Probabilistic approaches for modeling text structure and their application to text-to-text generation

Citation: Barzilay, Regina. "Probabilistic Approaches for Modeling Text Structure and Their Application to Text-to-Text Generation." Empirical Methods in Natural Language Generation. Ed. Emiel Krahmer & Mariët Theune. Vol. 5790. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. 1-12.

As Published: http://dx.doi.org/10.1007/978-3-642-15573-4_1

Publisher: Springer-Verlag

Persistent URL: http://hdl.handle.net/1721.1/72468

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Probabilistic Approaches for Modeling Text Structure and their Application to Text-to-Text Generation

Regina Barzilay

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge MA 02139

Abstract. Since the early days of generation research, it has been acknowledged that modeling the global structure of a document is crucial for producing coherent, readable output. However, traditional knowledge-intensive approaches have been of limited utility in addressing this problem since they cannot be effectively scaled to operate in domain-independent, large-scale applications. Due to this difficulty, existing text-to-text generation systems rarely rely on such structural information when producing an output text. Consequently, texts generated by these methods do not match the quality of those written by humans – they are often fraught with severe coherence violations and disfluencies.

In this chapter,[1] I will present probabilistic models of document structure that can be effectively learned from raw document collections. This feature distinguishes these new models from traditional knowledge intensive approaches used in symbolic concept-to-text generation. Our results demonstrate that these probabilistic models can be directly applied to content organization, and suggest that these models can prove useful in an even broader range of text-to-text applications than we have considered here.

1 Introduction

Text-to-text generation aims to produce a coherent text by extracting, combining, and rewriting information given in input texts. Summarization, answer fusion in question-answering, and text simplification are all examples of text-to-text generation tasks. At first glance, these tasks seem much easier than the traditional generation setup, where the input consists of a non-linguistic representation. Research in summarization over the last decade has proven that texts generated for these tasks rarely match the quality of those written by humans. One of the key reasons is the lack of coherence in the generated text. As the

[1] This chapter is based on the invited talk given by the author at the 2009 European Natural Language Generation workshop. It provides an overview of the models developed by the author and her colleagues, rather than giving a comprehensive survey of the field. The original papers [3, 2, 26, 4] provide the complete technical details of the presented models.

He said Cathay Pacific was still studying PAL's financial records. Ailing Philippine Airlines and prospective investor Cathay Pacific Airways have clashed over the laying off of PAL workers, prompting PAL to revive talks with another foreign airline, an official said Tuesday. PAL resumed domestic flights Oct. 7 and started restoring international flights last month after settling its labor problems. PAL officials say Singapore Airlines is also interested in a possible investment. "As much as PAL is the flag carrier, we should see to it that PAL will always fly. But Philippine officials said Cathay and PAL had run into difficulties in two areas: who would manage PAL and how many workers would lose their jobs.

Table 1. An example of a summary generated by a multidocument summarization system participating in the 2003 Document Understanding Conference (DUC).

automatically-generated summary in Table 1 illustrates, coherence violations can drastically reduce the information value of the system output.

To illustrate current text-to-text generation methods, let us consider the task of automatic text summarization. Most current summarizers select content based on low-level features such as sentence length and word occurrence. The inability of summarizers to consider higher level contextual features leads to decreased performance and a need for greater amounts of training data. Similarly, when these systems render the selected content as a summary, they operate with no structural considerations. Therefore, they are ill-equipped to answer the questions of how the selected items fit together, what their optimal ordering is, or, in short, how to ensure that a fluent text is produced rather than a heap of sentences glued together randomly. What we are missing is an automatic mechanism that can score the well-formedness of texts and that can select the most fluent alternative among various text rendering strategies. The ability to automatically analyze the topical structure of actual and potential documents should also enable the development of more effective content selection algorithms that can operate on a more abstract and global level.

