Packing the Meeting Summarization Knapsack

Korbinian Riedhammer[1,3]*, Dan Gillick*[2,3]*, Benoit Favre*[3]*, Dilek Hakkani-Tür*[3]

[1] Computer Science Dept. 5, University of Erlangen-Nuremberg, GERMANY
[2] Computer Science Dept., University of California Berkeley, USA
[3] International Computer Science Institute, Berkeley, USA
{koried,dgillick,favre,dilek}@icsi.berkeley.edu

Abstract

Despite considerable work in automatic meeting summarization over the last few years, comparing results remains difficult due to varied task conditions and evaluations. To address this issue, we present a method for determining the best possible extractive summary given an evaluation metric like ROUGE. Our oracle system is based on a knapsack-packing framework, and though NP-Hard, can be solved nearly optimally by a genetic algorithm. To frame new research results in a meaningful context, we suggest presenting our oracle results alongside two simple baselines. We show oracle and baseline results for a variety of evaluation scenarios that have recently appeared in this field.

Index Terms — summarization, meetings, evaluation

1. Introduction

In general, both text and speech summarization can be either *abstractive* – expressing the content in newly formulated sentences, or *extractive* – a direct selection of relevant sentences, utterances, or dialog acts (DAs) from the input. Since abstractive summarization is very difficult, most work in text summarization has been extractive, and since 2005, various methods for sentence selection have been ported to speech and meeting summarization.

Text summarization evaluations have been coordinated by NIST since 2001, but no such standard evaluation exists in the speech domain. As a result, every report on meeting summarization uses a different experimental setup and different evaluation criteria. Moreover, the widely used evaluation scores have no clear upper bound like word error rate in automatic speech recognition (best: 0%) or speaker identification (best: no errors). By introducing a method for finding an oracle extractive summary, this paper addresses two issues: First, expressing a result as a percent of the best possible result for the given evaluation criteria dramatically improves comparability. Second, computing the oracle result sheds some light on the future of extractive summarization for meetings – how much room for improvement exists if we simply select utterances?

This paper proceeds as follows. We begin with descriptions of the primary data sets, automatic evaluation metrics, and other details of previous meeting summarization work. Next, we introduce our oracle summarization framework based on the well known knapsack packing problem, and present a genetic algorithm for giving near optimal solutions. Our experiments show oracle results, along with two baseline systems, for each of the setups used in previous work. We conclude with a discussion and some suggestions for future work.

2. Data

In this work, we use two meeting data sets that have manually annotated summaries: the AMI and ICSI meeting corpora.

2.1. ICSI Meeting Corpus

The ICSI meeting corpus [1] consists of 75 naturally occurring meetings (that is, they would have taken place regardless of the recording project), each around one hour long. They have been transcribed and annotated with dialog acts and abstractive and extractive summaries [2]. To generate the abstractive summaries, annotators were asked to write 200 word summaries and were given a graphical user interface to facilitate browsing that aligned audio with transcriptions. Afterwards, the annotators were asked to create extractive summaries using a different tool. Next, given the transcripts and their abstract, each annotator was instructed to select DAs from the original meeting which supported their abstract. They had no constraints on the number of DAs to choose. In a second pass, they were asked to create a many-to-many linking between their extracted DAs and abstractive sentences to show which DAs support the abstract.

For the ICSI meeting corpus, prior work has mainly used a test set of six meetings: *Bed*{*004,009,016*}, *Bmr*{*005,019*}, and *Bro018*. Although the number of annotations for each meeting varies, there are three complete annotations from the same subjects for each instance of the test set.

2.2. AMI Meeting Corpus

Whereas the ICSI meetings are natural or "non-scenario", the AMI meeting corpus [3] consists of both non-scenario and scenario meetings. For the meeting summarization experiments, only the scenario meetings were used. In these, four participants play different roles in a fictional company and talk about the design and realization of a new kind of remote control. Although the topic was predetermined, the speech and actions are considered to be spontaneous and natural as the actors were not given any special instructions. Analogous to the ICSI corpus, the meetings were transcribed and annotated with abstractive and extractive summaries.

For the AMI meeting corpus, a test set of 19 meetings was defined, namely the series *ES2004, ES2014, IS1009, TS3003* and *TS3007* (additional training and evaluation sets are specified in the AMI documentation). In contrast to the ICSI meeting corpus, the number of annotators varies from one meeting to the next.

