Improving Chronological Sentence Ordering by Precedence Relation

Naoaki OKAZAKI
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656,
Japan
okazaki@miv.t.u-tokyo.ac.jp

Yutaka MATSUO
AIST
2-41-6 Aomi, Koto-ku,
Tokyo 135-0064,
Japan
y.matsuo@carc.aist.go.jp

Mitsuru ISHIZUKA
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656,
Japan
ishizuka@miv.t.u-tokyo.ac.jp

Abstract

It is necessary to find a proper arrangement of sentences in order to generate a well-organized summary from multiple documents. In this paper we describe an approach to coherent sentence ordering for summarizing newspaper articles. Since there is no guarantee that chronological ordering of extracted sentences, which is widely used by conventional summarization system, arranges each sentence behind presupposed information of the sentence, we improve chronological ordering by resolving antecedent sentences of arranged sentences. Combining the refinement algorithm with topical segmentation and chronological ordering, we address our experiment to test the effectiveness of the proposed method. The results reveal that the proposed method improves chronological sentence ordering.

1 Introduction

The growth of computerized documents enables us to find relevant information easily owing to technological advances in Information Retrieval. Although it is convenient that we can obtain a great number of documents with a search engine, this situation also presents the information pollution problem: "Who is willing to take the tedious burden of reading all those text documents?" Automatic text summarization (Mani, 2001), is one solution to the problem, providing users with a condensed version of the original text.

Most existing summarization systems make use of sentence or paragraph extraction, which finds significant textual segments in source documents, and compile them in a summary. After we select significant sentences as a material for a summary, we must find a proper arrangement of the sentences and edit each sentence by deleting unnecessary parts or inserting necessary expressions. Although there has been a great deal of research on extraction since the early stage of natural language processing (Luhn, 1958), research on post-processing of automatic summarization is relatively small in number. It is essential to pay attention to sentence ordering in case of multi-document summarization. Sentence position in the original document, which yields a good clue to sentence arrangement for single-document summarization, is not enough for multi-document summarization because we must consider inter-document order at the same time.

In this paper we propose an approach to coherent text structuring for summarizing newspaper articles. We improve chronological ordering, which is widely used by conventional summarization system, complementing presupposed information of each sentence. The rest of this paper is organized as follows. We first review the sentence ordering problem and present our approach to generate an acceptable ordering in the light of coherence relation. The subsequent section (Section 3) addresses evaluation metrics and experiment results. In Section 4 we discuss future work and conclude this paper.

2 Sentence Ordering

2.1 Sentence ordering problem

Our goal is to determine the most probable permutation of given sentences and to generate a well-structured text. When a human is asked to make an arrangement of sentences, he or she may perform this task without difficulty just as we write out thoughts in a text. However, we must consider what accomplishes this task since computers are unaware of order of things by nature. Discourse coherence as typified by rhetorical relation (Mann and Thompson, 1988) and coherence relation (Hobbs, 1990) is of help to this question. Hume (Hume, 1748) claimed that qualities from which association arises and by which the mind is conveyed from one idea to another are three: *resemblance*; *contiguity in time or place*; and *cause and effect*. That is to say we should organize a text from frag-

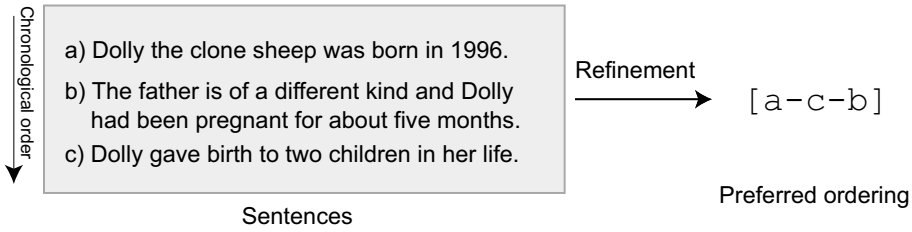

Figure 1: A chronological ordering is not enough.

mented information on the basis of topical relevancy, chronological sequence, and cause-effect relation. It is especially true in sentence ordering of newspaper articles because we must arrange a large number of time-series events concerning several topics.