This current state of affairs is rather surprising, given that NLP researchers have developed elaborate discourse models that capture various facets of text structure. These models encompass textual aspects ranging from the morphological [27] to the intentional [9], and include a characterization of documents in terms of domain-independent rhetorical elements, such as schema items [22] or rhetorical relations [19, 20]. In fact, for concept-to-text generation systems, encoding these structural theories has contributed significantly to output quality, making them almost indistinguishable from human writing. In contrast, for text-to-text generation systems, these theories are hard to incorporate as they stand: they rely on handcrafted rules, are valid only for limited domains, and have no guarantee of scalability or portability. This difficulty motivates the development of novel approaches for document organization that can rely exclusively on information available in textual input.

In this chapter, I will present models of document structure that can be effectively used to guide content selection in text-to-text generation. First, I will focus on unsupervised learning of domain-specific content models. These models capture the topics addressed in a text and the order in which these topics appear; they are similar in their functionality to the content planners traditionally used in concept-to-text generation. I will present an effective method for learning content models from unannotated domain-specific documents utilizing hierarchical Bayesian methods. Incorporation of these models into information ordering and summarization applications yields substantial improvements over previously proposed methods.

Next, I will present a method for assessing the coherence of a generated text. The key premise of this method is that the distribution of entities in coherent texts exhibits certain regularities. The models I will be presenting operate over an automatically-computed representation that reflects distributional, syntactic, and referential information about discourse entities. This representation allows us to induce the properties of coherent texts from a given corpus, without recourse to manual annotation or a predefined knowledge base. I will show how these models can be effectively employed for content organization in text-to-text applications.

2 Content Models

A content model is a structural representation that describes the main topics and their organization within a given domain of discourse. Modeling content structure is particularly relevant for domains that exhibit recurrent patterns in content organization, such as news and encyclopedia articles. These models aim to induce, for example, that articles about cities typically contain information about History, Economy, and Transportation, and that descriptions of History usually precede those of Transportation. Note that computing content models for an arbitrary domain is a challenging task due to the lack of explicit, unambiguous structural markers (e.g., subsections with the corresponding titles). Moreover, texts within the same domain may exhibit some variability in topic selection and ordering; this variability further complicates the discovery of content structure.

Since most concept-to-text generation systems operate in limited domains, using a set of deterministic rules is a feasible way to capture patterns in content organization. [25, 15, 22]. Despite substantial effort involved in such induction, these rules proved to be essential for successful content planning — they govern the selection of relevant material, its grouping into topically homogeneous units, and its subsequent ordering. The success of this architecture in concept-to-text generation systems motivates exploring it in the context of text-to-text generation. Clearly, the first step in this direction is automating the computation of content models, as manual crafting is infeasible in large, complex domains that a text-to-text generation system should handle.

Addis Ababa also has a railway connection with Djibouti City, with a picturesque French style railway station.
The rail network, connecting the suburbs in the tri-state region to the city, consists of the Long Island Rail Road, Metro-North Railroad and New Jersey Transit.
The most important project in the next decade is the Spanish high speed rail network, Alta Velocidad Espaola AVE.
Rail is the primary mode of transportation in Tokyo, which has the most extensive urban railway network in the world and an equally extensive network of surface lines.
The backbone of the city's transport, the Mumbai Suburban Railway, consists of three separate networks: Central, Western, and Harbour Line, running the length of the city, in a north-south direction.

Table 2. A sample of Wikipedia sentences assigned the same section heading by the editors.

2.1 Computational Modeling

Fortunately, recent research has demonstrated the feasibility of acquiring content models automatically from raw texts. These approaches adapt a *distributional view*, learning content models via analysis of word-distribution patterns across texts within the same domain. This idea dates back at least to Harris [11], who claimed that "various types of [word] recurrence patterns seem to characterize various types of discourse."

The success of automatic induction greatly depends on how much variability is exhibited by the texts in the underlying domain. In formulaic domains, a simple word-based clustering can be sufficient for computing content models [26]. More commonly, however, the same topic can be conveyed using very different wordings. In such cases, word distribution on its own may not be a sufficient predictor of a segment topic. For instance, consider sentences that represent the *Railway Transportation* topic in several Wikipedia articles about cities (see Table 2). The first two sentences clearly discuss the same topic, but they do not share any content words in common. To properly handle such cases, more elaborate algorithms for learning content models are needed.