3. Evaluation and Related Work

Evaluating the quality of a summary is non-trivial. Human evaluations at NIST's Document Understanding Conferences (DUC[1]) involve multiple judges scoring summaries in a variety of categories. Such manual assessment is slow, costly, and not necessarily reproducible since the criteria are subjective. Automatic evaluation, then, is a valuable commodity, both for

[1] details at: http://duc.nist.gov

text and speech summarization. Two of the most common such methods, as used in meeting summarization, are described here.

3.1. ROUGE

Recall Oriented Understudy for Gisting Evaluation (ROUGE) compares a summary to a set of human-generated, typically abstractive, summaries and counts overlapping n-grams (see [4]). ROUGE-2, with $n = 2$ (bigrams), is the current standard for DUC. ROUGE-1, ROUGE-2, and other variants, have been shown to produce summary rankings highly correlated with human rankings for multi-document newswire text cases[2]. ROUGE-2 is specified as

$$\mathrm{R} = \frac{\sum_t |\{b_h\} \cap \{b_t\}|}{\sum_t |\{b_t\}|} \quad (1)$$

where t is the reference summary, $\{b_t\}$ is the set of *gold bigrams*, that is bigrams appearing in the reference summaries, and $\{b_h\}$ is the set of bigrams appearing in the hypothesis summary. A related method using manually labeled summary content units instead of data driven ngrams is the Pyramid measure [5]. Both ROUGE and Pyramid are designed to compare abstractive summaries.

3.2. Weighted Precision

As extractive summarization can also be seen as a classification problem, recall, precision and F-measure can be appropriate metrics. However, each extracted segment could be linked to the human generated abstracts more than one time. Murray et. al. introduced *weighted precision* (WP) [6], described in detail in [7] as:

$$\mathrm{WP} = \frac{\sum_d w_d}{A \cdot D} \quad (2)$$

where w_d is the number of links for DA d, A is the number of annotators and D the number of extracted DAs. Both recall and precision can be weighted by the number of links between the extracted DA and the abstractive summary, and are normalized by the number of human annotators.

3.3. Previous Work

Despite considerable research in extractive meeting summarization, results remain hard to compare due to different experimental setups and evaluation criteria. Table 1 gives an overview of the varied evaluation conditions used in recent work.

In general, length constraints are either defined in terms of words or DAs. Word constraints can be somewhat unnatural near the cutoff, but DA constraints are more problematic because they provide an advantage to selection of long DAs.

As meetings have different lengths and some may contain more information than others, a variable summary length (a percent of the meeting length) is preferable to a fixed length summary for all meetings. Previous work employing variable summary length uses the ratio of meeting length to human extractive summary length to determine a word-count or DA-count goal.

Given the definitions of ROUGE and WP, choosing the target summary length constraint and the reference summaries is not a simple task as it directly affects the evaluation scores. For ROUGE, which is designed to compare (abstractive) summaries of about the same length, longer generated summaries have higher recall as more ngrams can be included. Similarly for WP, the maximum score is determined by the maximum number of DAs allowed, which can be greater than 1 due to the weighting. Future work might involve automatic selection of summary length.

When comparing the automatic extractive summaries to the human extracts one could argue that recall, precision and F-measure (and the related WP) are sufficient to measure the system performance. However, there might be DAs which contain similar information but are not selected by the human annotators. Therefore, the ROUGE scores of the extractive summaries should always be considered as it aims for salient *content* rather than salient *DAs*.

Article	Data	Ref. type	Length	Evaluation
[8, 9]	ICSI	E	10% DA	R
[6]	ICSI	L	350 W	WP
[6]	ICSI	A	350 W	R
[10]	ICSI	E	12.7% W	R,P
[11]	ICSI	L	700 W	WP
[11]	AMI	L	700 W	WP
[7]	AMI	L	$>$ 700 W	WP
[12]	ICSI	E	16% W	R
[12]	ICSI	E	4.2% DA	R
[13]	ICSI	E	4.2% DA	R

Table 1: *Evaluation in previous works ("A" – abstractive summary, "E" – extractive, "L" – linked subset of extracted DAs, "R" – ROUGE, "P" – Pyramid, "WP" – weighted precision).*

4. Summarization as a Knapsack Problem

Despite varied task conditions, every summarization approach can be roughly formulated as an attempt to maximize some score over possible summaries S subject to a length constraint L:

$$\begin{aligned} &\text{maximize} && score(S) \\ &\text{subject to} && length(S) \leq L \end{aligned} \quad (3)$$

Extractive summarization then, in which we assign value to each DA (with number of words as weight or cost), can be seen as a classic NP-hard knapsack problem (see [14]). Obviously, the optimum solution to this optimization problem is the best extractive summary according to the given evaluation measure. In recent work, McDonald uses a similar approach in text summarization [15].

