Sentence Ordering Techniques in Multi Document Summarization

Hari Om Sharan[1], Garima[1], Md. Haroon[1], and Rajeev Kumar[2]

[1]Deptt. of Computer Science,
COE, Teerthankar Mahaveer University, Moradabad, (India).
[2]Department of Computer Application
Teerthanker Mahaveer University Moradabad (U.P.) India
Email ID: rajeev2009mca@gmail.com,

Abstract: Extracting salient information to include in a summary has been researched extensively in the field of automatic text summarization. However, coherent arrangement of the extracted information has not received much attention. Specially, in the case of extractive multi-document text summarization, sentences that convey important information are selected from a set of documents. There is no guarantee that this set of extracted sentences will form a coherent summary by itself. The order of presentation of information extracted is an important factor and affects the coherence of a summary. This paper focuses on the various techniques for generating a coherent summary from a given set of documents by ordering the extracted sentences. In our previous paper "Approaches to Summarize Multi Documents Using Information Extraction" we discussed the approaches for information extraction, in this paper we are introducing various approaches for sentence ordering of extracted information in multi document summarization.

[Hari Om Sharan, Garima, Md. Haroon, and Rajeev Kumar. **Sentence Ordering Techniques in Multi Document Summarization.** Researcher. 2011;3(11):29-35]. (ISSN: 1553-9865). http://www.sciencepub.net

Key words: Multi-document summarization; sentence ordering; information extraction.

Introduction

Most often, the extractive summaries produced from multiple source documents suffer from an array of problems with respect to text coherence and readability, like dangling references, irrelevant context cue information, etc. Many approaches have been proposed to deal with problems, including co-reference resolution, temporal information recovery and removal of contextual phrases by sentence compression. But after these post processing steps, even if each individual sentence might be interpretable in isolation, it still does not mean that sentences gathered from different sources as a whole will be easy to understand. Interdependence between sentences greatly affects reader's understanding. Therefore, it is important to consider sentence ordering of extracted sentences in order to reconstruct discourse structure in a summary. Sentence ordering, which determines the sequence in which to represent a set of pre-selected sentences, is a critical task both for text summarization and natural language generation. The problem of how to structure the selected information to form a fluent summary has received very little attention until recently.

Several approaches have been taken in solving the information ordering task in multiple document summarization, all of which follow the assumption that the summary structure also follows the structure of the original document set, since multi-document summary captures the main contents among the document clusters.

Multi-document Summarization

In multi document summarization the information is distributed over multiple source documents. The multi document summarization task has turned out to be much more complex than summarizing a single document, even a very large one. These documents can be in different languages, written by different authors having different background knowledge and different document formats. A good summarization technology aims to combine the main themes with completeness, readability, and conciseness. An ideal multi-document summarization system does not simply shorten the source texts but presents information organized around the key aspects to represent a wider diversity of views on the topic. When such quality is achieved, an automatic multi-document summary is perceived more like an over view of a given topic [1, 2, 3].

Sentence Ordering

The problem of how to structure the selected information to form a fluent summary has received very little attention until recently. In single document summarization, summary sentences are typically arranged in the same order as they were in the original full document, although it was found that human summarizers do sometimes change the original order [2, 18]. In multi-document summarization, sentences are selected from multiple documents and no complete ordering from a single document is available, so most common approaches involve ordering by the original article publishing time or ordering sentences based on

their content importance score from the extraction stage.

Information ordering is also a critical task for natural language generation and has been extensively investigated [19, 12, 15] by the generation community. Sentence ordering for text generation was mostly studied in a domain dependent framework [11], where a priori ordering strategies can be identified through corpus analysis. Many approaches have been proposed on sentence ordering in generic multi-document summarization.

Sentence Ordering Techniques

When producing a summary, any multi-document summarization system has to choose in which order to present the output sentences. In single document summarization sentences in the summary usually arranged in the same order as they were present in the original document. This problem becomes complicated in multi document summarization. In multi document summarization sentences are extracted from multiple documents and no single document can give the complete ordering. So most common approaches involve the ordering by their publishing time or based on the content importance feature. Let us briefly discuss some approaches for sentence ordering.