One of the early instances of such approaches is an algorithm for learning content models using Hidden Markov Models (HMMs) [3]. In these models, states correspond to types of information characteristic of the domain of interest (e.g., earthquake magnitude or previous earthquake occurrences) and state transitions capture possible information-presentation orderings within that domain. Like clustering algorithms, these models capture the intuition that sentences that belong to the same topic use similar vocabulary. In addition, HMM-based models can exploit regularities in ordering to further refine the induction process.

Our recent work on content model induction has focused on modeling *global* constraints on discourse organization which cannot be easily expressed in Markovian models. An example of such a global constraint is *topic continuity* – it posits that each document follows a progression of coherent, nonrecurring topics [10]. Following the example above, this constraint captures the notion that a single topic, such as History, is expressed in a contiguous block within the document, rather than spread over disconnected sections. Another global constraint concerns similarity in *global ordering*. This constraint guides the model toward selecting sequences with similar topic *ordering*, such as placing History before Transportation.

To effectively capture these global constraints, a content model posits a single distribution over the *entirety* of a document's content ordering [4]. Specifically, the model represents content structure as a *permutation* over topics. This naturally enforces the first constraint since a permutation does not allow topic repetition. Despite apparent intractability, this distribution over permutations can be effectively learned using the *Generalized Mallows Model* (GMM) [6].

By design, GMMs concentrate the most probability mass on a single permutation, the *canonical permutation*. While other permutations are plausible, their likelihood decays exponentially with their distance from the the canonical permutation. In comparison to HMMs, GMMs greatly restrict a set of possible topic orderings predicted for a given domain as permutations drawn from this distribution are likely to be similar. However this restriction actually enables the compact parametrization of GMMs, supporting effective inference of its parameters in a Bayesian framework.

We position the GMM within a larger hierarchical Bayesian model that explains how a set of related documents is generated. Figure 1 pictorially summarizes the steps of the generative process. At a high level, the model first selects how frequently each topic is expressed in the document, and how the topics are ordered. These topics then determine the selection of words for each paragraph. More specifically, for each document, the model posits that a topic ordering is drawn from the GMM, and that a set of topic frequencies is drawn from a multinomial distribution. Together, these draws specify the document's entire content structure, in the form of topic assignments for each textual unit. As with traditional topic models, words are then drawn from language models indexed by topic. Model parameters are estimated using Gibbs sampling.

2.2 Applications to Text-to-text Generation

One of the important advantages of the automatically induced content models is that they can easily be integrated into existing generation applications. For instance, consider the task of information ordering, where the goal is to determine the sequence in which a pre-selected set of items is presented to the user. This is an essential step in concept-to-text generation, multi-document summarization, and other text-synthesis problems.

To apply a content model to this task, we assume we are provided with well structured documents from a single domain as training examples; once trained, the model is used to induce orderings of previously unseen collections of paragraphs from the same domain. The implementation of the ordering algorithms

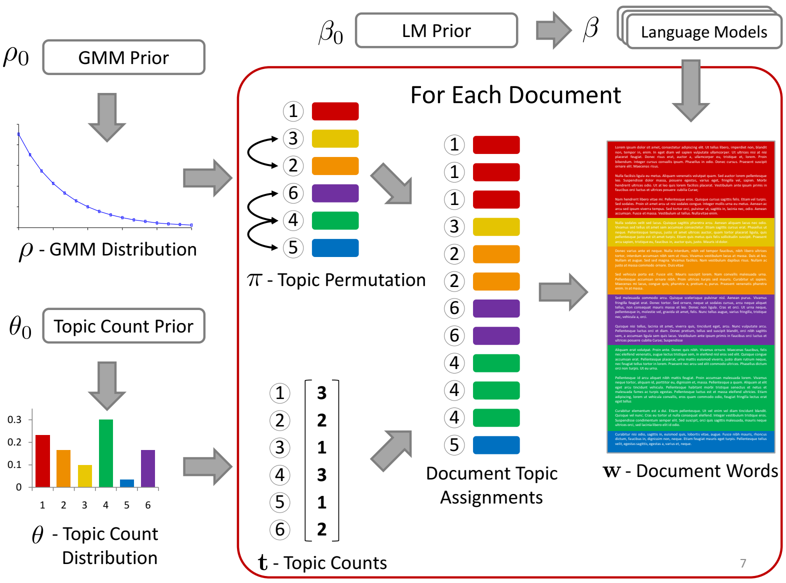

Fig. 1. The generative process for permutation-based content models.

