Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains

Klaus Zechner
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
zechner@cs.cmu.edu

ABSTRACT

Automatic summarization of open domain spoken dialogues is a new research area. This paper introduces the task, the challenges involved, and presents an approach to obtain automatic extract summaries for multi-party dialogues of four different genres, without any restriction on domain. We address the following issues which are intrinsic to spoken dialogue summarization and typically can be ignored when summarizing written text such as newswire data: (i) detection and removal of speech disfluencies; (ii) detection and insertion of sentence boundaries; (iii) detection and linking of cross-speaker information units (question-answer pairs). A global system evaluation using a corpus of 23 relevance annotated dialogues containing 80 topical segments shows that for the two more informal genres, our summarization system using dialogue specific components significantly outperforms a baseline using TFIDF term weighting with maximum marginal relevance ranking (MMR).

Keywords

summarization, spoken dialogue summarization

1. INTRODUCTION

While the field of summarizing written texts has been explored for many decades, gaining significantly increased attention in the last five to ten years, summarization of spoken language is a comparatively recent research area. As the amount of spoken audio databases is growing rapidly, however, we predict that the need for high quality summarization of information contained in this medium will rise substantially. Summarization of spoken dialogues, in particular, may aid the archiving, indexing, and retrieval of various records of oral communication, such as corporate meetings, sales interactions, or customer support conversations. The purpose of this paper is to present the main issues intrinsic to spoken dialogue summarization and to describe and evaluate an implementation which addresses these issues. In particular, we will show that while a baseline system using a state-of-the-art written text summarization technique (MMR) can generate good summaries, an enhanced system, using three additional components aimed at the core challenges of spoken dialogue summarization, can improve summarization accuracy significantly for two genres of informal conversations.

Intrinsic evaluations of text summaries typically use sentences as their basic units. For our data, however, sentence boundaries are not available in the first place. So we devise a word based evaluation metric based on an average relevance score from human relevance annotations (section *5.4.3*).

The organization of this paper is as follows: Section 2 introduces and discusses the main challenges of spoken dialogue summarization, followed by a section on related work (section 3). Section 4 describes the corpus we use to develop and evaluate our system, along with the description of the corpus annotation. The dialogue summarization system and its components are described in detail in section 5, along with evaluations thereof. Section 6 presents the global evaluation of our approach, before we conclude the paper with a summary of our contributions and results, as well as future directions in this field (section 7).

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.

2. MAIN CHALLENGES

In addition to the numerous challenges of written text summarization, work on spoken dialogue summarization has to address at least the following additional issues:

- coping with speech disfluencies
- identifying the units for extraction
- maintaining cross-speaker coherence
- coping with speech recognition errors

In the following, we shall discuss the nature of these challenges and indicate which approaches we take in our summarization system to address them. For the scope of this paper, we will only focus on the first three challenges and abstract away from the issue of speech recognition errors, which we addressed in previous work [34]. Thus, we exclusively use manually created dialogue transcriptions as the input for our spoken dialogue system in the context of this paper.

2.1 Disfluency detection

The two main negative effects speech disfluencies have on summarization are that they (i) decrease the readability of the summary and (ii) increase its non-content noise. In particular for informal conversations, the percentage of disfluent words is quite high, typically around 15-25% of the total words spoken. An example of a highly disfluent sentence, where the removal of disfluencies would enhance readability and conciseness of a summary, is given here:

```
A : well I um I think we should
    discuss this you know with her
A': I think we should discuss this with her
```

In section 5.2 we shall present a multi-stage approach for identifying the major classes of speech disfluencies in the input of the summarization system, such as filled pauses, repetitions, and false starts.

2.2 Sentence boundary detection

Unlike written texts, where punctuation or hypertext markers indicate sentence boundaries, spoken language is generated as a sequence of streams of words, where pauses (silences between words) do not always match linguistically meaningful segments: a speaker can pause in the middle of a sentence or even a phrase, or, on the other hand, might not pause at all after the end of a sentence or a clause. If an audio stream is segmented into smaller units (e.g., *speaker turns*[1]) by means of using a silence heuristic, one speaker's turn may contain multiple sentences, or, on the other hand, a speaker's sentence might span more than one turn, as demonstrated in the following example:

```
1 A: That's true / I suggest
2 A: you talk to him /
```

The main problem for a summarizer would thus be (i) the lack of coherence and readability of the output because of incomplete sentences and (ii) extraneous information due to extracted units consisting of more than one sentence. In section *5.2.2* we describe a component for sentence segmentation which addresses both of these problems.

