A KEYPHRASE BASED APPROACH TO INTERACTIVE MEETING SUMMARIZATION

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

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

ABSTRACT

Rooted in multi-document summarization, maximum marginal relevance (MMR) is a widely used algorithm for meeting summarization (MS). A major problem in extractive MS using MMR is finding a proper query: the centroid based query which is commonly used in the absence of a manually specified query, can not significantly outperform a simple baseline system. We introduce a simple yet robust algorithm to automatically extract keyphrases (KP) from a meeting which can then be used as a query in the MMR algorithm. We show that the KP based system significantly outperforms both baseline and centroid based systems. As human refined KPs show even better summarization performance, we outline how to integrate the KP approach into a graphical user interface allowing interactive summarization to match the user's needs in terms of summary length and topic focus.

Index Terms— meeting summarization, keyword generation, user interaction

1. INTRODUCTION

With an increasing amount of text and speech data available, techniques to assess this huge source of information gain more and more attention. Beside searching, indexing and categorization, summarization can help to find the important bits without reading or listening to the complete document or recording. A summary can be either *abstractive* – expressing the content in newly formulated sentences or *extractive* – by selecting relevant parts. For this work we focus on extracting important utterances of spontaneous meeting speech.

However generating a summary which actually fits the needs of the user is a non-trivial task. In addition to general summary features like length, level of abstraction (i.e. general or detailed) or topic focus which might all be different per user (consider for example an unfamiliar reader compared to a meeting participant), meetings (unlike news, for example) often contain unimportant chit chat, e.g. about the current weather or the latest movie, and dialogue phenomena, e.g. "could you say that again, please" which (usually) do not convey important information. In extractive summarization, this problem can be approached from two sides, by training a system to ignore these kind of utterances on the one hand and to identify important utterances on the other hand. Unfortunately, this approach requires carefully annotated training data and is then limited to the domain it was trained on, e.g. a certain kind of meetings.

In contrast to classification, a popular meeting summarization approach is query based summarization where one tries to find sentences or utterances similar to a specified question or topic which can be done without prior training. One of the simplest yet powerful algorithms is *maximum marginal relevance* (MMR) [1]. A query however requires the user to know in advance what he or she is looking for in the meeting, which is sometimes counter-intuitive: Usually users request a summary because they do *not* know what is in the data. To get a general summary not requiring a manual query, many systems construct a *centroid* representing an "average" sentence or utterance, so the resulting summary covers most of the information of the document or recording without a certain focus. Unfortunately, for meeting speech, such an average utterance also includes previously mentioned off-topic information and dialogue phenomena, which is the starting point for this work: We seek an easy and intuitive way to provide an alternative query that not only improves the general summarization performance but also, in a next step, allows easy user interaction to generate a summary that matches the user's expectations in length and topicality.

After introducing the data set, we describe the summarization systems used in this work. Beside a simple baseline and an oracle, we present a centroid based MMR summarizer. We propose a simple yet robust algorithm to extract keyphrases (KP) from a meeting which can then be used as a query for MMR summarization. We show that given these KPs, a significantly better summary can be extracted. Furthermore, our experiments indicate that human refined KP yield even better performance. Extending this idea, we outline how the KP approach can be used to allow interactive summarization to match the user's needs in terms of length and topic focus. We end with a discussion and an outlook on future work.

2. DATA

The ICSI meeting corpus [2] consists of 75 naturally occurring meetings (that is, they would have taken place regardless of the recording project), each around 45 minutes long. They have been transcribed and annotated with dialog acts and abstractive (i.e. freely formulated) summaries [3]. For the latter, annotators were given a graphical user interface which allowed to browse the aligned audio and transcriptions, and were asked to write summaries about the meeting in general as well as progress, decisions and problems discussed in the meeting.

We follow prior work on the ICSI corpus, and use 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. For this test set, the average length of an abstractive human summary is about 400 words.