depends on the underlying content model. An HMM-based model searches the space of possible orderings based on the likelihood of a sequence as predicted by the model. Typically, A* and other heuristic search algorithms are used. The GMM-based model takes a very different approach for computing an ordering: it predicts the most likely topic for each paragraph independently using topic-specific language models. Because the GMM distribution concentrates probability mass around one known topic ordering, these topic assignments determine the best ordering of the paragraphs.

These content modeling algorithms have been tested on multiple domains, ranging from product reviews to Wikipedia articles. The results consistently indicate that the ordering retrieved by the model are close to the ordering of a human writer. For instance, the GMM-based algorithm achieves *Kendall's*[2] τ of 0.678 on ordering documents in the domain of cell phone reviews [4].

[2] Kendall's τ measures how much an ordering differs from the reference order. Specifically, for a permutation π of the sections in an N-section document, $\tau(\pi)$ is computed as $\tau(\pi) = 1 - 2\frac{d(\pi,\sigma)}{\binom{N}{2}}$, where $d(\pi, \sigma)$ is the number of swaps of adjacent textual units necessary to rearrange π into the reference order. The metric ranges from -1 (inverse orders) to 1 (identical orders).

Moreover, the empirical results demonstrate the advantages of encoding global structural constraints into probabilistic content models [4]. In fact, the difference between the HMM-based and GMM-based content models is substantial: the former achieves *Kendall's* τ of only 0.256, when tested on the same domain.

In our recent work, we have considered a new application of content models — automatic generation of overview articles. These multi-paragraph texts are comprehensive surveys of a subject, generated by composing information drawn from the Internet [26]. Examples of such overviews include actor biographies from IMDB and disease synopses from Wikipedia. An example of our system's output[3] is shown in Figure 2.

In this application, content models are employed for both selecting and ordering the material in a generated article. Given a corpus of human-authored texts with a corresponding content model, the algorithm learns a content extractor targeted for each topic of the content model, such as *diagnosis* and *treatment*. This extractor specifies a query for selecting candidate material from the web and provides a ranker that assigns a relevance score for each retrieved passage. To produce a new article, the algorithm employs these extractors for each topic of the content model and then jointly selects the best passages based on local and global constraints.

This method is related to content selection methods developed for extractive summarization. In fact, the design of individual extractors is similar to supervised methods for sentence extraction [18]. The difference, however, is in the extraction criteria. Traditional summarization methods filter sentences based on the generic notion of "importance," while our selection criteria is more focused: it is driven by the topic of the candidate passage. This architecture ensures that the overview article will have the breadth expected in a comprehensive summary, with content drawn from a wide variety of Internet sources.

We evaluate the quality of the generated articles by comparing them with the corresponding human-authored articles in Wikipedia. System performance is measured using ROUGE-1, a standard measure used in the summarization community [17]. We employ the system to generate articles in two domains – American Film Actors and Diseases. For each domain, we randomly select 90% of Wikipedia articles in the corresponding categories for training and test on the remaining 10%. To measure the impact of the content model-based architecture for content selection, we consider a baseline that does not use a template to specify desired topics. Instead, we train a single classifier that learns to extract excerpts that are likely to appear in the Wikipedia article, without explicitly capturing the topic they convey. The results convincingly show the advantages of the content model-based approach: in both domains, this approach outperforms the baseline by a statistically significant margin. For instance, in the Disease domain the full model achieves an F-measure of 0.37, while the baseline yields an F-measure of 0.28.

