as computed using Formula 14 is 0.5 (i.e. $(1/(5-4+1))$). There are no segments with five or more consecutive sentences. We then use Formula 16 to compute average continuity between T_{eval} and T_{ref} as follows,

$$\mathrm{AC} = \exp\left(\frac{1}{4-1} \times (\log(P_2+\alpha) + \log(P_3+\alpha) + \log(P_4+\alpha))\right).$$

Substituting values for P_2, P_3, and P_4 computed as above, and setting $\alpha = 0.001$, we obtain an average continuity of 0.63 between T_{eval} and T_{ref}.

The underlying idea of Formula 16 is similar to that of BLEU metric (Papineni et al., 2002), which was developed for the semi-automatic evaluation of machine-translation (MT) systems. BLEU compares the overlap between a translation produced by an MT system and a translation created by a human using n-grams (of words). If the two translations (by the MT system and by the human) share a large number of n-grams, then the MT system receives a high BLEU score. BLEU first computes the precision for n-grams with different lengths and then computes the geometric mean of all precision values using a formula similar to Formula 16 used to compute average continuity. Between two orderings of the same set of sentences, the set of words that appear in both orderings is identical; only the ordering of sentences is different. Therefore, average continuity measures the overlap between a summary produced by a system and a human-made ordering of those sentences using segments of continuous sentences instead of n-grams of words.

4.5 Using multiple reference orderings to evaluate sentence orderings

There can be more than one way to order a set of sentences to create a coherent summary. Therefore, to evaluate a sentence ordering produced by an algorithm, we must compare it with multiple reference orderings produced by different human annotators. However, all evaluation measures described in Section 4.4 compare a system ordering against one reference ordering. In this section, we modify the evaluation measures introduced in Section 4.4 to handle more than one reference ordering.

When comparing a system ordering against multiple reference orderings using the Kendall's τ, we consider a sentence pair (a, b) in the system ordering to be concordant if at least one of the reference orderings has sentence a before sentence b. For example, let us consider the case shown in Figure 12 in which we compare a system ordering, T_{eval}, against two reference orderings T_A and T_B. If we only use T_A as the reference, then pair (e, a) in T_{eval} will not be a concordant pair. However, if we compare T_{eval} against T_B, we find that e is followed by a. Therefore, we conclude that (e, a) is a concordant pair. Kendall's τ between a system ordering π and a set $\{\sigma_1, \ldots, \sigma_t\}$ of t multiple reference orderings is defined as,

$$\tau = \frac{4C(\pi, \sigma_1, \ldots, \sigma_t)}{N(N-1)} - 1 \tag{17}$$

$$\text{where, } C(\pi, \sigma_1, \ldots, \sigma_t) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \vee_{k=1}^{t} (\pi(i) < \sigma_k(j)).$$

Here, N is the number of sentences being ordered, and $\vee_{k=1}^{t}$ is a disjunctive function that returns the value 1 if at least one of the $(\pi(i) < \sigma_k(j))$ inequalities is satisfied. It returns the value zero if none of the inequalities are satisfied.

We modify the Spearman's rank correlation coefficient (Formula 13) by replacing the term under summation by using the minimum distance between the corresponding ranks of a system ordering, and any one of the reference orderings. The revised formula is given by,

$$r_s = 1 - \frac{6}{N(N+1)(N-1)} \sum_{i=1}^{N} \min_{1 \leq j \leq t} (\pi(i) - \sigma_j(i))^2. \tag{18}$$

We consider continuous sentence sequences between a system ordering and multiple reference orderings, to extend the average continuity (Formula 16). Specifically, when computing m (the number of continuous sentences that appear in both a system and reference ordering) in Formula 14, we compare the continuous sequences of sentences in the system ordering with all reference orderings. If at least one of the reference orderings contains the sequence under consideration, then it is accepted. For example, in Figure 12, the sequence $e \succ a$ in T_{eval} appears in T_B. It is counted as a continuous sequence of length 2, when computing P_2 along with $a \succ b$, $b \succ c$, and $c \succ d$. Once P_n values are computed as describe above, we use Formula 16 to compute the average continuity.

4.6 Results of semi-automatic evaluation

Table 2 reports the resemblance of orderings produced by six algorithms to the two human-made ones, using the modified versions of Spearman's r_s, Kendall's τ, and average continuity described in Section 4.5. We experimentally determine the optimum parameter values (i.e. C = 32 and gamma = 0.5) for the radial basis

Table 2
Comparison with human-made ordering

Method	Spearman	Kendall	Average Continuity
RND	-0.127	-0.069	0.011
TOP	0.414	0.400	0.197
PRE	0.415	0.428	0.293
SUC	0.473	0.476	0.291
CHR	0.583	0.587	0.356
AGL	0.603	0.612	0.459

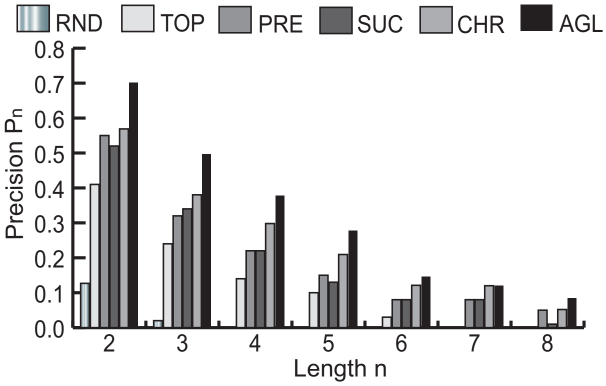


Fig. 13. Precision vs unit of measuring continuity.

functions (RBF) kernel in SVM. RBF kernel with those parameter values are used for all experiments. libSVM[6] was used as the SVM implementation.

As seen from Table 2, the proposed method (AGL) outperforms the rest in all evaluation metrics. Among the different baselines compared, chronological ordering (CHR) appeared to play the most major role. Succession criterion (SUC) performs slightly better than precedence criterion when compared using the Kendall coefficient and the Spearman coefficient. However, the two methods are indistinguishable using average continuity. Topical-relevance criterion (TOP) reports the lowest performance among the four criteria introduced in the paper. Random ordering gained almost zero in all three evaluation metrics. The one-way analysis of variance (ANOVA) verified the effects of different algorithms for sentence orderings with all metrics ($p < 0.01$). We performed the Tukey Honest Significant Differences (HSD) (Tukey, 1977) test to compare the differences among these algorithms. The Tukey HSD test revealed that AGL was significantly better than the rest. Even though we could not compare our experiment with the probabilistic approach (Lapata, 2003) directly due to the difference of the text corpora, the Kendall coefficient reported a higher agreement than Lapata's experiment (Kendall=0.48 with lemmatized nouns and Kendall=0.56 with verb-noun dependencies).

