Measuring Variability in Sentence Ordering for News Summarization

Nitin Madnani[a,b] **Rebecca Passonneau**[c] **Necip Fazil Ayan**[a,b] **John M. Conroy**[d]
Bonnie J. Dorr[a,b] **Judith L. Klavans**[e] **Dianne P. O'Leary**[a,b] **Judith D. Schlesinger**[d]
[a]Department of Computer Science
[b]Institute for Advanced Computer Studies
University of Maryland, College Park
{nmadnani,nfa,bonnie,oleary}@cs.umd.edu
[c]Center for Computational Learning Systems, Columbia University
becky@cs.columbia.edu
[d]IDA/Center for Computing Sciences
{conroy,judith}@super.org
[e]College of Information Studies, University of Maryland, College Park
jklavans@umd.edu

Abstract

The issue of sentence ordering is an important one for natural language tasks such as multi-document summarization, yet there has not been a quantitative exploration of the range of acceptable sentence orderings for short texts. We present results of a sentence reordering experiment with three experimental conditions. Our findings indicate a very high degree of variability in the orderings that the eighteen subjects produce. In addition, the variability of reorderings is significantly greater when the initial ordering seen by subjects is different from the original summary. We conclude that evaluation of sentence ordering should use multiple reference orderings. Our evaluation presents several metrics that might prove useful in assessing against multiple references. We conclude with a deeper set of questions: (a) what sorts of independent assessments of quality of the different reference orderings could be made and (b) whether a large enough test set would obviate the need for such independent means of quality assessment.

1 Introduction

The issue of ordering content in a multi-document extractive summary is an important problem that has received little attention until recently. Sentence ordering, along with other factors that affect coherence and readability, is of particular concern for multi-document summarization, where different source articles contribute sentences to a summary. We conducted an exploratory study to determine how much variation humans would produce in a reordering task under different experimental conditions, in order to assess the issues for evaluating automated reordering.

While a good ordering is essential for summary comprehension (Barzilay et al., 2002), and recent work on sentence ordering (Bollegala et al., 2006) does show promise, it is important to note that determining an optimal sentence ordering for a given summary may not be feasible. The question for evaluation of ordering is whether there is a single best ordering that humans will converge on, or that would lead to maximum reading comprehension, or that would maximize another extrinsic summary evaluation measure. On texts of approximately the same length as summaries we look at here, Karamanis et al. (2005) found that experts produce different sentence orderings for expressing database facts about archaeology. We find that summaries of newswire have a relatively larger set of coherent orderings.

We conducted an experiment where human subjects were asked to reorder multi-document summaries in order to maximize their coherence. The summaries used in this experiment were originally produced by a different set of human summarizers as part of a multi-document summarization task that was conducted by NIST in 2004. We present quantitative results that show that there is a large amount of variability among the reorderings considered coherent. On this basis, we suggest that evaluation of sentence ordering should use multiple references.

For each such summary in the experiment, we create three initial sentence orderings: (a) original order (b) random order, and (c) the output of an automated ordering algorithm. We show that:

- The initial orderings presented to the human subjects have a statistically significant impact on the reorderings that they create.
- The set of individual human reorderings ex-

hibits a significant amount of variability.

The next section provides some background for the sentence ordering task and presents the automated sentence ordering algorithm used in our experiments. Section 3 describes the experimental design. Sections 4 and 5 present quantitative analyses of the results of the experiment. Section 6 discusses related work. We discuss our results in Section 7 and conclude in Section 8.

2 Sentence Ordering Algorithms

A number of approaches have been applied to sentence ordering for multi-document summarization (Radev and McKeown, 1999). The first techniques exploited chronological information in the documents (McKeown et al., 1999; Lin and Hovy, 2002). Barzilay et al. (2002) were the first to discuss the impact of sentence ordering in the context of multi-document summarization in the news genre. They used an augmented chronological ordering algorithm that first identified and clustered related sentences, then imposed an ordering as directed by the chronology. Okazaki et al. (2004) further improved the chronological ordering algorithm by first arranging sentences in simple chronological order, then performing local reorderings.

More recent work includes probabilistic approaches that try to model the structure of text (Lapata, 2003) and algorithms that use large corpora to learn an ordering and then apply it to the summary under consideration (Bollegala et al., 2005).

