Towards Abstractive Speech Summarization: Exploring Unsupervised and Supervised Approaches for Spoken Utterance Compression

Fei Liu, *Member, IEEE*, and Yang Liu, *Senior Member, IEEE*

***Abstract*—Most previous studies on speech summarization focus on the extractive approaches. Yet directly concatenating the extracted speech utterances may not form a good summary due to the presence of disfluencies and redundancy in the unplanned spontaneous speech. In this paper, we proposed to generate compressed speech summaries by coupling the sentence level compression and summarization approaches, as a viable step towards generating abstractive summaries. We compared two utterance compression approaches: an unsupervised approach based on the Integer Linear Programming (ILP) framework, and a supervised method using conditional random fileds (CRF) that formulates the utterance compression problem as a sequence labeling task. We evaluated the compression performance using both human and ASR transcripts from the ICSI meeting corpus, and performed both automatic and human evaluation. Our results show that we can achieve reasonable utterance compression performance, and that the CRF-based method generally performs better. By coupling the compression and summarization approaches, we generated compressed speech summaries that cover more important information within the given length limit, yielding 5% absolute performance gain on both human and ASR transcripts as evaluated by the ROUGE-1 F-scores.**

***Index Terms*—Conditional random fields, ICSI meeting corpus, integer linear programming, speech summarization, spoken utterance compression.**

I. Introduction

Speech summarization identifies the key information from a collection of speech recordings, providing an efficient way for users to quickly browse through the lengthy multimedia contents, such as broadcast news, lectures, meetings, voice mails, etc. Most traditional summarization systems focus on the extractive approaches, which aim to extract the important sentences from the input text or audio recordings and concatenate them to form a summary. This approach has been performing well on the written text domain such as the news documents, since the extracted sentences themselves are usually well-formed, self-explainable, and have good sentence and discourse structure. On the contrary, directly concatenating the transcribed speech utterances may not result in high-quality summaries due to the large amount of redundancies and disfluencies in the conversational speech.

TABLE I
Human Compressed Summary Sentences for an Example Meeting Dialogue Segment. Dialogue Act Indices (Based on the Entire Meeting) Are Shown in the First Column

Original Extractive Summary	
423	there there are a variety of ways of doing it
433	so it's possible that we could do something like a summary node of some sort that
444	so what i was gonna say is is maybe a good at this point is to try to informally
446	i mean not necessarily in th- in this meeting but to try to informally think about what the decision variables are
450	and the other trick which is not a technical trick it's kind of a knowledge engineering trick is to make the n- -pau- each node sufficiently narrow that you don't get this combinatorics
Human Compressed Summary	
423	there are ways of doing it
433	it's possible we could do a summary node
444	good at this point is to try informally
446	to informally think about what the decision variables are
450	make each node sufficiently narrow that you don't get this combinatorics

In Table I, we show an example of the extractive summary and its compressed variant for a meeting dialogue segment. The "Original Extractive Summary" was formed by directly concatenating the extracted summary sentences (using human transcripts). The "Compressed Summary" was generated by manually compressing the extractive summary at the utterance level. We can see that the quality of the original extractive summary is not very good. In contrast, the compressed summary removes many unnecessary words from the original extractive summary. It effectively highlights the main content and its readability is much better. In this sense, the compressed speech summary is also closer to the abstractive summaries. For abstractive summarization, we may first apply the utterance level compression techniques to the extracted summary sentences, followed by further sentence merging, compaction, and generation.

Previous approaches to sentence compression mainly focus on the well-structured sentences from professional writers, while little research has been performed on compressing the

Manuscript received July 22, 2012; revised January 02, 2013; accepted March 17, 2013. Date of publication March 27, 2013; date of current version April 11, 2013. This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract HR0011-12-C-0016 and NSF Award IIS-0845484. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or NSF. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pascale Fung.