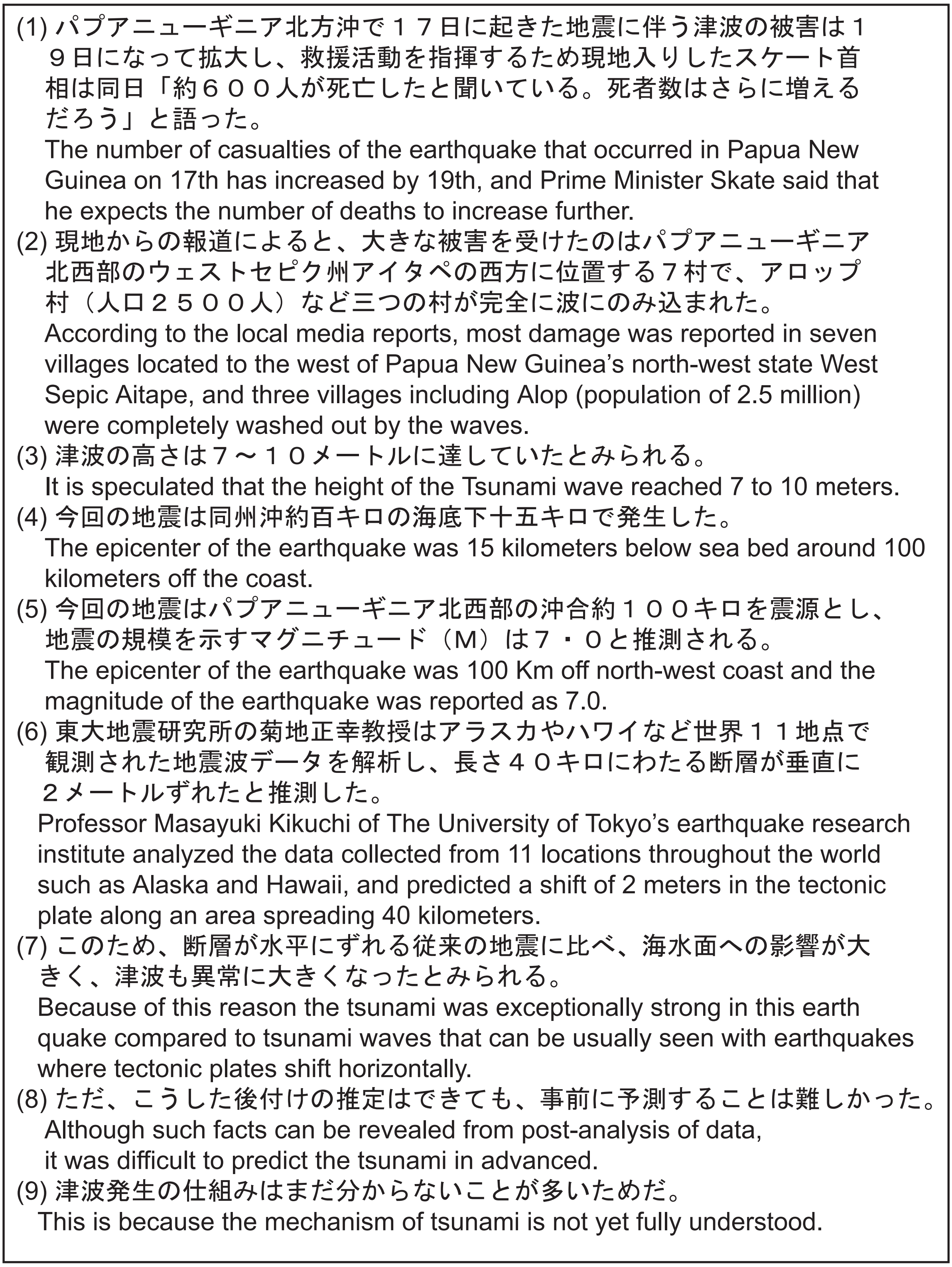
[6] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

If two sentences a and b appear next to each other in both a reference sentence ordering and in a sentence ordering under evaluation, then we say that sentences a and b are *continuous*. A coherent sentence ordering must share many segments of continuous sentences with a reference sentence ordering. Average continuity measure assigns high scores to sentence orderings that share many segments with a reference ordering. Figure 13 illustrates the behavior of precision P_n (given by Formula 14) with the length of the segment n (measured by the number of sentences in a segment) for the six methods compared in Table 2. The number of segments with continuous sentences becomes sparse (i.e. lesser) for a higher length n value. Therefore, the precision values decrease as the length n increases. Although RND ordering reported some continuous sentences for lower n values, no continuous sentences could be observed for the higher n values. The four criteria described in Section 3 (i.e., CHR, TOP, PRE, SUC) produce segments of continuous sentences at all values of n. AGL ordering obtained the highest precision for any length n, while CHR obtained the second highest precision values. The drop of P_n with n is super linear. Therefore, we used geometric mean of P_n instead of arithmetic mean to compute average continuity in Formula 16. A sentence ordering produced by the proposed algorithm is shown in Figure 14. English translations are given for the extracted original Japanese sentences.

4.7 Correlation between subjective gradings and semi-automatic evaluation measures

In section 4.4, we defined three evaluation measures: Kendall's τ, Spearman's rank correlation coefficient (r_s), and average continuity. Ideally, a semi-automatic evaluation measure should have a good agreement with subjective gradings. For this purpose, we measure the correlation between the elicited gradings in section 4.3, and the values reported by the semi-automatic measures described in section 4.4.

First, we order each set of extracted sentences using three different methods: random ordering, chronological ordering, and the proposed method. For the 30 topics in our dataset, this procedure yields 90 (30×3) summaries. We then compute Kendall's τ, Spearman's r_s, and average continuity (AC) for each of the 90 summaries, using two reference summaries for each topic as described in section 4.5. To compare the subjective grades with semi-automatic evaluation measures, we assign scores to each grade as follows: *Unacceptable* = 0.25, *Poor* = 0.50, *Acceptable* = 0.75, and *Perfect* = 1.00. Because all three judges graded each of these 90 summaries individually, each summary is assigned with three subjective grades. The final subjective score (grade) for a summary is computed as the average of the score (grade) given by each judge for that summary For example, the score of a summary with grades *Poor*, *Acceptable*, and *Unacceptable*, is computed as $(0.25 + 0.75 + 0.50)/3 = 0.5$.