Barzilay et al. (Barzilay et al., 2002) address the problem of sentence ordering in the context of multi-document summarization and the impact of sentence ordering on readability of a summary. They proposed two naive sentence-ordering techniques such as majority ordering (examines most frequent orders in the original documents) and chronological ordering (orders sentence by the publication date). Showing that using naive ordering algorithms does not produce satisfactory orderings, Barzilay et al. also investigates through experiments with humans, how to identify patterns of orderings that can improve the algorithm. Based on the experiments, they propose another algorithm that utilizes chronological ordering with topical segmentation to separate sentences referring to a topic from ones referring to another.

Lapata (Lapata, 2003) proposes another approach to information ordering based on a probabilistic model that assumes the probability of any given sentence is determined by its adjacent sentence and learns constraints on sentence order from a corpus of domain specific texts. Lapata estimates transitional probability between sentences by some attributes such as verbs (precedence relationships of verbs in the corpus), nouns (entity-based coherence by keeping track of the nouns) and dependencies (structure of sentences).

2.2 Improving chronological ordering

Against the background of these studies, we propose the use of antecedence sentences to arrange sentences. Let us consider an example shown in Figure 1. There are three sentences `a`, `b`, and `c` from which we get an order `[a-b-c]` by chronological ordering. When we read these sentences in this order, we find sentence `b` to be incorrectly positioned. This is because sentence `b` is written on the presupposition that the reader may know that Dolly had a child. In other words, it is more fitting to assume sentence `b` to be an elaboration of sentence `c`. As one may easily imagine, there are some precedent sentences prior to sentence `b` in the original document. Lack of presupposition obscures what a sentence is saying and confuses the readers. Hence, we should refine the chronological order and revise the order to `[a-c-b]`, putting sentence `c` before sentence `b`.

We show a block diagram of our ordering algorithm shown in Figure 2. Given nine sentences denoted by `[a b ... i]`, for example, the algorithm eventually produces an ordering, `[a-b-f-c-i-g-d-h-e]`. We consider topical segmentation and chronological ordering to be fundamental to sentence ordering as well as conventional ordering techniques (Barzilay et al., 2002) and make an attempt to refine the ordering. We firstly recognize topics in source documents to separate sentences referring to a topic from ones referring to another. In Figure 2 example we obtain two topical segments (clusters) as an output from the topical clustering. In the second phase we order sentences of each segment by the chronological order. If two sentences have the same chronological order, we elaborate the order on the basis of sentence position and resemblance relation. Finally, we refine each ordering by resolving antecedent sentences and output the final ordering. In the rest of this section we give a detailed description of each phase.

2.3 Topical clustering

The first task is to categorize sentences by their topics. We assume a newspaper article to be written about one topic. Hence, to classify topics in sentences, we have only to classify articles

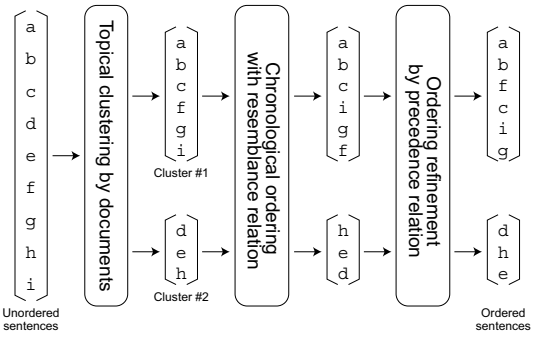

Figure 2: The outline of the ordering algorithm.