2.3 Distributed information

Since we have multi-party conversations as opposed to monologues, sometimes the crucial information is found in a sequence of sentences from several speakers — the prototypical case being a question-answer pair. If the summarizer were to extract only the question or only the answer, the lack of the corresponding answer or question would often cause a severe reduction of coherence in the summary. In some cases, either the question or the answer is very short and does not contain any words with high relevance, resulting in a very small relevance weight within an automatic summarizer, e.g.:

```
A: Are you inviting all of your friends?
B: Yes.
```

In order not to lose these short sentences at a later stage, when only the most relevant sentences are extracted, we need to identify matching question-answer pairs ahead of time, so that the summarizer can output these matching pairs during summary generation. We describe our approach to cross-speaker information linking in section 5.3.[2]

[1] A speaker turn is a contiguous part of the dialogue where one speaker is active.

2.4 Other issues

We see the work reported in this paper as the first in depth analysis and evaluation in the area of open domain spoken dialogue summarization. Given the large scope of this undertaking, we had to restrict ourselves to the aforementioned issues which are, in our opinion, the most salient for the task at hand. A number of other important issues for summarization in general and for speech summarization in particular are either simplified or not addressed in this paper and left for future work in this field. These issues would include topic segmentation (implemented, but not relevant in the context of this paper), handling of speech recognition errors (see our related work in [34]), integration of more prosodic information, resolution of anaphora, and automatic analysis of the discourse structure.

3. RELATED WORK

The vast majority of summarization research in the past clearly has been focusing exclusively on written text. A good selection of both early seminal papers and more recent work can be found in [15]. Two areas are exceptions to this general trend and we will briefly discuss them in the following: (i) summarization of task oriented dialogues in restricted domains, and (ii) summarization of spoken news in unrestricted domains.

In the context of task-oriented natural language understanding systems, several spoken dialogue summarization systems were developed, whose goal it was to capture the essence of the task based dialogues at hand. The MIMI System [9] dealt with the travel reservation domain and used a cascade of finite state pattern recognizers to find the desired information. Within the VERBMOBIL project [27], a more knowledge-rich approach was used for summarization [20]. In addition to finite state transducers for content extraction and statistical dialogue act recognition, there also is a dialogue processor and a summary generator which have access to a world knowledge database, a domain model, and a semantic database.

Within the context of the TREC spoken document retrieval conferences [3], as well as the recent DARPA Broadcast News workshops, a number of research groups have been developing multi-media browsing tools for text, audio, and video data, which should facilitate the access to news data, combining different modalities [26, 30, 28]. Other applications are considered by [7] who describe an approach to summarize closed captions to about 60-70% of the original text length in Japanese broadcast news, and by [10] who use summarization techniques to convert voicemail messages, transcribed by a speech recognizer, to the Short Message (SM) format (about 160 ASCII characters in length).

This paper is closer related to the news (or voicemail) summarization task, since we are not restricting the domain which makes the use of world knowledge prohibitively expensive. But unlike news data, the corpus we use for this system is *dialogical* in nature, and some of its genres are

[2] A more detailed discussion of this cross-speaker information linking component in the context of our spoken dialogue summarizer, along with various additional evaluations thereof, can be found in [32].

also rather informal in style. This clearly sets apart our focus from the previous related work in the broader area of spoken language summarization.

Related work to the dialogue specific components of our system can be found in [6, 24] (disfluency and sentence boundary detection), and [23] (speech act detection). A preliminary dialogue summarization system, whose architecture inspired our current approach (but whose components are substantially different), is described in [33]. A much more detailed account on our system than can be provided within the limits of this paper will be given in [31].

4. DIALOGUE CORPUS

4.1 Corpus characteristics

Table 1 provides the statistics on the corpus used for the development and evaluation of our system. It comprises 23 dialogues with about 47,000 words total which corresponds to about four hours of recorded speech. We use data from four different genres, two being more informal, two more formal:

- English CallHome and CallFriend: from the Linguistic Data Consortium (LDC) collections [12], 8 dialogues for the `devtest`-set (8E-CH) and 4 dialogues for the `eval`-set (4E-CH).[3] These are recordings of phone conversations between two family members or friends, typically about 30 minutes in length; the excerpts we used were matched with the transcriptions which typically represent 5–10 minutes of speaking time.
- 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 in the Interactive Systems Labs (ISL) at CMU.

Furthermore, we used the Penn Treebank distribution of the SwitchBoard corpus, annotated with disfluencies, to train the major components of the system [13].

From Table 1 we can see that the two more formal corpora, NewsHour and CrossFire, have longer sentences, more sentences per turn, and fewer disfluencies than English CallHome and the Group Meetings. This means that their flavor is more like that of written text, and not so close to conversational speech typically found in the SwitchBoard or CallHome corpora.

4.2 Corpus annotation

4.2.1 Topical boundaries and relevant spans

All the annotations are performed on human generated transcriptions of the dialogues. There were six human annotators performing the corpus annotation; of those, four completed the entire set of dialogues.