Probabilistic Approach

An unsupervised probabilistic model has been suggested by Lapata [9] for text structuring that learns ordering constraints from sentences represented by a set of lexical and structural features. It assumes the probability of any given sentence is determined by its previous sentence and learns the transition probability from one sentence to the next from the BLLIP corpus based on the Cartesian product between two sentences defined using the following features: verbs and their precedent relationships; nouns (entity-based coherence by keeping track of the nouns); and dependencies (structure of sentences). The overall ordering of the sentences in the summary is learned by greedily searching for a maximal weighted path through the graph. Based on the experimental results, she finds that entity-based coherence and the verb-noun structure features are significantly better than any other features.

Lapata shows that the lexical and structural information is very important for the ordering task, but learning those interesting lexical features requires a large corpus. She uses the BLLIP corpus which contains 30 million words. The query based summarization corpus we are using in the thesis is comparatively very small, so the probability calculation for feature learning will encounter the sparse data problem. Although we could train our model on a different and bigger corpus and then test on our own corpus, we are more interested in exploring the relations between queries and sentences in the summary genre. Such query and summary information are not provided by the BLLIP corpus.

Lapata presented an experimental setting which employs the distance between two orderings to estimate automatically how close a sentence ordering produced by her probabilistic model stands in comparison to orderings provided by several human judges. The task is to recover the originally human authored text. She is the first person who attempted to evaluate sentence ordering in text summarization quantitatively using an automatic performance measure. The automatic evaluation metric she proposed is Kendall's τ. Given an unordered set of sentences and two possible orderings, τ is used to calculate the distance between them.

The model is trained and tested on the BLLIP corpus, which contains a complete, Treebank-style, parsing of the three-year Wall Street Journal (WSJ) collection, approximately 30 million words as mentioned earlier. The average article length is 15.3 sentences. The model generated orders are compared with the original text order using the τ evaluation method. A random order is generated as the baseline for the lower bound of the τ value. The upper bound of the r value is determined by conducting an experiment to compare the model's performance with humans, where human subjects were invited to order the scrambled sentences of 12 texts from the test set and create an additional 33 orderings per text.

This method of using human agreement as the upper bound in a corpus-based evaluation provides an alternative to the view of the corpus text as an absolute gold standard. She also tests her methods on the multi-document summarization corpus that Barzilay et al. has created and achieved competitive results. She performed Post-hoc Tukey tests 3 to examine the significant differences among the different features and between models on above experiments.

It is sensitive to the fact that some sentences may be always ordered next to each other even though their absolute orders might differ. It also penalizes inverse rankings, which seems appropriate given that flipping the sentences that answer the second question with sentences that answer first question would seriously disrupt coherence. In our research, we adopt their evaluation metric and use τ for the ordering score.

Content Based Approach

Barzilay and Lee [8] have proposed domain-specific content models to represent topics and topic transitions for sentence ordering. They learn the content structure directly from unannotated texts via analysis of word distribution patterns based on the idea that "various types of [word] recurrence patterns seem to characterize various types of discourse" [6]. The

content models are Hidden Markov Models (HMMs) where in states correspond to types of information characteristic to the domain of interest, and state transitions capture possible information-presentation orderings within that domain.

The success of the distributional approach depends on the existence of recurrent patterns. Domain specific texts tend to exhibit high similarity, while in our task, the news articles came from different domains, which lack the recurrent property. But we follow their assumption that formulaic text structure facilitates readers' comprehension [14]. Instead of content patterns, we propose a method using question order to capture the overall text structure. Barzilay and Lee capture the topic clusters via complete-link clustering and measure sentence similarity with the cosine metric using word bigrams as features. In our case, we create topic clustering based on the semantic similarity between question and sentences. The topic clusters are then ordered based on the order of the questions.

To evaluate their content model, Barzilay and Lee created corpora from five domains: earthquakes, clashes between armies and rebel groups, drug-related criminal offenses, financial reports and summaries of aviation accidents 4. The average length of articles is 12 sentences. The corpora are domain specific and no queries are involved, so we can not use them for our task.

Three measures are given to evaluate the system performance. The first measure is the average original sentence order (OSO) rank. Since the content models compute the probability of generating each of the permutations of a given document's sentences, it is easy to get the OSO rank among all the alternative orderings. The best possible rank is 0 and the worst is $N! - 1$, here N is the number of sentences in the document. To compare their system with Lapata's [13], they also report the OSO prediction rate, which measures the percentage of test cases in which the model gives highest probability to the OSO from among all possible permutations, as they expect that a good content model should predict the OSO a fair fraction of the time. To assess the quality of the predicted orderings themselves, they follow Lapata's approach in employing Kendall's τ [13] as discussed above. Barzilay and Lee also compute the learning curve for different domains and show that the model performance improves as the size of training sets increases. But they do not report any statistical tests to verify that the observed differences are significant.

