Sentence Ordering Driven by Local and Global Coherence for Summary Generation

Renxian Zhang
Department of Computing
The Hong Kong Polytechnic University
csrzhang@comp.polyu.edu.hk

Abstract

In summarization, sentence ordering is conducted to enhance summary readability by accommodating text coherence. We propose a grouping-based ordering framework that integrates local and global coherence concerns. Summary sentences are grouped before ordering is applied on two levels: group-level and sentence-level. Different algorithms for grouping and ordering are discussed. The preliminary results on single-document news datasets demonstrate the advantage of our method over a widely accepted method.

1 Introduction and Background

The canonical pipeline of text summarization consists of topic identification, interpretation, and summary generation (Hovy, 2005). In the simple case of extraction, topic identification and interpretation are conflated to sentence selection and concerned with summary informativeness. In comparison, summary generation addresses summary readability and a frequently discussed generation technique is sentence ordering.

It is implicitly or explicitly stated that sentence ordering for summarization is primarily driven by coherence. For example, Barzilay et al. (2002) use lexical cohesion information to model local coherence. A statistical model by Lapata (2003) considers both lexical and syntactic features in calculating local coherence. More globally biased is Barzilay and Lee's (2004) HMM-based content model, which models global coherence with word distribution patterns.

Whilst the above models treat coherence as lexical or topical relations, Barzilay and Lapata (2005, 2008) explicitly model local coherence with an entity grid model trained for optimal syntactic role transitions of entities.

Although coherence in those works is modeled in the guise of "lexical cohesion", "topic closeness", "content relatedness", etc., few published works simultaneously accommodate coherence on the two levels: local coherence and global coherence, both of which are intriguing topics in text linguistics and psychology. For sentences, local coherence means the well-connectedness between adjacent sentences through lexical cohesion (Halliday and Hasan, 1976) or entity repetition (Grosz et al., 1995) and global coherence is the discourse-level relation connecting remote sentences (Mann and Thompson, 1995; Kehler, 2002). An abundance of psychological evidences show that coherence on both levels is manifested in text comprehension (Tapiero, 2007). Accordingly, an apt sentence ordering scheme should be driven by such concerns.

We also note that as sentence ordering is usually discussed only in the context of multi-document summarization, factors other than coherence are also considered, such as time and source sentence position in Bollegala et al.'s (2006) "agglomerative ordering" approach. But it remains an open question whether sentence ordering is non-trivial for single-document summarization, as it has long been recognized as an actual strategy taken by human summarizers (Jing, 1998; Jing and McKeown, 2000) and acknowledged early in work on sentence ordering for multi-document summarization (Barzilay et al., 2002).

In this paper, we outline a grouping-based sentence ordering framework that is driven by the concern of local and global coherence. Summary sentences are grouped according to their conceptual relatedness before being ordered on two levels: group-level ordering and sentence-level ordering, which capture global coherence and local coherence in an integrated model. As a preliminary study, we applied the framework to single-

Proceedings of the ACL-HLT 2011 Student Session, pages 6–11,
Portland, OR, USA 19-24 June 2011.

document summary generation and obtained interesting results.

The main contributions of this work are: (1) we stress the need to channel sentence ordering research to linguistic and psychological findings about text coherence; (2) we propose a grouping-based ordering framework that integrates both local and global coherence; (3) we find in experiments that coherence-driven sentence ordering improves the readability of single-document summaries, for which sentence ordering is often considered trivial.

In Section 2, we review related ideas and techniques in previous work. Section 3 provides the details of grouping-based sentence ordering. The preliminary experimental results are presented in Section 4. Finally, Section 5 concludes the whole paper and describes future work.

2 Grouping-Based Ordering

Our ordering framework is designed to capture both local and global coherence. Globally, we identify related groups among sentences and find their relative order. Locally, we strive to keep sentence similar or related in content close to each other within one group.

2.1 Sentence Representation

As summary sentences are isolated from their original context, we retain the important content information by representing sentences as concept vectors. In the simplest case, the "concept" is equivalent to content word. A drawback of this practice is that it considers every content word equally contributive to the sentence content, which is not always true. For example, in the news domain, entities realized as NPs are more important than other concepts.

