USING N-BEST RECOGNITION OUTPUT FOR EXTRACTIVE SUMMARIZATION AND KEYWORD EXTRACTION IN MEETING SPEECH

Yang Liu, Shasha Xie, Fei Liu

The University of Texas at Dallas, Richardson, TX, USA
{yangl,shasha,feiliu}@hlt.utdallas.edu

ABSTRACT

There has been increasing interest recently in meeting understanding, such as summarization, browsing, action item detection, and topic segmentation. However, there is very limited effort on using rich recognition output (e.g., recognition confidence measure or more recognition candidates) for these downstream tasks. This paper presents an initial study using n-best recognition hypotheses for two tasks, extractive summarization and keyword extraction. We extend the approach used on 1-best output to n-best hypotheses: MMR (maximum marginal relevance) for summarization and TFIDF (term frequency, inverse document frequency) weighting for keyword extraction. Our experiments on the ICSI meeting corpus demonstrate promising improvement using n-best hypotheses over 1-best output. These results suggest worthy future studies using n-best or lattices as the interface between speech recognition and downstream tasks.

***Index Terms*—** summarization, keyword extraction, n-best hypotheses

1. INTRODUCTION

Meetings happen all the time in the world. If we can record this data and apply speech and language technology to these recordings, it will greatly help efficient information management. Recently there have been many efforts on various meeting understanding tasks, such as automatic summarization, meeting browsing, detecting decision parts and action items, topic segmentation, keyword extraction, and dialog act tagging. However, most previous work used reference transcripts. Some studies used speech recognition (ASR) output, but so far they have only used 1-best ASR hypothesis.

ASR errors are known to degrade performance for many downstream tasks, therefore research has been conducted to use richer information from speech recognizers (such as n-best list, lattices or confusion network, and confidence measure from recognition output) for various tasks, including speech translation, spoken document retrieval, named entity recognition, among others. Compared to those tasks, there is no prior work as yet for the meeting understanding tasks mentioned above.

In this paper, we investigate using n-best output for two meeting processing tasks: extractive summarization and keyword extraction. We extend the approach used for 1-best to n-best hypotheses. Our experiments show that (1) using more hypotheses yields better performance for both tasks, but performance levels off after some hypotheses, and (2) the best result using n-best lists is still significantly worse than using human transcripts. These suggest more future studies are needed to incorporate information from multiple candidates as well as confidence measures from recognizers.

2. RELATED WORK

Many techniques have been proposed for meeting summarization. Some used unsupervised approaches and relied on textual information only, such as Maximum Marginal Relevance (MMR), latent semantic analysis, and integer linear programming [1, 2, 3]. Others were based on supervised methods, such as maximum entropy model, SVM, conditional random fields, using lexical, structural, and acoustic features [4, 5, 6]. Some previous work only used human transcripts. Others used ASR output and typically reported performance degradation due to recognition errors. Note that a lot of work on speech summarization has been performed using other domains, such as lectures, broadcast news, voice mails, etc. We only mentioned some work above in meeting domain.

Most of previous keyword extraction work has been done on written text domain, often based on information such as frequency, word association, sentence/document structure or position, and linguistic knowledge. These are either modeled using unsupervised or supervised methods. Compared to text domain, there have been very limited studies on speech data. [7] evaluated the performance of the tool "Extractor" on broadcast news transcripts with various quality. [8] compared two lexical resources, WordNet and EDR electronic dictionary, to extract simple noun keywords from multiparty meeting corpus. [9, 10] investigated unsupervised and supervised approaches for keyword extraction on the meeting domain, and showed using ASR output hurts system performance compared to human transcripts.

Various studies have been conducted on coupling ASR and language processing tasks. Loose coupling is the most widely used interface. It is a one-way pipeline, where ASR output is used as input to subsequent language processing components. The interface can be 1-best, n-best, lattices, or confusion network. For machine translation, [11] used n-best list for reranking by optimizing interpolation weights for ASR and translation, and [12] used confusion network, but without much improvement over n-best. Lattices have been studied intensively for spoken document retrieval and indexing [13, 14], with reported better performance than just using 1-best ASR output. For named entity recognition on speech data, [15] used n-best lists and obtained small improvement. There is also study trying to more tightly couple ASR and language processing components. For example, [16] proposed a joint decoding approach for speech translation. However, these systems are often too complex and hard to optimize, and do not always outperform those using loose coupling.