The authors are with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: feiliu@hlt.utdallas.edu; yang.liu@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2255279

unstructured spoken utterances, and even less studies were conducted on the automatic speech recognition output. There are many challenges in compressing the spoken utterances: ill-formed sentence structure is common among the spoken utterances; incomplete or ungrammatical sentences are abundant; sentence boundary is not clearly specified; spoken utterances contain lots of disfluencies, redundancies, and colloquial expressions; automatic speech recognizers (ASR) often yield high word error rate (WER) on the spontaneous conversations. These speech-specific characteristics significantly affect the existing sentence structure based compression approaches and the abstractive summarization approaches, which rely heavily on correctly parsing the sentences into syntactic constituents, then perform tree transduction or other language generation or paraphrasing techniques [1].

In this paper, we propose to generate compressed speech summaries by coupling the spoken utterance compression system with the extractive summarization system. This allows us to take advantage of the robust extractive summarization framework while generating more condensed speech summaries. Specifically, we raise the following questions: (1) is it possible to perform sentence compression on the noisy spoken utterances? (2) what is the performance difference between the unsupervised and supervised spoken utterance compression approaches? (3) what is the best setup to generate compressed speech summaries? should summarization be performed on the compressed sentences, or on the original ones and then followed by compression? (4) what is the impact of ASR errors on the compression and summarization systems? To address these questions, we compared the unsupervised Integer Linear Programming (ILP)-based approach with the supervised Conditional Random Fields (CRF)-based approach for spoken utterance compression. For the CRF-based approach, we further conducted feature analysis to capture the most effective feature categories in compressing the spoken utterances. We investigated possible ways of combing multiple reference compressions to train a better compression system. In evaluating the compressed spoken utterances, we compare against the human reference compressions on both word- and sentence-level. We also employed human annotators to judge the informativeness, grammaticality, and succinctness of the system and human compressions. We evaluated the piped compression and summarization systems on both the human transcripts and ASR output. In addition, we constructed a spoken utterance compression data set with multiple human reference compressions for each transcribed spoken utterance, which is comparable in size to other sentence compression corpora for the written-text domain.

II. Related Work

Extractive summarization approaches gained a lot of popularity in the past decades. Both unsupervised and supervised approaches have been explored for speech summarization. [2] applied the maximal marginal relevance (MMR) approach to extract salient sentences from the dialogue segments. [3], [4] proposed a concept-based integer linear programming (ILP) framework for meeting summarization. Given a desired summary length, this approach extracts sentences that cover as many important concepts as possible while within the length limit. [5] proposed a graph-based submodular selection approach, where summary sentence selection is formulated as optimizing the submodular functions defined on the semantic graph built from the given document. With respect to the supervised approaches, [6]–[10] incorporated lexical, structural, acoustic, and prosodic information (such as pitch, duration, energy, and pause) in the supervised framework for speech summarization. Different classification algorithms have been explored, including the hidden Markov model (HMM) [11], maximum entropy (ME) [12], support vector machines (SVM) [9], conditional random fields (CRF) [8], etc. [13], [14] proposed a rhetorical-state hidden Markov model (RSHMM) for summarizing the lecture speech. In addition, [15]–[18] explored semi-supervised approaches for extractive speech summarization, including active learning, co-training, probabilistic generative framework and risk minimization, hybrid of unsupervised and supervised approaches, etc. There are also some efforts that used multiple speech recognition candidates in order to address the problem due to recognition errors. [5], [19], [20] used n-best recognition output and the confusion networks for speech summarization.

In recent years, there is some work on abstractive speech summarization that focuses on generating condensed representation of summaries. [21] proposed to extract a set of words from automatically transcribed speech that maximizes a summarization score consisting of word significance measure, confidence score, linguistic likelihood, and a word concatenation probability. [22], [23] proposed to generate abstracts of meeting conversations based on the conversation ontology. They also showed that users prefer abstract-style summaries over extracts. [24] experimented with ILP and lexicalized Markov grammar based approaches to compress human transcribed speech utterances. [25] further developed an automatic summarizer to combine the sentence compression and summarization modules for meeting summarization.