(1) パプアニューギニア北方沖で１７日に起きた地震に伴う津波の被害は１９日になって拡大し、救援活動を指揮するため現地入りしたスケート首相は同日「約６００人が死亡したと聞いている。死者数はさらに増えるだろう」と語った。
The number of casualties of the earthquake that occurred in Papua New Guinea on 17th has increased by 19th, and Prime Minister Skate said that he expects the number of deaths to increase further.

(2) 現地からの報道によると、大きな被害を受けたのはパプアニューギニア北西部のウェストセピク州アイタペの西方に位置する７村で、アロップ村（人口２５００人）など三つの村が完全に波にのみ込まれた。
According to the local media reports, most damage was reported in seven villages located to the west of Papua New Guinea’s north-west state West Sepic Aitape, and three villages including Alop (population of 2.5 million) were completely washed out by the waves.

(3) 津波の高さは７～１０メートルに達していたとみられる。
It is speculated that the height of the Tsunami wave reached 7 to 10 meters.

(4) 今回の地震は同州沖約百キロの海底下十五キロで発生した。
The epicenter of the earthquake was 15 kilometers below sea bed around 100 kilometers off the coast.

(5) 今回の地震はパプアニューギニア北西部の沖合約１００キロを震源とし、地震の規模を示すマグニチュード（Ｍ）は７・０と推測される。
The epicenter of the earthquake was 100 Km off north-west coast and the magnitude of the earthquake was reported as 7.0.

(6) 東大地震研究所の菊地正幸教授はアラスカやハワイなど世界１１地点で観測された地震波データを解析し、長さ４０キロにわたる断層が垂直に２メートルずれたと推測した。
Professor Masayuki Kikuchi of The University of Tokyo’s earthquake research institute analyzed the data collected from 11 locations throughout the world such as Alaska and Hawaii, and predicted a shift of 2 meters in the tectonic plate along an area spreading 40 kilometers.

(7) このため、断層が水平にずれる従来の地震に比べ、海水面への影響が大きく、津波も異常に大きくなったとみられる。
Because of this reason the tsunami was exceptionally strong in this earth quake compared to tsunami waves that can be usually seen with earthquakes where tectonic plates shift horizontally.

(8) ただ、こうした後付けの推定はできても、事前に予測することは難しかった。
Although such facts can be revealed from post-analysis of data,
it was difficult to predict the tsunami in advanced.

(9) 津波発生の仕組みはまだ分からないことが多いためだ。
This is because the mechanism of tsunami is not yet fully understood.

Fig. 14. A sentence ordering produced by the proposed algorithm.

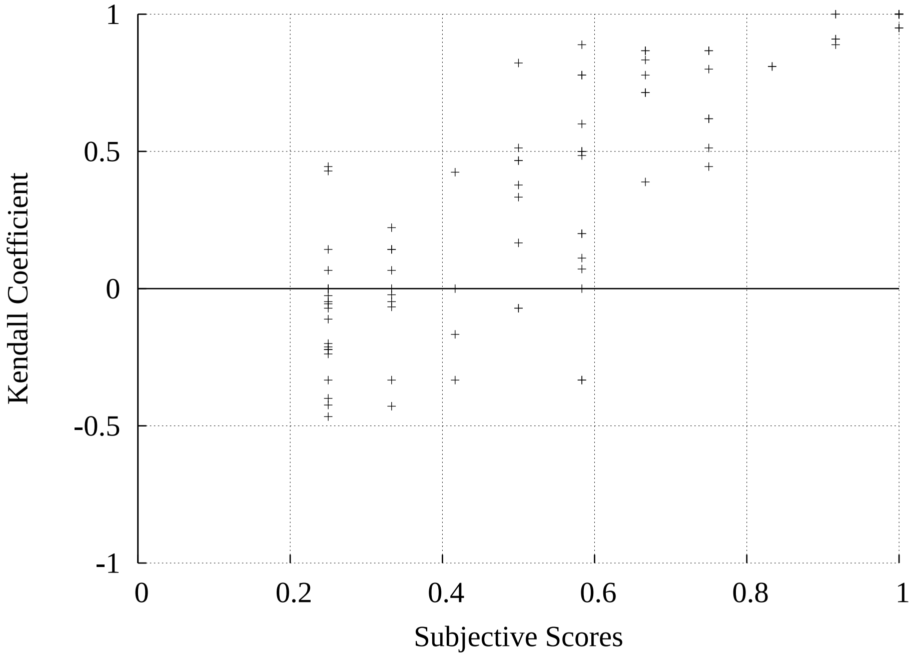


Fig. 15. Kendall's τ vs subjective scores (correlation = 0.8501)

In Figures 15-17, we plot semi-automatic evaluation measure against their corresponding subjective scores for the summaries. We compute Pearson correlation coefficient between each semi-automatic evaluation measure and subjective scores. Pearson correlation coefficient ranges from $[0, 1]$. The highest correlation with subjective scores is observed for Kendall's τ (0.8501), followed by Spearman's r_s (0.7889), and the average continuity (0.7643). We believe the higher correlation obtained for the Kendall's τ justifies the use of it as a measurement for sentence ordering for multi-document summarization. Our experimental results are in line with the proposal made by Lapata (2006) to use Kendall's τ to evaluate the information ordering tasks. In her experiments, she elicited judgments for 64 orderings (corresponding to 8 texts) from 179 human subjects. The Pearson correlation coefficient between Kendall's τ and subjective gradings was 0.45.

4.8 Effect of individual criterion on the proposed algorithm

The proposed sentence ordering method uses four criteria: chronology, topical-relateness, precedence, and succession. To evaluate the contribution of each criterion to the final ordering, we trained and tested by holding out each criterion at a time. The difference in performance when a particular criterion is removed from the algorithm, can be considered as a measure of the importance of that criterion. Table 3 shows the performance of the proposed method when each of the criteria is removed. We show the difference in performance within brackets, as measured by the evaluation metrics described in Section 4.5. As seen from Table 3, removing the

