Minimizing Word Error Rate in Textual Summaries of Spoken Language

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

Abstract

Automatic generation of text summaries for spoken language faces the problem of containing incorrect words and passages due to speech recognition errors. This paper describes comparative experiments where passages with higher speech recognizer confidence scores are favored in the ranking process. Results show that a relative word error rate reduction of over 10% can be achieved while at the same time the accuracy of the summary improves markedly.

1 Introduction

The amount of audio data on-line has been growing rapidly in recent years, and so methods for efficiently indexing and retrieving non-textual information have become increasingly important (see, e.g., the TREC-7 branch for "Spoken Document Retrieval" (Garofolo et al., 1999)).

One way of compressing audio information is the automatic creation of textual summaries which can be skimmed much faster and stored much more efficiently than the audio itself. There has been plenty of research in the area of summarizing *written* language (see (Mani and Maybury, 1999) for a comprehensive overview). So far, however, very little attention has been given to the question how to create and evaluate a summary of spoken audio based on automatically generated transcripts from a speech recognizer. One fundamental problem with those summaries is that they contain incorrectly recognized words, i.e., the original text is to some extent "distorted".

Several research groups have developed interactive "browsing" tools, where audio (and possibly video) can be accessed together with various types of textual information (transcripts, summaries) via a graphical user interface (Waibel et al., 1998; Valenza et al., 1999; Hirschberg et al., 1999). With these tools, the problem of misrecognitions is alleviated in the sense that the user can always easily listen to the audio recording corresponding to a passage in a textual summary. In some instances, however, this approach may not be feasible or too expensive to pursue, and a short, stand-alone textual representation of the spoken audio may be preferred or even required. This paper addresses in particular this latter case and (a) explores means of making textual summaries less distorted (i.e., reducing their word error rate (WER)), and (b) assesses how the accuracy of the summaries changes when methods for word error rate reduction are applied. Summary accuracy will be a function of how much relevant information is present in the summary.

Our results from experiments on four television shows with multiple speakers show that it is possible to reduce word error rate while at the same time also improving the accuracy of the summary. Furthermore, this paper presents a novel method for evaluation of textual summaries from spoken language data.

The paper is organized as follows: In the next section, we review related work on spoken language summarization. In section 3 we describe our summarizer. Next, we present and discuss our proposal for an audio summarization evaluation metric (section 4). In section 5 we describe the corpus that we use for our experiments and how it was annotated. Sections 6 and 7 describe experiments on both human and machine generated transcripts of the audio data. Finally, we discuss and summarize the results in sections 8 and 9.

2 Related work

(Waibel et al., 1998) report results of their summarization system on automatically transcribed SWITCHBOARD data (Godfrey et al., 1992), the word error rate being about 30%. In a question-answer test with summaries of five dialogues, subjects could identify most of the key concepts using a summary size of only five turns. However, the results vary widely across five different dialogues tested in this experiment (between 20% and 90% accuracy).

(Valenza et al., 1999) went one step further and report that they were able to reduce the word error rate in summaries (as opposed to full texts) by using speech recognizer confidence scores. They combined inverse frequency weights with confidence scores for each recognized word. Using summaries composed of one 30-gram per minute (approximately 15% length of the full text), the WER dropped from 25% for the full text to 10% for these summaries. They also conducted a qualitative study where human subjects were given summaries of n-grams of different length and also summaries with speaker utterances as minimal units, either giving a high weight to the inverse frequency scores or to the confidence scores. The utterance summaries were considered best, followed closely by 30-gram summaries, both using high confidence score weights. This suggests that not only does the WER drop by extracting passages that are more likely to be correctly recognized but also do summaries seem to be "better" which are generated that way.