To represent sentences as entity vectors, we identify both common entities (as the head nouns of NPs) and named entities. Two common entities are equivalent if their noun stems are identical or synonymous. Named entities are usually equated by identity. But in order to improve accuracy, we also consider: 1) structural subsumption (one is part of another); 2) hypernymy and holonymy (the named entities are in a superordinate-subordinate or part-whole relation).

Now with summary sentence S_i and m entities e_{ik} ($k = 1 \ldots m$), $S_i = (wf(e_{i1}), wf(e_{i2}), \ldots, wf(e_{im}))$, where $wf(e_{ik}) = w_k \times f(e_{ik})$, $f(e_{ik})$ is the frequency of e_{ik} and w_k is the weight of e_{ik}. We define $w_k = 1$ if e_{ik} is a common entity and $w_k = 2$ if e_{ik} is a named entity. We give double weight to named entities because of their significance to news articles. After all, a news story typically contains events, places, organizations, people, etc. that denote the news theme. Other things being equal, two sentences sharing a mention of named entities are thematically closer than two sentences sharing a mention of common entities.

Alternatively, we can realize the "concept" as "event" because events are prevalent semantic constructs that bear much of the sentence content in some domains (e.g., narratives and news reports). To represent sentences as event vectors, we can follow Zhang et al.'s (2010) method at the cost of more complexity.

2.2 Sentence Grouping

To meet the global need of identifying sentence groups, we develop two grouping algorithms by applying graph-based operation and clustering.

Connected Component Finding (CC)

This algorithm treats grouping sentences as finding connected components (CC) in a text graph $TG = (V, E)$, where V represents the sentences and E the sentence relations weighted by cosine similarity. Edges with weight $< t$, a threshold, are removed because they represent poor sentence coherence.

The resultant graph may be disconnected, in which we find all of its connected components, using depth-first search. The connected components are the groups we are looking for. Note that this method cannot guarantee that every two sentences in such a group are directly linked, but it does guarantee that there exists a path between every sentence pair.

Modified K-means Clustering (MKM)

Observing that the CC method finds only *coherent groups*, not necessarily *groups of coherent sentences*, we develop a second algorithm using clustering. A good choice might be K-means as it is efficient and outperforms agglomerative clustering methods in NLP applications (Steibach et al., 2000), but the difficulty with the conventional K-means is the decision of K.

Our solution is modified K-means (MKM) based on (Wilpon and Rabiner, 1985). Let's denote

cluster i by CL_i and cluster similarity by $Sim(CL_i) = \underset{S_{im}, S_{in} \in CL_i}{Min}(Sim(S_{im}, S_{in}))$, where $Sim(S_{im}, S_{in})$ is their cosine. The following illustrates the algorithm.

1. CL_1 = all the sentence vectors;
2. Do the 1-means clustering by assigning all the vectors to CL_1;
3. While at least 1 cluster has at least 2 sentences and $Min(Sim(CL_i)) < t$, do:
3.1 If $Sim(S_m, S_n) = Min(Sim(CL_i))$, create two new centroids as S_m and S_n;
3.2 Do the conventional K-means clustering until clusters stabilize;

The above algorithm stops iterating when each cluster contains all above-threshold-similarity sentence pairs or only one sentence. Unlike CC, MKM results in more strongly connected groups, or groups of coherence sentences.

2.3 Ordering Algorithms

After the sentences are grouped, ordering is to be conducted on two levels: group and sentence.

Composed of closely related sentences, groups simulate high-level textual constructs, such as "central event", "cause", "effect", "background", etc. for news articles, around which sentences are generated for global coherence. For an intuitive example, all sentences about "cause" should immediately precede all sentences about "effect" to achieve optimal readability. We propose two approaches to group-level ordering. 1) If the group sentences come from the same document, group (G_i) order is decided by the group-representing sentence (g_i) order ($\prec$ means "precede") in the text.

$$g_i \prec g_j \Rightarrow G_i \prec G_j$$

2) Group order is decided in a greedy fashion in order to maximize the connectedness between adjacent groups, thus enhancing local coherence. Each time a group is selected to achieve maximum similarity with the ordered groups and the first ordered group (G_1) is selected to achieve maximum similarity with all the other groups.

$$G_1 = \arg\max_{G} \sum_{G' \neq G} Sim(G, G')$$

$$G_i = \arg\max_{G \in \{\text{unordered groups}\}} \sum_{j=1}^{i-1} Sim(G_j, G) \quad (i > 1)$$

where $Sim(G, G')$ is the average sentence cosine similarity between G and G'.