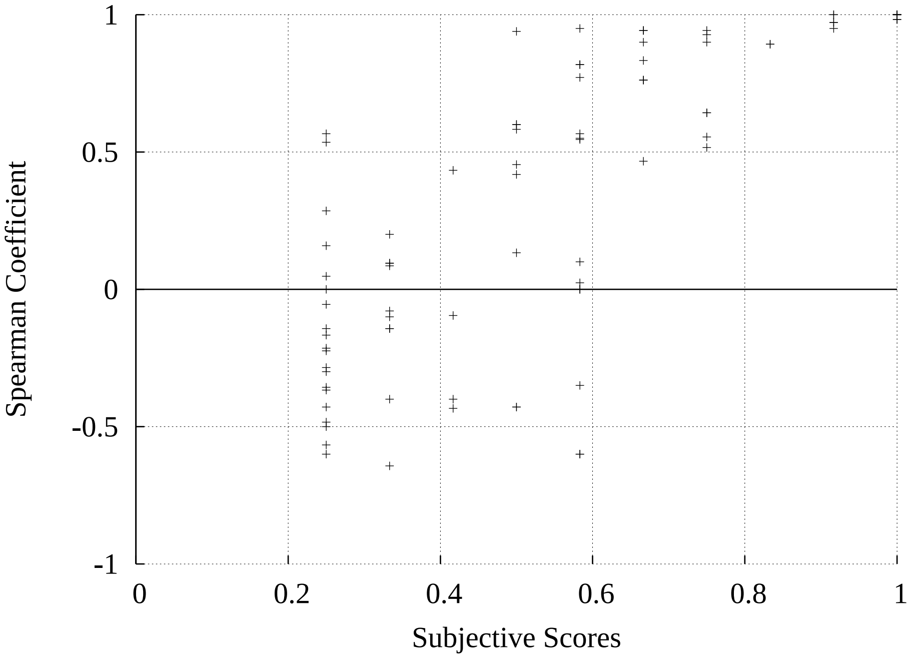


Fig. 16. Spearman coefficient vs subjective scores (correlation = 0.7889)

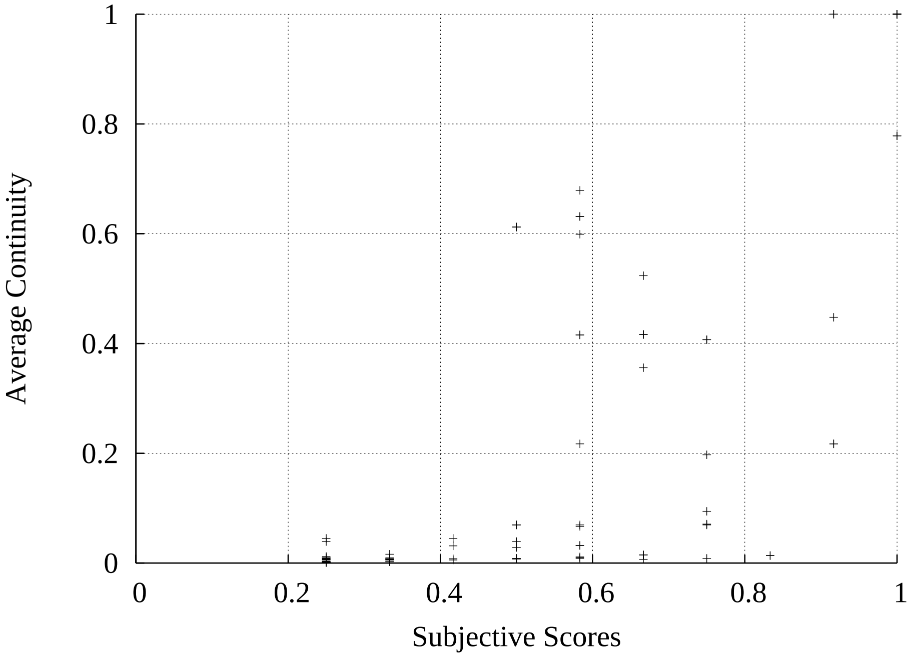


Fig. 17. Average continuity vs subjective scores (correlation = 0.7643)

chronology criterion results in the largest decline in performance. Removal of the topical-relatedness criterion has a higher impact on the average continuity, because this criterion clusters sentences that discuss the same topic, thereby improving the continuity of information related to a topic. Overall, from Table 3, it can be seen that all four criteria discussed in the paper positively contribute to the proposed

Table 3
Effect of removing a criterion

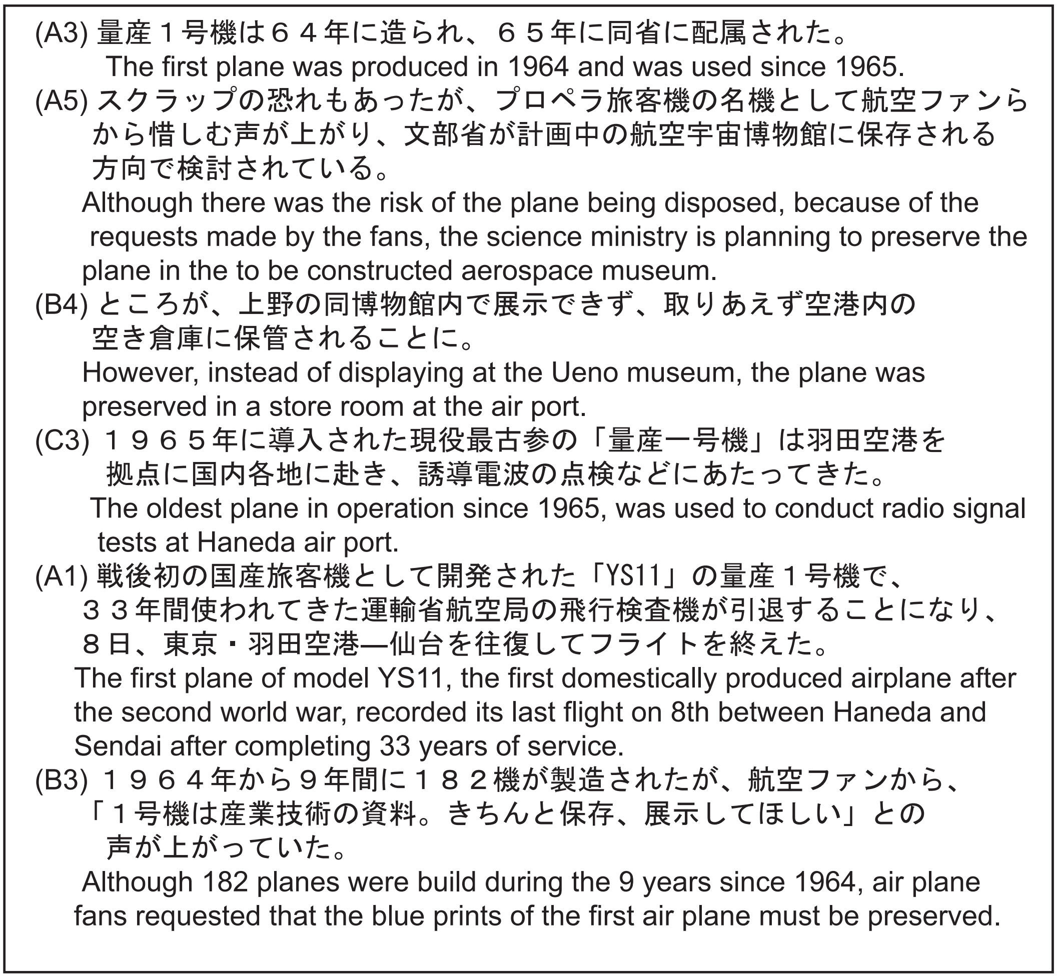
Criterion removed	Spearman	Kendall	Average Continuity
chronology	$0.398\ (-0.205)$	$0.363\ (-0.249)$	$0.076\ (-0.383)$
topical-relatedness	$0.532\ (-0.071)$	$0.517\ (-0.095)$	$0.295\ (-0.164)$
precedence	$0.520\ (-0.083)$	$0.502\ (-0.110)$	$0.311\ (-0.148)$
succession	$0.524\ (-0.079)$	$0.507\ (-0.105)$	$0.294\ (-0.165)$

sentence ordering algorithm.