There are a few prior studies related to speech summarization and keyword extraction using information beyond just 1-best ASR output, but none of the work is on meeting domain. [7] reported some improved results using additional hypotheses in n-best list for keyword extraction. [17] performed topic clustering using confidence scores, which resulted in better clusters and indirectly helped summarization. Very recently, [18] used confusion networks and expected word counts for speech summarization in Mandarin broadcast news, and achieved better performance than 1-best ASR output. In this paper, our goal is to leverage more ASR hypotheses for summarization and keyword extraction on the meeting data.

3. DATA

We use the ICSI meeting corpus [19], which consists of naturally-occurring meeting recordings, each about an hour long. These are mainly research discussions in the area of natural language processing, artificial intelligence, speech, and networking. All the meetings have been transcribed and annotated with dialogue acts (DAs), topic boundaries, and abstractive and extractive summaries. For extractive summary annotation, the annotators were asked to select and link DAs from the transcripts that are related to each of the sentences in the provided abstractive summaries (see [2] for more information on annotation). Following previous studies on extractive summarization using the ICSI meeting corpus [2, 5], we used the same 6 meetings as the test set. We randomly selected another 6 meetings as the development set [1]. The human agreement on summary annotation is quite low. The average Kappa coefficient among the three annotators on the test set ranges from 0.211 to 0.345. The lengths of the reference summaries are not fixed and vary across annotators and meetings. The average word compression ratio for the test set is 14.3%, with a deviation of 2.9%.

For keyword extraction, we used 26 meetings. They were chosen because they have been used in previous work for topic segmentation, summarization, and keyword extraction [2, 9, 20]. In total, there are 134 topic segments in the 26 meetings. We recruited two undergraduate computer science students to annotate keywords and topic categories for each topic segment. See [9] for more annotation details. The Kappa coefficient for keyword annotation is about 0.41. Note that keywords selected by human annotators are not restricted to single words. In fact, 66.06% of the total selected keywords are unigrams, 31.17% of them are bigrams, 2.25% are trigrams, and the rest are keywords with more than 3 words (only 0.52%). In the annotation, the average number of selected words is 5.92 per topic segment, with a deviation of 2.18.

We use n-best hypotheses from the SRI recognizer. The average WER for the 1-best output is about 41% for the 26 meetings used for keyword extraction experiments (the 6 meetings used in the summarization task are a subset of these). Table 1 shows some statistics for these 26 meetings. These scores are generated and averaged over all the topic segments. The speech recognizer typically performs some pause-based segmentation, therefore each resulting transcript segment in the ASR output does not correspond to a DA. The average "sentence" length in ASR output is generally longer than that on human transcripts, and the variance is also much larger.

Statistics of corpus	Human	1-best
Avg. num of words	1,867	1,861
Avg. num of sentences	269	202
Avg. sentence length	6.94	9.20
S.D. of sentence length	7.91	12.48

Table 1. Statistics of the 26 meetings used for keyword extraction.

4. USING N-BEST LIST FOR SUMMARIZATION AND KEYWORD EXTRACTION

4.1. Extractive Summarization

The task of extractive summarization is to select important sentences (or other segmentation units) with the predefined compression ratio constraint. MMR [21] has been widely used in text summarization because of its simplicity and efficacy. It is also a reasonable system for speech summarization. It selects the most relevant sentences at the same time avoiding redundancy. The final score of a sentence S_i in MMR is calculated as follows:

$$MMR(S_i) = \lambda \times Sim_1(S_i, D) - (1 - \lambda) \times Sim_2(S_i, Sum) \quad (1)$$

where *D* is the document vector, and *Sum* represents the sentences that have been extracted into the summary. The two similarity functions (Sim_1 and Sim_2) calculate the similarity of a sentence to the entire document (measures relevance) and to the selected summary (representing redundancy), respectively. λ is used to balance the two components. The sentences with the highest MMR scores will be iteratively chosen into the summary until the summary reaches a predefined proper size.

