Increasing the Coherence of Spoken Dialogue Summaries by Cross-Speaker Information Linking

Klaus Zechner and **Alon Lavie**
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
{zechner,alavie}@cs.cmu.edu

Abstract

This paper addresses the question of how to increase local coherence in summaries of multiparty conversations. Due to the interactive nature of dialogues, local regions of coherence often stretch across different speakers, as for instance in question-answer pairs. We present an approach to automatically detect those regions of local coherence and evaluate the resulting summaries' fluency and informativeness. Our approach increases summary fluency significantly while not compromising informativeness.

1 Introduction

Summary coherence is one of the major challenges for extract based summarization methods: in this case, summaries are not generated from some abstract semantic representation, resulting from a deeper analysis of the original text, but rather composed of passages from the original text (usually sentences or clauses) which are deemed to be relevant by some kind of relevance metric. When automatically summarizing well structured written texts, such as newswire data or scientific papers, the following strategies have been used to increase local coherence of such summaries, sometimes used in combination with each other: (a) lead based summaries (extracting the contiguous header of the text) (Brandow et al., 1995; Wasson, 1998); (b) paragraph based summaries (using paragraphs as minimal extraction units) (Mitra et al., 1997; Salton et al., 1997); (c) inclusion of sentences which likely contain antecedents to anaphora in the current summary sentences (Johnson et al., 1993); (d) replacement of pronouns and definite descriptions by their antecedents (by means of automatic anaphora resolution) (Boguraev and Kennedy, 1997). Other methods, such as automatic analysis of discourse structure (Marcu, 2000), are more aimed at increasing the global coherence of a summary.

Spoken dialogue summarization introduces at least one additional dimension of coherence which is absent from written text generated by a single author: local cross-speaker coherence. Speakers accept or deny requests from each other, pose and answer questions, or acknowledge or comment on what was said by another dialogue participant. This paper focuses on the challenge of how to identify passages with cross-speaker coherence and to what extent this cross-speaker information linking helps to increase the local coherence of spoken dialogue summaries. In particular, it addresses the important subset of the detection of question-answer regions within a dialogue. This information is used when generating extract based summaries to output complete question-answer regions, rather than portions of them. A user study is conducted to look at how our approach (as well as an optimal oracle) affects both the fluency and the informativeness of the resulting summaries.

The organization of this paper is as follows: The next section motivates the paper, then we briefly discuss related work in section 3. Section 4 describes the corpus we use for the experiments, followed by the description of our question and answer detection components (sections 5 and 6), including evaluations thereof. The effects of Q-A-detection on informativeness and fluency are evaluated in sections 7 and 8. We discuss the results of the experiments, as well as future work in section 9, before we conclude the paper (section 10).

2 Motivation

The focus of this work is the issue of local coherence in spoken dialogue summaries in unrestricted domains. The work is performed in the context of a maximum marginal relevance (MMR) summarizer (Carbonell and Goldstein, 1998), which is based on TFIDF[1] weights, and which uses manually determined topical segments from spoken dialogues as input. The summarizer also performs automatic disfluency detection and removal, but uses manually marked sentence boundaries.[2] For the purpose of this paper, we use manual transcriptions of the spoken dialogues. To generate the summary, the MMR summarizer iteratively creates a ranked list of sentences based on their saliency, while minimizing their redundancy with sentences already present in the summary. The generation process is stopped when the length threshold of the summary is reached (e.g., 15% of the words in a segment), and then the sentences from the ranked list are presented in text order.