---

[3] This system output was added to Wikipedia at `http://en.wikipedia.org/wiki/3-M_syndrome` on June 26, 2008. The page's history provides examples of changes performed by human editors to articles created by our system.

Diagnosis . . . No laboratories offering molecular genetic testing for prenatal diagnosis of 3-M syndrome are listed in the GeneTests Laboratory Directory. However, prenatal testing may be available for families in which the disease-causing mutations have been identified in an affected family member in a research or clinical laboratory.

Causes Three M syndrome is thought to be inherited as an autosomal recessive genetic trait. Human traits, including the classic genetic diseases, are the product of the interaction of two genes, one received from the father and one from the mother. In recessive disorders, the condition does not occur unless an individual inherits the same defective gene for the same trait from each parent. . . .

Symptoms . . . Many of the symptoms and physical features associated with the disorder are apparent at birth (congenital). In some cases, individuals who carry a single copy of the disease gene (heterozygotes) may exhibit mild symptoms associated with Three M syndrome.

Treatment . . . Genetic counseling will be of benefit for affected individuals and their families. Family members of affected individuals should also receive regular clinical evaluations to detect any symptoms and physical characteristics that may be potentially associated with Three M syndrome or heterozygosity for the disorder. Other treatment for Three M syndrome is symptomatic and supportive.

Fig. 2. A fragment from the automatically created article for 3-M Syndrome.

3 Coherence Models

While a content model is a powerful abstraction of document structure, by definition this representation is domain-specific. As many text-to-text applications are domain-independent, we need a model that can operate in such a context. In this section, we introduce a *coherence model* that captures text relatedness at the level of sentence-to-sentence transitions [7, 21, 1, 14, 16].

The key premise of this approach is that the distribution of entities in locally coherent texts exhibits certain regularities. This assumption is not arbitrary — some of these regularities have been recognized in Centering Theory [8] and other entity-based theories of discourse. Previous research has demonstrated that direct translation of these linguistic theories into a practical coherence metric is difficult: one has to determine ways of combining the effects of various constraints and to instantiate parameters of the theory that are often left unspecified [23, 12, 13, 24].

3.1 Computational Modeling

The model is based on an expressive linguistic representation called the *entity-grid*, a two-dimensional array that captures the distribution of discourse entities across text sentences. The rows of the grid correspond to sentences, while the columns correspond to discourse entities. By *discourse entity* we mean a class of coreferent noun phrases. For each occurrence of a discourse entity in the

1	[The Justice Department]$_S$ is conducting an [anti-trust trial]$_O$ against [Microsoft Corp.]$_X$ with [evidence]$_X$ that [the company]$_S$ is increasingly attempting to crush [competitors]$_O$.
2	[Microsoft]$_O$ is accused of trying to forcefully buy into [markets]$_X$ where [its own products]$_S$ are not competitive enough to unseat [established brands]$_O$.
3	[The case]$_S$ revolves around [evidence]$_O$ of [Microsoft]$_S$ aggressively pressuring [Netscape]$_O$ into merging [browser software]$_O$.
4	[Microsoft]$_S$ claims [its tactics]$_S$ are commonplace and good economically.
5	[The government]$_S$ may file [a civil suit]$_O$ ruling that [conspiracy]$_S$ to curb [competition]$_O$ through [collusion]$_X$ is [a violation of the Sherman Act]$_{]O}$.
6	[Microsoft]$_S$ continues to show [increased earnings]$_O$ despite [the trial]$_X$.

Table 3. Summary augmented with syntactic annotations for grid computation.

text, the corresponding grid cell contains information about its presence or absence in the sentences. In addition, for entities present in a given sentence, grid cells contain information about their syntactic role reflecting whether the corresponding entity is a subject (S), object (O), or neither (X). Entities absent from a sentence are signaled by gaps (–). Table 4 illustrates a fragment of an entity grid constructed for the text in Table 3.[4] Grid representation can be automatically computed using standard text processing algorithms such as a syntactic parser and a tool for coreference resolution.