Within the ordered groups, sentence-level ordering is aimed to enhance local coherence by placing conceptually close sentences next to each other. Similarly, we propose two approaches. 1) If the sentences come from the same document, they are arranged by the text order. 2) Sentence order is greedily decided. Similar to the decision of group order, with ordered sentence S_{pi} in group G_p:

$$S_{p1} = \arg\max_{S \in G_p} \sum_{S' \neq S} Sim(S, S')$$

$$S_{pi} = \arg\max_{S \in \{\text{unordered sentences in } G_p\}} \sum_{j=1}^{i-1} Sim(S_{pj}, S) \quad (i > 1)$$

Note that the text order is used as a common heuristic, based on the assumption that the sentences are arranged coherently in the source document, locally and globally.

3 Experiments and Preliminary Results

Currently, we have evaluated grouping-based ordering on single-document summarization, for which text order is usually considered sufficient. But there is no theoretical proof that it leads to optimal global and local coherence that concerns us. On some occasions, e.g., a news article adopting the "Wall Street Journal Formula" (Rich and Harper, 2007) where conceptually related sentences are placed at the beginning and the end, sentence conceptual relatedness does not necessarily correlate with spatial proximity and thus selected sentences may need to be rearranged for better readability. We are not aware of any published work that has empirically compared alternative ways of sentence ordering for single-document summarization. The experimental results reported below may draw some attention to this taken-for-granted issue.

3.1 Data and Method

We prepared 3 datasets of 60 documents each, the first (D400) consisting of documents of about 400 words from the Document Understanding Conference (DUC) 01/02 datasets; the second (D1k) consisting of documents of about 1000 words manually selected from popular English journals such as *The Wall Street Journal*, *The Washington Post*, etc; the third (D2k) consisting of documents of about 2000 words from the DUC 01/02 dataset. Then we generated 100-word summaries for D400 and 200-word summaries for D1k and D2k. Since sentence selection is not our

focus, the 180 summaries were all extracts produced by a simple but robust summarizer built on term frequency and sentence position (Aone et al., 1999).

Three human annotators were employed to each provide reference orderings for the 180 summaries and mark paragraph (of at least 2 sentences) boundaries, which will be used by one of the evaluation metrics described below.

In our implementation of the grouping-based ordering, sentences are represented as entity vectors and the threshold $t = Avg(Sim(S_m, S_n)) \times c$, the average sentence similarity in a group multiplied by a coefficient empirically decided on separate held-out datasets of 20 documents for each length category. The "group-representing sentence" is the textually earliest sentence in the group. We experimented with both CC and MKM to generate sentence groups and all the proposed algorithms in 2.3 for group-level and sentence-level orderings, resulting in 8 combinations as test orderings, each coded in the format of "Grouping (CC/MKM) / Group ordering (T/G) / Sentence ordering (T/G)", where T and G represent the text order approach and the greedy selection approach respectively. For example, "CC/T/G" means grouping with CC, group ordering with text order, and sentence ordering with the greedy approach.

We evaluated the test orderings against the 3 reference orderings and compute the average (Madnani et al., 2007) by using 3 different metrics.

The first metric is Kendall's τ (Lapata 2003, 2006), which has been reliably used in ordering evaluations (Bollegala et al., 2006; Madnani et al., 2007). It measures ordering differences in terms of the number of adjacent sentence inversions necessary to convert a test ordering to the reference ordering.

$$\tau = 1 - \frac{4m}{N(N-1)}$$

In this formula, m represents the number of inversions described above and N is the total number of sentences.

The second metric is the Average Continuity (*AC*) proposed by Bollegala et al. (2006), which captures the intuition that the quality of sentence orderings can be estimated by the number of correctly arranged continuous sentences.

$$\mathrm{AC} = exp(1/(k-1)\sum_{n=2}^{k}\log(P_n + \alpha))$$

In this formula, k is the maximum number of continuous sentences, α is a small value in case $P_n = 1$. P_n, the proportion of continuous sentences of length n in an ordering, is defined as $m/(N - n + 1)$ where m is the number of continuous sentences of length n in both the test and reference orderings and N is the total number of sentences. Following (Bollegala et al., 2006), we set $k = Min(4, N)$ and $\alpha = 0.01$.