Let us now look at the following MMR extract summary of a conversation transcript (15% length of the original; **a**, **a1**, and **b** are speaker labels):

```
254 b : It's just it's very hard to pass
        this test right now he is stressed
        it's awful
267 b : He's taking three hours of classes
269 a : What then what will happen with
        the roommate situation
278 a1: Are you planning to move closer
        to the university
291 a : I guess at school you just kind [...]
301 a : Because sometimes I perceive especially
        when we first start talking that you're
        sort of struggling to speak english
304 b : That's only because I just spent the
        weekend with four people I that only
        know spanish
365 a1: Do you are you taking lots of photos
        with your trips on
```

It is quite obvious that part of the reason for the impression of incoherence in this example is related to the fact that several questions in this segment of the dialogue lack their corresponding answers (sentences 269, 278, 365). To increase the local summary coherence, we would like to be able to (a) identify the questions, (b) identify their answers, and (c) link them together such that whole question-answer regions can appear in the summary. We now present, as a comparison, a 15%-summary of the same underlying text as above, but adding complete Q-A regions whenever the MMR algorithm encounters a question or its answer, assuming an oracle knowing the location of the Q-A regions (**Q** and **A** denote questions and their answers):

```
   254 b : It's just it's very hard to
           pass this test right now he
           is stressed it's awful
   267 b : He's taking three hours of classes
Q  269 a : What then what will happen with
           the roommate situation
A  271 b : No
A  272 b : I'm moving
Q  278 a1: Are you planning to move closer
           to the university
A  279 b : Yes
   291 a : I [...]
   301 a : Because sometimes I perceive
           especially when we first start
           talking that you're sort of
           struggling to speak english
   304 b : That's only because I just spent
           the weekend with four people I
           that only know spanish
Q  365 a1: Do you are you taking lots of
           photos with your trips on
A  366 b : Yeah
```

We can see that the effect of an approach for Q-A detection would be an increase in local coherence and text fluency. At the same summary size, these Q-A regions will cause some other sentences (or parts thereof) to disappear from the summary, so its overall informativeness could be affected negatively. However, we conjecture that this effect will be compensated by the higher informativeness of the locally coherent Q-A regions. Thus, our initial hypothesis for this paper is that when adding a question-answer detection component to the summarizer, (a) the coherence of the resulting summaries will be increased, which will be reflected in higher fluency and readability, and (b) the informativeness of the summary (its accuracy or relevance) will not decrease significantly.

[1] TFIDF: term frequency times inverse document frequency

[2] A detailed description of the dialogue summarization system is provided in the first author's dissertation (Zechner, forthcoming).

3 Related Work

While some amount of research has been done in the area of knowledge based summarization of spoken dialogues in narrow domains (Kameyama et al., 1996; Reithinger et al., 2000), extract based dialogue summarization in unrestricted domains has been explored very little so far. Research on audio summarization in general (without focusing on dialogues) has been conducted in the context of the TREC spoken document retrieval conferences and the DARPA Broadcast News workshops in recent years (Valenza et al., 1999; Hirschberg et al., 1999). Our own previous research described a first prototype of a dialogue summarization system in (Zechner and Waibel, 2000) and its embedding in an architecture for transcribing, archiving, retrieving, and summarizing of multi-party meetings in (Waibel et al., 2001).

Detecting a question and its corresponding answer can be seen as a sub task of the speech act (SA) detection and classification task. Most systems described in the literature have been operating in narrow and well-defined domains, but recently, (Stolcke et al., 2000) presented a comprehensive approach to speech act modeling with statistical techniques that extends to corpora of potentially unrestricted domains. A good overview and comparison of recent related work can also be found in this article. Results from their evaluations on SWITCHBOARD data (Godfrey et al., 1992) show that word based speech act classifiers usually perform better than prosody based classifiers, but that a model combination of the two approaches can yield to an improvement in classification accuracy.

4 Corpus

The corpus we use for our Q-A detection experiments reported in this paper is characterized in Table 1. We use human generated transcripts of spoken dialogues from four different genres:

- English CALLHOME and CALLFRIEND: 12 dialogues from the Linguistic Data Consortium (LDC) collections, of which we used 8 for general system tuning and development of the answer detection component (8E-CH) and 4 for an independent evaluation set (4E-CH) (LDC, 1996). These are recordings of phone conversations between two family members or friends.
- NEWSHOUR (NHOUR): Excerpts from PBS's NEWSHOUR TV show with Jim Lehrer (recorded in 1998).
- CROSSFIRE (XFIRE): Excerpts from CNN's CROSSFIRE TV show with Bill Press and Robert Novak (recorded in 1998).
- GROUP MEETINGS (G-MTG): Excerpts from recordings of scientific project group meetings[3] (recorded in 1998).