Entity Based Approach

Barzilay and Lapata focus on the evaluation of sentence order quality rather than generating a sentence order directly. Inspired by Centering Theory [7], Barzilay and Lapata [4] introduce an entity-based representation of discourse and treat coherence assessment as a ranking problem based on different discourse representations. A discourse entity is a class of co referent noun phrases. They use a grid to represent a set of entity transition sequences that reflect distributional, syntactic, and referential information about discourse entities. A fundamental assumption for this method is that the coherence on the level of local entity transitions is essential for generating globally coherent texts. They then take as input a set of alternative renderings of the same article and rank them based on the local coherence. The ranking problem is solved using the search techniques on a Support Vector Machine constraint optimization problem.

Their algorithm outperforms another coherence model based on Latent Semantic Analysis (LSA). They also conduct an experiment to show the contribution of various linguistic features, like syntax, coreference and salience on the model's performance. They judge the linguistic importance based on syntactic features, but not on semantic features. We use WordNet [16] to capture semantic similarity and local coherence in our work.

The data Barzilay and Lapata use for the evaluation task 5 is the DUC 2003 multi-document summaries produced by human writers and by automatic summarization systems. The training materials contain 96 pair wise rankings with an average summary length of 4.8 sentences. Coherence ratings were obtained during an elicitation study by 177 native speaker volunteers using a seven point scale rating. The ranking accuracy was measured as the fraction of correct pair wise rankings in the test set.

Since summary contents from different systems are different, this could introduce some bias to the judgment of coherence and the Kendall's τ evaluation method can not be used here. But this method might lead to the final automatic evaluation of coherence in the DUC task considering summaries produced by different systems are different. For our thesis task, the content of a summary is predetermined, so we will just follow the Kendall's τ approach as discussed earlier.

Hybrid Approach

The first systematic research on sentence ordering was done by Barzilay, et al. [17] they provided a corpus based approach to study ordering and conducted experiments which show that sentence ordering significantly affects the reader's comprehension. They also evaluated two ordering strategies: majority ordering which orders sentences by their most frequent orders across input documents and chronological ordering which orders sentences by their original article's publishing time. They then introduced

an augmented chronological ordering with topical relatedness information that achieves the best results. The augmented strategy used majority and chronological constraints to define the pair wise relations between sentences. Barzilay then identified the final order of sentences by finding a maximal weighted path in a precedence graph.

Majority ordering is critically linked to the level of similarity of information organization across the input texts. In the news genre for query-based summarization task, articles often come from different sources and provide different aspects of answers to the questions, so there is not a high level of similarity across texts. Chronological ordering could produce good results when the information is event-based, and therefore, is temporally sequenced.

A very important observation from the corpus analysis by Barzilay et al is that although there are many acceptable orderings given one set of sentences, topical related sentences always share an adjacency relation. They also point out that the notion of grouping topically related sentences is known as cohesion. As defined by Hasan (1984), cohesion is the device for "sticking together" different parts of the text. Good orderings are cohesive; this is what makes the summary readable. This approach requires a robust segmentation algorithm to identify themes which are clusters of similar sentences across different documents. Barzilay approximates theme segmentation by calculating the proportion of the number of sentence pairs which appear in the same text and same segment in the original text over the number of sentence pairs appearing in the same text. In this thesis, such themes are not identified, as no original article information is available, but we follow their insights and treat topical relatedness as one of important criteria for choosing the neighboring sentences.

Barzilay et al. collected 25 sets of articles for their experiment and evaluation 1. Each set consisted of two to three news articles reporting the same event. The extracted sentences for the summary were manually selected, simulating MULTIGEN2. The average summary length is 8.8 sentences. Among them, 10 summaries were given another 9 alternative orderings for each set for the study of patterns of summary ordering.

To evaluate different strategies, they ask human judges to manually rank each summary as Poor, Fair, or Good, which are defined as follows.