While the results of (Valenza et al., 1999) are *indicative* for their approach, we want to investigate the benefits of using speech recognizer confidence scores in more detail and particularly find out about the *trade-off* between WER and summarization accuracy when we vary the influence of the confidence scores. To our knowledge, this paper addresses this trade-off for the first time in a clear, numerically describable way. To be able to obtain numerical values for summary accuracy, we had our corpus annotated for relevance (section 5) and devised an evaluation scheme that allows the calculation of summary accuracy for both human and machine generated transcripts (section 4).

3 Summarization system

Prior to summarizing, the input text is cleaned up for disfluencies, such as hesitations, filled pauses, and repetitions.[1] In the context of multi-topical recordings we use for our experiments, summaries are generated for each topical segment separately. The segment boundaries were determined to be at those places where the majority (at least half) of the human annotators agreed (see section 5). Intercoder agreement for topical boundaries is fairly good (and higher than the agreement on relevant words or passages).[2]

To determine the content of the summaries, we use a "maximal marginal relevance" (MMR) based summarizer with speaker turns as minimal units (cf. (Carbonell and Goldstein, 1998)).

The MMR formula is given in equation 1. It generates a list of turns ranked by their relevance and states that the next turn to be put in this ranked list will be taken from the turns which were not yet ranked (t_{nr}) and has the following properties: it is (a) maximally similar to a "query" and (b) maximally dissimilar to the turns which were already ranked (t_r). As "query" we use a frequency vector for all content words within a topical segment. The λ-parameter ($0.0 \leq \lambda \leq 1.0$) is used to trade off the influence of (a) vs. (b).

Both similarity metrics (sim_1, sim_2) are inner vector products of (stemmed) term frequencies (see equations 2 to 4); $\vec{\mathrm{tf}}_t$ is a vector of stem frequencies in a turn; f_s are in-segment frequencies of a stem; f_{smax} are maximal segment frequencies of any stem in the topical segment. sim_1 can be normalized or not. The formulae for tf_s (equation 4) are inspired from Cornell's SMART system (Salton, 1971); we will call these parameters "smax", "log", and "freq", respectively.

$$\begin{aligned} nextturn &= \arg\max_{t_{nr,j}}(\lambda sim_1(t_{nr,j}, query) \\ &\quad -(1-\lambda)\max_{t_{r,k}} sim_2(t_{nr,j}, t_{r,k})) \end{aligned} \tag{1}$$

$$sim_1 = \vec{\mathrm{tf}}_s \vec{\mathrm{tf}}_t \quad \text{or} \quad \frac{\vec{\mathrm{tf}}_s \vec{\mathrm{tf}}_t}{\left|\vec{\mathrm{tf}}_s\right| \left|\vec{\mathrm{tf}}_t\right|} \tag{2}$$

$$sim_2 = \frac{\vec{\mathrm{tf}}_{t1} \vec{\mathrm{tf}}_{t2}}{\left|\vec{\mathrm{tf}}_{t1}\right| \left|\vec{\mathrm{tf}}_{t2}\right|} \tag{3}$$

$$\mathrm{tf}_{i,s} = 0.5 + 0.5\frac{f_{i,s}}{f_{smax}} \quad \text{or} \quad 1 + \log f_{i,s} \quad \text{or} \quad f_{i,s} \tag{4}$$

Using the MMR algorithm, we obtain a list of ranked turns for each topical segment. We compute this both for human and machine generated transcripts of the audio files ("reference text" vs. "hypothesis text").[3]

4 Evaluation metrics

The challenge of devising a meaningful evaluation metric for the task of audio summarization is that it has to be applicable to both the reference (human transcript) and the hypothesis transcripts (automatic speech recognizer (ASR) transcripts). We want to be able to assess the quality of the summary with respect to the relevance markings of the human annotators (see section 5), as well as to relate this "summary accuracy" to the word error rate present in the ASR transcripts.

The approach we take is to *align* the words in the summary with the words in the reference transcript (w_a). For ASR transcripts, word substitutions are aligned with their "true original" and word insertions are aligned with a NIL dummy. That way,

[1] More details about this component and other parts of the summarization system can be found in (Zechner and Waibel, 2000).