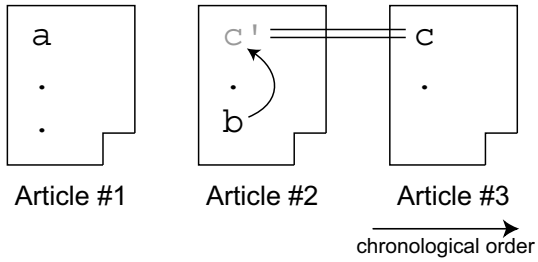

Figure 3: Background idea of ordering refinement by precedence relation.

by their topics. Given l articles and we found m kinds of terms in the articles. Let D be a document-term matrix ($l \times m$), whose element D_{ij} represents frequency of a term $\#j$ in document $\#i$, We use D_i to denote a term vector (i-component row vector) of document $\#i$. After measuring distance or dissimilarity between two articles $\#x$ and $\#y$:

$$\text{distance}(D_x, D_y) = 1 - \frac{D_x \cdot D_y}{|D_x||D_y|}, \qquad (1)$$

we apply the nearest neighbor method (Cover and Hart, 1967) to merge a pair of clusters when their minimum distance is lower than a given parameter $\alpha = 0.3$ (determined empirically). At last we classify sentences according to topical clusters, assuming that a sentence in a document belonging to a cluster also belongs to the same cluster.

2.4 Chronological ordering

It is difficult for computers to find a resemblance or cause-effect relation between two phenomena while we do not have conclusive evidence whether a pair of sentences gathered arbitrarily from multiple documents has some relation. A newspaper usually deals with novel events that have occurred since the last publication. Hence, publication date (time) of each article turns out to be a good estimator of resemblance relation (i.e., we observe a trend or series of relevant events in a time period), contiguity in time, and cause-effect relation (i.e., an event occurs as a result of previous events). Although resolving temporal expressions in sentences (e.g., *yesterday, the next year, etc.*) (Mani and Wilson, 2000; Mani et al., 2003) may give a more precise estimation of these relations, it is not an easy task. For this reason we order sentences of each segment (cluster) by the chronological order, assigning a time stamp for each sentence by its publication date (i.e., the date when the article was written).

When there are sentences having the same time stamp, we elaborate the order on the basis of sentence position and sentence connectivity. We restore an original ordering if two sentences have the same time stamp and belong to the same article. If sentences have the same time stamp and are not from the same article, we arrange a sentence which is more similar to previously ordered sentences to assure sentence connectivity.

2.5 Ordering refinement by precedence relation

After we obtain an ordering of a topical segment by chronological ordering, we improve it as shown in Figure 1 based on antecedence sentences. Figure 3 shows the background idea of ordering refinement by precedence relation. Just as in the example in Figure 1, we have three sentences `a`, `b`, and `c` in chronological order. At first we get sentence `a` out of the sentences and check its antecedent sentences. Seeing that there are no sentences prior to sentence `a` in article #1, we accept to put sentence `a` here. Then we get sentence `b` out of remaining sentences and check its antecedent sentences. We find several sentences before sentence `b` in article #2 this time. Grasping what the antecedent sentences are saying, we confirm first of all whether what they are saying is mentioned by previously arranged sentences (i.e., sentence `a`). If it is mentioned, we put sentence `b` here and extend the ordering to `[a-b]`. Otherwise, we search a substitution for what the precedence sentences are saying from the remaining sentences (i.e., sentence `c` in this example). In the Figure 3 example, we find out that sentence `a` is not referring to what sentence `c'` is saying but sentence `c` is approximately referring to that.