Prior to the relevance annotations, the annotators had to mark topical boundaries, because we want to be able to define and then create summaries for each topical segment separately (as opposed to a whole conversation consisting of multiple topics). For each topical segment, each annotator had to identify the most relevant information units (IUs), called *nucleus IUs*, and somewhat relevant IUs, called *satellite IUs*. IUs are often equivalent to sentences, but can span longer or shorter contiguous segments of text, dependent on the annotator's choice. The overall goal of this relevance mark-up was to create a concise and readable summary, containing the main information present in the topical segment. We also asked that the human annotators stay within a preset target length for their summaries (about 10–20% of the length of a topical segment).

After the first annotation phase, where each coder worked independently, we devised a second phase, in which two coders (from the initial group) were asked to create a common ground annotation, based on the majority opinion of the whole group.

To calculate inter-coder agreement, we only used the annotations of those four annotators who completed the entire corpus, and computed F_1-scores[4] of matching annotations. For topical boundaries, a *match* means that the two compared boundaries have to fall within a close window of text (+-3 turns). For relevance annotations, a *match* means that a word was considered relevant by both annotators. We then averaged the scores for all six annotator pairs. The results were $F_1 = .45$ for topical boundaries and $F_1 = .36$ for relevance annotations.

4.2.2 Disfluency annotation

In addition to the annotation for topic boundaries and relevant text spans, the corpus was also annotated for speech disfluencies in the same style as the Penn Treebank SwitchBoard corpus [13]. One coder manually tagged the corpus for disfluencies, sentence boundaries, as well as question speech-acts (and their corresponding answers), following the SwitchBoard disfluency annotation style book [17].

5. SUMMARIZATION SYSTEM

5.1 System overview

The global system architecture of the spoken dialogue summarization system presented in this paper is a sequence of the following components:

1. Part-of-speech tagger
2. Sentence boundary detection
3. False start detection
4. Question and answer detection
5. Repetition filter
6. Topic boundary detection
7. Sentence ranking and selection

The input data for the system is a time ordered sequence of speaker turns with the following quadruple of information: start time, end time, speaker label, and word sequence.

[3] We used the `devtest`-set corpus for system development and tuning, and set aside the `eval`-set for the final global system evaluation. For the other three genres, two dialogues each were used for the `devtest`-set, the remainder for the `eval`-set.

[4] $F_1 = \frac{2PR}{P+R}$ with P=precision and R=recall.

Table 1: Data characteristics for the corpus (average over dialogues).

data set	8E-CH	4E-CH	NHOUR	XFIRE	G-MTG
formal/informal	informal	informal	formal	formal	informal
topics pre-determined?	no	no	yes	yes	yes
dialogues (total)	8	4	3	4	4
topical segments (total)	28	23	8	14	7
different speakers	2.1	2	2	6	7.5
turns	242	276	25	96	140
sentences	280	366	101	281	304
sentences per turn	1.2	1.3	4.1	2.9	2.2
questions (in %)	3.7	6.4	6.3	9.8	4.0
words	1685	1905	1224	3165	2360
words per sentence	6.0	5.2	12.1	11.3	7.8
disfluent (in %)	16.0	16.3	11.8	4.2	23.9
disfluencies	222	259	48	95	265
disfluencies per sentence	0.79	0.71	0.48	0.34	0.87

All human and non-human noises, as well as all incomplete words were eliminated from the input transcripts. Further, we eliminated all information on case and punctuation, since we want to emulate the typical speech recognizer output format in that regard which does not provide this information. Contractions such as `don't` or `I'll` are expanded and treated as two separate words — in these examples we would obtain: `do n't` and `I 'll`.

The summarization system is embedded in an architecture with a graphical user interface ("Meeting Browser") which enables the recording, archiving, summarizing, indexing, and retrieval of meetings of multiple participants in near real time [29].

The following subsections describe the components of the system in more detail. Since our evaluations are using the manually marked topical segments from the human gold standard, the topic segmentation component is not relevant for the work reported in this paper. (We implemented a variant of Hearst's TextTiling algorithm [5].) The three components involved in disfluency detection are the part of speech (POS) tagger, the false start detection module, and the repetition filter (1, 3, 5). They, together with the sentence boundary detection module, are discussed in subsection 5.2. The question-answer pair detection is described in subsection 5.3, and the sentence selection module, performing relevance ranking, is outlined in subsection 5.4.

5.2 Disfluency and sentence boundary detection

Conversational, informal spoken language is quite different from written language in that a speaker's utterances are typically much less well-formed than a writer's sentences. We can observe a set of disfluencies such as false starts, hesitations, repetitions, filled pauses, and interruptions. Additionally, in speech there is no good match between linguistically motivated sentence boundaries and turn boundaries or recognition hypotheses from automatic speech recognition. This component can be seen to serve these two main functions of text normalization: (i) disfluency detection, and (ii) sentence boundary detection.

