Discourse Factors in Multi-Document Summarization

Ani Nenkova
Columbia University
Computer Science Department
New York, New York 10027
ani@cs.columbia.edu

Introduction

The over-abundance of information today, especially online, has established the need for natural language technologies that can help the user find relevant information; multi-document summarization (MDS) and question answering (QA) are two examples. The requirement in MDS and open-ended QA to produce multi-sentential answers imposes the extra demand that the output of such systems be a coherent discourse. The problem of generating appropriate referring expressions to entities in these texts is non-trivial, because different sentences are taken from their original context and put together to form a text. The new context of the summary often requires changes in surface realization of the references, demanding the inclusion of additional information or removal of redundant information. Such changes can be implemented by gathering a collection of possible references to an entity from the input documents and then rewriting the references in the sentences selected for inclusion in the summary. A question arises how to determine which attributes or descriptions of the referent would be appropriate for the context of the summary.

Information status

Newswire reports (and consequently, news summaries), often center around people—for example, national and international news revolve around politicians' decisions and acts. This means that it is important for newswire summarization systems to have a theory on referring to people, specifying how they are introduced in the story and how they are evoked subsequently.

A significant volume of linguistic literature has been devoted to the study of information status of discourse entities and the way this status influences how a speaker would refer to these entities. The information status of an entity is determined by what the hearer of an utterance might know about the entity, from previous discourse, from the situation or as prior knowledge. One obvious distinction is *discourse-new* (first mention in the current discourse) vs *discourse-old* (subsequent mention). Consider:

a. *Federal Reserve Chairman Alan Greenspan* made an official statement today.

b. *Greenspan* suggested that the Senate make the tax-cut permanent.

Examples (**a**) and (**b**) are (abbreviated) consecutive sentences in a newspaper report in which sentence (**a**) is the first sentence and `Greenspan` is a discourse-new entity, while in (**b**) it is discourse-old. The syntactic difference in the referring expression realization indicates (and is dictated by) this difference in information status.

In another possible context, where the target audience is known to be knowledgeable about U.S. internal affairs, a reference like that in (**b**) can felicitously be used as the initial reference. In this case, `Greenspan` would be discourse-new, but *hearer-old* since a reference by last name presupposes that the hearer already knows who the person is. However, when the entity is expected to be *hearer-new*, a reference similar to the one in (**a**) is more appropriate. Foundational work on generating references discusses how the above distinctions in information status can play a role in deciding on a suitable way to refer to an entity in a relatively limited domain of cooking recipes, where the knowledge of the user can be neatly modeled within a formal framework. Unfortunately, in open domain tasks like summarization, such description is practically impossible. One possible way around the problem is tailoring summarization via a user modeling component that contains information on entities that the user has previously read about. But currently the goal in summarization is to develop generic systems rather than custom-tailored systems and thus a more scalable approach is desirable. In particular, multi-document summarization is a suitable testbed for the development of text-to-text generation applications, where human authored texts are automatically rewritten. Summary rewriting for open domain newswire requires robust models of discourse flow, as well as models of the intended audience. The focus of the work outlined here is to demonstrate that such models can be automatically constructed and applied to improve summary quality.

Summary rewriting

We conducted a corpus study focusing on identifying the syntactic properties of first and subsequent mentions to people in newswire (Nenkova & McKeown 2003). The resulting statistical model of the flow of referential expressions, a Markov chain, suggests a set of rewrite rules that can

transform the summary back to a more natural and readable text. The process of reformulating sentences from the original new report is called *summary rewriting* and it corrects summaries to reflect the discourse-new or discourse-old status of entities. The approach is based on a simple idea—in has been suggested in linguistic literature that the syntactic form of a reference depends on the syntactic forms of previous mentions. Such a property can be captured well by a Markov chain model in which the states represent the different possibilities for syntactic realizations to references to people, such as the form of the name (full name vs. last name only), premodification, and postmodification. The model was trained on a corpus of news stories containing 651,000 words drawn from six different newswire agencies. The highest probability paths in the model were coded as rewrite rules that provide the full name and a description of a person at the first mention and a short reference by last name at subsequent mentions. The rewritten summaries were preferred by human readers. The rewrite modules has been successfully integrated with two summarization systems (Schiffman, Nenkova, & McKeown 2002), (Siddharthan, Nenkova, & McKeown 2004) and has been used as a starting point for other text-to-text generation approaches that have been shown to improve content selection in summarization.

