DiaSumm: Flexible Summarization of Spontaneous Dialogues in Unrestricted Domains

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

Abstract

In this paper, we present a summarization system for spontaneous dialogues which consists of a novel multi-stage architecture. It is specifically aimed at addressing issues related to the nature of the texts being spoken vs. written and being dialogical vs. monological. The system is embedded in a graphical user interface and was developed and tested on transcripts of recorded telephone conversations in English and Spanish (Callhome).

1 Introduction

Summarization of written documents has recently been a focus for much research in NLP (e.g., (Mani and Maybury, 1997; AAAI, 1998; Mani et al., 1998; ACL, 2000), to name some of the major events in this field in the past few years). However, very little attention has been given so far to the summarization of *spoken language*, even less of *conversations* vs. monological texts. We believe that summarization of speech will become increasingly more important, as the amount of online audio data grows and demand for rapid browsing, skimming, and access of speech data increases. Another application which particularly pertains to our interest in spoken *dialogue* summarization would be the generation of meeting minutes for archival purposes and/or to update participants joining at later stages on the progress of the conversation so far.

Summarization of dialogues within *limited domains* has been attempted within the context of the Verbmobil project ("protocol generation", (Alexandersson and Poller, 1998)) or by SRI's MIMI summarizer (Kameyama et al., 1996). Recent work on spoken language summarization in unrestricted domains has focused almost exclusively on Broadcast News, mostly due to the spoken language track of recent TREC evaluations (Garofolo et al., 1997; Garofolo et al., 1999). (Waibel et al., 1998) describe a Meeting Browser where summaries can be generated using technology established for written texts. (Valenza et al., 1999) go one step further and incorporate knowledge from the speech recognizer (confidence scores) into their summarization system, as well.

We argue that the nature of spoken dialogues, together with their textual representations as speech recognizer hypotheses, requires a set of specific approaches to make summarization feasible for this text genre.

As a demonstrable proof of concept, we present the multi-stage architecture of the summarization system DiaSumm which can flexibly deal with spoken dialogues in English and Spanish, without any restrictions of domain. Since it cannot rely on any domain specific knowledge base, it uses shallow statistical approaches and presents (possibly modified) *extracts* from the original text as summary.

We present results of several evaluations of our system using human transcripts of spontaneous telephone conversations in English and Spanish from the Callhome corpus ((LDC), 1996), in particular the accuracy of the topic segmentation and information condensing components (sections 6 and 7). Also, for the purpose of a global evaluation, a user study was performed which addressed information access time and accuracy of retained information comparing different versions of summaries (section 10).

This paper is organized as follows: In the next section, we provide an overview about the main issues for summarization of spoken dialogues and indicate the approaches we are taking in our system. We then present the system architecture (section 3), followed by a detailed description of the major building blocks (sections 4 to 8). After a brief characterization of the GUI (section 9) we describe a user study for global system evaluation in section 10. We conclude the paper with a summary and a brief outlook in section 11.

2 Issues and Approaches: Overview

In this section, we give an overview about the main issues that any summarization system for spoken dialogues has to address and indicate the approach we are taking for each of these in DiaSumm.

In a general sense, when dealing with written texts, usually there is plenty of information available which can be used for the purpose of summarization, such as capitalization, punctuation marks, titles, passage headers, paragraph boundaries, or other mark-ups. Unfortunately, however, *none* of this holds for speech data which arrives as a stream of word tokens from a recognizer, cut into "utterances" by using a silence heuristic.

2.1 Lack of clause boundaries

One of the most serious issues is the lack of sentence or clause boundaries in spoken dialogues which is particularly problematic since sentences, clauses, or paragraphs are considered the "minimal units" in virtually all existing summarization systems. When humans speak, they sometimes pause *during* a clause, and not always at the end of a clause, which means that the output of a recognizer (which usually uses some silence-heuristics to cut the segments) frequently does *not* match logical sentence or clause boundaries. Looking at five English CALLHOME dialogues with an average number of 320 utterances each, we find on average 30 such "continuations" of logical clauses over automatically determined acoustic segment boundaries. In a summary, this can cause a reduction in coherence and readability of the output.

