

Automatic Summarization of Conversational Multi-Party Speech

Michel Galley

Department of Computer Science
Columbia University
New York, NY 10027, USA
galley@cs.columbia.edu

Introduction

Document summarization has proven to be a desirable component in many information management systems, complementing core information retrieval and browsing functionalities. The use of document summarization techniques is especially important with speech documents such as meeting transcriptions, since they are particularly difficult to navigate. As opposed to their written counterparts, spoken documents lack structural elements like title, headers, and paragraphs that can ease the task of the information seeker. Document summarization can reduce the overhead of navigating large collections of speech data. For example, summarization techniques can be used to extract or highlight relevant passages in a full transcription. Alternatively, similar techniques might be used to generate short text substitutes, or abstracts, that capture the “aboutness” of the meeting, while discarding disfluencies and other unimportant elements.

In my thesis work, I address the problem of creating abstractive summaries of meetings, a task that has to date only received limited attention. Abstractive summarization differs from extraction in that it does not simply concatenate input utterances, but alters the extracted material in order to produce fluent and concise summaries. Speech summarization faces many challenges not found in texts, in particular high word error rates (WER on this data is 34.8%), absence of punctuation, and sometimes lack grammaticality and coherent ordering.

My approach to speech summarization divides the problem into two subtasks that are admittedly quite independent: *content selection*, i.e. identifying a set of salient utterances that is a practical substitute to the entire meeting transcription; *utterance revision*, i.e. correcting the various non-fluencies typical of conversational speech, and operating below the sentence level to further remove unimportant lexical material.

Content Selection

My approach to utterance selection incorporates two processing stages: since unstructured summaries can be hard to read, the first stage is to divide the meeting transcription by topics. This accounts for the fact that meetings under

investigation have pre-set agendas and are structured as sequences of research issues to be discussed,¹ and that, even though topical segments may seem chaotic or underpinned by complex interplays, their segmentation generally seems to be well-defined, and human labelers reached a marked agreement in identifying them (Galley *et al.* 2003).

The second stage is treated as a binary sequence classification task that labels each sentence as either SUMMARY or NON-SUMMARY. While I was able to devise many lexical, acoustic, durational, and structural predictors for the task that seem to correlate well with human summary annotation (including many features new to summarization research), an interesting research question in this task relates to more computationally and discourse oriented considerations: do interactions between participants and inter-sentential dependencies influence the selection of summary sentences? For example, it may seem reasonable to believe that a given utterance that elicits multiple responses or reactions from other participants in the meeting is quite likely to be included in the reference summary. Statistical correlation tests have shown that this is indeed the case, and that meetings present a great wealth of such inter-sentential dependencies.

These findings gave me solid grounds for modeling inter-sentential correlations in content selection with dynamic Bayesian networks (DBN). While it is quite beneficial to model first-order Markov dependencies between adjacent utterances, other types of dynamic dependencies were also found useful, in particular *speaker-addressee* relations.²

A main contribution of my work in content selection lies in its use of a probabilistic model of interpersonal interaction to automatically determine the structure of the DBN, i.e. probabilistically determine at each point in time the set of parent nodes of each dynamic variable, in particular *speaker-addressee* edges (Galley *et al.* 2004). While this kind of “switching-parent” DBN structure has been the ob-

¹The ICSI meetings used in my work consists of 75 meetings primarily concerned with speech and AI research.

²For example, let’s consider the following case: the first utterance is an offer from speaker A, the second a rejection from speaker B, and the third an acceptance of speaker C (addressed to A). It seems reasonable to believe that the inclusion or non-inclusion of utterance 1 will affect the chances of utterance 3 to be included in the summary—i.e., it is quite unlikely the acceptance will appear in the reference summary without the actual offer—and this independently of utterance 2.

ject of recent studies in machine learning (see, e.g., (Bilmes 2000)), it is quite new to natural language and particularly summarization research, and seems to have a broad range of applications in problems pertaining to multi-party discourse.

Utterance Revision

Purely extractive approaches to summarization may work reasonably well with written texts, but typically fail to produce good summaries when applied to spoken documents, particularly multi-party speech. Because of their conversational, informal, and competitive nature, transcripts tend to be fragmentary, disfluent, and riddled with errors. These phenomena are particularly prevalent in meetings, since participants are generally not professional speakers, as opposed to other speech domains such as broadcast news.