Our analysis revolves around patterns of local entity transitions. These transitions, encoded as continuous subsequences of a grid column, represent entity occurrences and their syntactic roles in n adjacent sentences. According to Centering Theory, in a coherent text some transitions are more likely than others. For instance, grids of coherent texts are likely to have some dense columns (i.e., columns with just a few gaps such as *Microsoft* in Table 4) and many sparse columns which will consist mostly of gaps (see *markets*, *earnings* in Table 4). One would further expect that entities corresponding to dense columns are more often subjects or objects. These characteristics will be less pronounced in low-coherence texts.

Therefore, by looking at the likelihood of various transition types in a text, we can assess the degree of its coherence. To automatically uncover predictive patterns, we represent a text by a fixed set of transition sequences using a feature vector notation. Given a collection of coherent texts and texts with coherence violations, we can employ standard supervised learning methods to uncover entity distribution patterns relevant for coherence assessment.

[4] These two tables are borrowed from the Computational Linguistics article that introduced the grid representation [3].

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	S	O	S	X	O	–	–	–	–	–	–	–	–	–	–	1
2	–	–	O	–	–	X	S	O	–	–	–	–	–	–	–	2
3	–	–	S	O	–	–	–	–	S	O	O	–	–	–	–	3
4	–	–	S	–	–	–	–	–	–	–	–	S	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	S	O	–	5
6	–	X	S	–	–	–	–	–	–	–	–	–	–	–	O	6

Table 4. A fragment of the entity grid. Noun phrases are represented by their head nouns.

3.2 Application to Text-to-text Generation

The main utility of a coherence model is to automatically assess the quality of a generated output. When a text-to-text generation system is equipped with such a measure, it can select the most fluent candidate among possible output realizations. These possible realizations may correspond to different orders of the output sentences, to different sets of selected sentences, or to different ways entities are realized within each sentence.

To validate the usefulness of the coherence model, we consider two evaluation scenarios. In the first scenario, the model is given a pair of texts consisting of a well-formed document and a random permutation of its sentences. The task is to select the more coherent document, which in this case corresponds to the original document. While this evaluation setup allows us to generate a large-scale corpus for training and testing the method, it only partially approximates the degrees of coherence violation observed in the output of text-to-text generation systems. Therefore, in our second evaluation scenario, we apply the model to assess coherence of automatically generated summaries. In particular, we use summaries collected for the 2003 Document Understanding Conference. These summaries are generated by different multidocument summarization systems, and therefore they exhibit a range of disfluencies.

In both experiments, the grid model achieves notably high performance. For instance, the algorithm can distinguish a coherent ordering from a random permutation with an accuracy of 87.3% when applied to reports from the National Transportation Safety Board Accident Database, and 90.4% when applied to Associated Press articles on Earthquakes. The task of coherence assessment turned out to be more challenging: the best configuration of the model achieves an accuracy of 81.3%. The results also demonstrate that incorporating salience and syntactic features, sources of information featured prominently in discourse theories, leads to a consistent increase in accuracy. For example, eliminating syntactic information decreases the ordering performance of the model by 10%. We also compare the performance of coherence models against content models. While these two types of models capitalize on different sources of discourse information, they achieve comparable performances — content models yield an accuracy of 88.0% and 75.8% on the two datasets, compared with the accuracy of 87.3% and 90.4% obtained by coherence models. Recent work has demonstrated that further improvements in performance can be achieved by combining coherence and content models [5].

3.3 Conclusions

In this chapter, I demonstrated that automatically-induced models of text structure advance the state of the art in text-to-text generation. Our experiments show that incorporating discourse constraints leads to more effective information selection and increases fluency and coherence of the system output. The key strength of the proposed discourse models is that they can be derived with minimal annotation effort, in some cases learning from only the raw text. The performance of these models validates the long-standing hypothesis stated by Harris about the connection between high-level discourse properties and distributional patterns at the word level.