We address this issue by linking adjacent turns of the same speaker together if the silence between them is less than a given constant (section 4).

2.2 Distributed information

Since we have multi-party conversations as opposed to monological texts, sometimes the crucial information is found in a question-answer-pair, i.e., it involves more than one speaker; extracting only the question or only the answer would be meaningless in many cases. We found that on average about 10% of the speaker turns belong to such question-answer pairs in five examined English CALLHOME dialogues. Often, either the question or the answer is very short and does not contain any words with high relevance. In order not to "lose" these short turns at a later stage, when only the most relevant turns are extracted, we link them to the matching question/answer ahead of time, using two different methods to detect questions and their answers (section 4).

2.3 Disfluent speech

Speech disfluencies in spontaneous conversations — such as fillers, repetitions, repairs, or unfinished clauses — can make transcripts (and summary extracts) quite hard to read and also introduce an unwanted bias to relevance computations (e.g., word repetitions would cause a higher word count for the repeated content words; words in unfinished clauses would be included in the word count.)

To alleviate this problem, we employ a clean-up filter pipeline, which eliminates filler words and repetitions, and segments the turns into short clauses (section 5). We also remove incomplete clauses, typically sentence-initial repairs, at this stage of our system. This "cleaning-up" serves two main purposes: (i) it increases the readability (for the finally extracted segments); and (ii) it makes the text more tractable by subsequent modules.

The following example compares a turn before and after the clean-up component:

```
before: I MEAN WE LOSE WE LOSE I CAN'T I
        CAN'T DO ANYTHING ABOUT IT SO
 after: we lose / i can't do anything
        about it
```

2.4 Lack of topic boundaries

CALLHOME speech data is multi-topical but does not include mark-up for paragraphs, nor any topic-informative headers. Typically, we find about 5–10 different topics within a 10-minute segment of a dialogue, i.e., the topic changes about every 1–2 minutes in these conversations. To facilitate browsing and summarization, we thus have to discover topically coherent segments automatically. This is done using a TextTiling approach, adapted from (Hearst, 1997) (section 6).

2.5 Speech recognizer errors

Last but not least, we face the problem of imperfect word accuracy of speech recognizers, particularly when dealing with spontaneous speech over a large vocabulary and over a low bandwidth channel, such as the CALLHOME databases which we mainly used for development, testing, and evaluation of our system. Current recognizers typically exhibit word error rates for these corpora in the order of 50%. In DIASUMM's information condensation component, the relevance weights of speaker turns can be adjusted to take into account their word confidence scores from the speech recognizer. That way we can reduce the likelihood of extracting passages with a larger amount of word misrecognitions (Zechner and Waibel, 2000). In this paper, however, the focus will be exclusively on results of our evaluations on human generated transcripts. No information from the speech recognizer nor from the acoustic signal (other than inter-utterance pause durations) are used. We are aware that in particular prosodic information may be of help for tasks such as the detection of sentence boundaries, speech acts, or topic boundaries (Hirschberg and Nakatani, 1998; Shriberg et al., 1998; Stolcke et al., 2000), but the investigation of the integration of this additional source of information is beyond the scope of this paper and left for future work.

3 System Architecture

The global system architecture of DIASUMM is a pipeline of the following four major components:

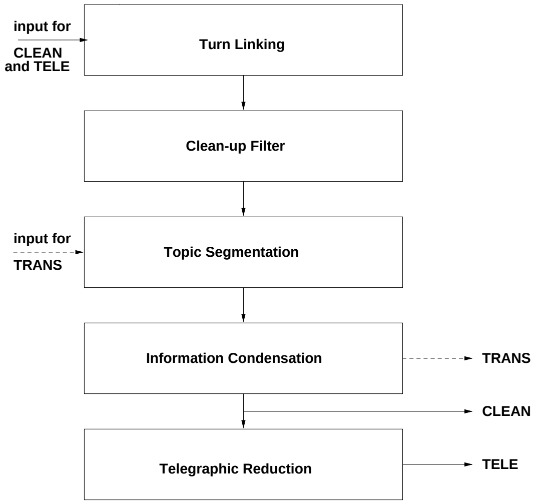

Figure 1: System architecture

turn linking; clean-up filter; topic segmentation; and information condensation. A fifth component is added at the end for the purpose of telegraphic reduction, so that we can maximize the information content in a given amount of space. The system architecture is shown in Figure 1. It also indicates the three major types of summaries which can be generated by DiaSumm: TRANS ("transcript"): not using the linking and clean-up components; CLEAN: using the main four components; TELE ("telegraphic" summary): additionally, using the telegraphic reduction component.

The following sections describe the components of DiaSumm in more detail.

4 Turn Linking

The two main objectives of this component are: (i) to form turns which contain a set of full (and *not* partial) clauses; and (ii) to form *turn-pairs* in cases where we have a question-answer pair in the dialogue.

To achieve the first objective, we scan the input for adjacent turns of one speaker and link them together if their time-stamp distance is below a pre-specified threshold θ. If the threshold is too small, we don't get most of the (logical) turn continuations across utterance boundaries, if it is too large, we run the risk of "skipping" over short but potentially relevant fragments of the speaker on the other channel. We experimented with thresholds between 0.0 and 2.0 seconds and determined a local performance maximum around $\theta = 1.0$.

For the second objective, to form turn-pairs which comprise a question-answer information exchange between two dialogue participants, we need to detect wh- and yes-no-questions in the dialogue. We tested two approaches: (a) a HMM based speech act (SA) classifier (Ries, 1999) and (b) a set of part-of-speech (POS) based rules. The SA classifier was trained on dialogues which were manually annotated for speech acts, using parts of the Switchboard corpus (Godfrey et al., 1992) for English and Callhome for Spanish. The corresponding answers for the detected questions were hypothesized in the first turn with a different speaker, following the question-turn. Table 1 shows the results of these experiments for 5 English and 5 Spanish Callhome dialogues, compared to a baseline of randomly assigning n question speech acts, n being the number of question-turns marked by human annotators. We report F_1-scores, where $F_1 = \frac{2PR}{P+R}$ with P=precision and R=recall. We note that while the results for the SA-classifier and the rule-based approach are very similar for English, the rule-based approach yields better results for Spanish. The much higher random baseline for Spanish can be explained by the higher incidence of questions in the Spanish data (14.9% vs. 5.3% for English).

	English	Spanish
Annotated Data		
turns	1603	1185
Wh-questions	42	78
yes-no-questions	43	98
questions total	85 (5.3%)	176 (14.9%)
Automatic Detection Results (F_1)		
SA classifier	0.24	0.22
POS rules	0.22	0.37
random baseline	0.02	0.13

Table 1: Q-A-pair distribution in the data and experimental results for automatic Q-A-detection

5 Clean-up Filter

The clean-up component is a sequence of modules which serve the purposes of (a) rendering the transcripts more readable, (b) simplifying the input for subsequent components, and (c) avoiding unwanted bias for relevance computations (see section 2). All this has to happen without losing essential information that could be relevant in a summary. While other work (Heeman et al., 1996; Stolcke et al., 1998) was concerned with building classifiers that can detect and possibly correct various speech disfluencies, our implementation is of a much simpler design. It does not require as much manual annotated training data and uses individual components for every major category of disfluency.[1]

[1] While we have not yet numerically evaluated the performance of this component, its output is deemed very natural to read by system users. Since the focus and goals of this component are somewhat different than previous work in that area, meaningful comparisons are hard to make.

Single or multiple word repetitions, fillers (e.g., "uhm"), and discourse markers without semantic content (e.g., "you know") are removed from the input, some short forms are expanded (e.g., "we'll" → "we will"), and frequent word sequences are combined into a single token (e.g., "a lot of" → "a_lot_of").