Conroy et al. (2006) treat sentence ordering as a Traveling Salesperson Problem (TSP), similar to Althaus et al. (2004). Starting from a designated first sentence, they reorder the other sentences so that the sum of the *distances* between adjacent sentences is minimized. The distance (c_{jk}) between any pair of sentences j and k is computed by first obtaining a similarity score (b_{jk}) for the pair, and then normalizing this score:

$$c_{jk} \;=\; 1 - \frac{b_{jk}}{\sqrt{b_{jj}}\sqrt{b_{kk}}}\;,\; (c_{jj} = 0) \qquad (1)$$

Because a typical multi-document extractive summary usually contains a small number of sentences, a near-optimal solution to this TSP can be found either by exhaustive search or by random sampling. In this paper, we use this TSP ordering algorithm to construct one of the three experimental conditions.

3 Experimental Design

We designed an experiment to test two hypotheses: (1) that the initial orderings presented to the human subjects have a statistically significant impact on the reorderings that they create, and (2) that the set of individual human reorderings exhibits a significant amount of variability.

For our experiment, we randomly chose nine 100-word human-written summaries[†] out of 200 human written summaries produced by NIST; they were used as references to evaluate extractive multi-document summaries in 2004 (Harman, 2004). We later retrieved the quality judgments performed by NIST assessors on seven of the summaries; the remaining two were used as a reference model for assessors and had no quality judgments. The seven summaries for which we had judgments were all given high ratings of 1 or 2 (out of 5) on seven questions such as, *Does the summary build from sentence to sentence to a coherent body of information about the topic?*

The nine summaries were evenly divided into three different groups: S_{1-3}, S_{4-6} and S_{7-9}. For each summary, we used three orderings:

- **O**: the original ordering of sentences in the summary, as written by the author of the summary.
- **R**: a random ordering of the sentences
- **T**: an ordering created by applying the TSP ordering algorithm described in the previous section.

We constrained the random and the TSP orderings so that the first sentence of the human summary appeared first.

Eighteen human subjects were divided into three groups (I, II, and III), 6 subjects per group. We presented each subject with each of the nine summaries, in either its original ordering (condition $\mathrm{C_O}$), random ordering (condition $\mathrm{C_R}$), or TSP ordering (condition $\mathrm{C_T}$), as described in the Latin square design of Figure 1. For example, the six subjects in group II were presented with summaries $1-3$ in random order, $4-6$ in original order, and $7-9$ in TSP order. Thus the experiment produced 18 reorderings for each of the nine summaries, six per initial order.

[†]D30024-C (Document set D30024, NIST author ID C), D31022-F, D31008-E, D30048-C, D30037-A, D30001-A, D30051-D, D30015-E, D31001-C

	$\mathbf{S_{1-3}}$	$\mathbf{S_{4-6}}$	$\mathbf{S_{7-9}}$
$\mathbf{C_O}$	I	II	III
$\mathbf{C_R}$	II	III	I
$\mathbf{C_T}$	III	I	II

Figure 1: Latin Square Design

The human subjects chosen for the experiment were all native English speakers. Subjects accessed the task on a website, including the instructions, which explained that they would be reading a document on the screen and could reorder the sentences in that document so as to make the document more coherent. In order to prevent the introduction of any bias, the order of presentation of summaries was randomized for every subject. The instructions clearly specified the possibility that summaries might need little or no reordering. It would be difficult to measure whether the instructions led subjects to believe that all summaries could be improved by reordering. We do not have objective criteria to identify a control set of summaries that cannot be improved by reordering; in fact, this is a subjective judgement that is likely to vary between individuals. Because the three experimental conditions had the same instructions, we believe the significant differences in *amount* of reordering across conditions is a real effect rather than an artifact.

4 Variability across Experimental Conditions

To measure the variability across the experimental conditions, we developed two methods that assign a global score to each set of reorderings by comparing them to a particular reference point.

4.1 Method 1: Confusion Matrices and κ

In NLP evaluation, *confusion matrices* have typically been used in annotation tasks (Bruce and Wiebe (1998), Tomuro (2001)) where the matrix represents the comparison of two judges, and the κ inter-annotator agreement metric value (Cohen, 1960) gives a measure of the amount of agreement between the two judges, after factoring out chance. However, κ has been used to quantify the observed distribution in confusion matrices of other types in a range of other fields and applications (e.g., assessing map accuracy (Hardin, 1999), or optical recognition (Ross et al., 2002)). Here we use it to quantify variability within sets of reorderings for a summary.