[2] For details see (Zechner, 2000).

[3] The human reference is considered to be an "optimal" or "ideal" rendering of the words which were actually said in a conversation. Human transcription errors do occur, but are marginal and hence ignored in the context of this paper.

we can determine *for each* individual word w_a in the summary (a) whether it occurs in a "relevant phrase" and (b) whether it is correctly recognized or a recognition error (for ASR transcripts).

We define word error rate as $\text{WER} = (S + I + D)/(S + I + C)$ (I=insertion, D=deletion, S=substitution, C=correct).

Each word's relevance score r is the average number it occurs in the human annotators' relevant phrases ($0.0 \leq r \leq 1.0$). Relevance scores for insertions and substitutions are always 0.0.

We choose to define the summary accuracy sa ("relevance") as the sum of relevance scores of all n aligned words $\sum_{w_a} r_{w_a}$ divided by the *maximum achievable relevance score* with the same number of n words *somewhere* in the text (i.e., $0.0 \leq sa \leq 1.0$). Word deletions obviously do not show up in the summary, but are accounted for, as well, to make the WER computation sound.

To better illustrate how these metrics work, we demonstrate them on a simplified example of only two speaker turns (Figure 1). The first line represents the relevance score r for each word (the number this word was within a "relevant phrase" divided by the number of annotators for that text). In turn 1, "this is to illustrate" was only marked relevant by two annotators, whereas "the idea" by 3 out of 4 annotators. The second line provides the reference transcript, the third line the ASR transcript. Line 4 gives the type of word error, and line 5 the confidence score of the speech recognizer (between 0.0 and 1.0, 1.0 meaning maximal confidence).

Now let us assume that turn 2 shows up in the summary. The scores are computed as follows:

- When summarizing the *reference*: Here, the word error rate is trivially 0.0; the summary accuracy sa is the sum of all relevance scores (=6.0) divided by the *maximal achievable score* with the same number of words ($n = 7$). Turn 2 has 6 words which were marked relevant by all coders ($r = 1.0$), turn 1's highest score is $r = 0.75$. Therefore: $sa_2 = 6.0/(6.0 + 0.75) = 0.89$. This is higher than the summary accuracy for turn 1: $sa_1 = 3.5/6.0 = 0.58 (n = 6)$.
- When summarizing the *ASR transcript* ("hypothesis"): Selecting turn 2 will give $sa_2 = 0.0/2.25 = 0.0$ ($n = 5$). For turn 1, $sa_1 = 2.25/(0.75 + 0.5 + 0.5 + 0.5 + 0.0 + 0.0) = 1.0$ ($n = 6$; the sum in the denominator can only use relevance scores based on the *aligned* words w_a which were *correctly* recognized, therefore the 1.0-scores in turn 2 cannot be used). Turn 2 has WER=6/5=1.2, turn 1 has WER=3/6=0.5.

Obviously, when summarizing the ASR output, we would rather have turn 1 showing up in the summary than turn 2, because turn 2 is completely off from the truth and turn 1 only partially. The fact that turn 2 was considered to be more relevant by human coders cannot, in our opinion, be used to favor its inclusion in the summary. An exception would be a situation where the user has immediate access to the audio as well and is able to listen to selected passages from the summary (see section 1). In our case, where we focus on text-only summaries to be used stand-alone, we have to minimize their word error rate.

Given that, turn 1 has to be favored over turn 2, both because of its lower WER and because of its higher accuracy with respect to the relevance annotations.

In order to increase the likelihood that turns with lower WER are selected over turns with higher WER, we make use of the speech recognizer's confidence scores which are attached to every word hypothesis and can be viewed as probabilities: they are in [0.0,1.0], high values reflecting a high confidence in the correctness of the respective word.[4] Following (Valenza et al., 1999) we conjecture that we can use these confidence scores to increase the probability of passages with lower WER to show up in the summary. To test how far this assumption is justified, we correlated the WER with various metrics of confidence scores: (i) sum of scores, (ii) average of scores, (iii) number of scores above a threshold, (iv) the latter normalized by the number of all scores, and (v) the geometric mean of scores. Table 1 shows the correlation coefficients (Pearson r) for the four ASR transcripts we used in our experiments (see section 5). To prevent the influence of large differences in turn length, those computations were done for subsequent "buckets" of 50 words each.