5 Conclusion

We presented a bottom-up approach to arrange sentences extracted for multi-document summarization. We defined four sentence ordering criteria based on previously proposed ordering heuristics, and introduced *succedence*: a novel sentence ordering criterion that we proposed for this task. Each criterion expresses the strength and direction of the association between two text segments. We utilized support vector machines to integrate the four criteria. Using the trained SVM, we agglomeratively clustered the extracted sentences to produce a total ordering.

The subjective evaluation of the sentence orderings produced by the proposed method outperformed all baselines including previously proposed heuristics, such as the chronological ordering of sentences. In fact, 64% of the sentence orderings produced using the proposed method were graded as either perfect or acceptable by human judges. Moreover, our semi-automatic evaluation using Spearman, Kendall rank correlation coefficients, and average continuity, showed that the proposed method performs significantly better than the baselines. In particular, the proposed method reported a Kendall's τ coefficient of 0.612 with human-made sentence orderings, whereas the Kendall's τ coefficient reported by the previously proposed chronology ordering was only 0.587. Among the four sentence ordering criteria considered in the proposed method, chronology ordering was the most influential. The removal of chronology criterion reduced the performance of the proposed method by 0.249 measured in Kendall's τ coefficient. We compared the scores produced by the semi-automatic evaluation measures with human gradings, and found that Kendall's τ showed a high correlation with human gradings. The correlation between subjective scores and Kenall's τ coefficient was 0.8501, whereas the same respectively between Spearman coefficient and average continuity were 0.7889 and 0.7643. As a future direction of this study, we intend to explore the possible applications of the proposed method in concept-to-text generation (Reiter and Dale, 2000b).



(A3) 量産１号機は６４年に造られ、６５年に同省に配属された。
The first plane was produced in 1964 and was used since 1965.

(A5) スクラップの恐れもあったが、プロペラ旅客機の名機として航空ファンらから惜しむ声が上がり、文部省が計画中の航空宇宙博物館に保存される方向で検討されている。
Although there was the risk of the plane being disposed, because of the requests made by the fans, the science ministry is planning to preserve the plane in the to be constructed aerospace museum.

(B4) ところが、上野の同博物館内で展示できず、取りあえず空港内の空き倉庫に保管されることに。
However, instead of displaying at the Ueno museum, the plane was preserved in a store room at the air port.

(C3) １９６５年に導入された現役最古参の「量産一号機」は羽田空港を拠点に国内各地に赴き、誘導電波の点検などにあたってきた。
The oldest plane in operation since 1965, was used to conduct radio signal tests at Haneda air port.

(A1) 戦後初の国産旅客機として開発された「YS11」の量産１号機で、３３年間使われてきた運輸省航空局の飛行検査機が引退することになり、８日、東京・羽田空港—仙台を往復してフライトを終えた。
The first plane of model YS11, the first domestically produced airplane after the second world war, recorded its last flight on 8th between Haneda and Sendai after completing 33 years of service.

(B3) １９６４年から９年間に１８２機が製造されたが、航空ファンから、「１号機は産業技術の資料。きちんと保存、展示してほしい」との声が上がっていた。
Although 182 planes were build during the 9 years since 1964, air plane fans requested that the blue prints of the first air plane must be preserved.

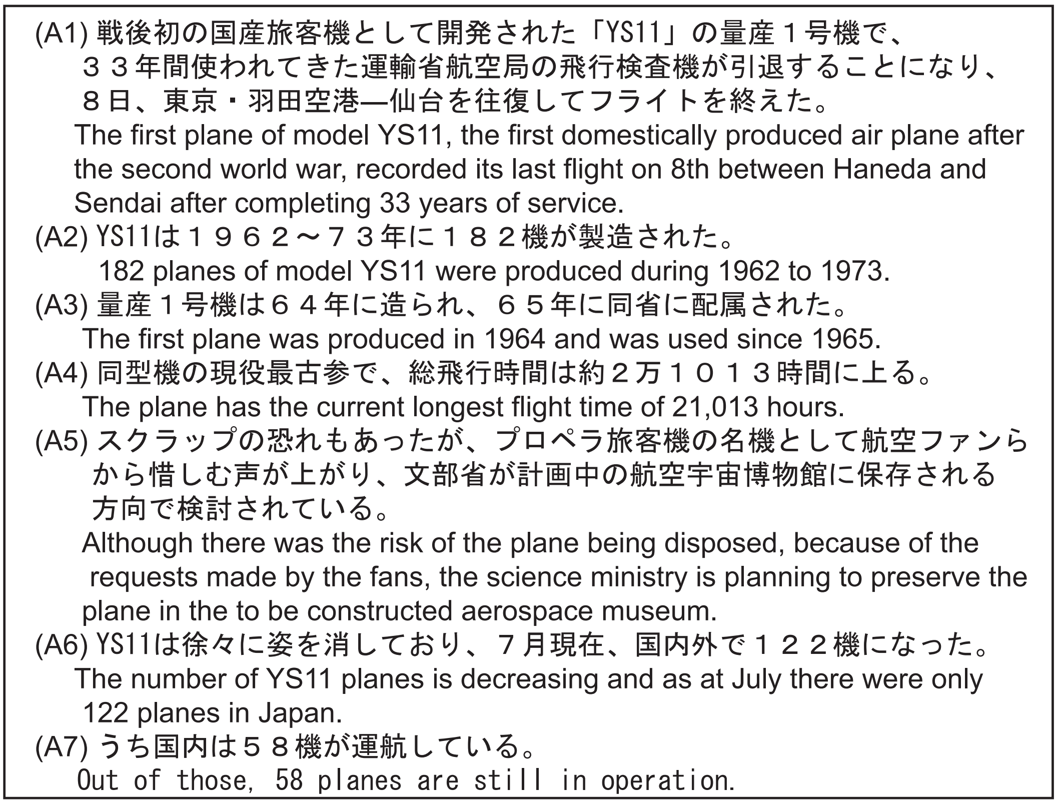
Fig. 18. A set of extracted sentences from document A, B, and C. (in random order)

Acknowledgment

We used Mainichi Shinbun and Yomiuri Shinbun newspaper articles, and the TSC-3 test collection. We would like to thank NTCIR workshop organizers for providing us with the data.