Given a representation of each summary as a set of sequentially indexed sentences (e.g., 1, 2, 3, 4, 5), and of each reordering as a corresponding sequence of positional indices, we can create confusion matrices as in Figure 2. The rows represent the sequential positions of the summary sentences in one of the three initial orders that subjects were presented with, the columns represent the sentence indices, and each cell value m_{ij} indicates how often sentence j occurred in position i. Figure 2 shows the confusion matrix obtained by comparing the 6 reorderings for summary 1, obtained under the $\mathrm{C_O}$ experimental condition, to the original order (O) for that summary. The first column of the figure indicates that among the reorderings under consideration, five have sentence 1 in the first position, and one has sentence 1 in the second position. If all reorderings reproduced the original order O, the five cells of the matrix on the main diagonal would all have the value 6, and all other cells would have the value 0. The column headings of the corresponding confusion matrix for the random order (R) correspond to the sequence R, and similarly for (T).

	1	2	3	4	5
1	5	0	0	1	0
2	1	3	1	0	1
3	0	2	2	1	1
4	0	0	2	2	2
5	0	1	1	2	2

Figure 2: Confusion matrix for a set of reorderings (summary 1, condition=$\mathrm{C_O}$, reference=O; κ=0.33)

In the general case (any agreement matrix), κ measures whether the distribution of values within a matrix differs from the distribution that would be predicted by chance; the ratios of column and row marginals to the matrix total provide estimates of the expected values within each cell. Given a confusion matrix where the cells on the matrix diagonal are denoted as n_{ii}, the row marginals as n_{i+}, the column marginals as n_{+i} and the matrix total as n_{++}, the formula for κ is:

$$\kappa = \frac{\sum_i P_{ii} - \sum_i P_{i+} P_{+i}}{1 - \sum_i P_{i+} P_{+i}} \qquad (2)$$

where $P_{ii} = \frac{n_{ii}}{n_{++}}$ (the proportion of cases where a sentence ended up in its original slot), $P_{+i} = \frac{n_{+i}}{n_{++}}$, and $P_{i+} = \frac{n_{i+}}{n_{++}}$. If all cells on the diagonal are 6, κ is equal to 1. The question of interest to us is how closely a given matrix approximates this degree of agreement with the initial order.

For each summary 1-9 and for each condition,

we construct three confusion matrices: one with each initial order O, R, and T as the *target* of comparison. We denote the corresponding κ values as κ_O, κ_R, and κ_T. In principle, κ values range from 1 to values approaching -1, with 1 indicating perfect agreement, 0 indicating no difference from chance, and negative values indicating disagreements greater than expected by chance. Here, because all row and column marginals necessarily sum to 6, κ ranges from 1 to 0, with 1 indicating that the set of reorderings all reproduce the initial ordering, and 0 indicating that the set of reorderings conforms to chance.

4.2 Method 2: Means Vectors and Three Correlation Metrics

Our second method for measuring the amount of variability in a set of reorderings is based on the observation that each reordering is the same length as the initial ordering, and that each sentence index must occur exactly once per reordering. Each set of reorderings can be represented by a *means vector*, where each element of the vector is the mean sentence index for all reorderings in that set. We use three correlation metrics to give different measures of how well a means vector correlates with the initial orderings O, R and T.

The mean of the indices in each sentence position will more nearly approximate the original sentence index in that position when there are fewer instances of substituting a different sentence, and when substitutions involve sentences that were originally closer to the given slot. Figure 3 gives a hypothetical example in which each of the 6 subjects used the same ordering shifted by one: half the subjects shifted the summary by starting with the last sentence, then continuing from sentence 1 through 5 in sequence; the other half shifted the summary by starting with sentence 2 and continuing in sequence, with sentence 1 in the last position. Comparison of the means vector to the original order (O) indicates that this set of reorderings is quite similar to the original for the second through fifth positions and different in the first and last positions.

There are many distributions of sentences within a set of reorderings that can lead to the same means vector, thus we lose the power to identify some of the differences between individual reorderings within an experimental condition. However, the question we want to assess is whether the pattern given by a set of reorderings taken as a whole correlates well with the initial presentation order. We

O	1	2	3	4	5	6
S_1	2	3	4	5	6	1
S_2	6	1	2	3	4	5
S_4	6	1	2	3	4	5
S_3	2	3	4	5	6	1
S_5	2	3	4	5	6	1
S_6	6	1	2	3	4	5
Mean	4	2	3	4	5	3

Figure 3: A hypothetical example illustrating Means Vectors

compute means vectors for each condition for each summary, giving 27 such vectors. We compare each means vector representing a set of reorderings to each initial ordering O, R and T using three correlation coefficients: Pearson's r, Spearman's ρ, and Kendall's τ (Lapata, 2006).

