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
regina@csail.mit.edu

Text-to-text generation aims to produce a coherent text by extracting, combining and rewriting information given in input texts. Examples of its applications include summarization, answer fusion in question-answering and text simplification. At first glance, text-to-text generation seems a much easier task than the traditional generation set-up where the input consists of a non-linguistic representation. Research in summarization over the last decade proved that the opposite is true — texts generated by these methods rarely match the quality of those written by humans. One of the key reasons is the lack of coherence in the generated text.

In contrast to the traditional set-up in concept-to-text generation, these applications do not have access to semantic representations and domain-specific communication knowledge. Therefore, traditional approaches for content selection cannot be employed in text-to-text applications. These considerations motivate the development of novel approaches for document organization that can exclusively rely on information available in textual input.

In this talk, 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 close 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 improvement over previously proposed methods.

Next, I will present a method for assessing the coherence of a generated text. The key premise of our work 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 integrated in text-to-text applications such as summarization and answer fusion.

This is joint work with Branavan, Harr Chen, Mirella Lapata and Lillian Lee.

Proceedings of the 12th European Workshop on Natural Language Generation, page 33,
Athens, Greece, 30 – 31 March 2009.