If we take ROUGE-2 (Eq. 1) as an evaluation measure, we can reformulate Eq. 3 as

$$\begin{aligned} &\text{maximize} && \sum_i w_i \cdot \max_j b_{ij} x_j \\ &\text{subject to} && \sum_j l_j \cdot x_j \leq L \end{aligned} \quad (4)$$

where x_j indicates whether DA j is included in the summary, b_{ij} is an indicator for the presence of gold bigram i in DA j, w_i is the normalized document frequency of gold bigram i, l_j is the length of DA j, and L, as before, is the maximum summary length in words. The max in Eq. 4 ensures that each bigram is only accounted for once, regardless of the number of DAs it appears in.

In short, we value each input bigram according to the number of reference summaries it appears in. Then, we are looking for the set of DAs that gives the maximum total value subject to the summary length constraint, with the caveat that each included bigram only adds its value once. The solution to this problem directly optimizes ROUGE-2.

[2] ROUGE, of course, can only measure content. Whether a summary is coherent is a different matter, though sentence extraction ensures somewhat reasonable grammar and sense.

In the case of weighted precision (Eq. 2) as an evaluation measure, we can reformulate Eq. 3 as:

$$\begin{aligned} \text{maximize} \quad & \frac{1}{A} \cdot \frac{\sum_j x_j \cdot w_j}{\sum_j x_j} \\ \text{subject to} \quad & \sum_j l_j \cdot x_j \leq L \end{aligned} \tag{5}$$

where x_j is a binary indicator for whether DA j is extracted, l_j is the length of DA j and w_j indicates how many times DA j was linked to the abstractive summaries by the A different annotators.

Unfortunately, these formulations, except in degenerate cases, have greater complexity than the classic knapsack problem. To see this, consider the special case of the ROUGE formulation in which each gold bigram appears exactly once in the input. Then, the total value of each DA is simply the sum of the values of its bigrams and independent from the value of any other DA in the summary. In this case, our problem reduces to the classic knapsack problem and can be solved with a dynamic program that takes advantage of optimal substructure. However, in the general case where the bigrams in the DAs interact, there is no inductive procedure for building an optimal summary step by step. We affectionately refer to this kind of packing problem as a *knapsack of knapsacks*, since we are trying to pack a summary full of DAs, each of which contains a fixed set of valued items. We cannot solve this problem exhaustively since there are too many possible summaries so in the following section, we outline a local search method for approximating optimal solutions.

5. Experiments for a Common Ground

Table 1 shows a variety of task conditions and evaluations, making explicit the difficulty in meaningful comparison. To address this issue, we suggest two simple, easily computable baselines, as well as a genetic algorithm for approximating oracle solutions. Presenting new results alongside these standards as computed for the desired task conditions will aid progress in the field of meeting summarization.

5.1. Baselines

Perhaps the simplest way to perform extractive summarization is to select DAs at random. Besides being relatively difficult to beat (a random sentence selector, on average, outperformed at least five of the systems submitted to DUC 2007), it is intuitively useful to compare results to a random baseline. In our implementation, we select DAs at random (without replacement) until the length constraint is satisfied. If a selected DA is too long, we try again, and stop if no DA fits in the remaining space. We generate 30 summaries for each meeting in this fashion and report the average score.

In creating a second baseline, we consider the question: what is the simplest system that might work? Since DA length is correlated with relevance, our baseline greedily selects the longest remaining DA until the length constraint is satisfied. Longer DAs tend to include more information, however, the much shorter backchannels and other dialog information are more unlikely to be chosen.