The three correlation coefficients test the closeness of two series of numbers, or two variables x and y, in different ways. Pearson's r is a parametric test of whether there is a perfect linear relation between the two variables. Spearman's ρ and Kendall's τ are non-parametric tests. Spearman's ρ is computed by replacing the variable values by their rank and computing the correlation. Kendall's τ is based on counting the number of pairs x_i, x_{i+1} and y_i, y_{i+1} where the deltas of both pairs have the same sign. In sum, the three metrics test whether x and y are in a linear relation, a rank-preserving relation, or an order-preserving relation. Since we are comparing a set of reorderings to an initial order, rather than two sequences, it is unclear to us what grounds there would be for preferring one correlation over another. Given the exploratory nature of this method, we chose to compare results across metrics in order to determine empirically whether they support the same conclusions.

4.3 Results

The two scoring methods yield 12 global scores[‡] per summary per experimental condition, or twenty seven observations per score. We computed ANOVAs (analysis of variance) for each of the twelve scores in turn, with the score as the dependent variable and with summary number and condition as factors. Condition had a significant effect for all three κ metrics, and for r_O, ρ_O, τ_O and τ_T. This indicates that the mean score for the nine summaries differs, depending on the condition, for these seven

[‡]4 metrics (κ, r, ρ, τ) and three initial orders (O, R and T) as targets of comparison

Metric	p-value	HSD	$\mathbf{D_1}$	δ_1	$\mathbf{D_2}$	δ_2
κ_O	0.004027	0.1370	$\mathbf{C_O > C_R}$	0.2768	$\mathbf{C_O > C_T}$	0.2143
κ_R	0.0001682	0.0858	$\mathbf{C_R > C_O}$	0.2279	$\mathbf{C_R > C_T}$	0.1929
κ_T	0.002455	0.1356	$\mathbf{C_T > C_O}$	0.2848	$\mathbf{C_T > C_R}$	0.2325
r_O	0.00004985	0.1556	$\mathbf{C_O > C_R}$	0.4646	$\mathbf{C_T > C_R}$	0.3643
r_R	0.7604	-	-	-	-	-
r_T	0.1135	-	-	-	-	-
ρ_O	0.0003774	0.2006	$\mathbf{C_O > C_R}$	0.5103	$\mathbf{C_T > C_R}$	0.3983
ρ_R	0.931	-	-	-	-	-
ρ_T	0.09643	-	-	-	-	-
τ_O	0.0004306	0.2129	$\mathbf{C_O > C_R}$	0.5338	$\mathbf{C_T > C_R}$	0.4209
τ_R	0.8394	-	-	-	-	-
τ_T	0.03532	0.2482	$\mathbf{C_O > C_R}$	0.3685	$\mathbf{C_T > C_R}$	0.2957

Table 1: Analysis of variance (ANOVA) using Tukey's HSD for twelve global scores

of the twelve scores we computed. In all cases, summary was a non-significant factor, meaning that the means of the specified metric (e.g., κ_O) do not vary depending on the summary.

Table 1 presents the p-values for the analysis of variance of each metric with condition as a factor. A significant p-value indicates that there is a significant effect of condition on the mean, but does not indicate whether the means for each condition were significantly different from all others. We use Tukey's Honest Significant Difference (HSD) method to examine the significant differences in more detail. For each score, Table 1 shows Tukey's HSD (the delta at which two means become significantly different), the pairs of conditions whose difference in means was greater than the HSD, and the actual deltas. For example, row 1 of the table indicates that the mean κ_0 is higher in condition O (C_O) than in condition R (C_R) by 0.2768, which is approximately twice the HSD of 0.1370. In all cases where condition was significant, two out of the three possible differences were statistically significant.