One most commonly used similarity measure is cosine similarity. In this approach, each document (or a sentence) is represented using a vector space model. The cosine similarity between two vectors (D_1, D_2) is:

$$sim(D_1, D_2) = \frac{\sum_i t_{1i} t_{2i}}{\sqrt{\sum_i t_{1i}^2} \times \sqrt{\sum_i t_{2i}^2}} \quad (2)$$

where t_i is the term weight for a word w_i, for which we use the TF-IDF value (term frequency, inverse document frequency). The IDF weighting is used to represent the specificity of a word: a higher weight means a word is specific to a document, and a lower weight means a word is common across many documents. IDF values are generally obtained from a large corpus. One widely used method for the IDF value for a word w_i is $IDF(w_i) = log(N/N_i)$ where N_i is the number of documents containing w_i in a collection of N documents. We compute IDF values for words using the ICSI corpus.

We extend the above MMR approach to n-best lists. The task now is to determine the segments to include in the summary, as well as the hypothesis in the segments to use in the summary. To use Eq 1, we form D, the entire document, by using the 1-best hypothesis from each segment. We treat each hypothesis separately in each segment, that is, we calculate the salience score for each hypothesis. In the iterative selection process, when one hypothesis is selected into the summary, the other hypotheses for that segment will not be considered in the iterative MMR approach. The final summary is composed of all the selected hypotheses.

One approximation we used in the MMR approach for human transcripts and 1-best ASR output is to consider only a subset of sentences (about 50% in our experiments), rather than all the sentences in the document. This subset of candidate sentences is selected based on their similarity score to the entire document (first term in MMR formula). [1] showed that this does not degrade performance, but can significantly speed up the extraction process. For n-best, we use the highest score (similarity to the entire document) from all the hypotheses for a segment to represent the segment's score, and perform similar approximation to preselect some segment candidates for summarization.

4.2. Keyword Extraction

For keyword extraction, we use a unigram based method, that is, we only consider single word candidates. The core part of keyword extraction is to assign a salience score to each word, such that the system selects top ranked words as keywords. We use an unsupervised method based on TF-IDF scores to rank words. TFIDF is calculated in similar ways as above for summarization. We only consider content words as candidate keywords. When using n-best hypotheses, we put all the hypotheses together for each segment, and then apply the TFIDF weighting to the expanded transcripts for each document.

Note that we chose these keyword extraction and summarization methods since it is a natural extension from 1-best ASR output to n-best lists. The supervised methods perform better than these unsupervised methods for both tasks [6, 9]; however, it is not straightforward to generalize the features used for 1-best to n-best lists. We will investigate using supervised methods in future work.

5. EXPERIMENT RESULTS

5.1. Evaluation Metrics

To evaluate summarization performance, we use ROUGE [22], which has been used in previous studies of speech summarization. ROUGE compares the system-generated summary with reference summaries (there can be more than one reference summary), and measures different matches, such as N-gram, longest common sequence, and skip bigrams. In this paper, we use ROUGE-2 F-measure as it has been shown to correlate better with human evaluation compared to ROUGE-1 unigram matches.

For keyword extraction, we use each annotation as a reference, and then compute the average F-measure as the final result. We perform the evaluation on a lenient unigram basis, that is, both the system hypotheses and human annotated keywords are first lemmatized, then compared to each other on a unigram basis.

5.2. Keyword Extraction Results

The keyword extraction results are shown in Figure 1 (left Y-axis). We use n-best hypotheses for $n = 1, 2, ..., 10, 15, 20$. The system generates 5 keyword hypotheses for all of the experiments. The right Y-axis shows the the coverage of reference keywords in the n-best, i.e., the percentage of the human annotated keywords that appear in the n-best list.

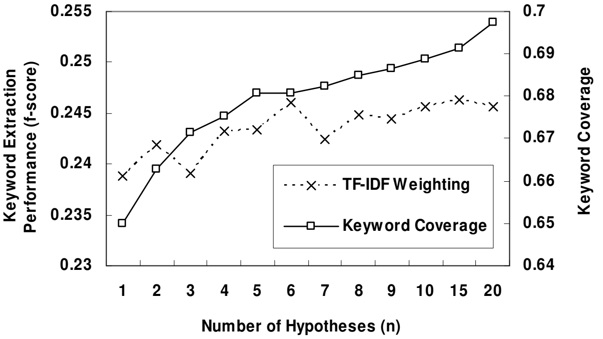

Fig. 1. Keyword extraction performance (left Y-axis) and reference keyword coverage (right Y-axis) using n-best output.