5.2. Oracle Summaries

As discussed above, finding optimal extractive summaries is prohibitively difficult to do exactly. To approximate these solutions, we use a genetic algorithm, shown in Figure 1. Starting from a greedy initialization using the evaluation measure, we found that setting $N = 5000000$, $H = 5000$ and $k = 1$ leads to stable results. While the results are nearly identical from one run to another across all problems, suggesting that we are near the global maximum, we have no way of knowing for sure. Future work can address the topography of summarization space.

Input: DAs U, constraint L, evaluation measure E
Output: near-optimum summary
Greedy initialize a list of H good hypotheses
for N *iterations* **do**
  Randomly select and copy hypothesis and save as h
  Randomly remove k DAs from h
  Randomly add DAs from U to h w.r.t. L
  Add h to H
  Remove worst hypothesis from H according to E
end
return *Best hypothesis according to* E

Figure 1: *Genetic program for an oracle solution*

5.3. Data

We use the AMI and ICSI test sets as described in Section 2. From the officially available data, we used the manual transcription, its segmentation into DAs and the annotated summaries (abstractive, extractive and linking information). We use a basic Perl routine to remove disfluencies from the DAs[3], which is similar to preprocessing used in much of the work in this field.

5.4. Experimental Results

Tables 2 and 3 show the results obtained using the three different systems and the length constraints from Table 1. Table 2 is ROUGE-based, and Table 3 is based on weighted precision.

#	Data	RT	Len	Rand	Base	MaxR
1	ICSI	A	350 W	0.04	0.06	0.16
2		E	12.7% W	0.23	0.34	0.46
3			16% W	0.28	0.40	0.55
4			4.2% DA	0.08	0.41	0.50
5			10% DA	0.18	0.64	0.80

Table 2: *Random (30 trials, column "Rand"), baseline ("Base") and maximum ROUGE ("MaxR") results using ROUGE-2 (recall) under different references ("RT",* A *– abstractive,* E *– extractive) and summary constraints ("Len").*

The low ROUGE scores for experiment 1 are due to the evaluation scenario: The extracted DAs are spontaneous speech but the references are the abstractive summaries of human annotators which often use words that never appear in the input. The random system performance is weak as expected since many of the DAs are backchannels. This effect is particularly striking when the length constraint is on DAs instead of words. The DA-length baseline gives good performance, a result somewhat analogous to the simple yet efficient method in multi-document summarization of selecting the first sentences of each document until the length constraint is satisfied.

The maximum result of 1.75 in experiment 8 and the decrease from experiment 6 (1.42) to 7 (1.20) can be explained by revisiting Eq. 2. The weighted precision can be greater than 1 if for example every DA was correctly classified and had more than one link to the abstract. Also, relaxing the length constraint from 350 to 700 words allows more DAs to be selected.

[3] `http://icsi.berkeley.edu/~koried/rd.pl`

#	Data	RT	Len	Rand	Base	MaxWP
6	ICSI	L	350 W	0.10	0.49	1.42
7			700 W	0.10	0.49	1.20
8	AMI	L	700 W	0.21	0.71	1.75

Table 3: *Random (30 trials, column "Rand"), baseline ("Base") and maximum weighted precision ("MaxWP") results using weighted precision under different summary constraints ("Len") and the linked subset of the extractive summaries.*

For example, consider two summary lengths (1 and 4 DAs) and 4 DAs, 1 with 5 links and 3 with 1 link. The optimal weighted precision result for 1 DA is obtained by selecting the DA with link count 5, thus yielding WP $= 5$. However, when selecting 4 DAs, the weighted precision drops to WP $= 2$ without decreasing summary quality.

6. Conclusion

We have addressed two key issues in extractive meeting summarization: First, the oracle results in Section 5 suggest that there is still considerable room for improvement in extractive meeting summarization given the current evaluations. Although our maximum ROUGE and maximum WP scores are quite high, some human evaluation will be necessary to see whether these are, in fact, good summaries. Second, results of previous work in extractive meeting summarization are hard to compare due to different experimental setups and evaluation criteria (see Section 3.3). Using the baselines and oracle system we described, new results can be published alongside upper and lower boundary values, thus providing both intuition and meaningful comparison with other research.

However, this is only a first step. System performance depends strongly on a few addition factors. Some work uses manual transcriptions while other work uses automatic speech recognition output. Sometimes DAs are manually annotated, while in other cases, the DAs are tagged automatically. Beyond these initial steps, there are many ways to preprocess the data (e.g. stopword and disfluency removal). All of these variations make it difficult to reproduce and compare results.

