A GLOBAL OPTIMIZATION FRAMEWORK FOR MEETING SUMMARIZATION

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

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

ABSTRACT

We introduce a model for extractive meeting summarization based on the hypothesis that utterances convey bits of information, or *concepts*. Using keyphrases as concepts weighted by frequency, and an integer linear program to determine the best set of utterances, that is, covering as many concepts as possible while satisfying a length constraint, we achieve ROUGE scores at least as good as a ROUGE-based oracle derived from human summaries. This brings us to a critical discussion of ROUGE and the future of extractive meeting summarization.

Index Terms— meeting summarization, integer linear programming, summarization evaluation

1. INTRODUCTION

Meetings have always been too long. Meeting summarization attempts to distill the most important information from a recorded meeting into a short textual passage for the benefit of both participants and non-participants. Most systems perform some selection of relevant segments, which has proven successful in document summarization, and is considerably easier than abstractive language generation.

Extractive summarization is typically expressed as a combination of two simultaneous goals: maximizing the information covered and minimizing redundancy. Previous approaches tend to rank utterances by relevance, selecting as many as possible within the length constraint. Redundancy is addressed by pruning out utterances too similar to those already selected. *Maximal marginal relevance* (MMR) [1, 2, 3] is a good example of a greedy approach of this kind. Other work, such as [4], does not consider redundancy and only addresses the problem of selecting relevant utterances.

One main problem with MMR is non-optimality. During the greedy search, the selection of the next utterance depends strongly on those chosen so far, and the more utterances available, the more sub-optimal the greedy approximation is likely to be. [5] studies the prospect of replacing this greedy search with an optimal formulation. Given some general definition of relevance and redundancy, the basic MMR framework can be expressed as a quadratic knapsack packing problem. An *integer linear program* (ILP) solver can be used to maximize the resulting objective function, which searches efficiently over the large space of possible summaries for an optimal solution.

Here, we consider a technique for utterance selection that is based on the hypothesis that utterances contain independent units of information, or *concepts*. These are defined so that the quality of a summary, at least in terms of its content, can be measured by the total value of unique concepts it contains. Redundancy is limited implicitly by the length constraint. Most prior work, including [5], uses utterance-level relevance and an explicit redundancy model. More specifically, we show how to implement the proposed model using an ILP and how to use keyphrases (KP) as concepts in this framework. Experiments on the AMI meeting corpus show that the new model significantly outperforms MMR. Moreover, the ILP formulation can be intuitively extended to account for meeting-specific constraints.

2. CONCEPT-BASED SUMMARIZATION

Summarization models commonly assign value to a summary as the sum of the values of utterances it contains. Such an approach assumes that utterances are independent in terms of informativeness, but in reality, utterances often share information in the form of pronoun coreference, repetitions, and re-statements, for example. The idea of assigning a score to a summary as the sum of independent pieces is not bad in itself, but using utterances as an atomic unit is problematic. The model we present here defines *concepts* as minimal independent pieces of information. Summing the values of a unique concept set gives a global summary score. Utterances can refer to multiple concepts and concepts can be referred to by multiple utterances. To fully specify this model, we need only define a function that maps the input to valued concepts. For the sake of generality, we withhold this specification until the next section. According to our model, we seek a summary that maximizes a global objective function:

$$\text{maximize} \qquad \sum_i w_i c_i \tag{1}$$

where w_i is the weight of concept i and c_i is a binary variable indicating the presence of that concept in the summary. The score of a summary is the weighted sum of the concepts it contains. This function gives a selection over concepts while we are interested in a selection over utterances. Thus, we introduce u_j, a binary variable representing the selection of utterance j for the summary. Next, we add a length constraint:

$$\text{subject to} \qquad \sum_j l_j u_j < L \tag{2}$$

where l_j is the length of utterance j and L is the desired summary length. Now we need to tie utterances and concepts together to maintain consistency. A concept can be selected only if it is referred to in at least one selected utterance and an utterance can be selected only if all concepts it refers to are selected. Formally, this can be

represented as two types of constraints:

$$\sum_j u_j o_{ij} \geq c_i \forall i \quad (3)$$

$$u_j o_{ij} \leq c_i \forall i, j \quad (4)$$

where o_{ij} represents the occurrence of concept i in utterance j. While this can lead to $O(n^2)$ constraints, in practice, $o_{ij} = 0$ for most of the concept-utterance pairs, keeping the number of effective constraints quite low. Lastly, we formalize the variables introduced above, c_i and u_j:

$$c_i = 0 \text{ or } 1, \forall i \qquad u_j = 0 \text{ or } 1, \forall j \quad (5)$$

This formulation is an *integer linear program*, a single maximization term subject to a number of linear constraints on integer-valued variables. While the ILP problem is NP-complete, considerable optimization research has produced software for solving instances efficiently[1].