Detecting Hearer-old vs Hearer-New

Detecting hearer-old vs hearer-new status of entities would allow another type of summary rewrite—it would be possible to omit descriptions for hearer-old entities altogether, even at the first mention. Because of journalistic conventions, people, even if they are well known, will always be introduced in news articles, but such restrictions do not apply in the summary genre. An interesting question arises—is it possible to automatically determine which people mentioned in news reports are already known to the large readership. Is it at all feasible to try to model the intended audience in general rather a specific user? In order to answer this question, we gave four America graduate students a questionnaire consisting of a list of names of people appearing in the human written summaries provided for the Document Understanding Conference, and asked them to write down for each person, their country/state/organization and their role (writer/president/attorney-general). We considered a person hearer-old to a subject if they correctly identified both role and affiliation for that person. For the 258 people in the summaries, the four subjects demonstrated 87% agreement (kappa = 0.74), which indicates that there is good agreement and that determining hearer status is a reasonable task. We trained a support vector machine classifier for the task, using frequency and syntactic realization features from the input documents that achieved 76% accuracy, compared to majority class prediction baseline of 54%.

Evaluating summarization systems

In recent years the development of new reliable methods for summarization evaluation has received considerable attention. While the focus of our summary rewriting task is to improve the summary readability, it has been shown that input modifications, such as removing parenthetical constructions, improves content selection (Siddharthan, Nenkova, & McKeown 2004). Thus, summary rewrite needs to be evaluated for its impact both on content selection and on readability. The problem in evaluation arises from the fact that summarization is quite a subjective task and different humans would make different decisions on what content to include in a summary, thus making it impossible to produce a single gold-standard. We developed an empirically motivated evaluation method, based on comparisons of *several* human written summaries, that produces more reliable and diagnostic scores than previous methods (Nenkova & Passonneau 2004). The idea behind the approach is to assign an importance weight to information units—the more human summarizers chose a unit for inclusion in their summary, the more important it can be considered to be. The proposed evaluation procedure incorporates the idea that no single best model summary exists. An effort is under way to manually annotate the multiple human summaries per topic from the Document Understanding Conference. Such a corpus of summaries and reliable scores for content selection will be immensely useful for future work in automating summarization evaluation—a good automatic evaluation procedure would be one that produces scores that correlate well with the manual scores for the summaries.

Future Work

The work that I need to still complete for my dissertation includes the development of methods for evaluation of linguistic quality of summaries, such as fluency and readability. The evaluation of linguistic quality in summarization have not been the focus of research up to date, with more effort devoted to the evaluation of content selection. Methods for assessing readability become more important with the development of text-to-text generation methods such as the summary rewrite method that we proposed. Another part of my future work is to extend the framework for reference generation and use the derived Markov chain model directly to stochastically suggest syntactic realizations, as well as the merging of descriptions from the input articles to form new, unseen in the input, descriptions.

References

Nenkova, A., and McKeown, K. 2003. References to named entities: a corpus study. In *Proceedings of HLT/NAACL 2003*.

Nenkova, A., and Passonneau, R. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT/NAACL 2004*.

Schiffman, B.; Nenkova, A.; and McKeown, K. 2002. Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference*.

Siddharthan, A.; Nenkova, A.; and McKeown, K. 2004. Syntactic simplificantion for improving content selection in multi-document summarization. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*.