Beside improving extractive summarization, it is important to note that whereas these summaries for text are usually quite readable, the same is not necessarily true for speech. Selecting DAs, at least within the ROUGE or weighted precision frameworks, takes place without consideration of context, and often they make little sense on their own. Also, as Table 2 shows, it is more difficult to compare speech extracts to human abstracts as is done for text summarization since spontaneous speech and planned, composed sentences are fundamentally different.

Although extractive meeting summaries are not particularly satisfying in of themselves, they have been successfully integrated into a meeting browser (e.g. [16]) where the extracts can be used to find areas of interest within a meeting. A user can thus use extracts as part of a tool for reviewing meetings. However, we could also use extractive summaries as an intermediate step toward better abstractive summarization. The extracted DAs, while not a finished product, indicate important pieces to include in a summary. On this basis, the application of abstractive summarization techniques like automatic reformulation to generate indirect discourse, or textual entailment can be applied to produce coherent summaries.

Finally, the experiments based on the knapsack problem required manual annotations to derive the weights used in the optimization. But these weights might be automatically estimated from the data using textual or prosodic features. In future experiments, we would like to use such features to estimate weights automatically.

7. Acknowledgments

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811).

8. References

[1] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," in *Proc. ICASSP, Hong Kong*, 2003.

[2] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus," in *Proc. 5th SIGDAL*, 2004.

[3] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A Pre-Announcement," in *Proc. MLMI*, ser. LNCS, S. Renals and S. Bengio, Eds., no. 3869. Springer-Verlag, 2005, pp. 28–39.

[4] C. Lin, "ROUGE: a Package for Automatic Evaluation of Summaries," in *Proc. ACL Text Summarization Workshop*, 2004.

[5] A. Nenkova and R. Passonneau, "Evaluating Content Selection in Summarization: The Pyramid Method," in *Proc. HTL/NAACL*, 2004.

[6] G. Murray, S. Renals, J. Moore, and J. Carletta, "Incorporating Speaker and Discourse Features into Speech Summarization," in *Proc. HLT/NAACL, New York City, USA*, 2006.

[7] G. Murray and S. Renals, "Towards Online Speech Summarization," in *Proc. Interspeech, Antwerp, Belgium*, 2007, pp. 2785–2788.

[8] G. Murray, S. Renals, and J. Carletta, "Extractive Summarization of Meeting Recordings," in *Proc. Interspeech, Lisboa, Portugal*, 2005.

[9] G. Murray, S. Renals, J. Carletta, and J. Moore, "Evaluating Automatic Summaries of Meeting Recordings," in *Proc. 43rd Annual Meeting of the ACL, Ann Arbor, USA*, 2005.

[10] M. Galley, "A Skip-Chain Conditional Random Field for Ranking Meeting Utterances by Importance," in *Proc. ACL/EMNLP*. Sydney, Australia: Association for Computational Linguistics, 2006, pp. 364–372.

[11] G. Murray and S. Renals, "Term-Weighting for Summarization of Multi-Party Spoken Dialogues," in *Machine Learning for Multimodal Interaction IV*, ser. LCNS, A. Popescu-Belis, S. Renals, and H. Bourlard, Eds. Springer, 2007, vol. 4892, pp. 155–166.

[12] Y. Liu and S. Xie, "Impact of Automatic Sentence Segmentation on Meeting Summarization," in *to appear in Proc. ICASSP, Las Vegas, USA*, 2008.

[13] S. Xie and Y. Liu, "Using Corpus and Knowledge-Based Similarity Measure in Maximum Marginal Relevance for Meeting Summarization," in *to appear in Proc. ICASSP, Las Vegas, USA*, 2008.

[14] R. Karp, "Reducibility Among Combinatorial Problems," *Complexity of Computer Computations*, vol. 43, pp. 85–103, 1972.

[15] R. McDonald, "A study of global inference algorithms in multi-document summarization," in *Proc. ECIR 2006*, 2007.

[16] G. Murray, P. Hsueh, S. Tucker, J. Kilgour, J. Carletta, J. Moore, and S. Renals, "Automatic Segmentation and Summarization of Meeting Speech," in *NAACL/HTL Demonstration Program, Rochester, NY, USA*, 2007, pp. 9–10.