3. SUMMARIZATION APPROACH

3.1. Baseline and Oracle Systems

In our previous work on meeting summarization, we suggested to place results of new systems in context to results of well specified baseline and oracle systems. Following the setup in [4], the *baseline* (system 1) selects the longest possible utterance until the length constraint is satisfied. The *oracle* (system 5) is the maximum ROUGE

oracle selecting the best possible utterances according to ROUGE-1 (see Section 4.1) and the human abstracts.

3.2. Maximum Marginal Relevance (MMR)

MMR [1] is widely used in extractive speech and meeting summarization [5, 6, 7]. In an iterative process, MMR selects utterances most relevant to a given query q while avoiding utterances redundant to the already selected ones. In each iteration, every utterance u_i is assigned a score c_i which is a weighted combination of its similarity to the query and to the pool of already selected utterances S:

$$c_i = L \cdot sim_1(u_i, q) - (1 - L) \cdot sim_2(u_i, S) \quad (1)$$

where L is the relevance parameter determining the trade-off between query and pool similarity and sim is a function to measure similarity between utterances. The highest scoring utterance is then selected and the process repeated until the desired length is reached.

In the implementation used in this work, we normalize query and utterance similarities to be within $[0; 1]$. To minimize the effect of L on the comparison of the different systems, we report results with the best L value for each algorithm (determined on the test set). Furthermore, we set $sim_2(u_i, S) = max_{u_j \in S}\, sim_2(u_i, u_j)$. The two similarity functions used in this article are introduced below.

3.3. Cosine Similarity and Centroid Vector

A common approach is to build up a feature vector for each utterance and then define a distance measure to compare these. As mentioned in the introduction, one drawback of query based summarization is that defining a query already requires a certain knowledge about the data. If the user is not familiar with the data or is interested in a more general summary, a widely used approach is to generate an artificial query vector by computing an average over all utterance vectors, the so called *centroid* vector. Intuitively, utterances similar to the centroid contribute to the tenor of the overall meeting.

In a similar setup to [6] and other related work, *MMR-centroid* (system 2) applies MMR using cosine similarity (for sim_1 and sim_2) with the centroid vector as query. To obtain the best performance, preliminary experiments on the test set suggested to set $L = 0.1$ and use term frequencies normalized by the overall number of words as term weights (while no stopword list was used). No IDF was used as it turned out to be a non-trivial task to estimate proper IDFs for the ICSI meeting data. Additionally, utterances shorter than 10 words are discarded for all MMR summaries.

3.4. Keyphrases (KP) and Keyphrase Similarity

Although the centroid method is useful for getting a gist of what happened in a meeting, it is distorted by conversational speech artifacts (e.g. "sort of", "like"), off-topic utterances (e.g. "sorry, could you repeat that please") and disfluencies. Hence, a better way of choosing the query would clearly help to improve the quality of the summary. Other than a centroid, keywords or keyphrases (i.e. n-grams of words, e.g. "presidential campaign") can represent the tenor of a document or meeting very well. Prior work on keyword extraction and related summarization in the text domain includes approaches mainly based on frequency information [8], latent semantic and topic related approaches [9], and lexical chains [10, 11]. On speech data (e.g. broadcast news, meetings), recent works combine simple textual features with more speech specific ones (prosody, recognition confidence) and semantic verification [12] or genetic algorithms trained to identify KPs of variable length [13]. For meeting data, [14] use a POS tagger to extract simple nouns and compute frequency and semantic similarity measures (using WordNet and EDR) to refine their results. However, systems using more than simple frequency information to extract keywords or KPs rely on annotated data for analysis and training which is not available for conversational meeting speech. Therefore, our approach is more related to [8, 13], combining a score derived from simple KPs of variable length with a redundancy score, which is explained in detail below.