An important future direction lies in designing statistical models of text structure that match the representational power of traditional discourse models. Admittedly, the models described in this chapter constitute a relatively impoverished representation of discourse structure. While this contributes to the ease with which they can be learned, it limits their potential to improve the quality of text-to-text generation systems. For instance, the models described above do not capture the hierarchical structure of discourse, which has been shown to be important for content planning. Another limitation of the above models is that they capture just one aspect of discourse structure rather than modeling text structure in a holistic manner. I believe that recent advances in machine learning (e.g., discriminative structure prediction algorithms and unsupervised Bayesian methods) would enable us to further refine statistical discourse models and consequently improve the output of text-to-text generation systems.

3.4 Bibliographic Note and Acknowledgments

This paper builds on several previously published results [3, 2, 26, 4]. I would like to thank my co-authors – Branavan, Harr Chen, David Karger, Mirella Lapata, Lillian Lee and Christina Sauper – for their generous permission to use this material in this chapter. I thank Branavan and Harr Chen for their thoughtful comments and suggestions for improving this manuscript. This work was funded by the National Science Foundation (NSF) CAREER grant IIS-0448168 and the Microsoft Research New Faculty Fellowship. Any opinions, findings, conclusions or recommendations expressed in this article do not necessarily reflect the views of NSF or Microsoft.

References

1. Ernst Althaus, Nikiforos Karamanis, and Alexander Koller. Computing locally coherent discourses. In *Proceedings of the ACL*, pages 399–406, 2004.
2. Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.
3. Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL*, pages 113–120, 2004.
4. Harr Chen, S.R.K. Branavan, Regina Barzilay, and David R. Karger. Content modeling using latent permutations. *JAIR*, pages 129–163, 2009.
5. Micha Elsner, Joseph Austerweil, and Eugene Charniak. A unified local and global model for discourse coherence. In *Proceedings of HLT-NAACL*, pages 436–443, 2007.
6. M.A. Fligner and J.S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society, Series B*, 48(3):359–369, 1986.
7. Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. Textual coherence using latent semantic analysis. *Discourse Processes*, 25(2&3):285–307, 1998.
8. Barbara Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.
9. Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.
10. M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1976.
11. Zellig Harris. Discourse and sublanguage. In Richard Kittredge and John Lehrberger, editors, *Sublanguage: Studies of Language in Restricted Semantic Domains*, pages 231–236. Walter de Gruyter, Berlin; New York, 1982.
12. Laura Hasler. An investigation into the use of centering transitions for summarisation. In *Proceedings of the 7th Annual CLUK Research Colloquium*, pages 100–107, University of Birmingham, 2004.
13. Nikiforos Karamanis. Exploring entity-based coherence. In *Proceedings of CLUK4*, pages 18–26, Sheffield, UK, 2001.
14. Nikiforos Karamanis, Massimo Poesio, Chris Mellish, and Jon Oberlander. Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus. In *Proceedings of the ACL*, pages 391–398, 2004.
15. Richard Kittredge, Tanya Korelsky, and Owen Rambow. On the need for domain communication language. *Computational Intelligence*, 7(4):305–314, 1991.
16. Mirella Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the ACL*, pages 545–552, 2003.
17. Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of ACL*, pages 74–81, 2004.
18. Inderjeet Mani and Mark T. Maybury. *Advances in Automatic Text Summarization*. The MIT Press, 1999.
19. William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *TEXT*, 8(3):243–281, 1988.
20. Daniel Marcu. The rhetorical parsing of natural language texts. In *Proceedings of the ACL/EACL*, pages 96–103, 1997.
21. Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, 2000.

22. Kathleen R. McKeown. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, Cambridge, UK, 1985.
23. Eleni Miltsakaki and Karen Kukich. The role of centering theory's rough-shift in the teaching and evaluation of writing skills. In *Proceedings of the ACL*, pages 408–415, 2000.
24. Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. Centering: a parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363, 2004.
25. Owen Rambow. Domain communication knowledge. In *Fifth International Workshop on Natural Language Generation*, pages 87–94, 1990.
26. Christina Sauper and Regina Barzilay. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the ACL/IJCNLP*, pages 208–216, 2009.
27. Richard Sproat. *Morphology and Computation*. MIT Press, 1993.