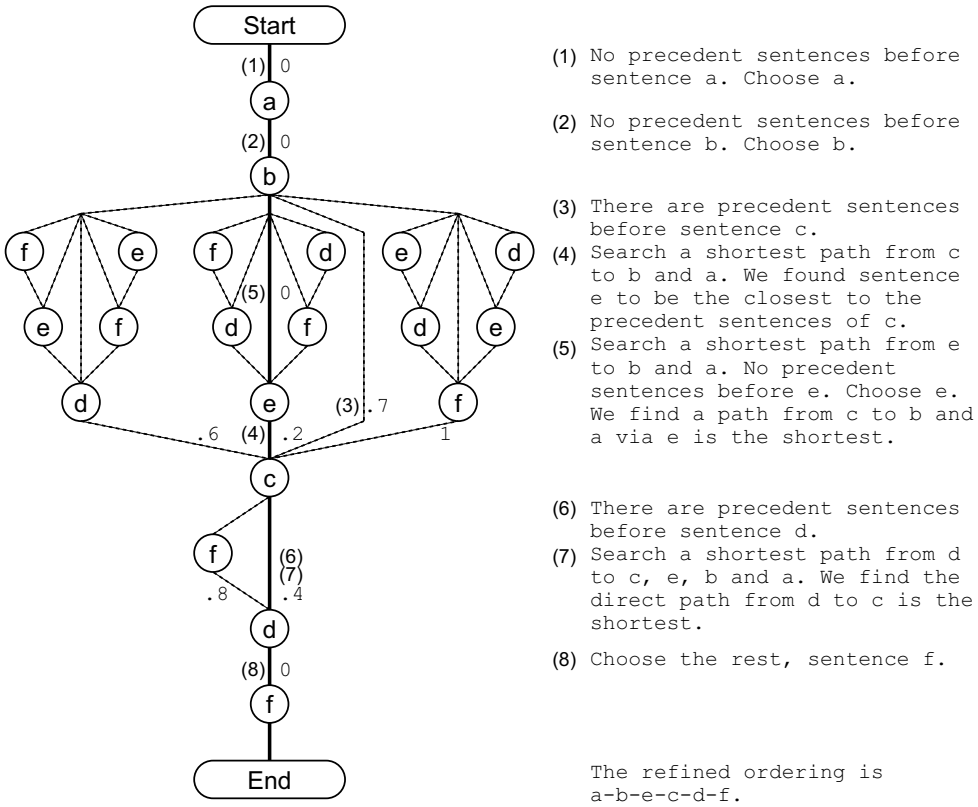

Figure 4: Ordering refinement by precedence relation as a shortest path problem.

Putting sentence `c` before `b`, we finally get the refined ordering `[a-c-b]`.

Supposing that sentence `c` mentions similar information as `c'` but expresses more than `c'`, it is nothing unusual that an extraction method does not choose sentence `c'` but sentence `c`. Because a method for multi-document summarization (e.g., MMR (Carbonell and Goldstein, 1998)) makes effort to acquire information coverage and refuse redundant information at the same time, it is quite natural that the method does not choose both sentence `c'` and `c` in terms of redundancy and prefers sentence `c` as `c'` in terms of information coverage.

Figure 4 illustrates how the algorithm refines a given chronological ordering `[a-b-c-d-e-f]`. We define *distance* as a dissimilarity value of precedent information of a sentence. When a sentence has antecedent sentences and their content is not mentioned by previously arranged sentences, this *distance* will be high. When a sentence has no precedent sentences, we define the *distance* to be 0. In the example shown in Figure 4 example we do not change position of sentences `a` and `b` because they do not have precedent sentences (i.e., they are lead sentences). On the other hand, sentence `c` has some precedent sentences in its original document. Preparing a term vector of the precedent sentences, we calculate how much the precedent content is covered by other sentences using *distance* defined above. In Figure 4 example the *distance* from sentence `a` and `b` to `c` is high ($distance = 0.7$). We search a shortest path from sentence `c` to sentences `a` and `b` by best-first search in order to find suitable sentences before sentence `c`. Given that sentence `e` in Figure 4 describes similar content as the precedent sentences of sentence `c` and is a lead sentence, we trace the shortest path from sentence `c` to sentences `a` and `b` via sentence `e`. We extend the resultant ordering to `[a-b-e-c]`, inserting sentence `e` before sentence `c`. Then we consider sentence `d`, which is not a lead sentence again ($distance = 0.4$). Preparing a term vector of the precedent sentences of sentence `d`, we search a shortest path from sentence `d` to sentences `a`, `b`, `c`, and `e`. The search result shows that we should leave sentence `d` this time because the precedent content seems to be described in sentences `a`, `b`, `c`, and `e` better than `f`. In this way we get the final ordering, `[a-b-e-c-d-f]`.