To extract the KPs, we follow a combined heuristic and semantic approach:

1. Extract all n-grams g_i for $n = 1, 2, 3$ but only allow instances which are made up by *content words* and do not contain *stopwords*. As *content words*, we consider adjectives and nouns from the WordNet database [15]. For *stopwords*, we use a manually edited stopword list consisting of 501 words, covering common text stopwords as well as conversational speech phenomena (e.g. "hm", "ahm", "ooops").
2. To reduce noise, remove n-grams which appear only once or are fully enclosed by longer n-grams sharing the same frequency, e.g. remove "manager" in presence of "dialogue manager" if their frequencies match.
3. To ensure a fair weighting of longer n-grams, re-weight all remaining n-grams by $w_i = \text{frequency}(g_i) \cdot n$, where w_i is the final weight of n-gram i and n is the n-gram length.

We choose this rather simple approach of extracting keyphrases for two reasons. First, the extraction turns out to be fairly robust against spontaneous speech artifacts and second, it allows us to come up with simply generated KPs of variable lengths without the need of extra annotations or training.

Once the KP are extracted, we need to find utterances which contain these. In preliminary experiments on the test set, word overlap and cosine similarity did not yield convincing results, mainly due to the fact that the first does not account for weighted KPs and the second relies on the quality of the term weights. To compute a more characteristic measure, we propose the following KP similarity sim_{kp} which can be computed as

$$sim_{\text{kp}}(u) = \sum_i \text{occ}(g_i, u) \cdot w_i \quad (2)$$

where $\text{occ}(g_i, u)$ is the number of occurrences of n-gram g_i in utterance u, and w_i is the weight of the n-gram. Thus, the more and better KPs an utterance contains, the more important it is. For inter-utterance similarity sim_{u}, we choose the number of overlapping words normalized by the maximum utterance length

$$sim_{\text{u}}(u, v) = \frac{|U \cap V|}{max(|U|, |V|)} \quad (3)$$

where $|U|$ and $|V|$ are the sets of non-stopword words the utterances u and v are composed of.

MMR-autoKP (system 3) applies MMR using $sim_1 = sim_{\text{kw}}$ and $sim_2 = sim_{\text{u}}$. Again, to obtain best performance, preliminary experiments on the test set suggested to set $L = 0.1$ and use the 50 best automatically generated KPs.

For *MMR-refinedKP* (system 4), three human annotators each manually refined the list of 50 KPs. After reading the abstracts for each meeting, they were asked to remove useless KPs, e.g. "sort" (rooting from "sort of") or "planner" in presence of "action planner". Note that the annotators were not explicitly asked to remove keywords which do not appear in the human summaries. Their multi-rater agreement was $\kappa = 0.44$. Preliminary experiments suggested to set $L = 0.5$ which seams reasonable as the human refined KPs are supposed to be more accurate. We display the average system performance using the three sets of KPs.

	#	ROUGE-1			ROUGE-1/sw		
		F	R	P	F	R	P
baseline	1	.31	.35	.29	.15	.13	.20
MMR-centroid	2	.32	.36	.30	.17	.15	.21
MMR-autoKP	3	.33	.38	.32	.20	.19	.24
MMR-refinedKP	4	**.34**	**.39**	**.33**	**.21**	**.20**	**.25**
oracle	5	.38	.43	.32	.31	.30	.24

Table 1. ROUGE-1 and ROUGE-1/sw (stopwords active) (F)-measure, (R)ecall and (P)recision for the five competing systems.

	ROUGE-1			ROUGE-1/sw		
	2	3	4	2	3	4
1	no	**yes**	**yes**	no	**yes**	**yes**
2		no	**yes**		**yes**	**yes**
3			no			no

Table 2. Table of significant improves, read as "column system significantly outperforms row system".

4. EXPERIMENTAL SETUP AND EVALUATION

Using the five systems introduced above, we generate extractive summaries with a word count of about 5% of the words of the original meeting (that is on average about 400 words as the human abstracts). To evaluate the performance of the different systems, we use the ROUGE toolkit [16] which was shown to correlate with human rankings [17]. We consider ROUGE-1 (uni-gram overlap) as we compare extractive summaries of spontaneous speech to written-language abstracts which are – by construction – expected to show little overlap in bi-grams. Still, the computed scores may be distorted by overlap in non-content words. Therefore, the ROUGE toolkit allows to ignore stopwords and provides a list of these assembled for multi-document summarization. We present both raw scores (ROUGE-1) and scores with stopwords removed (ROUGE-1/sw).