For each κ row, the initial order that is used as the *target* is also the order defining the condition whose mean κ scores are significantly greater than both other conditions; thus for κ_O (comparison to the original order O), the $\mathrm{C_O}$ condition (where subjects reordered O) is the one that has statistically significant differences from the other two. In other words, no matter what the target is, analysis of variance of the mean κ scores shows that the reorderings created under condition $\mathrm{C_O}$ have a non-chance similarity to the O ordering that is significantly greater than the other two reordering conditions; the reorderings under condition $\mathrm{C_R}$ have a non-chance similarity to the R ordering that is significantly greater than the other two reordering conditions; the reorderings under condition $\mathrm{C_T}$ have a non-chance similarity to the T ordering that is significantly greater than the other two orderings. The sizes of the δs are roughly the same. There are no other significant differences.

The κ scores provide one type of evidence showing that taken as a set, the initial order that a set of subjects are presented with has a statistically significant effect on the reorderings that they produce; the reorderings they produce will always be significantly closer to the initial order they are presented with, with chance similarity factored out, than to any of the other two initial orders.

For the three correlation coefficients that compare the means vectors to one of the three possible targets, the experimental condition has a significant effect on the means of the coefficient only for the cases where the target is the original order (O). In other words, there is no significant difference, depending on experimental condition, in the mean r, ρ, or τ scores when the sets of reorderings are compared with the random (R) or TSP (T) order, with one exception. There is also a significant difference in the means of τ when the sets of reorderings are compared with the TSP ordering. However, the effect is less significant than when compared with the O ordering, and the HSD is greater.

The results for r scores indicate that sets of reorderings created under conditions $\mathrm{C_O}$ and $\mathrm{C_T}$ have significantly higher mean correlations with the original order (O) than those under the $\mathrm{C_R}$ condition. The other correlation analyses have parallel results: both the $\mathrm{C_O}$ and $\mathrm{C_T}$ conditions have mean correlations with the O ordering (and with the T ordering, for τ) that are significantly higher than the $\mathrm{C_R}$ condition. These results suggest that not only is there

a significant effect of the initial order on the range of reorderings produced, but also that under the C_R condition (subjects see a random order), the orderings produced are far more variable (less correlated with anything) than under the C_O and C_T conditions. r seems more sensitive than ρ or τ in that the HSDs are smaller.

5 Variability among Individual Subjects

The second analysis we performed was to measure the amount of variability among individual subjects. For this analysis, we use a variant of a method used by Karamanis et al. (2005) (based on (Lapata, 2003)). They first computed the average distance between each pair of expert pairs among 4 experts (6 pairs). Experts were then compared based on their average τ value, $\overline{\tau}$. When the $\overline{\tau}$ for a pair of experts is high, the experts are quite similar on all reorderings. For three of the experts, the $\overline{\tau}$ scores for all pairwise combinations were found to be rather high, and not significantly different, while the fourth expert was different from the other three.

Where they had 16 observations for which they computed $\overline{\tau}$ scores (each expert performed 16 orderings of the same items), we have more observations overall, but far fewer on which to make a comparison among pairs of subjects. Within a given condition, we have 6 subjects (15 pairs) but only 3 summaries, which is not enough to justify comparing the mean τ scores. Thus, we measure the similarity of two subjects' orderings using Kendall's τ.

5.1 Results

For 18 subjects, there are $C(18, 2) = 153$ unique pairwise comparisons among subjects. We computed Kendall's τ for every pair of subjects and then applied analysis of variance to the τ scores. With τ score as the dependent variable, and summary number, pair id and condition as factors, all factors were highly significant ($p \simeq 0$).

From the fact that summary number is a significant factor in predicting mean τ scores, we can conclude that the 9 summaries differ from each other in terms of the variability among individuals. As in the earlier ANOVA presented in Table 1, we use Tukey's HSD to determine the magnitude of the difference in means that is necessary for statistical significance, and use this to identify which summaries have significant differences in the amount of similarity among subjects' reorderings.

Applying Tukey's method to summary number as a factor yields the differences shown in Table 2.

S_+	S_-
8, 9, 6, 5, 3, 1, 2, 4	> 7
8, 9, 6	$> 4, 2$
8, 9	$> 1, 3, 5$

Table 2: Tukey analysis of summary number as a factor on Kendall's τ scores between individual subjects' reorderings

Among the 36 pairs of comparisons, twenty were significantly different. Here we present only the significant comparisons, not the size of the HSD nor the deltas for each comparison. The column on the left (S_+) shows the summary numbers where the mean τ values for pairs of individuals were significantly greater than for the summary numbers in the right column (S_-). Each row summarizes $|S_+| \times |S_-|$ comparisons. With summary 7, there were lower τ values than for all other summaries, meaning individuals' orderings were least alike. There were two other sets of comparisons with significant differences: summaries 8, 9 and 6 had significantly higher τ values than 4 and 2, and summaries 8 and 9 had significantly higher τ values than 1, 3 and 5. No other comparisons among summaries had significantly distinct mean τ values.