Note that there is no explicit redundancy term in this formulation. Instead, redundancy is limited implicitly by the fact that concept values are only counted once, combined with a length constraint that prefers utterances with high concept density. Moreover, the solver usually finds an exact solution to the problem very quickly, depending on the choice of concepts.

3. KEYPHRASE EXTRACTION

Concepts should represent pieces of information, such as a decisions made in a meeting or the opinion of a participant on a given topic. However, such abstract concepts are difficult to extract automatically, so we experiment with a much simpler set of concepts: content words. The ILP formulation above can find the summary that maximizes value, given some function mapping words to weights. We have shown in previous work [3] that simple n-grams often overlap with discourse markers ("sort of", "you know") which can add noise to the process. Thus we have proposed a keyphrase extraction algorithm that is quite successful at detecting word sequences representative of content. The algorithm and improvements compared to [3] are detailed below.

1. Extraction: All content word n-grams g_i for $n = 1, 2, 3$
2. Noise reduction: Remove n-grams appearing only once or as often as enclosing ones, e.g. remove "manager" if frequency matches "dialogue manager".
3. Bigram and trigram re-weighting: $w_i = \text{frequency}(g_i) \cdot n$, where w_i is the final weight and n is the n-gram length.

Though rather simple, this algorithm does not require additional annotation and training data to find n-grams of variable length and turned out to be fairly robust in the presence of spontaneous speech phenomena. In previous work, the content words were limited to adjectives and nouns included in the WordNet database [6] minus a list of 501 stopwords. This idea, though a reasonable attempt to exclude irrelevant words that often appear in the meeting domain and focus on topic-related noun phrases, lacks word sense disambiguation (e.g. "change" can be used as a noun or a verb). Instead, we use a part of speech (PoS) tagger based on a Hidden Markov Model, trained on broadcast news [7, 8] and modify the keyphrase algorithm given above to allow only words tagged as numbers (*CD*), foreign words (*FW*), adjectives (*JJ, JJR, JJS*) and nouns (*NN, NNS, NNP, NNPS*). As shown in Figure 1, the tagger works fairly well on spontaneous meeting speech, resulting in disambiguated keyphrases.

using WordNet:
Especially the important buttons, if you want to switch channel, change your volume, use teletext, it— it has to work at once.

using PoS tags:
Especially/RB the/DT important/JJ buttons/NNS if/IN you/PRP want/VBP to/TO switch/VB channel/NN, change/VB your/PRP volume/NN, use/VBD teletext/RB it/PRP— it/PRP has/VBZ to/TO work/VB at/IN once/RB.

Fig. 1. Example from *TS3007b* showing the benefit of using a PoS tagger in contrast to WordNet (extracted keyphrases underlined). Note that "teletext" is mis-tagged as RB (adverb).

As a side effect of our modeling choice, we can easily modify the concept weighting algorithm to produce maximum ROUGE "oracle" summaries. ROUGE [9] approximates summary quality by measuring n-gram overlap with a set of reference summaries[2]. Oracle summarization simply involves replacing the input frequency heuristic with n-gram concepts weighted by the number of human-generated reference summaries in which they appear. This method is proposed in [10] for defining ROUGE performance boundaries, though a non-optimal search technique was used, which we replace with the ILP formulation.

4. EXPERIMENTAL SETUP AND EVALUATION

4.1. Data

For our experiments, we use the AMI corpus test set [11] consisting of 20 meetings in these series: *ES2004, ES2014, IS1009, TS3003* and *TS3007*. In each meeting, 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. All meetings were transcribed and annotated with an abstractive summary of an average of about 290 words (roughly 6% of the words) covering the general intent of the meeting, issues discussed, actions to be taken, and decisions made.

4.2. Systems

We show results for the keyphrase systems and the oracle, along with a baseline (selecting the longest utterances until the length constraint is satisfied) and MMR [12] based systems. To confirm the gains from using keyphrases instead of a document centroid for MMR for this corpus, we conduct experiments using cosine and keyphrase similarity measures as detailed in [3].