The corpus was manually annotated in the following three dimensions:

1. *topical boundaries:* First, four to six human annotators marked topical boundaries independently, and in a second phase, a sub-group of two annotators constructed a "gold standard", based on boundary positions selected by at least half of the annotators in the first phase. The resulting corpus comprises 80 topical segments in total. For all experiments in this paper, the summarization system operates on these topical segments individually.
2. *relevant text spans:* The same group of 4-6 annotators marked text spans[4] comprising about 10-20% of the words in each topical segment which should form their summary-relevant "cores".[5]
3. *questions and their answers:* Questions and answers were marked up in the following way: Every sentence which is a question was marked as either a Yes-No-question or a Wh-question. Exceptions were back-channel questions, such as "Is that right?", rhetorical questions, such as "Who would lie in public?", and other questions which do not refer to a propositional content. These were not marked

[3] Interactive Systems Laboratories, Carnegie Mellon University

[4] Typically, these correspond to speaker turns or sentences, but they can be longer or shorter, as well.

[5] A second phase for the "gold standard" creation was also conducted here; this paper, however, does not make use of the gold standard annotations, but rather of the average of the individual annotations in the first phase (see section 7.1).

Data set	8E-CH	4E-CH	NHour	XFire	G-Mtg
dialogues (total)	8	4	3	4	4
topical segments (total)	28	23	8	14	7
different speakers (average)	2.1	2	2	6	7.5
sentences (average)	280	366	101	281	304
questions (in % of sentences)	3.7	6.4	6.3	9.8	4.0
disfluent (in % of words)	16.0	16.3	11.8	4.2	23.9

Table 1: Data characteristics for the corpus.

(even if they have an apparent answer), since we see the latter class of questions as irrelevant for the purpose of increasing the local coherence within summaries. For each Yes-No-question and Wh-question which has an answer, the answer was marked with its relative offset to the question it belongs to. Some answers are continued over several sentences, but only the core answer (which usually consists of a single sentence) is marked. This decision was made to bias the answer-detection module towards brief answers, and to avoid the question-answer regions getting too lengthy, at the expense of summary conciseness.

5 Automatic Question Detection

For training of the question detection component, we use the manually annotated set of roughly 200,000 SwitchBoard speech acts.[6] We used two different methods: (a) a speech act tagger,[7] and (b) a decision tree based on trigger word and part-of-speech information.

The speech act tagger tags one speech act at a time and hence can only make use of speech act unigram information. Within a speech act, it uses a language model based on parts of speech (POS) and the 500 most frequent word/POS pairs. It was not specifically optimized for the task of question detection. Its typical runtime for speech act classification is about 10 speech acts per second.

The decision tree classifier (C4.5, (Quinlan, 1992)) uses the following set of features: (a) POS and trigger word information for the first and last five tokens of each speech act[8]; (b) SA length; and (c) occurrence of POS bigrams. The set of trigger words is determined by collecting words which frequently occur close to sentence boundaries.[9] To obtain the most discriminative POS bigrams, i.e., bigrams whose distribution varies heavily across Q-SAs vs. non-Q-SAs, we used the following procedure:

1. for a balanced set of Q-SAs and non-Q-SAs (about 9000 SAs each, see Table 3): count all the POS bigrams in SA positions 1..5 and $(n-4)..n$ (using Start and End for the first and last bigrams, respectively) and memorize position type (beginning or end of SA) and SA type (Q-SA vs. non-Q-SA)
2. for all bigrams:
   (a) add 1 to the count (to prevent division by zero)
   (b) divide the Q-SA-count by the non-Q-SA-count
   (c) if the ratio is smaller than 1, invert it (ratio:=1/ratio)
   (d) multiply the ratio with the total frequency of Q-SA-count and non-Q-SA-count combined[10]
3. extract the 100 bigrams with the highest value

We trained two versions of the decision tree: (a) with an unbalanced training set of about 20,000 SAs from the SwitchBoard training data which reflects the true distribution of SAs

[6]From the Johns Hopkins University LVCSR Summmer Workshop 1997, available from http://www.colorado.edu/ling/jurafsky/ws97/.