3 Evaluation

In this section we describe our experiment to test the effectiveness of the proposed method.

3.1 Experiment and evaluation metrics

We conducted an experiment of sentence ordering through multi-document summarization to test the effectiveness of the proposed method. We utilized the TSC-3 (Hirao et al., to appear in 2004) test collection, which consists of 30 sets of multi-document summarization tasks. For more information about TSC-3 task, see the workshop proceedings. Performing an important sentence extraction (Okazaki et al., to appear in 2004) up to the specified number of sentences (approximately 10% of summarization rate), we made a material for a summary (i.e., extracted sentences) for each task. We order the sentences by six methods: *human-made ordering (HO)* as the highest anchor; *random ordering (RO)* as the lowest anchor; *chronological ordering (CO)* (i.e., phase 2 only); *chronological ordering with topical segmentation (COT)* (i.e., phases 1 and 2); *proposed method without topical segmentation (PO) (i.e., phases 2 and 3)*; and *proposed method with topical segmentation (POT))*. We asked human judges to evaluate sentence ordering of these summaries.

The first evaluation task is a subjective grading where a human judge marks an ordering of summary sentences on a scale of 4: 4 (*perfect*), 3 (*acceptable*), 2 (*poor*), and 1 (*unacceptable*). We give a clear criterion of scoring to the judges as follows. A perfect summary is a text that we cannot improve any further by re-ordering. An acceptable summary is a one that makes sense and is unnecessary to be revised even though there may be some room for improvement in terms of readability. A poor summary is a one that loses a thread of the story at some places and requires minor amendment to bring it up to the acceptable level. An unacceptable summary is a one that leaves much to be improved and requires overall restructuring rather than partial revision. Additionally, we inform the judges that summaries were made of the same set of extracted sentences and only sentence ordering made differences between the summaries in order to avoid any disturbance in rating.

In addition to the rating, it is useful that we examine how close an ordering is to an acceptable one when the ordering is regarded as *poor*. Considering that several sentence-ordering patterns are acceptable for a given summary, we think that it is valuable to measure the degree of correction because this metric virtually requires a human corrector to prepare a correct answer for each ordering in his or her mind. Therefore, a human judge is supposed to illustrate how to improve an ordering of a summary when he or she marks the summary with *poor* in the rating task. We restrict applicable operations of correction to move operation so as to keep the minimum correction of the ordering. We define a move operation here as removing a sentence and inserting the sentence into an appropriate place (see Figure 5).

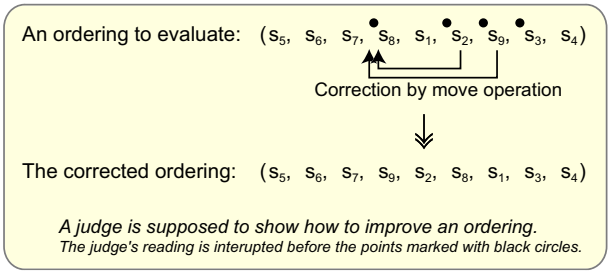

Figure 5: Correction of an ordering.

Supposing a sentence ordering to be a rank, we can calculate rank correlation coefficient of a permutation of an ordering π and a permutation of the reference ordering σ. Let $\{s_1, ..., s_n\}$ be a set of summary sentences identified with index numbers from 1 to n. We define a permutation $\pi \in S_n$ to denote an ordering of sentences where $\pi(i)$ represents an order of sentence s_i. Similarly, we define a permutation $\sigma \in S_n$ to denote the corrected ordering. For example, the π and σ in Figure 5 will be:

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 6 & 8 & 9 & 1 & 2 & 3 & 4 & 7 \end{pmatrix}, \quad (2)$$