Preliminary experiments suggested using the top 50 keyphrases for MMR and all keyphrases for the ILP-based system. This difference further indicates the disadvantages of MMR, which requires more fine-tuning–there is also an α parameter that balances query relevance with redundancy, and must be tuned manually. We used the α that gave the best test set performance to make the comparison with the ILP system as competitive as possible.

[1] We use the open source solver from http://www.gnu.org/software/glpk/

[2] ROUGE-1: unigrams, ROUGE-2: bigrams

ROUGE-1	R	P	F
baseline	0.12	0.22	0.15
MMR/centroid	0.18	0.27	0.21
max. ROUGE	0.27	0.33	0.29

Table 1. ROUGE-1 scores (Recall, Precision, F-measure) for the baseline, centroid based MMR, and the maximum ROUGE oracle.

	WordNet			PoS tags		
ROUGE-1	R	P	F	R	P	F
MMR/cosine	0.21	0.32	0.25	0.23	0.33	0.26
MMR/kp-sim	0.21	0.30	0.24	0.22	0.32	0.25
ILP/unique	0.25	0.32	0.28	**0.27**	**0.33**	**0.29**
ILP/each-spkr	**0.26**	**0.33**	**0.28**	0.26	0.33	0.29
ILP/all	0.22	0.28	0.24	0.22	0.28	0.24

Table 2. ROUGE-1 scores for MMR and ILP summarizers using keyphrases based on WordNet and PoS tags.

To analyze the performance of the ILP system, we study 3 variations allowing different amounts of redundancy:

1. As described above, award points for including a keyphrase only on its first occurrence (system "ILP/unique").
2. As important keyphrases are common, award points for including a keyphrase once for every speaker (system "ILP/each-spkr").
3. As a keyphrase might be persistent over the whole meeting, award points for every inclusion, thus ignoring the most important constraint on redundancy (system "ILP/all").

4.3. Evaluation

Using the systems described above, we generate extracts with lengths limited to 6% of the number of words in the original meeting (around 290 words per summary, as in the human abstracts). To evaluate performance, we use the ROUGE toolkit [9] which correlates well with human rankings of summary quality [13]. We show ROUGE-1 scores (unigram overlap) since spontaneous speech tends to overlap with abstracts much more consistently in unigrams than in bigrams. We use the toolkit's built-in option to ignore stopwords to reduce the impact of non-content overlap.

4.4. Results

Table 1 shows the results for the baseline, the best centroid-based MMR system, and the maximum ROUGE oracle. While the centroid-based MMR clearly outperforms the baseline, it still is far from reaching the oracle results. Table 2 shows the ROUGE-1 scores achieved by the systems using both old and new keyphrase algorithms. As was the case for ICSI meeting data, MMR using keyphrases significantly outperforms the document centroid in terms of ROUGE. To help understand the performance gap, we note that the optimal relevance parameter for the centroid system is around $\alpha = 0.9$, compared with $\alpha = 0.5$ for the keyphrase systems. This suggests that the keyphrase query is focused enough to allow an even mixture of relevance and non-redundancy, whereas the centroid is too general to capture relevance.

The new ILP-based systems increase performance dramatically, so long as some notion of concept redundancy is maintained. This result neatly demonstrates the effectiveness of the implicit redundancy constraints built into the ILP. Without it, the resulting summaries repeat a few common keyphrases, providing poor coverage of the meeting, and low ROUGE scores.

The right hand side of Table 2 shows results using the revised keyphrase extraction based on PoS tags. The differences are not significant, but the new keyphrase algorithm is more intuitively satisfying and works at least as well as the WordNet version. Lastly and most remarkably, the *ILP/unique* system achieves ROUGE results indistinguishable from the maximum ROUGE oracle in recall, precision and F-measure.

Figure **??** shows examples for a human abstract and the generated ILP and oracle summaries. Note that due to stemming in ROUGE and sentences selected by the oracle might not show direct word overlap with the reference.

One important observation regarding these examples is that the extracts tend to have a much lower information density relative the human abstracts. This is because the meetings contain spontaneous speech which is unlikely to convey any information succinctly. Increasing the length constraint in order to improve coverage would be counterproductive as it also increases the time needed to read the summary. Deeper information analysis, fusion and reformulation are needed in order to achieve such density. For instance, a study of the structure of the argumentation between speakers could be used to isolate and emphasize important issues. Or, an analysis of dialogue types could distinguish action items in meetings. Such tasks are of course quite difficult even with pure text, and probably more challenging in the meeting domain.