We also go a step further by considering only the continuous sentences in a paragraph marked by human annotators, because paragraphs are local meaning units perceived by human readers and the order of continuous sentences in a paragraph is more strongly grounded than the order of continuous sentences across paragraph boundaries. So in-paragraph sentence continuity is a better estimation for the quality of sentence orderings. This is our third metric: Paragraph-level Average Continuity (*P-AC*).

$$\text{P-AC} = exp(1/(k-1)\sum_{n=2}^{k}\log(PP_n + \alpha))$$

Here $PP_n = m'/(N - n + 1)$, where m' is the number of continuous sentences of length n in both the test ordering and a paragraph of the reference ordering. All the other parameters are as defined in *AC* and P_n.

3.2 Results

The following tables show the results measured by each metric. For comparison, we also include a "Baseline" that uses the text order. For each dataset, two-tailed t-test is conducted between the top scorer and all the other orderings and statistical significance ($p < 0.05$) is marked with *.

	τ	*AC*	*P-AC*
Baseline	0.6573*	0.4452*	0.0630
CC/T/T	**0.7286**	**0.5688**	**0.0749**
CC/T/G	0.7149	0.5449	0.0714
CC/G/T	0.7094	0.5449	0.0703
CC/G/G	0.6986	0.5320	0.0689
MKM/T/T	0.6735	0.4670*	0.0685
MKM/T/G	0.6722	0.4452*	0.0674
MKM/G/T	0.6710	0.4452*	0.0660
MKM/G/G	0.6588*	0.4683*	0.0682

Table 1: D400 Evaluation

	τ	*AC*	*P-AC*
Baseline	0.3276	0.0867*	0.0428*
CC/T/T	0.3324	0.0979	0.0463*
CC/T/G	0.3276	0.0923	0.0436*
CC/G/T	0.3282	0.0944	0.0479*
CC/G/G	0.3220	0.0893*	0.0428*
MKM/T/T	**0.3390**	**0.1152**	**0.0602**
MKM/T/G	0.3381	0.1130	0.0588
MKM/G/T	0.3375	0.1124	0.0576
MKM/G/G	0.3379	0.1124	0.0581

Table 2: D1k Evaluation

	τ	*AC*	*P-AC*
Baseline	0.3125*	0.1622	0.0213
CC/T/T	**0.3389**	**0.1683**	**0.0235**
CC/T/G	0.3281	**0.1683**	0.0229
CC/G/T	0.3274	0.1665	0.0226
CC/G/G	0.3279	0.1672	0.0226
MKM/T/T	0.3125*	0.1634	0.0216
MKM/T/G	0.3125*	0.1628	0.0215
MKM/G/T	0.3125*	0.1630	0.0216
MKM/G/G	0.3122*	0.1628	0.0215

Table 3: D2k Evaluation

In general, our grouping-based ordering scheme outperforms the baseline for news articles of various lengths and statistically significant improvement can be observed on each dataset. This result casts serious doubt on the widely accepted practice of taking the text order for single-document summary generation, which is a major finding from our study.

The three evaluation metrics give consistent results although they are based on different observations. The P-AC scores are much lower than their AC counterparts because of its strict paragraph constraint.

Interestingly, applying the text order posterior to sentence grouping for group-level and sentence-level ordering leads to consistently optimal performance, as the top scorers on each dataset are almost all "__/T/T". This suggests that the textual realization of coherence can be sought in the source document if possible, after the selected sentences are rearranged. It is in this sense that the general intuition about the text order is justified. It also suggests that tightly knit paragraphs (groups), where the sentences are closely connected, play a crucial role in creating a coherence flow. Shuffling those paragraphs may not affect the final coherence[1].

The grouping method does make a difference. While CC works best for the short and long datasets (D400 and D2k), MKM is more effective for the medium-sized dataset D1k. Whether the difference is simply due to length or linguistic/stylistic subtleties is an interesting topic for in-depth study.

4 Conclusion and Future Work

We have established a grouping-based ordering scheme to accommodate both local and global coherence for summary generation. Experiments on single-document summaries validate our approach and challenge the well accepted text order by the summarization community.

Nonetheless, the results do not necessarily propagate to multi-document summarization, for which the same-document clue for ordering cannot apply directly. Adapting the proposed scheme to multi-document summary generation is the ongoing work we are engaged in. In the next step, we will experiment with alternative sentence representations and ordering algorithms to achieve better performance.