Table 1 shows the ROUGE scores for the five competing systems. On all measures, the systems' performance increases from baseline over centroid to KP based (and of course the oracle). It is however notable, that the *MMR-refinedKP* system 4 shows a higher precision than the *oracle* system 5 (which maximizes recall).

While Table 1 is merely shown to give an intuition for the numbers, Table 2 reveals the more interesting information, that is which system significantly outperforms others. Although better, the centroid based system 2 does not significantly outperform the baseline regardless of the ROUGE stopword option. However, using ROUGE-1/sw, the KP based systems 3 and 4 significantly outperform both systems 1 and 2. This is expected as the the KPs on the one hand do not contain stopwords as opposed to the centroid and are on the other hand designed to catch the important topics of the meeting. Note however that the summaries generated with a non-stopword centroid showed even lower ROUGE-1 scores ($F = .30$, $R = .34$, $P = .28$).

Although the KP based system tends to be better using the human refined KPs, the improvement is not significant in terms of ROUGE but might show a subjective improvement as bad KPs are removed. Also note that each of the human systems significantly outperforms systems 1 and 2 regardless of the stopword option. To illustrate the significant improvement in ROUGE-1/sw, Figure 1 gives an example of the sequential utterance extraction by the *MMR-centroid* and *MMR-autoKP* systems for the *Bed009* meeting as well as the ten highest weighted centroid words and KPs. It is easy to see that the KP based system is more likely to choose more informative utterances.

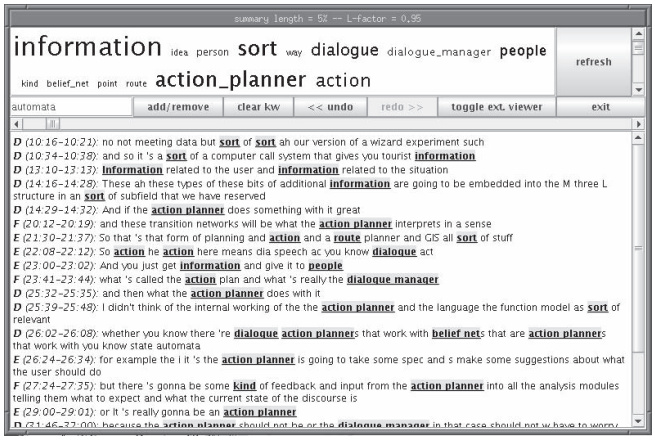

Fig. 2. Screenshot of the interactive meeting summarization tool.

5. INTERACTIVE SUMMARIZATION

The results presented in the previous section indicate that good keyphrases help with building a good summary. However, as mentioned in Section 1, a good summary should satisfy the individual needs of a user in terms of length and topic focus. Therefore, we believe that the user should be in control of these factors. To allow this, we created a graphical user interface (see Figure 2) to

- display KPs according to their weight to give the user an overview of the meeting (realized by different font sizes in the upper text area).
- customize the summary by adding/removing/re-weighting KPs (using mouse selection, buttons and wheel, or by typing in the small text field next to the "add/remove" button), and change the summary length (horizontal slider) and topicality/redundancy trade-off (relevance parameter L, small vertical slider on the top right).
- conveniently display of the summary (lower text area) after its generation ("refresh" button) by highlighting KPs and, in case of sequential summaries, indicate already seen utterances (gray background), and to allow undo/redo of user interactions (see above).

To allow future evaluations of the interface, we added a protocol function keeping track of every user interaction.