Longer turns are segmented into *short clauses*, which are defined as consisting of at least a subject and an inflected verbal form. While (Stolcke and Shriberg, 1996) use n-gram models for this task, and (Gavaldà et al., 1997) use neural networks, we decided to use a rule-based approach (using word and POS information), whose performance proved to be comparable with the results in the cited papers ($F_1 > 0.85$, $error < 0.05$).[2]

For several of the clean-up filter's components, we make use of Brill's POS tagger (Brill, 1994). For English, we use a modified version of Brill's original tag set, and the tagger was adapted and retrained for spoken language corpora (CALLHOME and SWITCHBOARD) (Zechner, 1997). For Spanish, we created our own tag set, derived from the LDC lexicon and from the CRATER project (León, 1994), and trained the tagger on manually annotated CALLHOME dialogues. Furthermore, a POS based shallow chunk parser (Zechner and Waibel, 1998) is used to filter out likely candidates for incomplete clauses due to speech repair or interruption by the other speaker.

6 Topic Segmentation

Since CALLHOME dialogues are always multi-topical, segmenting them into topical units is an important step in our summarization system. This allows us to provide "signature" information (frequent content words) about every topic to the user as a help for faster browsing and accessing the data. Furthermore, the subsequent information condensation component can work on smaller parts of the dialogue and thus operate more efficiently.

Following (Boguraev and Kennedy, 1997; Barzilay and Elhadad, 1997) who use TextTiling (Hearst, 1997) for their summarization systems of written text, we adapted this algorithm (its block comparison version) for speech data: we choose turns to be minimal units and compute block similarity between blocks of k turns every d turns. We use 9 English and 15 Spanish CALLHOME dialogues, manually annotated for topic boundaries, to determine the optimum values for a set of TextTiling parameters and at the same time to evaluate the accuracy of this algorithm. To do this, we ran an n-fold cross-validation ("jack-knifing") where all dialogues but one are used to determine the best parameters ("train set") and the remaining dialogue is used as a held-out data set for evaluation ("test set"). This process is repeated n times and average results are reported. Table 2 shows the set of parameters which worked best for most dialogues and Table 3 shows the evaluation results of the cross-validation experiment. F_1-scores improve by 18–24% absolute over the random baseline for *unseen* and by 23–35% for seen data, the performance for English being better than for Spanish. These results, albeit achieved on a quite different text genre, are well in line with the results in (Hearst, 1997) who reports an absolute improvement of about 20% over a random baseline for seen data.

	English	Spanish
blocksize k	25	15
sample distance d	2	2
rounds of smoothing r	2	1
smoothing width s	2	1

Table 2: Optimal TextTiling parameters for English and Spanish CALLHOME dialogues

	English	Spanish
number of dialogues	9	15
random baseline	0.34	0.35
test set avg. ("unseen data")	0.58	0.53
train set avg. ("seen data")	0.69	0.58

Table 3: Topic segmentation results for English and Spanish CALLHOME dialogues (F_1-scores)

7 Information Condensation

The information condensation component is the core of our system. Its purpose is to determine weights for terms and turns (or linked turn-pairs) and then to rank the turns according to their relevance within each topical segment of the dialogue.

For term-weighting, *tf*idf*-inspired formulae (Salton and Buckley, 1990) are used to emphasize words which are in the "middle range" of frequency in the dialogue and do not appear in a stop list.[3] For turn-ranking, we use a version of the "maximal marginal relevance" (MMR) algorithm (Carbonell and Goldstein, 1998), where emphasis is given to turns which contain many highly weighted terms for the current segment ("salience") and are sufficiently dissimilar to previously ranked turns (to minimize redundancy).

For 9 English and 14 Spanish dialogues, the "most relevant" turns were marked by human coders. We ran a series of cross-validation experiments to (a) optimize the parameters of this component related to *tf*idf* and MMR computation and to (b) determine

[2] The comparison was done on the same data set as used in (Gavaldà et al., 1997).

[3] For English, our stop list comprises 557 words, for Spanish, 831 words.

how well this information condensing component can match the human relevance annotations.