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 7 & 5 & 8 & 9 & 1 & 2 & 3 & 6 & 4 \end{pmatrix}. \quad (3)$$

Spearman's rank correlation $\tau_s(\pi, \sigma)$ and Kendall's rank correlation $\tau_k(\pi, \sigma)$ are known as famous rank correlation metrics.

$$\tau_s(\pi, \sigma) = 1 - \frac{6}{n(n+1)(n-1)} \sum_{i=1}^{n} (\pi(i) - \sigma(i))^2 \quad (4)$$

$$\tau_k(\pi, \sigma) = \frac{1}{n(n-1)/2} \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{sgn}(\pi(j) - \pi(i)) \cdot \mathrm{sgn}(\sigma(j) - \sigma(i)), \quad (5)$$

	4	3	2	1
RO	0.0	0.0	6.0	94.0
CO	13.1	22.6	63.1	1.2
COT	10.7	22.6	61.9	4.8
PO	16.7	38.1	45.2	0.0
POT	15.5	36.9	44.0	3.6
HO	52.4	21.4	26.2	0.0

Table 1: Distribution of rating score of orderings in percent figures.

where $\mathrm{sgn}(x) = 1$ for $x > 0$ and -1 otherwise. These metrics range from -1 (an inverse rank) to 1 (an identical rank) via 0 (a non-correlated rank). In the example shown in Equations 2 and 3 we obtain $\tau_s(\pi, \sigma) = 0.85$ and $\tau_k(\pi, \sigma) = 0.72$.

We propose another metric to assess the degree of sentence continuity in reading, $\tau_c(\pi, \sigma)$:

$$\tau_c(\pi, \sigma) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{eq}\left(\pi\sigma^{-1}(i), \pi\sigma^{-1}(i-1)+1\right), \tag{6}$$

where: $\pi(0) = \sigma(0) = 0$; $\mathrm{eq}(x, y) = 1$ when x equals y and 0 otherwise. This metric ranges from 0 (no continuity) to 1 (identical). The summary in Figure 5 may interrupt judge's reading after sentence S_7, S_1, S_2 and S_9 as he or she searches a next sentence to read. Hence, we observe four discontinuities in the ordering and calculate sentence continuity $\tau_c(\pi, \sigma) = (9-4)/9 = 0.56$.

3.2 Results

Table 1 shows distribution of rating score of each method in percent figures. Judges marked about 75% of human-made ordering (HO) as either perfect or acceptable while they rejected as many as 95% of random ordering (RO). Chronological ordering (CO) did not yield satisfactory result losing a thread of 63% summaries although CO performed much better than RO. Topical segmentation could not contribute to ordering improvement of CO as well: COT is slightly worse than CO. After taking an in-depth look at the failure orderings, we found the topical clustering did not perform well during this test. We suppose the topical clustering could not prove the merits with this test collection because the collection consists of relevant articles retrieved by some query and polished well by a human so as not to include unrelated articles to a topic.

On the other hand, the proposed method (PO) improved chronological ordering much better than topical segmentation. Note that the sum of perfect and acceptable ratio jumped up from 36% (CO) to 55% (PO). This shows the ordering refinement by precedence relation improves chronological ordering by pushing poor ordering to an acceptable level.

Table 2 reports closeness of orderings to the corrected ones with average scores (AVG) and the standard deviations (SD) of the three metrics τ_s, τ_k and τ_c. It appears that average figures shows similar tendency to the rating task with three measures: HO is the best; PO is better than CO; and RO is definitely the worst. We applied one-way analysis of variance (ANOVA) to test the effect of four different methods (RO, CO, PO and HO). ANOVA proved the effect of the different methods ($p < 0.01$) for three metrics. We also applied Tukey test to compare the difference between these methods. Tukey test revealed that RO was definitely the worst with all metrics. However, Spearman's rank correlation τ_S and Kendall's rank correlation τ_k failed to prove the significant difference between CO, PO and HO. Only sentence continuity τ_c proved PO is better than CO; and HO is better than CO ($\alpha = 0.05$). The Tukey test proved that sentence continuity has better conformity to the rating results and higher discrimination to make a comparison.