Since in most cases we achieve the highest correlation coefficient (absolute value) for method (iv = avgth) (average number of words whose confidence score is greater than a threshold of 0.95), we apply this metric to the computation of turn-query similarities (sim_1 in equation 1). We use the two following formulae to adjust the similarity-scores. (We shall call these adjustments MULT and EXP in the following.)

$$[mult] \quad sim'_1 = sim_1(1 + \alpha \text{avgth}) \tag{5}$$
$$[exp] \quad sim''_1 = sim_1 \text{avgth}^{\alpha} \tag{6}$$

For both equations it holds that if $\alpha = 0.0$, the scores don't change, whereas if $\alpha > 0.0$, we enhance the weights of turns with many high confidence scores ("boosting") and hence increase their likelihood of showing up earlier in the summary.[5]

Even though our evaluation method looks like it would "guarantee" an increase in summary accu-

[4] The speech recognizer computes these scores based on the acoustic stability of words during lattice rescoring.

[5] For EXP, we define $0^0 = 0$.

```
TURN 1:
rel:      0.5  0.5 0.5     0.5             0.75  0.75   ***
REF:      this is  to      illustrate      the   idea   ***
HYP:      this is  to      ILLUMINATE      ***   idea   AND
err:      C    C   C       S               D     C      I
con:      1    1   1       0.9             -     0.8    0.8

TURN 2:
rel:      0   1    1  1       1       1          1
REF:      and here we have    very    relevant   information
HYP:      and HE   ** BEHAVES ****    IRREVERENT FORMATION
err:      C   S    D  S       D       S          S
con:      0.8 0.7  -  0.8     -       0.8        0.9
```

Figure 1: Simplified example of two turns (for score computation)

	BACK	19CENT	BUCHANAN	GRAY
(i) sum	-0.43	-0.51	-0.12	-0.03
(ii) average	-0.53	-0.52	-0.43	-0.42
(iii) scores > 0.95	-0.55	-0.48	-0.35	-0.25
(iv) normalized (iii)	**-0.58**	-0.48	**-0.48**	**-0.44**
(v) geometric mean	-0.53	**-0.53**	-0.42	-0.38

Table 1: Pearson r correlation between WER and confidence scores

racy when the word error rate is reduced, this is *not* necessarily the case. For example, it could turn out that while we can reduce WER by "boosting" passages with higher confidence scores, those passages might have (much) fewer words marked relevant than those being present in the summary without boosting. This way, it would be conceivable to create low word error summaries that contain also very few *relevant* pieces of information. However, as we will see later, WER reduction goes hand in hand with an increase of summary accuracy.

5 Data characteristics and annotation

Table 2 describes the main features of the corpus we used for our experiments: we selected four audio excerpts from four television shows, together with human generated textual transcripts. All these shows are conversations between multiple speakers. The audio was sampled at 16kHz and then also automatically transcribed using a gender independent, vocal tract length normalized, large vocabulary speech recognizer which was trained on about 80 hours of Broadcast News data (Yu et al., 1999). The average word error rates for our 4 recordings ranged from 25% to 50%.

The reference transcripts of the four recordings were given to six human annotators who had to segment them into topically coherent regions and to decide on the "most relevant phrases" to be included in a summary for each topical region. Those phrases usually do not coincide exactly with speaker turns and the annotators were encouraged to mark sections of text freely such that they would form meaningful, concise, and informative phrases. Three annotators could listen to the audio while annotating the corpus, the other three only had the human generated transcripts available. 2 of the 6 annotators only finished the NewsHour data, so we have the opinion of 4 annotators for the recordings BUCHANAN and GRAY and of 6 annotators for BACK and 19CENT.