If we were to apply Tukey's HSD to the pair id factor, which was also highly significant as a predictor of τ values, it becomes difficult to summarize the significant differences. There are $C(153, 2) = 11,628$ pairwise comparisons of pairs of subjects; of these, $4,225$ were found to be statistically significant, using Tukey's method, and an analogous table to Table 2 would have 210 lines. This demonstrates that, overall, there is a large amount of variability among the individuals' reorderings.

6 Related Work on Evaluating Sentence Ordering

Karamanis and Mellish (2005) also measure the amount of variability between human subjects. However, there are several dimensions of contrast between our experiment and theirs: Their experiment operates in a very distinct domain (archaeology) and genre (descriptions of museum artifacts) whereas we use domain-independent multi-document summaries derived from news articles. We use ordinary, English-speaking volunteers as compared to the domain and genre experts that they employ (archaeologists trained in museum labeling). In terms of the experimental design, we use a Latin square design with three experimental condi-

tions whereas they have a single experimental condition. Another important difference is the nature of the ordering task itself—the task we chose was a simple text-to-text ordering task whereas their task was a modified fact-to-text ordering task, i.e., although their subjects saw sentences, it is not clear whether they were simply sentences corresponding to database facts and devoid of connectives, pronouns etc. We applied analysis of variance to all pairs of subjects' τ scores directly, rather than to the specialized scores that they compute, so we cannot directly compare results. However, the amount of variation we find seems far greater.

Barzilay et al. (2002) also conducted experiments that asked human subjects to create alternative orderings and showed that subjects rarely agreed on a single ordering for the given text. However, they did not conduct a detailed quantitative analysis of the amount of variability found in the set of human reorderings.

Okazaki et al. (2004) do ask the human judges to provide a corrected ordering for each ordering that they grade during evaluation. However, only *one* corrected ordering per summary is created. In addition, the number of humans subjects used for the evaluation task and the measures taken for circumventing bias, if any, are not reported. By contrast, our experiment uses a Latin Square with fully randomized presentation order to circumvent the introduction of any bias. Moreover, we create 18 corrected orderings for each summary and are, therefore, in a much better position to draw general conclusions about variability in sentence orderings for extractive news summaries.

Lapata (2003) required the human subjects to create multiple orderings so as to produce a coherent text but used all the human orderings solely for the purpose of comparing the proposed ordering technique and not for any form of variability analysis.

7 Discussion

Our most noteworthy finding is that for summaries of clusters of news articles, the degree of similarity of the reorderings to the original text is inversely related to the degree of randomness in the ordering that humans see. This gives us new insight into the sentence ordering task for humans. Our results suggest that humans are better at creating coherence from coherence than from incoherence. Even under the experimental condition with the lowest disorder ($\mathrm{C_O}$), there is a significant amount of variation. While there is no single *best ordering*, there are better and worse orderings, and TSP generally seems better than a random set of orderings. This is clearly apparent from Table 1— if we look at the cases where we use the original order O as the target of comparison (first, fourth, seventh and tenth rows), column five shows that in 3 out of 4 cases, the $\mathrm{C_T}$ condition (where subjects were presented with the TSP ordered sentences) has significantly higher scores than the $\mathrm{C_R}$ condition (where they were presented with randomly ordered sentences). We conclude from this that evaluation of sentence ordering should use multiple references.

To evaluate against multiple references, we suggest that a variation of the metrics we present here can be used in which test orderings are each compared to a *set* of target orderings comprised of the multiple references for a given text, in contrast to our method of comparing a set to a single target ordering. In confusion matrices for each text, the columns would represent the sentence indices of the multiple references, the rows would represent the sequential positions of the algorithm output for that text, and the cell values m_{ij} would represent how often the ith sentence of the output ordering occurred in the jth position across the multiple references.