Appendix I

In this appendix we illustrate the proposed method using a set of sentences extracted from actual newspaper articles. Consider the six sentences shown in Figure 18 extracted from the three documents shown in Figures 19 (document A), 20 (document B), and 21 (document C). All documents are selected from the TSC-3 dataset which was used in the experiments in Section 4.2. The documents are actual news-



(A1) 戦後初の国産旅客機として開発された「YS11」の量産１号機で、３３年間使われてきた運輸省航空局の飛行検査機が引退することになり、８日、東京・羽田空港—仙台を往復してフライトを終えた。
The first plane of model YS11, the first domestically produced air plane after the second world war, recorded its last flight on 8th between Haneda and Sendai after completing 33 years of service.

(A2) YS11は１９６２～７３年に１８２機が製造された。
182 planes of model YS11 were produced during 1962 to 1973.

(A3) 量産１号機は６４年に造られ、６５年に同省に配属された。
The first plane was produced in 1964 and was used since 1965.

(A4) 同型機の現役最古参で、総飛行時間は約２万１０１３時間に上る。
The plane has the current longest flight time of 21,013 hours.

(A5) スクラップの恐れもあったが、プロペラ旅客機の名機として航空ファンらから惜しむ声が上がり、文部省が計画中の航空宇宙博物館に保存される方向で検討されている。
Although there was the risk of the plane being disposed, because of the requests made by the fans, the science ministry is planning to preserve the plane in the to be constructed aerospace museum.

(A6) YS11は徐々に姿を消しており、７月現在、国内外で１２２機になった。
The number of YS11 planes is decreasing and as at July there were only 122 planes in Japan.

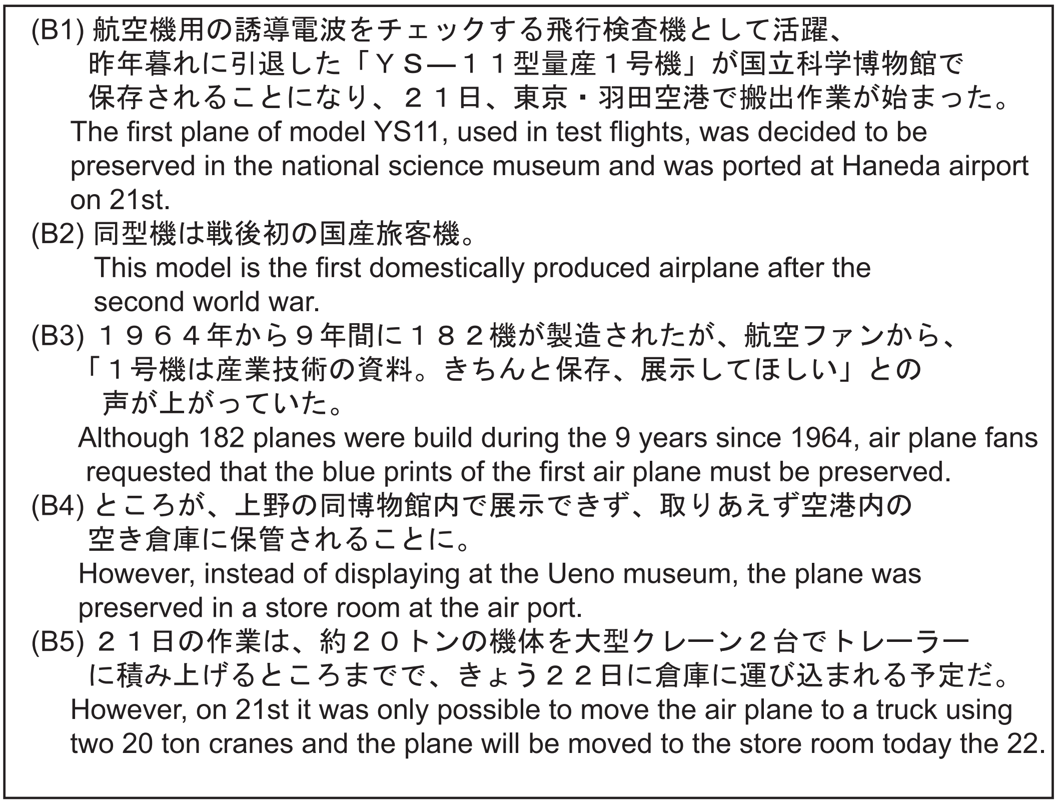
(A7) うち国内は５８機が運航している。
Out of those, 58 planes are still in operation.

Fig. 19. source document A. (publication date: 09-12-1998).

Table 4
Optimal sentence orderings at each step.

step	f_{chro}	f_{topic}	f_{pre}	f_{succ}	strength (f)	ordering
1	1.000	0.058	1.000	1.000	0.999	$(A1) \succ (A5)$
2	0	0.06	1.000	0.114	0.999	$(A1 \succ A5) \succ (A3)$
3	0	0.091	0.605	0.248	0.991	$(A1 \succ A5 \succ A3) \succ (C3)$
4	1.000	0.105	1.000	1.000	0.984	$(B3 \succ B4)$
5	1.000	0.058	0.3815	0.241	0.629	$(A1 \succ A5 \succ A3 \succ C3) \succ (B3 \succ B4)$

paper articles selected from Mainichi and Yomiuri newspapers. All documents are originally written in Japanese and we have provided English translations alongside with the original sentences. Moreover, we have assigned a sequentially numbered IDs to the sentences in each source document for the ease of reference. For ex-



(B1) 航空機用の誘導電波をチェックする飛行検査機として活躍、昨年暮れに引退した「ＹＳ—１１型量産１号機」が国立科学博物館で保存されることになり、２１日、東京・羽田空港で搬出作業が始まった。

The first plane of model YS11, used in test flights, was decided to be preserved in the national science museum and was ported at Haneda airport on 21st.

(B2) 同型機は戦後初の国産旅客機。

This model is the first domestically produced airplane after the second world war.