The revision system in my thesis is aimed at condensing utterances by removing non-fluencies typical to spontaneous speech, as well as semantically poor phrases (e.g., “I mean”), and grammatically optional constituents. More specifically, my current utterance revision model is based on general formalisms known as synchronous grammars (or transformational grammars), which are designed to generate two languages synchronously. In the case of revision, a rule of a synchronous grammar may for example correspond to the deletion of a prepositional phrase or an adverb. Such formalisms have generated substantial interest in other areas of NLP, particularly machine translation, but there are relatively few previous applications to summarization.

My model is fully trainable from any parallel corpus of syntactically parsed (utterance,revision) sentence pairs. The inference procedure obtains synchronous rules from aligned pairs of trees using an algorithm to determine the minimum tree-to-tree edit distance. The probabilistic model scores possible revisions according to various information sources: syntactic transformation rules; a syntax-based language model to promote revision hypotheses that are grammatical; a phrase-based model that was trained to identify phrases likely to be deleted (“you know”); a *tf·idf* model of word importance to promote the removal of constituents with low informative content; prosodic features to exploit existing acoustic correlates of words found in abstracts, e.g. prosodically stressed words. The different sources are integrated in a discriminative training framework where the available evidence of all models is combined to select a globally optimal analysis. A chart-based dynamic programming algorithm is used to ensure that the different possible analyses are scored and ranked in a relatively efficient manner.

While similar techniques have been used to perform sentence compression (Knight & Marcu 2000)—though not in speech domains, I believe that my proposed work encompasses significant contributions.

Firstly, I expect to improve previous work by not limiting revision operations to deletions of subtrees in the syntactic analysis of the input sentence. Such limitation not only prevents word insertions and substitution, but also disallows certain types of deletions, due to syntactic constraints. It was shown on actual summarization data that such a restricted deletion model does not account well for human abstraction behavior (and, in particular, that only 2% of the sentence

pairs in the Ziff-Davis corpus fit this deletion model), which make it difficult to apply corpus-based methods. In preliminary work, I experimented with various techniques for extracting synchronous grammars that are more expressive and representative of human abstraction behavior.

Secondly, I introduce a more robust grammar parameterization. Previous approaches use quite simple techniques to assign rule probabilities, i.e. relative frequencies. This presents major data sparseness issues, since normalization generally occurs over quite complex low-occurrence structures. Instead, I use a parameterization inspired by work in syntactic parsing, which factors each rule probability through reasonable linguistic independence assumptions. This factorization reduces data sparseness and allows better probability estimates by introducing lexical dependencies (e.g., this permits the distinction between the adverb “not”—obviously a bad deletion candidate—and “actually”).

Current Status and Plan for Completion

My research on content selection is in an advanced stage. A topic segmentation method has been implemented (Galley *et al.* 2003), which gives a performance that is state-of-the-art on the meeting data. I have finished constructing a set of fairly useful predictors for binary sequence classification in content selection, and I am currently exploring different parameter estimation and model selection techniques to build and train a switching-parent DBN. Recent experiments have shown that this type of models outperforms first-order Markov models in the utterance selection task.

The proposed research in utterance revision is largely future work. I already developed a baseline revision system that replicates previous work, and added my less restrictive rule extraction component, but I still need to further improve probabilistic modeling for the acquired rules. The decoding problem requires a significant amount of extra work, since the search to find the most likely revision sentence is currently quite slow and does not scale to long input sentences. Finally, after each component has been evaluated in isolation, I plan to work on overall system evaluation.

Acknowledgments

I would like to thank Kathleen McKeown and Julia Hirschberg for helpful discussions. This work was funded by the NSF under grant No. IIS-0121396.

References

- Bilmes, J. 2000. Dynamic bayesian multinets. In *Proc. of UAI*, 38–45.
- Galley, M.; McKeown, K.; Fosler-Lussier, E.; and Jing, H. 2003. Discourse segmentation of multi-party conversation. In *Proc. of ACL*, 562–569.
- Galley, M.; McKeown, K.; Hirschberg, J.; and Shriberg, E. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proc. of ACL*, 669–676.
- Knight, K., and Marcu, D. 2000. Statistics-based summarization—step one: Sentence compression. In *Proc. of AAAI*, 703–710.