Summarization results are computed using 11-pt-avg precision scores for ranked turn lists where the maximum precision of the list of retrieved turns is averaged in the 11 evenly spaced intervals between recall=[0,0.1),[0.1,0.2), ... [1.0,1.1) (Salton and McGill, 1983).[4] Table 4 shows the results from these experiments. Similar to other experiments in the summarization literature (Mani et al., 1998), we find a wide performance variation across different texts.

8 Telegraphic Reduction

The purpose of this component is to maximize information in a fixed amount of space. We shorten the output of the summarizer to a "telegraphic style"; that way, more information can be included in a summary of k words (or n bytes). Since we only use shallow methods for textual analysis that do not generate a dependency structure, we cannot use complex methods for text reduction as described, e.g., in (Jing, 2000). Our method simply excludes words occurring in the stop list from the summary, except for some highly informative words such as "I" or "not".

9 User Interface and System Performance

Since we want to enable interactive summarization which allows a user to browse through a dialogue quickly to search for information he is interested in, we have integrated our summarization system into a JAVA-based graphical user interface ("Meeting Browser") (Bett et al., 2000). This interface also integrates the output of a speech recognizer (Yu et al., 1999), and can display a wide variety of information about a conversation, including speech acts, dialogue games, and emotions.

For summarization, the user can determine the size of the summary and which topical segments he wants to have displayed. He can also focus the summary on particular content words ("query-based summary") or exclude words from consideration ("dynamic stop list expansion").

Summarizing a 10 minute segment of a CALLHOME dialogue with our system takes on average less than 30 seconds on a 167 MHz 320 MB Sun Ultra1 workstation.[5]

10 Human Study

10.1 Experiment Setup

In order to evaluate the system as a whole, we conducted a study with humans in the loop to be able to compare three types of summaries (TRANS, CLEAN, TELE, see section 3) with the full original transcript. We address these two main questions in this study: (i) how fast can information be identified using different types of summaries? (ii) how accurately is the information preserved, comparing different types of summaries?

We did not only ask the user "narrow" questions for a specific piece of information — along the lines of the Q-A-evaluation part of the SUMMAC conference (Mani et al., 1998) — but also very "global", non-specific questions, tied to a particular (topical) segment of the dialogue.

The experiment was conducted as follows: Subjects were given 24 texts each, accompanied by either a *generic* question ("What is the topic of the discussion in this text segment?") or three specific questions (e.g., "Which clothes did speaker A buy?"). The texts were drawn from five topical segments each from five English CALLHOME dialogues.[6] They have four different formats: (a) full transcripts (i.e., the transcript of the whole segment) (FULL); (b) summary of the raw transcripts (without linking and clean-up) (TRANS); (c) cleaned-up summary (using all four major components of our system) (CLEAN); and (d) telegram summary (derived from (c), using also the telegraphic reduction component) (TELE). The texts of formats (b), (c), and (d) were generated to have the same length: 40% of (a), i.e., we use a 60% reduction rate. All these formats can be accompanied by either a generic or three specific questions, hence there are eight types of tasks for each of the 24 texts.

We divided the subjects in eight groups such that no subject had to perform more than one task on the same text and we distributed the different tasks evenly for each group. Thus we can make unbiased comparisons across texts and tasks.

The answer accuracy vs. a pre-defined answer key was manually assessed on a 6 point discrete scale between 0.0 and 1.0.

10.2 Results and Discussion

Of the 27 subjects taking part in this experiment, we included 24 subjects in the evaluation; 3 subjects were excluded who were extreme outliers with respect to average answer time or score (not within $\mu + -2$stddev).

From the results in Table 5 we observe the following trends with respect to answer accuracy and response time:

[4] We are aware that this annotation and evaluation scheme is far from optimal: it does neither reflect the fact that turns are not necessarily the best units for extraction nor that the 11-pt-avg precision score is not optimally suited for the summarization task. We thus have recently developed a new word-based method for annotation and evaluation of spontaneous speech (Zechner, 2000).

[5] The average was computed over five English dialogues.

[6] One of the 25 segments was set aside for demonstration purposes.