We can see from the figure that the keyword coverage gradually improves as the number of hypotheses increases, compared to 65% from 1-best ASR output. There is a general trend of improved performance of keyword extraction when n increases from 1 up to 6, even though there is some fluctuation as n changes. After that, although the keyword coverage continues to increase, there is no improvement in keyword extraction performance. This shows that, more ASR hypotheses contain more reference keywords and can result in better keyword extraction performance. However, too many hypotheses may also introduce confusion to the keyword extraction task. Similar findings have been shown in [7]. We performed some manual analysis in order to understand if the improvement can be attributed to the better keyword coverage. We found that it is not simply because of the additional correct word hypotheses on n-best lists. Sometimes the correct keyword is already in the 1-best hypothesis, but using frequency-based keyword extraction approach is not able to select this word. When n-best lists are used, a keyword may appear in more hypotheses than some other candidate words, resulting in the higher rank for the correct keyword. In addition, we computed the oracle performance using ASR output by randomly selecting 5 keywords from those reference keywords that appear in ASR hypotheses (reference keywords are weighted based on the number of annotators who selected the keywords). We found that the oracle performance slightly increases from 0.583 to 0.592 when n increases from one to 20, with some fluctuation in between. Finally, for a comparison, we want to point out that using human transcripts, the F-measure is around 0.34. This indicates there is still a big gap between using human transcripts and ASR output (with WER as high as about 41%).

5.3. Extractive Summarization Results

Figure 2 shows the ROUGE-2 results using n-best for extractive summarization. The word compression ratio is 18% in these experiments. Two scores are presented in the graph. In the first one (ASR words), the summary is formed using the selected recognition candidates (one hypothesis on the n-best list). In the other one (mapped REF words), after the hypothesis is selected in MMR, we determine the corresponding segment in the summary, and then use the human transcripts for these selected summary segments to compute the ROUGE scores in comparison to the reference summary. We calculate this score to eliminate the effect of ASR errors in evaluation and focus more on the question whether using more hypotheses helps better determine which segments are the salient ones to include in the summary. This can be useful if speech summarization results are presented to users via speech segments, therefore there is no recognition error in this kind of summary rendering. We can see from the graph that both scores show improvement using more hypotheses, with improvement coming mostly from using two hypotheses, and that adding too many candidates does not yield further gain. As expected, using mapped REF words yields better ROUGE scores than using the ASR hypotheses to form the summary. Note again that these two results have the same selected summary segments. The difference is simply in the transcripts used in evaluation. Finally, similar to keyword extraction, we conduct experiments using human transcripts. The ROUGE score using human transcripts is 0.3613, significantly better than using ASR output, suggesting room for potential improvement.

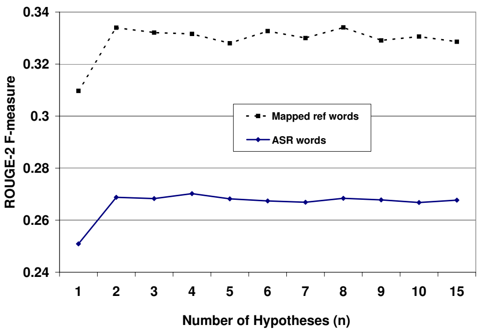

Fig. 2. Summarization results (ROUGE-2 F-measure) using n-best hypotheses.

When using the n-best list, we observe that often the selected hypotheses are not the top one ASR hypothesis. For example, using 15-best, only 2.86% of the selected hypotheses are the top one hypothesis. In another analysis, we first use the time information in the reference summary sentences to determine which ASR segments should be in the summary. Then for each summary segment, we compute the ROUGE scores for each recognition hypothesis on the n-best list compared to the reference summary sentence. Using 50 hypotheses in this oracle analysis, we found that 75% of the time, the best scoring hypothesis (in terms of ROUGE score) is not the 1-best hypothesis. This shows that considering other hypotheses can potentially help summarization.We will perform more analysis to understand the benefit of more hypotheses and in particular why the second hypothesis contributes significantly to the improvement.

We also explored another method to leverage n-best lists, where we use all the candidates in n-best to form a vector for a segment, instead of treating each hypothesis separately. The weight for each word in the vector is its frequency in n-best list, normalized by the number of hypotheses (n value). We also varied the contribution for each hypothesis based on their rank with a linear decay function. Then the same MMR approach is used to determine if a segment is in summary. To generate the summary, for this method, after one segment is selected, we use the corresponding human transcripts. We observe similar performance using this method compared to the one presented above when treating each hypothesis separately (and using the human transcripts for evaluation). This method of using n-best lists is similar to expected word counts from lattices that have been used in other tasks (e.g., spoken document retrieval, speaker identification). We will further investigate this in the future.