We address the following types of disfluencies, following the classification in [17, 22, 21]. We also briefly indicate the method we use in our system to detect these disfluencies.

- Filled pauses: We devise special tags for these and detect them using a POS tagger. The two variants are (i) non-lexicalized filled pauses (typically *uh, um*), and (ii) lexicalized filled pauses (e.g., *like, you know*).
- Restarts or repairs: Those are fragments which are resumed, but without completely abandoning the first attempt. There are insertions, substitutions, and repetitions, with repetitions being by far the most frequent phenomenon (more than 60% of all restarts). We filter repetitions using a repetition detection algorithm.
- False starts: These are abandoned, incomplete clauses. In some cases, they may occur at the end of a turn, and they can be due to interruption by another speaker. Detection of false starts is accomplished by a trained decision tree.

For training, we use a part of the SWITCHBOARD transcriptions which were manually annotated for sentence boundaries, POS, and the following types of disfluent regions or words [13]:

C : empty coordinating conjunctions (they act as links between sentences of one speaker but are semantically empty, e.g. *and then*)

D : lexicalized filled pauses (e.g., *you know*)

E : editing terms (within repairs, e.g., *I mean*)

F : non-lexicalized filled pauses (e.g., *uh, um*)

5.2.1 POS Tagger

We trained Brill's rule based POS tagger [1] on the Penn Treebank SWITCHBOARD corpus with POS and disfluency annotations [13]. We replaced the tags in the regions of [C], [D], [E], and [F] with the tags CO, DM, ET, and UH, respectively. The entire tag set comprises 42 different POS tags. We trained the POS tagger in three phases, using three different parts of the corpus with about 250,000 word-tag pairs each. The final tagging accuracy on an unseen test set was 94.1%, compared to a baseline of 84.8%, where each word is tagged with its most likely tag. Non-lexicalized fillers are almost perfectly tagged ($F_1 = .98$), whereas the hardest task for the tagger are the empty coordinating conjunctions

($F_1 = .88$): there are a few highly ambiguous words in that set, such as *and, so, or.*

5.2.2 Sentence boundary detection

The purpose of this component is to insert linguistically meaningful sentence boundaries in the text, given a POS tagged input. We consider all intra-turn and inter-turn boundary positions with respect to a single speaker. The frequency of sentence boundaries in SWITCHBOARD (with respect to the total number of words) is about 13.3%. If we would mark all inter-word positions with a non-boundary, this would thus yield a baseline error rate of 13.3%.

We use Release 8 of the C4.5 decision tree distribution [18]. To encode the feature vectors, we apply a context of four words before and after a hypothesized sentence boundary, motivated by the results of [4]. The input features for every word position are: (i) POS tag, (ii) trigger word[5], (iii) "turn boundary before this word?", and (iv) length of pause after the last turn of the same speaker (zero if not a turn boundary).

For the training and test sets for this component, we used a previously unused portion of the Penn Treebank corpus (50000 words, 80% for training, 20% for testing). The best decision tree yielded a test set classification error rate of 3.6% (and $F_1 = .89$ for all boundary and non-boundary positions combined), compared to the baseline of 13.3%. We determined that for good performance we need to know about at least either one of these two features: "is there a turn boundary before this word?" (iii) or "pause duration after last turn from same speaker" (iv). When using imperfect POS tags, the performance drops only minimally (1.1% relative). This shows that the decision tree is not very sensitive to the majority of POS tagger errors.

When comparing the performance of inter-turn with intra-turn boundary detection, we find, not unexpectedly, that for the two cases with higher frequency — the inter-turn boundaries and the intra-turn non-boundaries — the results are excellent ($F_1 > .94$), whereas for the two much rarer cases — inter-turn non-boundaries and intra-turn boundaries — the performance is around $F_1 \approx .65$.

5.2.3 Repetition detection

Repetition detection is done using an algorithm which identifies repetitions of word sequences of length 1 to 4 (longer repetitions are extremely rare [22]). Words which were marked as disfluent by the POS tagger are ignored when considering the repeated sequences. With just this simple module, we are able to detect and correct about 65% of all repairs in the SWITCHBOARD database, since the non-repetition types of repairs are comparatively rare (substitutions and insertions).

5.2.4 False start detection

False starts are quite frequent in spontaneous speech, occurring at a rate of about 10-15% of all sentences both in SWITCHBOARD and CALLHOME. They involve less than 10% of the total words of a dialogue; about 34% of the words in these incomplete sentences are part of some other disfluencies, such as filled pauses or repairs. (In complete sentences, only about 15% of the words are part of some disfluencies.) For CALLHOME, the average length of complete sentences is about 6 words, of incomplete sentences about 4.1 words (including disfluencies).