[7]Thanks to Klaus Ries for providing us with the software.

[8]Shorter SAs are padded with dummies.

[9]In this paper, we use the notions of *speech acts* and *sentences* interchangeably.

[10]Leaving out this step favors low frequency high discriminative bigrams too much and causes a slight reduction in overall Q-detection performance.

	SA tagger	unbalanced DTree	balanced DTree
precision	.57	.63	.26
recall	.61	.51	.84
F_1-score	.59	.56	.40
pr_{avg}	.59	.57	.55
typical classification speed (SAs/sec)	10	1000	1000

Table 2: Question detection performance on the 8E-CH corpus using three different methods.

in general and questions in particular; (b) with a balanced training set of about 18,000 SAs from the SWITCHBOARD training data which contains approximately the same number of questions and answers. The motivation for the latter decision tree was to enforce focus on the relatively infrequent Q-classes (see Table 3) and hence trying to boost recall at the expense of precision, since the classifier would overestimate the Q-classes on a non-skewed test set.

We evaluated the speech act tagger and the decision tree classifiers on the 8E-CH data set. Table 2 reports precision, recall, F_1-score[11], and precision-recall average[12] for these question detection experiments with the three different classifiers used. Note that while the decision trees' performance is only slightly below the speech act tagger's scores, their typical classification speed is two orders of magnitude higher. For the remainder of this paper, we use the unbalanced decision tree as question detection component, which yields the best compromise in overall performance (pr_{avg}) and runtime.

6 Automatic Answer Detection

After identifying which sentences are questions, the next step is to identify their answers. From the annotations, we observe that for the majority of the Yes-No- and Wh-questions, the answer is to be found in the first sentence of the speaker following the speaker uttering the question. In the remainder of cases, the majority of answers are in the second (instead of the first) sentence of the other speaker. Further, there are usually no (or only very few) sentences uttered by the speaker who posed a question *after* the question is being asked and before the next speaker starts talking. Unlike for the question detection task, we here use a set of heuristic rules whose parameters we train on the annotated corpus:

- if the first speaker change after the question occurs more than $maxChg$ sentences after the question, the search is stopped and no Q-A-pair is returned
- answer hypotheses are sought for maximally $maxSeek$ sentences after the first speaker-change after the question, but not over interruptions by any other speaker, i.e., we check within a single speaker region[13] (this is also the stopping criterion for the following two heuristics)
- answers have to be minimally $minAns$ words long; if they are shorter, we add the next sentence to the current answer hypothesis
- even if the minimum answer length is reached, the answer can be (optionally) *extended* if at least one word in the next contiguous sentences matches a word from the question (parameter *sim*); two different stop lists ($StopShort$, $StopLong$), or no stop list are used to remove function words from consideration (parameter *stop*)[14]

Our heuristics further handles simple embedded questions of the type `speaker_A-Q1, speaker_B-Q2, speaker_A-A2, speaker_B-A1`, while disallowing *crossovers* of the type `speaker_A-Q1, speaker_X-Q2, speaker_B-A1, speaker_Y-A2`, where A1 is the answer to Q1 and A2 the answer to Q2. The output of the algorithm is a list of triples $< Q, A_{start}, A_{end} >$, where Q is the sentence-ID of the question, A_{start} the first sentence, and A_{end} the last sentence of the answer. There are four possible outcomes for each hypothesized