6. CONCLUSION

This paper presents a study using n-best as the interface between ASR and two meeting understanding tasks, extractive summarization and keyword extraction. This is a most natural extension beyond using 1-best interface. For both tasks, we apply similar approaches as used on 1-best to n-best hypotheses. Specifically, we use MMR for summarization and TFIDF weighting for keyword extraction. Our experiments showed that there is a consistent improvement over using 1-best hypothesis when considering additional candidates for both tasks, suggesting promising future directions to explore. We also found that using too many hypotheses does not yield further gain and the performance flattens after certain n. In addition, the best result from using n-best hypotheses is still significantly worse than using human transcripts. These show the limitation of using n-best lists. For future work, we will evaluate using supervised methods for these tasks. In addition, we will investigate using more compact interface, for example, lattices or confusion networks, as well as take into account the confidence measures of the recognition hypotheses for term weighting.

7. ACKNOWLEDGMENT

We thank Dilek Hakkani-Tur and Benoit Favre for useful discussions. This work is supported by NSF award IIS-0845484. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of NSF.

8. REFERENCES

[1] S. Xie and Y. Liu, "Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization," in *Proceedings of ICASSP*, 2008.

[2] G. Murray, S. Renals, J. Carletta, and J. Moore, "Evaluating automatic summaries of meeting recordings," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, 2005.

[3] D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tur, "A global optimization framework for meeting summarization," in *Proceedings of ICASSP*, 2009.

[4] A. H. Buist, W. Kraaij, and S. Raaijmakers, "Automatic summarization of meeting data: A feasibility study," in *Proceedings of the 15th CLIN conference*, 2005.

[5] M. Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *Proceedings of EMNLP*, July 2006.

[6] S. Xie, Y. Liu, and H. Lin, "Evaluating the effectiveness of features and sampling in extractive meeting summarization," in *Proceedings of IEEE Workshop on Spoken Language Technology*, 2008, pp. 157–160.

[7] A. Désilets, B.D. Bruijn, and J. Martin, "Extracting keyphrases from spoken audio documents," *Information Retrieval Techniques for Speech Applications*, vol. 2273, pp. 36–50, 2002.

[8] L. 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 *Proceedings of LREC*, 2004, pp. 2205–2208.

[9] F. Liu, D. Pennell, F. Liu, and Y. Liu, "Unsupervised approaches to automatic keyword extraction using meeting transcripts," in *Proceedings of HLT-NAACL*, 2009, pp. 620–628.

[10] F. Liu, F. Liu, and Y. Liu, "Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion," in *Proceedings of IEEE Workshop on Spoken Language Technology*, 2008, pp. 181–184.

[11] V.H. Quan, M Federico, and M. Cettolo, "Integrated n-best re-ranking for spoken language translation," in *Proceedings of EuroSpeech*, 2005.

[12] N. Bertoldi and M. Federico, "A new decoder for spoken language translation based on confusion networks," in *Proceedings of IEEE ASRU Workshop*, 2005.

[13] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," in *Proceedings of ACL*, 2005.

[14] T.K. Chia, H. Li, and H. T. Ng, "A statistical language modeling approach to lattice-based spoken document retrieval," in *Proceedings of EMNLP*, 2007.

[15] L. Zhai, P. Fung, R. Schwartz, M. Carpuat, and D. Wu, "Using n-best lists for named entity recognition from chinese speech," in *Proceedings of NAACL*, 2004.

[16] H. Ney, "Speech translation: Coupling of recognition and translation," in *Proceedings of ICASSP*, 1999.

[17] S. Maskey, *Automatic broadcast news speech summarization*, Ph.D. thesis, Columbia University, 2008.

[18] S.H. Lin and B. Chen, "Improved speech summarization with multiple-hypothesis representations and Kullback-Leibler divergence measures," in *Proceedings of Interspeech*, 2009.

[19] A. Janin, D. Baron, J. Edwards, D. Ellis, G. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proceedings of ICASSP*, 2003, pp. 364–367.

[20] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proceedings of ACL*, 2003, pp. 562–569.

[21] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of SIGIR*, 1998.

[22] C. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of ACL Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.