5. WHAT'S NEXT FOR MEETING SUMMARIZATION?

Perhaps the most notable result presented in this work is that the proposed KP/ILP system actually achieves ROUGE-1 scores that match the oracle (though selected sentences are different). While this is a nice result, indicating the success of our model and the keyphrase algorithm, meeting summarization is far from perfect. As the example summaries in Figure **??** show, either the use of ROUGE as a performance measure or the use of extraction for summarization needs to be rethought. While the automatic and oracle summaries seem to refer to important information from the meeting, they both lack the structure, coherence, and abstraction of the summaries written by human subjects.

We submitted a similar ILP-based system [**?**] for multi-document text summarization to the Text Analysis Conference (TAC) evaluation. Though the TAC "update" task is different from meeting summarization, our system obtained the highest ROUGE-2 scores of all participanting systems. Manually evaluated Pyramid content scores were among the top ten, though linguistic quality scores were somewhat lower. These results are promising, suggesting that our model is useful for many types of summarization tasks.

Finally, we advocate for our particular version of the global optimization approach to summarization because it allows for a lot of flexibility. For instance, it was very easy in our experiments to introduce speaker-specific scoring. By pushing search complexity to the ILP solver, we lower the barrier for researchers new to the field and provide a high performance baseline easy to implement. Nevertheless, approximate solutions to the ILP might be necessary in time-constrained scenarios such as interactive summarization.

6. CONCLUSION

We have introduced a concept-based approach to summarization to overcome the drawbacks of the widely used MMR approach. Whereas MMR iteratively extracts utterances using a greedy search based on query similarity and non-redundancy, our ILP formulation finds the optimal set of utterances covering the most informative concepts. Redundancy is limited implicitly. When these concepts are n-grams weighted by their frequency in the human reference summaries, the resulting extracts correspond to a ROUGE oracle. When concepts and weights are selected using our keyphrase heuristic, the resulting summaries significantly outperform previous MMR summaries as measured by ROUGE. Furthermore, the ILP/KP approach is independent of a manual query and relevance parameter as required for MMR, and using keyphrases as concepts allows intuitive user interaction (as demonstrated in [3]). Still, the resulting summaries are far from perfect, we call for new ways of evaluating summarization and new approaches to supplement extraction.

As for future work, three main issues need to be addressed. First, possible improvements for the ILP system include a more sophisticated notion of concepts, selection of partial or compressed utterances, and improvements in readability through constraints on order. Second, the actual performance of the ILP summaries needs to be validated by human evaluators, and third, the reliability of ROUGE for measuring the quality of extractive meetings needs to be re-assessed.

7. ACKNOWLEDGMENTS

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-Party Interaction, FP6-506811) and DARPA CALO (NBCHD-030010). The opinions and conclusions are those of the authors and not necessarily endorsed by the sponsors.

8. REFERENCES

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

[2] S. Xie and Y. Liu, "Using Corpus and Knowledge-Based Similarity Measure in Maximum Marginal Relevance for Meeting Summarization," in *Proc. ICASSP, Las Vegas, Nv, USA*, 2008, pp. 4985–4988.

[3] K. Riedhammer, B. Favre, and D. Hakkani-Tür, "A Keyphrase Based Approach to Interactive Meeting Summarization," in *submission to SLT2008, Goa, India*, 2008.

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

[5] R. McDonald, "A Study of Global Inference Algorithms in Multi-Document Summarization," in *Proc. European Conference on Information Retrieval*, 2007.

[6] C.Fellbaum, Ed., *WordNet: an electronic lexical database*, MIT Press, 1998.

[7] S. Thede and M. Harper, "A Second-Order Hidden Markov Model for Part-of-Speech Tagging," in *Proc. 28th ACL, Baltimore MD, USA*, 1999, pp. 175–182.

[8] Z. Huang, M. Harper, and W. Wang, "Mandarin Part-of-Speech Tagging and Discriminative Reranking," in *Proc. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic*, 2007.

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

[10] K. Riedhammer, D. Gillick, B. Favre, and D. Hakkani-Tür, "Packing the Meeting Summarization Knapsack," in *Proc. Interspeech, Brisbane, Australia*, 2008.

[11] 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*, Steve Renals and Samy Bengio, Eds. 2005, number 3869 in LNCS, pp. 28–39, Springer-Verlag.

[12] J. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," *Research and Development in Information Retrieval*, pp. 335–336, 1998.

[13] F. Liu, Y .Liu, and B. Li, "Study on Correlation between ROUGE and Human Evaluation in Meeting Summarization," in *Proc. MLMI*, 2007.