[11] $F_1 = \frac{2PR}{P+R}$, where P=precision and R=recall.
[12] $pr_{avg} = \frac{P+R}{2}$

[13] A sequence of sentences spoken by the same speaker.
[14] *StopLong* contains 571 words, *StopShort* only 89 words, most of which are auxiliary verbs and filler words.

	unbalanced set	balanced set
Yes-No-questions	539	5569
Wh-questions	199	1989
other questions	178	1621
questions total	916 (4.6%)	9179 (50.3%)
non-Q speech acts	18784	9064
total speech acts	19700	18243

Table 3: Speech act frequencies for the 2 decision tree training sets.

triple: (a) irrelevant: a Q-A pair with a wrongfully hypothesized question (this is the fault of the question detection module, not of this heuristics); (b) missed: the answer was missed entirely; (c) completely correct: A_{end} coincides with the correct answer sentence ID; and (d) correct range: the answer is contained in the interval $[A_{start}, A_{end}]$ but does not coincide with A_{end}. For the calculation of precision, recall, and F_1-score, we count classes (c) and (d) as correct and use the sum of all classes for the denominator of precision and the total number of Q-A-pairs as the denominator of recall.

To determine the best parameters, we varied them across a reasonable set of values and ran the answer detection script for all combinations of parameters, using again the 8E-CH corpus. The following parameter setting yielded the best answer detection performance: $maxChg = 2, maxSeek = 4, minAns = 10, sim = on, stop = no$.

Table 4 presents the results of the Q-detection, using the decision tree trained on unbalanced data, and the Q-A-detection, using the output from the decision tree, for all 5 sub-corpora. Except for the rather small NewsHour corpus (with fewer than 20 questions or Q-A-pairs to identify), the typical Q-detection F_1-score is around .6 and the Q-A-F_1-score around .5. In two cases, the Q-A-detection performance is slightly better than the Q-detection performance. This seems puzzling at first glance but can be explained by the fact that the answer detection algorithm prunes away a number of Q-hypotheses, reducing the space for potential Q-A-hypotheses.

Summary generation using detected Q-A-regions

When we use the Q-A-detection component to aid summarization, the basic MMR algorithm stays the same. However, whenever a sentence which is part of a Q-A-region is put into the ranked list, the whole region is now added to the summary. This amounts to taking the maximum MMR score of the sentences within a Q-A-region to be its representative. Q-A regions are always described with the triple of sentence-IDs defined above: $< Q, A_{start}, A_{end} >$.

7 Influence on Summary Accuracy

This section uses a numeric score, *summary accuracy*, to represent the quality of a summary. It is based on human relevance annotations of the dialogues and reflects how close the summary represents the opinion of the majority of the annotators. We first define summary accuracy, and then present experiments and their results, using the MMR summarization system and various methods of Q-A detection.

7.1 Definition

For each topical segment s, for each annotator a, we define a boolean word vector of annotations $w_{s,a}$, each component $w_{s,a,i}$ being 1 if the word w_i is part of a relevant text span for that annotator and segment, and 0 otherwise. We then sum over all annotators' annotation vectors and normalize them by the number of annotators per segment (A) to obtain the average relevance vector for segment s, r_s:

$$r_{s,i} = \frac{\sum_{a=1}^{A} w_{s,a,i}}{A} \quad (1)$$

To obtain the summary accuracy score $sa_{s,N}$ for any segment summary with length N, we

	8E-CH	4E-CH	NHOUR	XFIRE	G-MTG
Q to detect	83	94	19	110	49
Q hypotheses	67	60	16	71	52
Q-detection (F_1)	.56	.58	.80	.60	.59
Q-A pairs to detect	68	69	18	79	32
Q-A pair hypotheses	54	54	14	54	33
Q-A-detection (F_1)	.51	.60	.81	.51	.51

Table 4: Performance comparison for Q- and Q-A-detection (Q-detection with unbalanced decision tree).