We trained a C4.5 decision tree on 8000 sentences of the SWITCHBOARD Penn Treebank. As features we use the first and last four trigger words and POS of every sentence, as well as the first and last four chunks from a POS based chunk parser. This chunk parser is based on a simple context-free POS grammar for English and uses a heuristic for maximal coverage of the input for ambiguity resolution. Its output are common phrases such as noun phrases or prepositional phrases. Further, we encode the length of the sentence in words and the number of the words not parsed by the chunk parser. We observed that while the chunk information itself does not improve performance over the baseline of using trigger words and POS information only, the derived feature of "number of not parsed words" actually does improve the results.

The evaluations were performed on an independent test set of about 3000 sentences. False start detection accuracy was $F_1 = .61$, non-false start detection accuracy was $F_1 = .96$. (Note that the latter case is much more frequent than the former and hence easier to learn.)

[5]Words which discriminate well between boundary and non-boundary positions. See [4] for the method of computing these trigger words.

Table 2: Frequency of different types of questions in the 8E-CH data set.

turns total	2211
Wh-questions total	20
... with immediate answers	15 (75%)
yes-no-questions total	48
... with immediate answers	38 (79%)
questions without answers	2
rhetorical and back-channel questions	13
questions total	83 (3.75%)

5.3 Cross-speaker information linking

One of the properties of multi-party dialogues is that shared information is created between dialogue participants. The most obvious interactions of this kind are question-answer pairs. The purpose of this component is to automatically create such coherent pieces of relevant information which can then be extracted together while generating the summary. The question-answer linking (Q-A-linking) task consists of the following two intuitive sub-tasks: (i) identifying questions; (ii) finding their corresponding answers.

5.3.1 Automatic question detection

We trained a decision tree classifier (C4.5) using about 20,000 manually annotated SWITCHBOARD speech acts[6] [8]. We encoded the following set of features: (i) POS and trigger word information for the first and last five tokens of each speech act, (ii) speech act length, and (iii) occurrence of POS bigrams which are highly discriminative between question and non-question speech acts. We evaluated the decision tree classifier on the 8E-CH data set (see Table 2); the classification score was $F_1 = .56$, compared to a baseline of $F_1 = .07$.

[6]In the context of this paper, speech acts correspond to sentences.

5.3.2 *Detecting the answers*

After identifying which sentences are questions, the next step is to identify their answers. From the 8E-CH-statistics of Table 2 we observe that for more than 75% of the yes-no-questions 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. We devised a search heuristic to detect answers, using the following features:

1. maximum distance of the first speaker change following the question
2. number of sentences to be included in the answer hypothesis (they have to be in a single speaker region)
3. minimum word length of answers
4. matching words between questions and answers

The search heuristic further handles embedded questions where the speaker expected to be answering a question is posing a question himself, as shown in this example (a1 is the answer to q1, a2 the answer to q2):

```
q1  A: When are we meeting?
 q2  B: You mean tomorrow?
 a2  A: Yes.
a1  B: At 4 p.m.
```

We optimized the parameters on the 68 yes-no-questions and Wh-questions of the 8E-CH data set, excluding rhetorical and back-channel questions and questions without answers. The best result for the question-answer pair detection task, using questions detected by the decision tree, was $F_1 = .51$.

5.4 Sentence ranking and selection

This component's purpose is to determine weights for terms and sentences, to rank the sentences according to their relevance within each topical segment of the dialogue, and finally to select the sentences for the summary output according to their rank, as well as to other criteria, such as question-answer linkages, established by previous components.

5.4.1 *Term and sentence weighting*

To determine the most relevant sentences within a topical segment, we use a TFIDF based version of the maximum marginal relevance algorithm (MMR) [2] which maximizes salience (cosine similarity between query q and sentence word vector s_{nr}) and minimizes redundancy (avoiding to select sentences with similar keywords to previously ranked sentences s_r). The algorithm is stated as an iterative formula:

$$\bar{s} = \arg\max_{s_{nr}}(\lambda sim_1(q, s_{nr}) - (1-\lambda)\max_{s_r} sim_2(s_{nr}, s_r)) \tag{1}$$

As query vector we use the vector of terms within the current topical segment. Terms are stemmed words not contained in a stop list of common words. The trainable λ-parameter ($0.0 \leq \lambda \leq 1.0$) is used to trade off the influence of salience vs. redundancy.

5.4.2 *Q-A linking*

While generating the output summary from the MMR-ranked list of sentences, whenever a question or an answer is encountered (detected before by the Q-A detection module), the corresponding answer (or: question) is linked to it and moved up the relevance ranking list to immediately follow the current question (answer). If the question-answer pair consists of more than two sentences, the linkages are repeated until no further questions or answers can be added to the current linkage cluster.