6. DISCUSSION AND OUTLOOK

The promising results presented in Section 4 suggest the following conclusions and future work. Though the centroid based system 2 could not significantly outperform the baseline, its performance strongly depends on the chosen term weights. In this work, we solely relied on normalized term frequencies as it is on the one hand difficult to come up with IDFs representative for the conversational speech in the meeting domain and on the other hand independent of additional training data. However, it is interesting to note

centroid words: *something, sort, stuff, problem, way, kind, net, system, decision, time, ...*
A: There's– this is maybe something that this module can do– something that this module can do.
C: So I'm gonna hafta think about it some more.
A: And there she was talking about looking at pictures that are painted inside a wall– on walls.

keyphrases: *action_planner, planner, information, sort, dialogue_manager action, belief_net, dialogue, people, tourist_domain, ...*
C: This is a way of playing with this abs source-path-goal trajector exp abstraction, and and sort of displaying it in a particular way.
C: Yeah so the pro– The immediate problem is what you are doing with the belief net.
A: And so we have for example some information there that the town hall is both a a building and it has doors and stuff like this.

Fig. 1. First three sentences chosen in sequence by the centroid system 2 (top) and the keyphrase system 3 (bottom) for the *Bed009* meeting. The "centroid words" and "keyphrases" lines display the ten highest weighted query elements starting highest first.

that the centroid including stopwords yields better ROUGE-1 scores than when excluding stopwords. This suggests that a proper stopword list accommodating the domain and topics of the meetings is strongly required to exclude spontaneous speech artifacts and off-topic words from the centroid and thus improve the summarization performance.

Generating keyphrases limited to n-grams of non-stopword adjectives and nouns approaches that problem from the opposite side: Using a simple and robust algorithm to extract representative KPs from the meeting allows to automatically generate a query which leads to significantly better summarization performance. Still, more semantic knowledge and machine learning might lead to more robust, less redundant and better weighted KPs. Although showing better performance than the cosine similarity, the KP similarity should be refined to increase the performance, for example by reweighting KPs after each MMR iteration to account for already covered KPs. Finally, the effect of ASR on both KP extraction and summarization has to be investigated.

As human refined KPs yield even better summarization results, in Section 5, we outlined how the KPs can be integrated in a graphical user interface for interactive and iterative summarization. We believe that allowing the user to take control of length and topic focus of the summary is a crucial step towards a satisfying summarization tool. However, a human evaluation is required to verify both usability and efficiency of the program. We suggest to study by what kind of actions (adding/removing KPs, change of summary length, L, KP weight), and in what time a user achieves a satisfying summary or is able to complete a certain task.

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] 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.

[2] 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*, 2003.

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

[4] K. Riedhammer, D. Gillick, B. Favre, and D. Hakkani-Tür, "Packing the Meeting Summarization Knapsack," in *Proc. Interspeech*, 2008, pp. 2434–2437.

[5] X. Zhu and G. Penn, "Summarization of Spontaneous Conversations," in *Proc. Interspeech*, 2006, pp. 1531–1534.

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

[7] 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.

[8] G. Carenini, R. Ng, and X. Zhou, "Summarizing Email Conversations with Clue Words," in *Proc. WWW*, 2007, pp. 91–100.

[9] M. Dredze, H. Wallach, D. Puller, and F. Pereira, "Generating Summary Keywords for Emails Using Topics," in *Proc. Intelligent User Interfaces*, 2008.

[10] G. Silber and K. McCoy, "Efficient Text Summarization Using Lexical Chains," in *Proc. Intelligent User Interfaces*, 2000, pp. 252–255.

[11] G. Ercan and I. Cicekli, "Using Lexical Chains for Keyword Extraction," *Information Processing and Management: An International Journal*, vol. 43, pp. 1705–1714, 2007.

[12] C.-H. Wu, C.-L. Huang, C.-S. Hsu, and K.-M. Lee, "Speech Retrieval Using Spoken Keyword Extraction and Semantic Verification," in *Proc. TENCON*, 2007, pp. 1–4.

[13] A. Desilets, B. de Bruijn, and J. Martin, "Extracting Keyphrases from Spoken Audio Documents," *Information Retrieval Techniques for Speech Applications*, pp. 339–342, 2002.

[14] L. van der Plas, V. Pallotta, M. Rajman, and H. Ghorbel, "Automatic Keyword Extraction from Spoken Text. A Comparison of two Lexical Resources: The EDR and WordNet.," in *Proc. LREC*, 2004.

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

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

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