• **Poor**: Readability would be significantly improved by reordering its sentences.
• **Fair**: A summary makes sense, but reordering of some sentences can yield a better readability.
• **Good**: A summary which cannot be further improved by any sentence reordering.

To assess the significance of improvement, they use the Fisher-exact test (p-value). Manual evaluation is more reliable than automatic evaluation if inter-human agreement is higher than a certain threshold. But it is often very expensive to construct and results can not be reproduced.

Bollegala, Okazaki and Ishizuka [5] provide a novel supervised learning framework to integrate different criteria. They also propose two new criteria precedence and succession developed from their previous work. A fundamental assumption for the precedence criteria is that each sentence in newspaper articles is written on the basis that pre-suppositional information should be transferred to the reader before the sentence is interpreted. The opposite assumption holds for the succession criteria. They define a precedence function between two segments (a sequence of ordered sentences) on different criteria and formulate the criteria integration task as a binary classification problem and employ a Support Vector Machine (SVM) as the classifier. After the relations between two textual segments are learned, they then repeatedly concatenate them into one segment until the overall segment with all sentences is arranged. Precedence and succession are interesting criteria, but as we use a human written summary, such information is not available. We adopt the topical relatedness criterion and propose another query-based criterion. Similar to their supervised learning framework, we could also use SVM to combine our criteria to learn the sentence order.

Bollegala et al evaluate the method by using the third Text Summarization Challenge (TSC-3) corpus, which contains 30 extracts, each consisting of unordered sentences extracted from Japanese newspaper articles relevant to a query. Each extract has around 15 sentences on average. Two human subjects then arrange the extracts and obtain 30topics × 2humans = 60sets. Although this corpus has queries associated with each topic, the articles are written in Japanese not English, so we can not use it for our task. Furthermore, their algorithm does not consider any query-related features.

Their system performance was evaluated both manually and automatically. Manual evaluation involves two judges rating the summaries using a four point scale rating: Perfect, Acceptable; Poor; Unacceptable. Automatic evaluation employs rank correlation coefficients including Spearman's rank correlation and Kendall's τ rank correlation. Bollegala et al. also propose a new metric: average continuity, which is equivalent to measuring the precision of continuous sentences in an ordering against the reference ordering. It is defined as

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

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

(1.1) and (1.2)

Where N is the number of sentences in the reference orderings; n is the length of continuous sentences on which we are evaluating; m is the number of continuous sentences that appear in both the evaluation and reference orderings. k and τ are control parameters. Average Continuity becomes 0 when evaluation and reference orderings share no continuous sentences and 1 when the two orderings are identical.

Average continuity shares the same concepts as Kendall's τ, so we will only choose Kendall's τ as our evaluation metric. Since manual grading of the system output requires a large amount of human time and effort, we are not able to reproduce this approach.

They also use the one-way analysis of variance (ANOVA) to verify the effects of different algorithms and performed Tukey Honest Significant Differences (HSD) test to compare differences among the algorithms.

Table 1.1[6] gives brief description of four sentence ordering techniques with features and scoring methods they used.

Table 1.1: Sentence Ordering Techniques [6]

	Brazilay 2002 [10]	**Lapata 2003 [9]**	**Brazilay & Lee 2004 [8]**	**Okazaki, et al 2006 [5]**
Hypothesis	1.sentence order to impact user comprehension 2.multiple acceptable ordering for one document 3. Topical related sentences share adjacency relation	Local coherence can be captured through the probability of lexical and syntactic features of sentences based on previous sentence. Learn text structure for a specific domain.	Word distributional patterns characterize various types of discourse (content Structure) which can be captured using HMM	Use the machine learning framework to incorporate the four ordering criteria to capture the contingency between two sentences
Rank/search	Search through weighted precedence graph	Simple weighted search	Ranking by HMM	Agglomerative hierarchical clustering with the ordering information remained
Features	Majority ordering, chronological ordering, topical relatedness augmented chronological ordering	Verbs, nouns, structure dependencies	State: topic clustering Transitional Pr: sentence position in the original article	Chronological sequence, topical relatedness, precedence and succession

Table 1.2 [6] gives the brief description about the data and evaluation methods of four sentence ordering techniques.