	Q-A pairs	no Q-A det.	automatic	oracle
8E-CH	68	0.569 (0.170)	0.568 (0.169)	0.559 (0.170)
4E-CH	69	0.605 (0.128)	0.608 (0.123)	0.599 (0.139)
NHOUR	18	0.457 (0.232)	0.476 (0.248)	0.453 (0.230)
XFIRE	79	0.603 (0.129)	0.621 (0.151)	0.598 (0.118)
G-MTG	32	0.572 (0.194)	0.595 (0.155)	0.572 (0.194)
total	266	0.574 (0.163)	0.582 (0.163)	0.568 (0.164)

Table 5: Average summary accuracy (with standard deviations in brackets) for 15% summaries, using three different Q-A-detection methods.

multiply the boolean summary vector summ_s[15] with the average relevance vector r_s, and then divide this product by the sum of the N highest scores within r_s (maximum achievable score), rsort_s being the vector r_s sorted by relevance weight in descending order:

$$sa_{s,N} = \frac{\text{summ}_s r_s}{\sum_{i=1}^{N} \text{rsort}_{s,i}} \tag{2}$$

It is easy to see that the summary accuracy score always is in the interval $[0.0, 1.0]$.

To illustrate the computation of *summary accuracy*, we present a simplified example in Figure 1, where the segment to be summarized consists only of one sentence and is annotated by two human coders.

Suppose we want to summarize the 10-word sentence in Figure 1 using 5 words (i.e., creating a 50%-summary). Suppose the summarizer yields the string "for the concept summary accuracy" as output. Since there are 5 words with an average relevance score of 1.0, the maximum achievable score for a 5-word summary would be 5.0, and the corresponding optimal summary would be the string "simplified example for summary accuracy". Summing the average scores of the 5 words in the actual summary yields 4.0. So the summary accuracy for this sub-optimal summary is 4.0/5.0=0.8.

7.2 Experiments

To get an idea about how the summary accuracy changes using Q-A-pair detection and linking, we first tuned the parameters of the MMR summarization system, using the 8E-CH sub-corpus only. For the Q-A-detection component, we use three different options: (1) no Q-A-detection (this is the baseline system for this experiment), (2) automatic Q-A-pair detection with the unbalanced Q-detection decision tree and the A-detection script, and (3) optimal Q-A-pair detection, using an oracle informed by the human annotators' mark-ups.

Table 5 shows the results of these experiments. While we note that in most cases, the differences are rather small (t-test: $t < .6$, no significant differences overall), we have to take into account the low number of Q-A-pairs in most of the dialogues. In dialogues with a larger number of Q-A-pairs, there is sometimes a noticeable improvement in summary accuracy, particularly for the automatic Q-A detection method. On average, the accuracy scores for the

[15]For every word: 1 if the word is in the summary, 0 otherwise.

```
TEXT:      this is a simplified example for the concept summary accuracy
ann1:               x          x       x                 x       x
ann2:               x          x       x   x   x         x       x
avg.score: 0    0  0 1          1       1   0.5 0.5       1       1
```

Figure 1: Example for the computation of the summary accuracy score.

oracle summaries are slightly below the baseline, while the summaries using our automatic detection module are slightly above the baseline. In short, this experiment shows that using Q-A-detection for summary generation does not significantly affect summary relevance.

8 User Study

For the purpose of testing whether Q-A-detection can increase the local coherence of summaries, we performed a user study. We picked the 15 dialogue segments with the highest number of questions, since we wanted to quantify the effect of Q-A detection on texts which are particularly rich in Q-A-regions. For each of these dialogue segments, we took the same three versions of summaries described in the preceding section, each of them again at 15% length of the original (by word count). We had to exclude four segments which did not change when using Q-A-detection, due to the fact that the top-ranked sentences did not belong to any Q-A regions.

