

Abstract

While the majority of summarization research so far has focused on written documents (mostly news articles or scientific papers), this thesis addresses for the first time the challenge of automatically summarizing spoken dialogues in a variety of genres and without any restriction on domain.

To achieve the goal of spoken dialogue summarization, we implement a system (DiaSumm) using a multi-stage architecture with trainable components which addresses the dialogue-specific issues of summarization and which involves (i) speech disfluency detection and removal, (ii) identification and insertion of sentence boundaries, (iii) identification and linking of question-answer regions, (iv) topical segmentation, and (v) information condensation (ranking of relevant pieces of information with the maximum marginal relevance technique (MMR)). We can also optionally reduce the summary content in an orthogonal dimension by rendering only a subset of the phrases within a relevant sentence (typically, noun phrases).

For system development and evaluation, we use a corpus of 23 dialogue excerpts from four different text genres, totalling 80 topical segments, about 47000 words, or about 4 hours of recorded speech: English CallHome (informal, colloquial style), Group Meetings (task oriented, rather informal, colloquial), and dialogue oriented television shows: NewsHour and CrossFire (more formal, potentially partially scripted). The corpus had been manually transcribed and was annotated for topical boundaries and relevant text spans by six human annotators. Further, it was annotated for speech disfluencies and questions and their corresponding answers. We devise a word-based evaluation criterion, relative summary accuracy, which reflects how well the summary captures passages that were placed in man-made summaries by the largest number of annotators.

The global evaluation, performed on human transcripts, shows that for the two more informal genres (CallHome and Group Meetings), DiaSumm significantly outperforms a baseline using TF*IDF term weighting with MMR ranking only, while tying with the MMR baseline for the two more formal genres. Furthermore, except for the NewsHour corpus, both the MMR baseline and our DiaSumm system are significantly better than a LEAD baseline (first N words of each segment). Finally, when using speech recognizer output, our system can make successful use of speech recognizer confidence scores to focus on sentences which are more likely to be correctly recognized; thereby, the word error rate in summaries can be reduced significantly while relative summary accuracy improves on average.