(B3) １９６４年から９年間に１８２機が製造されたが、航空ファンから、「１号機は産業技術の資料。きちんと保存、展示してほしい」との声が上がっていた。

Although 182 planes were build during the 9 years since 1964, air plane fans requested that the blue prints of the first air plane must be preserved.

(B4) ところが、上野の同博物館内で展示できず、取りあえず空港内の空き倉庫に保管されることに。

However, instead of displaying at the Ueno museum, the plane was preserved in a store room at the air port.

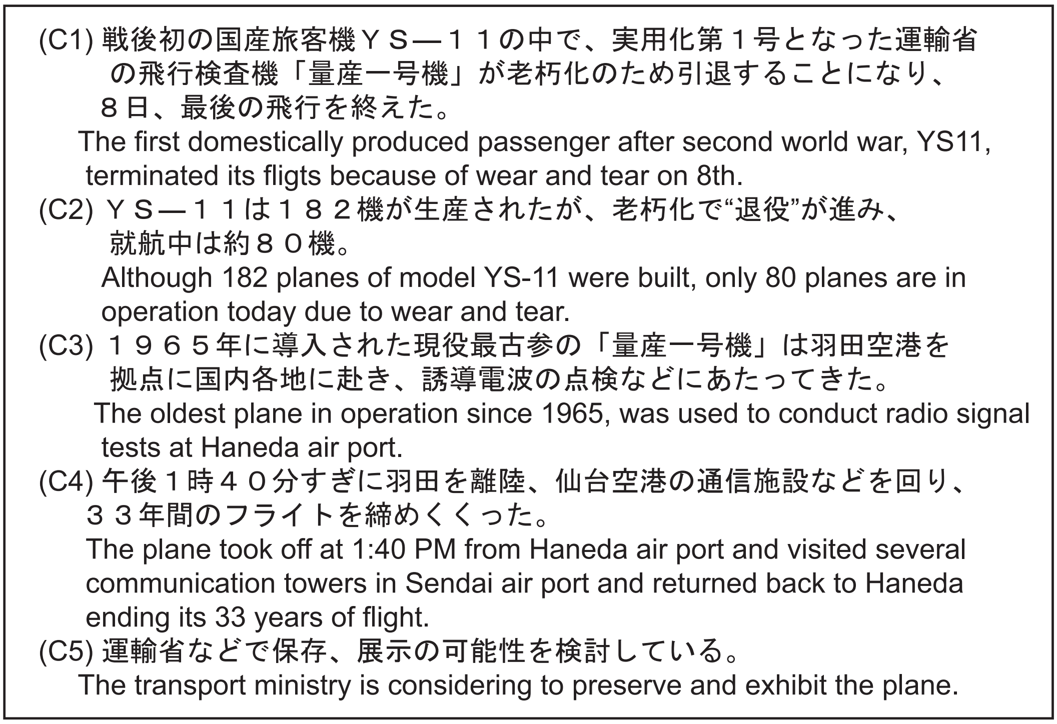
(B5) ２１日の作業は、約２０トンの機体を大型クレーン２台でトレーラーに積み上げるところまでで、きょう２２日に倉庫に運び込まれる予定だ。

However, on 21st it was only possible to move the air plane to a truck using two 20 ton cranes and the plane will be moved to the store room today the 22.

Fig. 20. source document B. (publication date: 22-08-1999).

ample, the sentences in document A are numbered A_1, A_2, A_3, $\ldots$. In Figure 18, sentences are ordered randomly and shown with the corresponding ID in the source documents. The articles are about the YS-11 airplane – the first passenger plane built in Japan after the second world war. The airplane after completing 33 years of service marked its final flight in December 1998. Documents A (Figure 19) and C (Figure 21) are published on the same date.

For the six sentences shown in Figure 18, the proposed sentence ordering algorithm first creates six segments with each segment containing exactly one sentence. We use the bracket notation to indicate a segment. For example, the segment created by sentence $A1$ is written as $(A1)$. We then compare all possible pair-wise orderings of those six segments using the four criteria: chronology (f_{chro}), topical-closeness (f_{topic}), precedence (f_{pre}), and succession (f_{succ}). For example, for the segment $(A3)$ the proposed method computes each of those criteria for the 10 orderings: $(A3) \succ (A5)$, $(A5) \succ (A3)$, $(A3) \succ (B4)$, $(B4) \succ (A3)$, $(A3) \succ (C3)$, $(C3) \succ (A3)$, $(A3) \succ (A1)$, $(A1) \succ (A3)$, $(A3) \succ (B3)$, and $(B3) \succ (A3)$. In total, for the six sentences in Figure 18 we generate 24 orderings of segment pairs. Each ordering is represented by a four dimensional feature vector as described in



(C1) 戦後初の国産旅客機ＹＳ—１１の中で、実用化第１号となった運輸省の飛行検査機「量産一号機」が老朽化のため引退することになり、８日、最後の飛行を終えた。

The first domestically produced passenger after second world war, YS11, terminated its fligts because of wear and tear on 8th.

(C2) ＹＳ—１１は１８２機が生産されたが、老朽化で“退役”が進み、就航中は約８０機。

Although 182 planes of model YS-11 were built, only 80 planes are in operation today due to wear and tear.

(C3) １９６５年に導入された現役最古参の「量産一号機」は羽田空港を拠点に国内各地に赴き、誘導電波の点検などにあたってきた。

The oldest plane in operation since 1965, was used to conduct radio signal tests at Haneda air port.

(C4) 午後１時４０分すぎに羽田を離陸、仙台空港の通信施設などを回り、３３年間のフライトを締めくくった。

The plane took off at 1:40 PM from Haneda air port and visited several communication towers in Sendai air port and returned back to Haneda ending its 33 years of flight.

(C5) 運輸省などで保存、展示の可能性を検討している。

The transport ministry is considering to preserve and exhibit the plane.

Fig. 21. source document C. (publication date: 09-12-1998).

Section 3.5. Using the trained SVM model (the training procedure is detailed in Section 3.5) we obtain the strength of association for each sentence ordering. The proposed sentence ordering algorithm then selects the ordering with the highest strength of association. We then fix the ordering between the two sentences in the selected ordering and form a new segment using those sentences. The above process is repeated with the newly formed segment and the remaining segments until we obtain a single segment with all six sentences ordered. This process can be seen as a hierarchical agglomerative clustering algorithm as described in Section 3.