We then asked 6 subjects to rank the three different versions of summaries of the remaining 11 texts for (a) informativeness and (b) fluency (the latter should reflect local coherence). To aid the ranking process, the subjects had to score the summaries first using a discrete scale from 1 to 5 (for both dimensions). Informativeness should measure how much information the summary contains ("dense" vs. "sparse" text); the criterion for fluency should be how easy it is to read the summary and how coherent it is.

The order of the texts, as well as the summary versions within each text, were randomized. The average summary length was 142 words, thus each subject had to read and evaluate a text corpus of about 4700 words, which took, on average, about 31 minutes to complete.

Table 6 presents the results of this study. Each number in the table is the average of 66 scores (11 texts times 6 subjects). For the rank scores, we gave 3 points to the first rank, 2 to the second, and 1 point to the last ranked summary version. In case of rank ties, we assigned 2.5 points (for rank 1=2) or 1.5 points (for rank 2=3), respectively. We observe that while the informativeness of the different summary versions does not change on average (no statistical difference), there is a significant improvement in fluency over the baseline for both summaries using automatic Q-A detection and oracle Q-A detection (significant at $\alpha < 0.05$).[16] Individual subjects' scores did not differ much in these overall trends.

9 Discussion and Future Work

Both the results from the automatic summary accuracy evaluation, as well as the results from the user study show that using Q-A detection does not significantly decrease the informativeness of the resulting summaries: neither evaluation showed a significant difference in information content or relevance for the three different versions of summaries. At the same time, as the user study clearly indicates, there is a significant benefit to be gained from including Q-A-regions in the summary in terms of summary fluency or local coherence. These results confirm our initial hypothesis which stated that while we expect summary coherence to increase with Q-A-region detection, we are sceptical about its benefit to increase overall summary informativeness and relevance.

We looked into the question why summary accuracy (on average) seems to be improving slightly (though not significantly) when we use our automatic Q-A detection module, while it stays at about the level of the baseline when using the oracle Q-A detection. When inspecting the summaries where this effect is most pronounced, we find that the main reason lies in the difference in Q-A region size between the automatic method and the oracle: While the oracle, derived from the human Q-A annota-

[16] Using t-tests.

evaluation dimension	Informativeness			Fluency		
Q-A detection method	no	auto	oracle	no	auto	oracle
average score	3.18	3.18	3.24	2.82	3.12	3.50
average rank score	2.01	2.00	1.99	1.68	2.02	2.30

Table 6: Results of the user study comparing three different versions of summaries (average across all subjects and texts; $n = 66$).

tions, typically generates short answers ("core answers"), the automatic method tends to produce somewhat longer answers, consisting of multiple sentences. Particularly in cases where the core answer consists of only one or very few words (e.g., "yes"), the gain for summary accuracy is negligible. To avoid this effect, the oracle answer regions would probably have to be designed longer than they currently are, but this might have an adverse effect on Q-A detection training and testing accuracy. Another side effect of the shorter Q-A regions of the oracle method is that there are some (albeit few) cases where the MMR ranking module misses a Q-A region because the oracle Q-A region does not include the current MMR-selected sentence in its answer-part; this sentence is part of an extended answer region, which is in fact detected by the automatic answer detection module.

We also want to note some mostly genre-specific phenomena, which pose problems for the Q-A detection component:

- In CallHome, we sometimes encounter quoted questions, as in "A: he said: do you like it?" — "A: and i said: yes". The answer detection module fails here since the answer is provided by the speaker posing the (quoted) question.
- A similar case, also mostly in CallHome, is self-answered questions, such as "A: what is my plan?" — "A: to graduate next spring".
- In CrossFire, we sometimes encounter questions with anaphoric reference to larger parts of the discourse, where the linking to their answers helps little for the local summary coherence (e.g., "would you accept that?", "do you agree to this?").