In using multiple references, we also need further research on how to assess the relative quality of each reference order, and how to assess whether differences in quality among the references affect the evaluation. We believe that the relative coherence of summaries depends on many factors besides sentence order. A raw comparison of the κ_O values for summary 8[‡‡], for example, gives unusual results: the $\mathrm{C_R}$ condition reorderings most closely reproduce the original summary, followed by the $\mathrm{C_T}$ condition. We had the impression on reading this summary that there is a relative lack of use of devices linking the sentences to one another, such as discourse connectives, lexical cohesion, or anaphora. This raises the question of whether such devices are less necessary to create readability in short texts. While the seven summaries for which we found quality ratings were of roughly equal quality (summary 8 had the highest possible quality ratings), the DUC quality assessments have been shown to differ between assessors (Passonneau et al., 2005).

In the DUC summarization evaluations, performance of systems is assessed by averaging over

[‡‡]Document Set: D30015, NIST Author ID: E

large numbers of conditions, e.g., different document sets with different characteristics. We believe our lack of knowledge about the range of factors affecting the ordering task, and the way they interact, can be partly compensated for by evaluating ordering algorithms over a wide range of inputs.

8 Conclusions and Future Work

We conducted a reordering experiment that aims to gauge the difficulty of the sentence ordering task in the context of short, domain-independent multi-document summaries. Our results indicate that the sets of reorderings produced by human subjects depend, in a statistically significant manner, on the initial orders that the subjects are shown. In addition, we also show the existence of significant variability among subjects' reorderings. Both facts support our claim that there are *multiple* coherent orderings for a given summary. We believe that this has a significant impact on the evaluation of automatic ordering algorithms—all such algorithms should be evaluated against multiple reference orderings.

Our experiment has quantified the range of variability in human generated orderings under three conditions. In our view, a second, extrinsic assessment of the quality of the various reorderings would be necessary in order to determine whether there are grounds for ranking different orderings of the same summary. An example of an extrinsic assessment would be reading comprehension under time constraints. In future work, we would like to extend our investigations to include extrinsic assessment.

9 Acknowledgments

We are indebted to Mirella Lapata for sharing resources useful for this work. This work has been supported under the GALE program of the Defense Advaned Research Projects Agency, Contract Nos. HR0011-06-2-001 and HR0011-06-C-0023. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA.

References

E. Althaus, N. Karamanis and A. Koller. 2004. Computing locally coherent discourses. In *Proceedings of ACL*.

R. Barzilay, N. Elhadad, and K. McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *JAIR*, 17:35–55.

D. Bollegala, N. Okazaki, and M. Ishizuka. 2005. A machine learning approach to sentence ordering for multi-document summarization and its evaluation. In *Proceedings of IJCNLP*.

D. Bollegala, N. Okazaki, and M. Ishizuka. 2006. A bottom-up approach to sentence ordering for multi-document summarization. In *Proceedings of COLING/ACL*.

R. Bruce and J. Wiebe. 1998. Word-sense distinguishability and inter-coder agreement. In *Proceedings of COLING/ACL*.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

J. M. Conroy, J. D. Schlesinger, D. P. O'Leary, and J. Goldstein. 2006. Back to basics: Classy 2006. In *Proceedings of DUC'06*.

P. Hardin. 1999. Comparing main diagonal entries in normalized confusion matrices: A bootstrapping approach. In *IEEE International GRS Symposium*.

D. Harman. 2004. *Proceedings of DUC'04*. Boston, MA.

N. Karamanis and C. Mellish. 2005. Using a corpus of sentence orderings defined by many experts to evalute metrics of coherence for text structuring. In *Proceedings of ENLG*.

M. Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL*.

M. Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

C-Y. Lin and E. Hovy. 2002. Automated multi-document summarization in neats. In *Proceedings of HLT*.

K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of AAAI/IAAI*.

N. Okazaki, Y. Matsuo, and M. Ishizuka. 2004. Improving chronological sentence ordering by precedence relation. In *Proceedings of COLING*.

R. Passonneau, A. Nenkova, K. McKeown, and S. Sigelman. 2005. Applying the pyramid method in DUC 2005. In *Proceedings of DUC'05*.

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

T. D. Ross, L. A. Westerkamp; R. L. Dilsavor, J. C. Mossing, and E. G. Zelnio. 2002. Performance measures for summarizing confusion matrices: The AFRL COMPASE approach. In *SPIE International Society for Optical Engineering Proceedings Series*.

N. Tomuro. 2001. Systematic polysemy and inter-annotator disagreement: Empirical examinations. In *Proceedings of the First International Workshop on Generative Approaches to Lexicon*.