Table 1.2: Data and evaluation for Sentence Ordering Techniques [6]

		Brazilay 2002 [10]	Lapata 2003 [9]	Brazilay & Lee 2004 [8]	Okazaki, et al 2006 [5]
Data	**Corpus**	25 sets of topics, each has 2-3 news articles reporting the same event	BILLIP corpus and Brazilay 2002 corpus	5 domains Each domain has 100 test, 100 training and 100 development sets	TSC-3 corpus containing 30sets of human ordered extracts of multiple document summarization relevant to questions
	Input	Manually selected sentences as extract	Human written articles	Human written articles	Automated extracted sentences for summary
Evaluation	**Human**	Three level grading Poor, fair, good	Human produced summary for upper bound of Kendall's Tau	No	4 scales Perfect, acceptable, Poor , unacceptable
	Automated	No	Kendall's Tau	Pair wise comparison	Spearmen's and Kendall's Tau correlation and continuity metrics

References

1) Hari Om Sharan, Rajeev Kumar, Garima Singh, Mohammad Haroon. Approaches To Summarize Multi Documents Using Information Extraction. Academia Arena. 2011;3(7):62-67] (ISSN 1553-992X).
2) Shanmugasundaram Hariharan , Extraction Based Multi Document Summarization using Single Document Summary Cluster, Int. J. Advance. Soft Comput. Appl., Vol. 2, No. 1, March 2 ; ISSN 2074-8523; ICSRS Publication, 2010
3) Shanmugasundaram Hariharan, "Merging Multi-Document Text Summaries- A Case Study", Journal of Science and Technology, Vol.5, No.4, pp.63-74, December 2009.
4) Barzilay, R. and Lapata, M. "Modeling local coherence: An entity-based approach". Comput. Linguist. 34, 1 (Mar. 2008), pp 1-34.
5) 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 the 44th Annual Meeting of the ACL (Sydney, Australia, July 17 -18, 2006). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 2006, pp 385-392.
6) Yang-Wendy Wang, "Sentence Ordering for Multi Document Summarization in Response to Multiple Queries". A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the School of Computing Science, Simon Fraser University, 2006.

7) D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, E. Drabek, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel and Z. Zhang, "MEAD -a platform for multidocument multilingual text summarization," in LREC 2004, 2004.
8) R. Barzilay and L. Lee. "Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization". In HLT-NAACL 2004: Proceedings of the Main Conference, 2004, pp 113-120.
9) Lapata, M. 2003. "Probabilistic text structuring: experiments with sentence ordering". In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics -Volume 1 (Sapporo, Japan, July 07 -12, 2003). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 2003, pp 545-552.
10) R Barzilay, N Elhadad, KR McKeown, "Inferring Strategies for Sentence Ordering in Multidocument News Summarization", Journal of Artificial Intelligence Research, 2002
11) N. Elhadad and K.McKeown. Towards generating patient specific summaries of medical articles. In Proceedings of NAACL Workshop on Automatic Summarization, 2001.
12) P.A. Duboue and K.R. McKeown. Empirically estimating order constraints for content planning in generation. In ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pages 172–179, Morristown, NJ, USA, 2001. Association for Computational Linguistics.
13) Barzilay, R., Elhadad, N., and McKeown, K. R. 2001. "Sentence ordering in multi document summarization". In Proceedings of the First international Conference on Human Language Technology Research (San Diego, March 18 -21, 2001). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 2001, pp 1-7.
14) Mani, I. "Automatic Summarization". John Benjamin's publishing Co., limited edition 2001
15) E. Reiter and R. Dale. Building Natual Language Generation Systems. Cambridge University Press, 2000.
16) Goldstein, J., Mittal, V., Carbonell, J., and Callan, J. "Creating and evaluating multi-document sentence extract summaries". In Proceedings of the Ninth international Conference on information and Knowledge Management (McLean, Virginia, United States, November 06 -11, 2000). CIKM '00. ACM, New York, NY, 2000, pp165-172.
17) Barzilay, R., McKeown, K. R., and Elhadad, M. "Information fusion in the context of multi-document summarization". In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (College Park, Maryland, June 20 -26, 1999). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 1999, pp 550-557.
18) H.Y. Jing. Summary generation through intelligent cutting and pasting of the input document. Technical report, Columbia University, 1998.
19) K. R. McKeown. Text generation: using discourse strategies and focus constraints to generate natural language text. Cambridge University Press, 1985.

11/12/2011