6 Experiments on human generated transcripts

We created summaries of the reference transcripts using different parameters for the MMR computation: For tf we used "freq", "log", and "smax"; further, we did or did not normalize these weights; finally, we varied the MMR-λ from 0.85 to 1.0. Summarization accuracy was determined at 5%, 10%, 15%, 20%, and 25% of the text length of each summarized topical segment and then averaged over all sample points in all segments. Since these were *word-based* lengths, words were added incrementally to the summary in the order of the turns ranked via MMR; turns were cut off when the length limit was reached. As explained in the example in section 4, the accuracy score is defined as the fraction of the sum of all individual word relevance scores (as de-

	BACK	19CENT	BUCHANAN	GRAY
TV show	NewsHour	NewsHour	Crossfire	Crossfire
number of speakers	5	2	4	5
speaker turns	24	27	69	70
words in transcript	1216	1281	3252	2205
length in minutes	8.6	8.6	17.3	11.9
topical segments	4	4	4	3
word error rate (in %)	25.6	32.6	32.5	49.8

Table 2: Characteristics of the corpus

BACK	19CENT	BUCHANAN	GRAY	average
0.533	0.596	0.513	0.443	0.522

Table 3: Reference summarization accuracy of MMR summaries

termined by human annotators) over the *maximum possible* score given the current number of words in the summary.

Table 3 shows the summary accuracy results for the best parameter setting (tf=log, no normalization)[6].

7 Experiments on automatically generated transcripts

Using the same summarizer as before, we now created summaries from ASR transcripts. Additionally to the summary accuracy, we evaluate now also the WER for each evaluation point. Again, we ran a series of experiments for different parameters of the MMR formula (tf=log, smax, freq; with/without normalization). As before, we achieved the best results for non normalized scores and tf=log. We varied α from 0.0 to 10.0 to see how much of an effect we would get from the "boosting" of turns with many high confidence scores (see equations 5 and 6).

The EXP formula yielded better results than MULT (Table 4), the optimum for EXP was reached for $\alpha = 3.0$ with a WER of 26.6%, an absolute improvement of over 8% over the average of WER=35.1% for the complete ASR transcripts (non-summarized). The summarization accuracy peaks at 0.47, a 9% absolute improvement over the $\alpha = 0.0$-baseline and only about 5% absolute lower than for reference summaries (Table 4 and Figure 2).

When we compare the baseline of $\alpha = 0.0$ (i.e., no "boosting" of high confidence turns) to the best result ($\alpha = 3.0$), we see that the WER drops markedly by about 12% relative from 30.1 to 26.6%. At the same time, the summarization accuracy *increases* by about 18% relative form 0.401 to 0.472.

[6] If we use non-normalized scores, the value of the MMR-λ does not have any measurable effect; we assigned it to be 0.95 for all subsequent experiments.

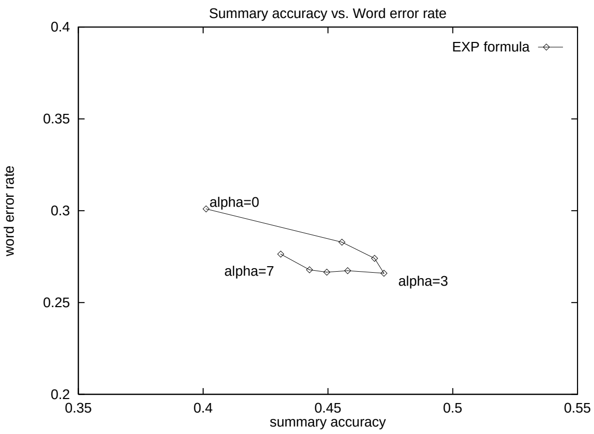

Figure 2: Summary accuracy vs. word error rates with EXP boosting ($0 \leq \alpha \leq 7$)

Results for the MULT formula confirm this trend, but it is considerably weaker: approximately 6% WER reduction and 14% accuracy improvement for $\alpha = 10.0$ over the $\alpha = 0.0$ baseline.