5.4.3 *Evaluation metric*

As other studies have shown [16, 19], the agreement between human annotators about which passages to choose to form a good summary is usually quite low. We also found this in our data (section *4.2.1*). We decided to minimize the bias that would result from selecting either a particular human annotator, or even the manually created gold standard as a reference for automatic evaluation, but instead weigh all annotations from all human coders equally. Intuitively, we want to reward summaries which contain a high amount of words considered to be relevant by most annotators.

Another consideration is related to the fact that we do not have a priori given sentence boundaries in our corpus. Thus, unlike for most text based evaluations which operate on the sentence level, we decided to use a word-based evaluation metric to be able to capture even subtle differences in the summaries. This has the further advantage of being easily extensible to evaluations of imperfect transcriptions generated by automatic speech recognizers [34].

All evaluations are based on topically coherent segments from the dialogues of our corpus. (The segment boundaries are chosen from the human gold standard.) For each 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 nucleus-IU or a satellite-IU 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:

$$r_{s,i} = \frac{\sum_{a=1}^{A} w_{s,a,i}}{A} \tag{2}$$

To obtain the summary accuracy score $sa_{s,N}$ for any segment summary with length N (automatically generated or produced by a human annotator), we multiply the boolean summary vector summ_s[7] 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:

$$sa_{s,N} = \frac{\mathrm{summ}_s r_s}{\sum_{i=1}^{N} \mathrm{rsort}_{s,i}} \tag{3}$$

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

5.4.4 *System tuning*

To arrive at an optimal parameter setting for each subcorpus of our four different genres (CALLHOME, NEWSHOUR, CROSSFIRE, GROUP MEETINGS), we first established a tuned

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

MMR-baseline. This we can then use for the global system evaluations, where we compare the baseline performance to the results of the entire system. Note that for this baseline tuning, we did not make use of any other system component, namely disfluency detection, sentence boundary detection, and question-answer linking. We only used the `devtest` sets for the 4 sub-corpora here: 8E-CH, DEVTEST-NH, DEVTEST-XF, and DEVTEST-MTG.

The tuning proceeded in three phases, where we optimized the TFIDF term weighting parameters, the MMR-λ, as well as a LEAD-emphasis-parameter. (The effect of this parameter is to increase the scores of turns within the first N% of a topical segment.) The results of this baseline tuning procedure are shown in the MMR column of Table 3.

Table 3: Average summary accuracy scores. `devtest`-set and `eval`-set sub-corpora on optimized parameters, comparing LEAD, MMR baseline, complete system, and the human gold standard (with nucleus-IUs).

sub-corpus	LEAD	MMR	complete	gold std.
8E-CH	0.463	0.545	0.597	0.709
DEVTEST-NH	0.391	0.617	0.516	0.744
DEVTEST-XF	0.516	0.595	0.541	0.764
DEVTEST-MTG	0.497	0.587	0.650	0.659
4E-CH	0.438	0.526	0.614	0.793
EVAL-NH	0.692	0.526	0.506	0.850
EVAL-XF	0.378	0.564	0.566	0.770
EVAL-MTG	0.324	0.449	0.583	0.704

6. EVALUATION

Traditionally, the evaluation of summarization systems has been performed in two major ways: (i) intrinsically, measuring the amount of the core information preserved from the original text [11, 25]; and (ii) extrinsically, measuring how much the summary can benefit in accomplishing another task, e.g., finding a document relevant to a query or classifying a document into a topical category [14]. In this paper, we focus on intrinsic evaluation exclusively. That is, we want to assess, how well the summaries preserve the essential information contained in the original text.

In this evaluation, we compare our complete system with a LEAD baseline and the MMR baseline system, which operates without any dialogue specific components, as described above (section *5.4.4*). For the complete system, using all the components except for the topic segmentation module, we used the optimized baseline MMR parameters and varied emphasis parameters for (i) false starts, (ii) lead factor, and (iii) Q-A sentences, to optimize the summaries. Again, we only used the `devtest`-set for this optimization. For each corpus in the `devtest`-set, we determined the optimal parameter setting and report the corresponding results also for the `eval`-set sub-corpora. Table 3 provides the comparison of the average scores for LEAD method (first N percent of the word tokens within a segment), baseline MMR, complete system, and the human gold standard (nucleus-IUs only[8]). Figure 1 provides a comparison of the four summary types for one topical segment from the CALLHOME `eval`-sub-corpus, at the same length of 14% of the original text.

We determined the statistical differences between the complete system and the two baselines (MMR and LEAD) for the `eval`-set, using the Wilcoxon rank sum test for each of the 4 sub-corpora. Comparisons were made for five summary sizes (5-25% length) within each topical segment. For the CALLHOME and GROUP MEETINGS sub-corpora, our system is significantly better than the MMR baseline ($p < 0.01$); for the two more formal sub-corpora, NEWSHOUR and CROSSFIRE, the difference is not significant. Except for the NEWSHOUR sub-corpus, both the MMR baseline and the complete system perform significantly better than the LEAD baseline ($p < 0.01$).