For the sentences shown in Figure 18, we show the best two segments selected at each step to form a new segment in Table 5. In the initial step, the ordering $(A1) \succ (A5)$ has the highest strength of association (i.e. 0.999). Therefore, first the segment $(A1 \succ A5)$ is formed. The values of individual criteria for the best ordering at each step are also shown in Table 5. Because both $A1$ and $A5$ are extracted from the same source document (i.e. document A), and $A1$ precedes $A5$ in document A, all three criteria chronology, precedence and succession report a value of 1. Topical-closeness criterion has a low value because there are few content words in common between sentences $A1$ and $A5$. Next, the newly formed segment $(A1 \succ A5)$ and the remaining four sentences (i.e. $A3$, $B4$, $C3$, and $B3$)

are compared. The algorithm selects the ordering $((A1 \succ A5) \succ A3)$ which has the highest strength of association (i.e. 0.999) in the second step. The chronology criterion has a value of 0 in this ordering because sentence $A5$, the last sentence in segment $(A1 \succ A5)$, appears after the sentence $A3$ in document A. However, precedence criterion reports a value of 1 in this ordering because sentence $A1$ precedes sentence $A3$ in document A. In step three, segment $(C3)$ gets merged into segment $(A1 \succ A5 \succ A3)$. The chronology criterion has a zero value in this ordering because the $A3$ (the last sentence in segment $(A1 \succ A5 \succ A3)$) and sentence $C3$ are both extracted from documents which are published on the same day (i.e. 09-12-1998). Chronology criterion is undefined and returns a value of zero as defined in Equation 8 when two sentences are extracted from different documents which are published on the same day. In the fifth step, sentences $B3$ and $B4$ are ordered into a single segment. Both $B3$ and $B4$ are extracted from the same source document (document B) and appear in that order in the source document. Therefore, all three criteria: chronology, precedence and succession report a value of 1 for the ordering $B3 \succ B4$. In the final step, segment $(B3 \succ B4)$ is attached to the end of segment $(A1 \succ A5 \succ A3 \succ C3)$ to form the final sentence ordering, $A1 \succ A5 \succ A3 \succ C3 \succ B3 \succ B4$. This algorithm is guaranteed to produce a unique sentence ordering in a finite number of steps, because at each step the total number of segments is reduced by exactly one, and only the ordering with the highest strength of association is considered.

References

Barzilay, R., Elhadad, N., McKeown, K., 2002. Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research 17, 35–55.

Barzilay, R., Lee, L., 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In: HLT-NAACL 2004: Proceedings of the Main Conference. pp. 113–120.

Carbonell, J., Goldstein, J., 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retreival. pp. 335–336.

Duboue, P., McKeown, K., 2001. Empirically estimating order constraints for content planning in generation. In: Proc. of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01). pp. 172–179.

Duboue, P., McKeown, K., 2002. Content planner construction via evolutionary algorithms and a corpus-based fitness function. In: Proc. of the second International Natural Language Generation Conference (INLG'02). pp. 89–96.

Elhadad, N., McKeown, K., 2001. Towards generating patient specific summaries of medical articles. In: Proceedings of the NAACL 2001 Workshop on Automatic Summarization.

Filatova, E., Hovy, E., 2001. Assining time-stamps to event-clauses. In: Proceedings of the 2001 ACL Workshop on Temporal and Spatial Information Processing.

Ji, P. D., Pulman, S., 2006. Sentence ordering with manifold-based classification in multi-document summarization. In: Proceedings of Empherical Methods in Natural Language Processing. pp. 526–533.

Karamanis, N., Manurung, H. M., 2002. Stochastic text structuring using the principle of continuity. In: Proc. of the second International Natural Language Generation Conference (INLG'02). pp. 81–88.

Karamanis, N., Mellish, C., 2005. Using a corpus of sentence orderings defined by many experts to evaluate metrics of coherence for text structuring. In: Proc. of the 10th European Workshop on Natural Language Generation. pp. 174–179.

Kendall, M. G., 1938. A new measure of rank correlation. Biometrika 30, 81–93.

Lapata, M., 2003. Probabilistic text structuring: Experiments with sentence ordering. Proc. of the annual meeting of ACL, 2003., 545–552.

Lapata, M., 2006. Automatic evaluation of information ordering. Computational Linguistics 32 (4).

Lapata, M., Lascarides, A., 2006. Learning sentence-internal temporal relations. Journal of Artificial Intelligence Research 27, 85–117.

Lin, C., Hovy, E., 2001. Neats:a multidocument summarizer. Proceedings of the Document Understanding Workshop(DUC).

Madnani, N., Passonneau, R., Ayan, N. F., Conroy, J., Dorr, B., Klavans, J., O'Leary, D., Schlesinger, J., 2007. Measuring variability in sentence ordering for news summarization. In: Proc. of 11th European Workshop on Natural Language Generation.

Mani, I., Schiffman, B., Zhang, J., 2003. Inferring temporal ordering of events in news. In: Proc. of North American Chapter of the ACL on Human Language Technology (HLT-NAACL 2003). pp. 55–57.

Mani, I., Wilson, G., 2000. Robust temporal processing of news. In: Proc. of the 38th Annual Meeting of ACL (ACL 2000). pp. 69–76.

Mann, W., Thompson, S., 1988. Rhetorical structure theory: Toward a functional theory of text organization. Text 8 (3), 243–281.

Marcu, D., 1997. From local to global coherence: A bottom-up approach to text planning. In: Proceedings of the 14th National Conference on Artificial Intelligence. Providence, Rhode Island, pp. 629–635.

Marcu, D., 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. Computational Linguistics 26 (3), 395–448.

McKeown, K., Klavans, J., Hatzivassiloglou, V., Barzilay, R., Eskin, E., 1999. Towards multidocument summarization by reformulation: Progress and prospects. AAAI/IAAI, 453–460.

Mori, T., Sasaki, T., 2002. Information gain ratio meets maximal marginal relevance – a method of summarization for multiple documents. In: Proc. of NTCIR Workshop 3 Meeting – Part V: Text Summarization Challenge 2 (TSC2). pp. 25–32.

Okazaki, N., Matsuo, Y., Ishizuka, M., 2004. Improving chronological sentence ordering by precedence relation. In: Proceedings of 20th International Conference on Computational Linguistics (COLING 04). pp. 750–756.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu:a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 311–318.

Platt, J., 2000. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. Advances in Large Margin Classifiers, 61–74.

Radev, D. R., McKeown, K., 1999. Generating natural language summaries from multiple on-line sources. Computational Linguistics 24 (3), 469–500.

Reiter, E., Dale, R., 2000a. Building Natural Language Generation Systems. Cambridge University Press.

Reiter, E., Dale, R., 2000b. Building Natural Language Generation Systems. Cambridge University Press.

Tukey, J. W., 1977. Exploratory Data Analysis. Addison-Wesley.

Vapnik, V., 1998. Statistical Learning Theory. Wiley, Chichester, GB.