An appendix (section 11) provides an example of actual summaries generated by our system for the first topical segment of the BACK conversation. It illustrates how WER reduction and summary accuracy improvement can be achieved by using our confidence boosting method.

8 Discussion

The most significant result of our experiments is, in our opinion, the fact that the trade-off between *word* and *summary* accuracy indeed leads to an optimal parameter setting for the creation of textual summaries for spoken language (Figure 2). Using a formula which emphasizes turns containing many high confidence scores leads to an average WER reduction of over 10% and to an average improvement in summary accuracy of over 15%, compared to the baseline of a standard MMR-based summary.

Comparing our results to those reported in (Valenza et al., 1999), we find that their relative

	BACK		19CENT		BUCHANAN		GRAY		average	
	acc	WER	acc	WER	acc	WER	acc	WER	acc	WER
$\alpha = 0.0$	0.411	26.2	0.501	26.7	0.412	30.6	0.280	36.9	0.401	30.1
EXP ($\alpha = 3.0$)	0.648	18.8	0.501	26.7	0.444	26.9	0.296	34.0	**0.472**	**26.6**
MULT ($\alpha = 10.0$)	0.575	21.5	0.501	26.7	0.429	29.6	0.317	35.7	0.456	28.3

Table 4: Effect of α on summary accuracy vs. WER (in %) transcripts with EXP and MULT boosting methods

	BACK	19CENT	BUCHANAN	GRAY
avgth	-0.79	-0.11	-0.43	-0.03

Table 5: Correlation between WER and confidence scores on a turn basis

WER reduction for summaries over full texts was considerably larger than ours (60% vs. 24%). We conjecture that reasons for this may be due to the different nature and quality of the confidence scores, and (not unrelated), to the different absolute WER of the two corpora (25% vs. 35%): in transcripts with higher WER, the confidence scores are usually less reliable (cf. Table 1).

Looking at the four audio recordings individually, we see that the improvements vary strongly across different recordings. We conjecture that one reason for this fact may be due to the high variation in the correlation between WER and confidence scores on a *turn* basis (Table 5). This would explain why, e.g., BACK's improvements are much stronger than those of the BUCHANAN recording or why there are no improvements for the 19CENT recording. However, GRAY *does* improve despite its very low absolute correlation.

9 Summary

In this paper, we presented experiments on summaries of both human and machine generated transcripts from four recordings of spoken language. We explored the trade-off of word accuracy vs. summary accuracy (relevance) using speech recognizer confidence scores to rank passages with lower word error rate higher in the summarization process.

Results comparing our approach to a simple MMR ranking show that while the WER can be reduced by over 10%, summarization accuracy improves by over 15% as measured against transcripts with relevance annotations.

10 Acknowledgements

We thank the six human annotators for their tedious work of annotating the corpus with topical segment boundaries and relevance information. We also want to thank Alon Lavie and the three anonymous reviewers for useful feedback and comments on earlier drafts of this paper.

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

References

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, 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.

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.

Inderjeet Mani and Mark T. Maybury, editors. 1999. *Advances in automatic text summarization.* MIT Press, Cambridge, MA.

Gerard Salton, editor. 1971. *The SMART Retrieval System — Experiments in Automatic Text Processing.* Prentice Hall, Englewood Cliffs, New Jersey.

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.*

α	relative *sa*	WER in %	turns in summary
0.0	0.428	29.2	2, 1[beginning]
3.0	0.885	11.8	1, 5[beginning]

Table 6: Relative summary accuracy, WER, and selected turns by the summarizer for (a) no boosting and (b) EXP boosting.

turn	avg. relevance score	WER in %	avgth
1	0.663	9.5	0.84
2	0.369	27.5	0.40
3	0.149	26.9	0.39
4	0.212	11.1	0.08
5	0.274	27.7	0.17

Table 7: Average relevance scores, WER, and confidence values for the five turns of BACK's first topical segment.