We can see two reasons why the complete dialogue summarization system does not outperform the MMR baseline for the more formal genres (NEWSHOUR, CROSSFIRE). First, the training corpus (SWITCHBOARD) is definitely more similar to CALLHOME and GROUP MEETINGS than it is to the more formal genres. The second reason is related to the observation we made in section 4.1, that the two more formal genres resemble written text more than the informal genres do (fewer disfluencies, longer and more complex sentences); our components added on top of the baseline MMR may be best suited for more informal, conversational, spontaneous types of dialogues.

[8]These gold standard summaries have a fixed length of about 15% of the original text.

7. CONCLUSION

The problem of how to automatically generate readable and concise summaries for spoken dialogues of unrestricted domains has many challenges that need to be addressed. Some of the research issues are similar or identical to those faced when summarizing written texts (such as topic segmentation, determining the most relevant information, anaphora resolution, summary evaluation), but other additional dimensions are added on top of this list, including speech disfluency detection, sentence boundary detection, cross-speaker information linking, and coping with imperfect speech recognition. The line of argument of this paper was that while using a traditional approach for written text summarization (such as the MMR based sentence selection component) may be a good starting point, addressing the dialogue specific issues is key for obtaining better summaries, particularly for informal dialogue genres.

Given the complexity of the task, we made a number of simplifying assumptions: (i) we only use perfect dialogue transcriptions by humans (and not output from automatic speech recognizers); (ii) we limit the use of prosodic information to start and end times of speaker turns; (iii) we only consider input which was pre-segmented into topically coherent regions; (iv) we limit ourselves to only one aspect of discourse related coherence, the question-answer pairs. Removing these limitations, as well as improving the system for more formal genres of spoken dialogues, can be seen as directions of future work in this area.

Our main contribution is that we have motivated, described, and evaluated an approach to automatically create extract summaries for open domain spoken dialogues in informal and formal genres of multi-party conversations. Our dialogue summarization system uses trainable components (i) to detect and remove speech disfluencies (making the output more readable and concise), (ii) to determine sen-

```
1- LEAD baseline:
   39  b :  Yeah well now get this we might go to live in switzerland
   40  a :  Oh really
   41  b :  Yeah because they've made him a job offer
             there and at first he's thinking nah he wasn't [...]

2- MMR baseline:
   40 b :  Yeah because they've made him a job offer there and at first
            he's thinking nah he wasn't going to take it but now he's like
   44 b :  And then you know the [...]

3- Complete system:
   56 b :  Now get this we might go to live in switzerland
   59 b :  They've made him a job offer there
   60 b :  At first he's thinking
   63 b :  Maybe he could get [...]
   65 b :  The swiss phone company whatever and telefonika

4- Human gold standard:
   39 b :  Might go to live in switzerland
   41 b :  They've made him a job offer there
   43 b :  Maybe he could get in his foot in the door with
            because they've united with a t and t
```

Figure 1: Four different summary versions for a topical segment in CallHome (all at 14% length): LEAD, MMR, complete system, human gold standard.

tence boundaries (creating suitable text spans for summary generation), and (iii) to link cross-speaker information units (allowing for increased summary coherence).

We used a corpus of 23 dialogues from four different genres (80 topical segments, about 47000 words total) for system development and evaluation and the disfluency annotated SwitchBoard corpus [13] for training of the three dialogue specific components. Our corpus was annotated by six human coders for topical boundaries and relevant text spans for summaries. Additionally, we had annotations made for disfluencies, question speech acts, and their corresponding answers.

In a global system evaluation we compared a LEAD baseline and a MMR baseline with the complete system using all of its components discussed in this paper. The results showed that (i) both the baseline MMR system as well as the complete system create better summaries than the LEAD baseline (except for NewsHour), and that (ii) the complete system performs significantly better than the baseline MMR system for the informal dialogue corpora (CallHome and Group Meetings).

8. ACKNOWLEDGMENTS

We are grateful to Alex Waibel, Alon Lavie, Jaime Carbonell, Vibhu Mittal, and Jade Goldstein for many discussions, suggestions, and comments regarding this work. We also thank the anonymous reviewers of this paper for their comments, and are indebted to the corpus annotators for their important work. Finally, we give credit to the people from the Interactive Systems Laboratories (ISL) at Carnegie Mellon University for providing a supportive working environment.

The research reported here was supported in part by grants from the US Department of Defense.

9. REFERENCES

[1] E. Brill. Some advances in transformation-based part of speech tagging. In *Proceeedings of AAAI-94*, 1994.

[2] J. Carbonell and J. Goldstein. 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*, 1998.

[3] J. S. Garofolo, E. M. Voorhees, C. G. P. Auzanne, and V. M. Stanford. Spoken document retrieval: 1998 evaluation and investigation of new metrics. In *Proceedings of the ESCA workshop: Accessing information in spoken audio*, pages 1–7. Cambridge, UK, Apr. 1999.