	English	Spanish
number of dialogues	9	14
turns per dialogue marked as relevant by human coders	12%	25%
11-pt-avg precision (average over topical segments)	0.45	0.59
score variation between dialogues	0.2–0.49	0.15–0.8

Table 4: Summarization results for English and Spanish CALLHOME

Format	full		trans		clean		tele	
Time vs. Acc.	Time	Acc.	Time	Acc.	Time	Acc.	Time	Acc.
generic ($q = 72$)	75.2	0.814	53.9	0.739	52.6	0.617	54.4	0.622
specific ($q = 216$)	109.1	0.834	82.2	0.624	88.0	0.593	91.6	0.665

Table 5: Average answer times (in sec) and accuracy scores ([0.0-1.0]) over eight different tasks (number of subjects=24; q=number of questions per task type).

summary type	trans	clean	tele
generic / indicative	90.8	75.8	76.4
specific / informative	74.8	71.0	79.7

Table 6: Relative answer accuracies in % for different summaries

- *generic* questions ("indicative summaries", the task being to identify the topic of a text): The two cleaned up summaries took about the same time to process but had lower accuracy scores than the version directly using the transcript.
- *specific* questions ("informative summaries", the task being to find specific information in the text): (1) The accuracy advantage of the raw transcript summaries (TRANS) over the cleaned up versions (CLEAN) is only small (*not* statistically significant: t=0.748)[7]. (2) There is a superiority of the TELE-summary to both other kinds (TELE is significantly more accurate than CLEAN for $p < 0.05$).

From this we conjecture that our methods for customization of the summaries to spoken dialogues is mostly relevant for *informative*, but not so much for *indicative* summarization. We think that other methods, such as lists of signature phrases would be more effective to use for the latter purpose.

Table 6 shows the answer accuracy for the three different summary types *relative* to the accuracy of the full transcript texts of the same segments ("relative answer accuracy"). We observe that the relative accuracy reduction for all summaries is markedly lower than the reduction of text size: all summaries were reduced from the full transcripts by 60%, whereas the answer accuracy only drops between 9% (TRANS) and 24% (CLEAN) for the generic questions, and between 20% (TELE) and 29% (CLEAN) for the specific questions. This proves that our system is able to retain most of the relevant information in the summaries.

As for average answer times, we see a marked reduction (30%) of all summaries compared to the full texts in the *generic* case; for the *specific* case, the time reduction is somewhat smaller (15%–25%).

One shortcoming of the current system is that it operates on turns (or turn-pairs) as minimal units for extraction. In future work, we will investigate possibilities to reduce the minimal units of extraction to the level of clauses or sentences, *without* giving up the idea of linking cross-speaker information.

11 Summary and Future Work

We have presented a summarization system for spoken dialogues which is constructed to address key differences of spoken vs. written language, dialogues vs. monologues, and multi-topical vs. mono-topical texts. The system cleans up the input for speech disfluencies, links turns together into coherent information units, determines topical segments, and extracts the most relevant pieces of information in a user-customizable way. Evaluations of major system components and of the system as a whole were performed. The results of a user study show that with a summary size of 40%, between 71% and 91% of the information of the full text is retained in the summary, depending on the type of summary and the types of questions being asked.

We are currently extending the system to be able to handle different levels of granularity for extraction (clauses, sentences, turns). Furthermore, we plan to investigate the integration of prosodic information into several components of our system.

12 Acknowledgements

We want to thank the annotators for their efforts and Klaus Ries for providing the automatic speech act

[7] In fact, in 2 of 5 dialogues, the CLEAN summary scores are higher than those of the TRANS summaries.

tagger. We appreciate comments and suggestions from Alon Lavie, Marsal Gavaldà, Jade Goldstein, Thomas MacCracken, and the anonymous reviewers on earlier drafts of this paper.

This work was funded in part by the VERBMOBIL project of the Federal Republic of Germany, ATR – Interpreting Telecommunications Research Laboratories of Japan, and the US Department of Defense.