Klaus Zechner and Alex Waibel. 2000. Diasumm: Flexible summarization of spontaneous dialogues in unrestricted domains. 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.

11 Appendix: Example summaries

This appendix provides summaries for the first topical segment of the BACK conversation. The contents of this conversation revolves around former Illinois congressman Dan Rostenkowski who had been released from prison and was ready to re-enter public life.

Figure 3 shows the human transcript of this segment which is about two minutes long and consists of 5 speaker turns. Figure 4 contrasts the machine generated summaries for this segment (a) without confidence boosting ($\alpha = 0.0$) and (b) using the optimal confidence boosting ($\alpha = 3.0$, method EXP). Insertions and substitutions are capitalized and marked with I- or S- prefixes. Table 6 compares the relative summary accuracies (*sa*) and word error rates (WER in %) for these two summaries (average over the 5 sample points from 5% to 25% summary length). Additionally, the turns that show up in the summaries are listed in their ranking order. Table 7 provides the average relevance scores, word error rates, and confidence scores ("avgth") for each turn of this topical segment.

We observe that the most relevant turn is turn 1 which has, incidentally, also the lowest WER. Whereas in case (a) ($\alpha = 0.0$), turn 2 is ranked first and therefore dominates the lower relevance and higher WER scores, case (b) ($\alpha = 3.0$) successfully ranks turn 1 first due to its higher confidence scores and hence both summary accuracy and WER scores improve.

```
 1  elizabeth: it has been eight months since dan rostenkowski walked out of a wisconsin federal
   prison six months since he left a halfway house in chicago the former chairman of the house ways and means
   committee is ready to step back into the public eye
2   elizabeth: the reception was warm the banquet hall packed with the city's movers and shakers
   the thirty five dollars a plate invitation referred to rostenkowski as mr. chairman rostenkowski
   made no reference to his conviction for misusing federal funds only a brief reference to his fifteen
   months of prison time
3   dan: i graduated from oxford and i really had a rhodes scholarship the past three years have been a
   constantly challenging time for me change never comes easily and given the circumstances
   of my situation that was particularly true for me at times things have been downright bleak and i
   wouldn't want to wish my experience on my worst enemy but there were some silver linings i've had an
   opportunity to read and reflect in a way that wasn't possible when i was
   in constant moment in these
   remarks today i'd like to share some of my conclusions
4   elizabeth: the conclusions did not dwell on the demise of dan rostenkowski's career but
   the demise of party politics
5   dan: those who say that the president's political power has been weakened by scandal have truly short
   memories the sad fact is that president clinton has never had a democratic base in congress a group
   of people whom one could support the white house on any given issue are not there
```

Figure 3: Human transcript of first topical segment (BACK)

```
1 elizabeth: has been eight months since dan rostenkowski walked out of
   wisconsin federal prison I-MAYBE
2 elizabeth: was S-ALARMED the banquet hall packed with the city's
   S-COMMUTERS S-IN S-CHAMBERS S-WHICH thirty five S-DOLLAR a plate
   S-IMITATION referred to rostenkowski as S-MR. chairman I-LET I-ME I-ASK
   rostenkowski made no reference to his conviction for I-MIS S-USING federal
   funds only a brief reference to S-IS fifteen months of prison time

1 elizabeth: has been eight months since dan rostenkowski
   walked out of wisconsin federal prison I-MAYBE six months since he left
   S-THE halfway house in chicago the former chairman of the house ways and
   means committee ready to step back into the public eye
5 dan: S-ALSO say that the president's political power has been weakened by
  scandal S-RIGHT S-ESPECIALLY short S-MEMORY S-THAT S-DISSATISFACTION that
  president clinton has never
```

Figure 4: Machine generated summaries for (a) $\alpha = 0.0$ and (b) $\alpha = 3.0$ (25% of text length)