[4] M. Gavaldà, K. Zechner, and G. Aist. High performance segmentation of spontaneous speech using part of speech and trigger word information. In *Proceedings of the 5th ANLP Conference, Washington DC*, pages 12–15, 1997.

[5] M. A. Hearst. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, March 1997.

[6] P. A. Heeman and J. F. Allen. Intonational boundaries, speech repairs and discourse markers: Modeling spoken dialog. In *Proceedings of the ACL/EACL-97, Madrid, Spain*, pages 254–261, 1997.

[7] C. Hori and S. Furui. Improvements in automatic speech summarization and evaluation methods. In *Proceedings of ICSLP-00, Beijing, China, October*, pages 326–329, 2000.

[8] D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. V. Ess-Dykema. SwitchBoard discourse

language modeling project, final report. Research Note 30, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 1998.

[9] M. Kameyama, G. Kawai, and I. Arima. A real-time system for summarizing human-human spontaneous spoken dialogues. In *Proceedings of the ICSLP-96*, pages 681–684, 1996.

[10] K. Koumpis and S. Renals. Transcription and summarization of voicemail speech. In *Proceedings of ICSLP-00, Beijing, China, October*, pages 688–91, 2000.

[11] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, pages 68–73, 1995.

[12] Linguistic Data Consortium. LDC. CallHome and CallFriend LVCSR databases, 1996.

[13] Linguistic Data Consortium. LDC. Treebank-3: CD-ROM containing databases of disfluency annotated Switchboard transcripts (LDC99T42), 1999.

[14] I. Mani, D. House, G. Klein, L. Hirschman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim. The TIPSTER SUMMAC text summarization evaluation. Mitre Technical Report MTR 98W0000138, October 1998, 1998.

[15] I. Mani and M. T. Maybury, editors. *Advances in automatic text summarization.* MIT Press, Cambridge, MA, 1999.

[16] D. Marcu. Discourse trees are good indicators of importance in text. In Mani and Maybury [15], pages 123–136.

[17] M. Meteer, A. Taylor, R. MacIntyre, and R. Iyer. Dysfluency annotation stylebook for the Switchboard corpus. Revised by Ann Taylor, June 1995, available on the LDC99T42 CD-ROM, published by LDC, 1995.

[18] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1992.

[19] G. J. Rath, A. Resnick, and T. R. Savage. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143, 1961.

[20] N. Reithinger, M. Kipp, R. Engel, and J. Alexandersson. Summarizing multilingual spoken negotiation dialogues. In *Proceedings of the 38th Conference of the Association for Computational Linguistics, Hongkong, China, October*, pages 310–317, 2000.

[21] R. L. Rose. *The communicative value of filled pauses in spontaneous speech.* PhD thesis, University of Birmingham, Birmingham, UK, 1998.

[22] E. E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies.* PhD thesis, University of Berkeley, Berkeley, CA, 1994.

[23] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, September 2000.

[24] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proceedings of the ICSLP-98, Sydney, Australia, December*, volume 5, pages 2247–2250, 1998.

[25] S. Teufel and M. Moens. Sentence extraction as a classification task. In *ACL/EACL-97 Workshop on Intelligent and Scalable Text Summarization, Madrid, Spain*, 1997.

[26] R. Valenza, T. Robinson, M. Hickey, and R. Tucker. Summarisation of spoken audio through information extraction. In *Proceedings of the ESCA workshop: Accessing information in spoken audio*, pages 111–116. Cambridge, UK, Apr. 1999.

[27] W. Wahlster. Verbmobil — translation of face-to-face dialogs. In *Proceedings of MT Summit IV, Kobe, Japan*, 1993.

[28] A. Waibel, M. Bett, and M. Finke. Meeting browser: Tracking and summarizing meetings. In *Proceedings of the DARPA Broadcast News Workshop*, 1998.

[29] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. Advances in automatic meeting record creation and access. In *Proceedings of ICASSP-2001, Salt Lake City, UT, May*, 2001.

[30] S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of the 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval, Berkeley, CA, August*, pages 26–33, 1999.

[31] K. Zechner. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains.* PhD thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, forthcoming.

[32] K. Zechner and A. Lavie. Increasing the coherence of spoken dialogue summaries by cross-speaker information linking. In *Proceedings of the NAACL-01 Workshop on Automatic Summarization, Pittsburgh, PA, June*, 2001.

[33] K. Zechner and A. Waibel. DiaSumm: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of COLING-2000, Saarbrücken, Germany, July/August*, pages 968–974, 2000.

[34] K. Zechner and A. Waibel. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics, NAACL-2000, Seattle, WA, April/May*, pages 186–193, 2000.