We are also considering adapting more sophisticated coherence-oriented models, such as (Soricut and Marcu, 2006; Elsner et al., 2007), to our problem so as to make more interesting comparisons possible.

Acknowledgements

The reported work was inspired by many talks with my supervisor, Dr. Wenjie Li, who saw through this work down to every writing detail. The author is also grateful to many people for assistance. You Ouyang shared part of his summarization work and helped with the DUC data. Dr. Li Shen, Dr. Naishi Liu, and three participants helped with the experiments. I thank them all.

The work described in this paper was partially supported by Hong Kong RGC Projects (No. PolyU 5217/07E).

[1] I thank an anonymous reviewer for pointing this out.

References

Aone, C., Okurowski, M. E., Gorlinsky, J., and Larsen, B. 1999. A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques. In I. Mani and M. T. Maybury (eds.), *Advances in Automatic Text Summarization*. 71–80. Cambridge, Massachusetts: MIT Press.

Barzilay, R., Elhadad, N., and McKeown, K. 2002. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research*, 17: 35–55.

Barzilay, R. and Lapata, M. 2005. Modeling Local Coherence: An Entity-based Approach. In *Proceedings of the 43rd Annual Meeting of the ACL*, 141–148. Ann Arbor.

Barzilay, R. and Lapata, M. 2008. Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*, 34: 1–34.

Barzilay, R. and Lee L. 2004. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *HLT-NAACL 2004: Proceedings of the Main Conference*. 113–120.

Bollegala, D, Okazaki, N., and Ishizuka, M. 2006. A Bottom-up Approach to Sentence Ordering for Multi-document Summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 385–392. Sydney.

Elsner, M., Austerweil, j. & Charniak E. 2007. "A Unified Local and Global Model for Discourse Coherence". In *Proceedings of NAACL HLT 2007*, 436-443. Rochester, NY.

Grosz, B. J., Aravind K. J., and Scott W. 1995. Centering: A framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225.

Halliday, M. A. K., and Hasan, R. 1976. *Cohesion in English*. London: Longman.

Hovy, E. 2005. Automated Text Summarization. In R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, pp. 583–598. Oxford: Oxford University Press.

Jing, H. 2000. Sentence Reduction for Automatic Text Summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, WA, pp. 310–315.

Jing, H., and McKeown, K. 2000. Cut and Paste Based Text Summarization. In *Proceedings of the 1st NAACL*, 178–185.

Kehler, A. 2002. *Coherence, Reference, and the Theory of Grammar*. Stanford, California: CSLI Publications.

Lapata, M. 2003. Probabilistic Text Structuring: Experiments with Sentence Ordering. In *Proceedings of the Annual Meeting of ACL*, 545–552. Sapporo, Japan.

Lapata, M. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):1–14.

Madnani, N., Passonneau, R., Ayan, N. F., Conroy, J. M., Dorr, B. J., Klavans, J. L., O'leary, D. P., and Schlesinger, J. D. 2007. Measuring Variability in Sentence Ordering for News Summarization. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, 81–88. Germany.

Mann, W. C. and Thompson, S. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8:243–281.

Rich C., and Harper, C. 2007. *Writing and Reporting News: A Coaching Method, Fifth Edition*. Thomason Learning, Inc. Belmont, CA.

Soricut, R. and Marcu D. 2006. Discourse Generation Using Utility-Trained Coherence Models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 803–810.

Steibach, M., Karypis, G., and Kumar V. 2000. A Comparison of Document Clustering Techniques. Technical Report 00-034. Department of Computer Science and Engineering, University of Minnesota.

Tapiero, I. 2007. *Situation Models and Levels of Coherence: Towards a Definition of Comprehension*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wilpon, J. G. and Rabiner, L. R. 1985. A Modified K-means Clustering Algorithm for Use in Isolated Word Recognition. In *IEEE Trans. Acoustics, Speech, Signal Proc*. ASSP-33(3), 587–594.

Zhang R., Li, W., and Lu, Q. 2010. Sentence Ordering with Event-Enriched Semantics and Two-Layered Clustering for Multi-Document News Summarization. In *COLING 2010: Poster Volume*, 1489–1497, Beijing.