Future work in this area includes the following:

- Improving the Q-A detection components by incorporating additional features into the decision tree (such as distance from speaker change, preceding hypothesized speech acts, prosodic information), and optimizing the length of the target answer span.
- Making the algorithm for selecting Q-A regions more sensitive to the overall region relevance (as opposed to the maximum sentence relevance).
- Learning heuristics to decide when the linking of Q-A regions helps both coherence and summary accuracy, and when it is better avoided.
- Combining different methods for creating local coherence in a sensible way, e.g., using trainable weights. For instance, one could combine weights for *lead* (segment initial region) with weights for Q-A regions (emphasizing the dialogical nature of a text).
- Exploring the effect of relative and absolute summary size on the benefit of Q-A detection.

10 Conclusion

We have presented and evaluated an approach to automatically detect questions and their answers in spoken dialogues in unrestricted domains. The output of this component is used to aid a MMR summarizer to create summaries with increased local coherence. Automatic evaluations as well as a user study show that the goal of increasing summary fluency can be achieved, without compromising summary accuracy.

Acknowledgements

The authors wish to thank Klaus Ries for providing the automatic speech act tagger, the text annotators who made the experiments possible, and the participants of the user study. We are also grateful for the detailed comments by the anonymous reviewers of this paper.

This research was funded in part by ATR — Interpreting Telecommunications Research Laboratories of Japan and the US Department of Defense.

References

Branimir Boguraev and Christopher Kennedy. 1997. Salience-based characterisation of text documents. In *ACL/EACL-97 Workshop on Intelligent and Scalable Text Summarization, Madrid, Spain.*

Ronald Brandow, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia.*

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of the ICASSP-92*, volume 1, pages 517–520.

Julia Hirschberg, Steve Whittaker, Don Hindle, Fernando Pereira, and Amit Singhal. 1999. Finding information in audio: A new paradigm for audio browsing/retrieval. In *Proceedings of the ESCA workshop: Accessing information in spoken audio*, pages 117–122. Cambridge, UK, April.

Frances C. Johnson, Chris D. Paice, William J. Black, and A.P. Neal. 1993. The application of linguistic processing to automatic abstract generation. *Journal of Document and Text Management*, 1(3):215–241.

Megumi Kameyama, Goh Kawai, and Isao Arima. 1996. A real-time system for summarizing human-human spontaneous spoken dialogues. In *Proceedings of the ICSLP-96*, pages 681–684.

Linguistic Data Consortium. LDC. 1996. CallHome and CallFriend LVCSR databases.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

Mandar Mitra, Amit Singhal, and Christ Buckley. 1997. Automatic text summarization by paragraph extraction. In *ACL/EACL-97 Workshop on Intelligent and Scalable Text Summarization, Madrid, Spain.*

J. Ross Quinlan. 1992. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Norbert Reithinger, Michael Kipp, Ralf Engel, and Jan Alexandersson. 2000. Summarizing multilingual spoken negotiation dialogues. In *Proceedings of the 38th Conference of the Association for Computational Linguistics, Hongkong, China, October*, pages 310–317.

Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. 1997. Automatic text structuring and summarization. *Information Processing and Management*, 33(2):193–207.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, September.

Robin Valenza, Tony Robinson, Marianne Hickey, and Roger Tucker. 1999. Summarisation of spoken audio through information extraction. In *Proceedings of the ESCA workshop: Accessing information in spoken audio*, pages 111–116. Cambridge, UK, April.

Alex Waibel, Michael Bett, Florian Metze, Klaus Ries, Thomas Schaaf, Tanja Schultz, Hagen Soltau, Hua Yu, and Klaus Zechner. 2001. Advances in automatic meeting record creation and access. In *Proceedings of ICASSP-2001, Salt Lake City, UT, May.*

Mark Wasson. 1998. Using leading text for news summaries: Evaluation results and implications for commercial summarization approaches. In *Proceedings of COLING/ACL-98, Montreal, Canada*, pages 1364–1368.

Klaus Zechner and Alex Waibel. 2000. DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of COLING-2000, Saarbrücken, Germany, July/August*, pages 968–974.

Klaus Zechner. forthcoming. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains.* Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.