Table 3 shows closeness of orderings to ones made by human (all results of HO should be 1 by necessity). Although we found RO is clearly the worst as well as other results, we cannot find the significant difference between CO, PO, and HO with all metrics. This result presents to the difficulty of automatic evaluation by preparing one correct ordering.

4 Conclusions

In this paper we described our approach to coherent sentence ordering for summarizing newspaper articles. We conducted an experiment of sentence ordering through multi-document summarization. The proposed method which utilizes precedence relation of sentence archived good results, raising poor chronological orderings to an acceptable level by 20%. We also proposed an evaluation metric that measures sentence continuity and a amendment-based evaluation task. The amendment-based evaluation outperformed the evaluation that compares an ordering with an answer made by a human. The sentence continuity metric applied to the amendment-based task showed more agree-

Method	Spearman AVG	Spearman SD	Kendall AVG	Kendall SD	Continuity AVG	Continuity SD
RO	0.041	0.170	0.035	0.152	0.018	0.091
CO	0.838	0.185	0.870	0.270	0.775	0.210
COT	0.847	0.164	0.791	0.440	0.741	0.252
PO	0.843	0.180	0.921	0.144	0.856	0.180
POT	0.851	0.158	0.842	0.387	0.820	0.240
HO	0.949	0.157	0.947	0.138	0.922	0.138

Table 2: Comparison with corrected ordering.

Method	Spearman AVG	Spearman SD	Kendall AVG	Kendall SD	Continuity AVG	Continuity SD
RO	-0.117	0.265	-0.073	0.202	0.054	0.064
CO	0.838	0.185	0.778	0.198	0.578	0.218
COT	0.847	0.164	0.782	0.186	0.571	0.229
PO	0.843	0.180	0.792	0.184	0.606	0.225
POT	0.851	0.158	0.797	0.171	0.599	0.237
HO	1.000	0.000	1.000	0.000	1.000	0.000

Table 3: Comparison with human-made ordering.

ments with the rating result.

We plan to do further study on the sentence ordering problem in future work, exploring how to apply our algorithm to documents other than newspaper or integrate ordering problem with extraction problem to improve each other. We also recognize the necessity to establish an automatic evaluation method of sentence ordering.

Acknowledgments

We made use of Mainichi Newspaper and Yomiuri Newspaper articles and summarization test collection of TSC-3.

References

R. Barzilay, E. Elhadad, and K. McKeown. 2002. Inferring strategies for sentence ordering in multidocument summarization. *Journal of Artifical Intelligence Research (JAIR)*, 17:35–55.

J. Carbonell and J. Goldstein. 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 Retrieval*, pages 335–336.

T. M. Cover and P. E. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13:21–27.

T. Hirao, T. Fukusima, M. Okumura, and H. Nanba. to appear in 2004. Text summarization challenge 3: text summarization evaluation at ntcir workshop4. In *Working note of the 4th NTCIR Workshop Meeting*.

J. Hobbs. 1990. *Literature and Cognition, CSLI Lecture Notes 21*. CSLI.

D. Hume. 1748. *Philosophical Essays concerning Human Understanding*.

M. Lapata. 2003. Probabilistic text structuring: experiments with sentence ordering. In *Proceedings of the 41st Meeting of the Association of Computational Linguistics*, pages 545–552.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

I. Mani and G. Wilson. 2000. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of ACL'2000*, pages 69–76.

I. Mani, B. Schiffman, and J. Zhang. 2003. Inferring temporal ordering of events in news. *Proceedings of the Human Language Technology Conference (HLT-NAACL) '03*.

I. Mani. 2001. *Audomatic Summarization*. John Benjamins.

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

N. Okazaki, Y. Matsuo, and M. Ishizuka. to appear in 2004. TISS: An integrated summarization system for TSC-3. In *Working note of the 4th NTCIR Workshop Meeting*.