References

AAAI, editor. 1998. *Proceedings of the AAAI-98 Spring Symposium on Intelligent Text Summarization, Stanford, CA.*

ACL. 2000. *Proceedings of the ANLP/NAACL-2000 Workshop on Automatic Summarization, Seattle, WA, May.*

Jan Alexandersson and Peter Poller. 1998. Towards multilingual protocol generation for spontaneous speech dialogues. In *Proceedings of the INLG-98, Niagara-on-the-lake, Canada, August.*

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *ACL/EACL-97 Workshop on Intelligent and Scalable Text Summarization.*

Michael Bett, Ralph Gross, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel. 2000. Multimodal meeting tracker. In *Proceedings of the Conference on Content-Based Multimedia Information Access, RIAO-2000, Paris, France, April.*

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

Eric Brill. 1994. Some advances in transformation-based part of speech tagging. In *Proceeedings of AAAI-94.*

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.*

John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, and Karen Sparck Jones. 1997. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 1997 TREC-6 Conference, Gaithersburg, MD, November*, pages 83–91.

John S. Garofolo, Ellen M. Voorhees, Cedric G. P. Auzanne, and Vincent M. Stanford. 1999. 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, April.

Marsal Gavaldà, Klaus Zechner, and Gregory Aist. 1997. 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.

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.

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

Peter A. Heeman, Kyung ho Loken-Kim, and James F. Allen. 1996. Combining the detection and correction of speech repairs. In *Proceedings of ICSLP-96.*

Julia Hirschberg and Christine Nakatani. 1998. Acoustic indicators of topic segmentation. In *Proceedings of the ICSLP-98, Sydney, Australia.*

Hongyan Jing. 2000. Sentence reduction for automatic text summarization. In *Proceedings of ANLP-NAACL-2000, Seattle, WA, May*, pages 310–315.

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.

Fernando Sánchez León. 1994. Spanish tagset for the CRATER project. http://xxx.lanl.gov/cmp-lg/9406023.

Inderjeet Mani and Mark Maybury, editors. 1997. *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain.*

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Leo Obrst, Therese Firmin, Michael Chrzanowski, and Beth Sundheim. 1998. The TIPSTER SUMMAC text summarization evaluation. Mitre Technical Report MTR 98W0000138, October 1998.

Klaus Ries. 1999. HMM and neural network based speech act detection. In *Proceedings of the ICASSP-99, Phoenix, Arizona, March.*

Gerard Salton and Chris Buckley. 1990. Flexible text matching for information retrieval. Technical report, Cornell University, Department of Computer Science, TR 90-1158, September.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval.* McGraw Hill, Tokyo etc.

Elizabeth Shriberg, Rebecca Bates, Andreas Stolcke, Paul Taylor, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3-4):439–487.

Andreas Stolcke and Elizabeth Shriberg. 1996. Automatic linguistic segmentation of conversational speech. In *Proceedings of the ICSLP-96*, pages 1005–1008.

Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Mari Ostendorf, Dilek Hakkani, Madeleine Plauche, Gökhan Tür, and Yu Lu. 1998. 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.

Andreas Stolcke, Elizabeth Shriberg, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2).

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, and Michael Finke. 1998. Meeting browser: Tracking and summarizing meetings. In *Proceedings of the DARPA Broadcast News Workshop.*

Hua Yu, Michael Finke, and Alex Waibel. 1999. Progress in automatic meeting transcription. In *Proceedings of EUROSPEECH-99, Budapest, Hungary, September.*

Klaus Zechner and Alex Waibel. 1998. Using chunk based partial parsing of spontaneous speech in unrestricted domains for reducing word error rate in speech recognition. In *Proceedings of COLING-ACL 98, Montreal, Canada.*

Klaus Zechner and Alex Waibel. 2000. 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.

Klaus Zechner. 1997. Building chunk level representations for spontaneous speech in unrestricted domains: The CHUNKY system and its application to reranking N-best lists of a speech recognizer. Master's thesis (project report), CMU, available from: `http://www.cs.cmu.edu/~zechner/publications.html`.

Klaus Zechner. 2000. A word-based annotation and evaluation scheme for summarization of spontaneous speech. Available from http://www.cs.cmu.edu/~zechner/publications.html.