Spoken utterance compression bears similarities to the traditional sentence compression task, which has been widely studied to shorten the long sentences in the newswire or broadcast news. [26] employed the decision tree model to learn rewriting rules that decide whether a syntactic constituent should be dropped within a given context. [1] formulated sentence compression as a discriminative tree-to-tree rewriting framework, where all possible rewrites are generated from a set of synchronous tree substitution grammar (STSG). [26], [27] applied the noisy-channel framework to predict the possibilities of translating a sentence to a shorter word sequence. [28] extended the noisy-channel approach and proposed a Markovization formulation of the synchronous context-free grammar (SCFG). [29], [30] proposed discriminative sentence compression with conditional random fields (CRF) model. Unlike these approaches that need a training corpus, [31] employed the integer programming approach to find a subset of words that maximize an objective function. There are also initial attempts to integrate sentence compression with text summarization system on news documents. [26], [32]–[34] used the sentence compression module to postprocess the extracted summary sentences or jointly learned to extract and compress sentences.

Some of the above compression approaches rely heavily on correctly parsing the syntactic structure of sentences, which may not be directly applicable to the noisy spoken utterances. Instead, we investigated supervised spoken utterance compression approaches that can effectively capture the speech-specific characteristics, including the word and part-of-speech (POS) n-grams, position features, transition features, shallow and deep syntactic features, distance to the same words, etc. We also investigated ways of leveraging multiple reference compressions in the training process. We show that the spoken utterance compression and summarization system can achieve satisfying performance on both human and ASR transcripts.

This paper is an extension of [24] and [25]. In this study, we compare the unsupervised and supervised approaches for spoken utterance compression, investigate the use of multiple references for compression, evaluate generation of compressed speech summaries using both human and ASR transcripts, and perform more analysis and provide more discussions about various aspects of the compression and summarization systems.

III. Corpus and Data Annotation

We use the ICSI meeting corpus [35], [36] for our experiments. They are naturally occurring meeting recordings that are mainly research discussions on natural language processing, artificial intelligence, speech, and networking. Each meeting is about an hour long. All the meetings have been manually transcribed and annotated with dialogue acts (DAs) [37], topic boundaries, extractive and abstractive summaries [7]. The ASR transcripts were generated from a state-of-the-art SRI recognizer [38], with a word error rate (WER) of about 38.2% on the entire corpus. For the ASR output, we obtained the DA boundary information by aligning the human annotations to the ASR words.[1]

26 meetings from the ICSI corpus were used for the spoken utterance compression and summarization task. Six of the meetings are the commonly used test set for meeting summarization [4], [7], [8], which contains 1088 extractive summary sentences from three annotators. The rest 20 meetings have only one summary annotation, with 1772 extractive summary sentences in total. We employed human annotators from Amazon Mechanical Turk (AMT)[2] to manually compress the summary sentences by dropping the unnecessary words. These sentences are grouped into 286 human intelligence tasks (HITs); each HIT contains 10 sentences that need to be compressed. The human transcripts were used for compression annotation, with filled pauses (e.g., “uh, um, eh”) and incomplete words (e.g., “h-”) removed in the preprocessing step to increase the sentence readability for human annotators.

We used a two-stage annotation scheme. In the first stage, each HIT was annotated by 8 mechanical turk workers. Each received $0.15 as compensation for every HIT. For each sentence that needs to be compressed, two sentences before and after it are displayed in the annotation interface in order to provide some context. The turkers can click on the unnecessary words and remove them from the original sentence. After this stage, we obtained 8 reference compressions for each summary sentence. In the second stage, turkers are asked to find the best compression among the 8 annotations from the first stage. We provide the same original summary sentence and its context to the annotators as in the first stage, list all the compression variants, and ask the turkers to select the best compression for each original summary sentence. In this annotation stage each sentence is annotated by 6 turkers. Their majority vote is used as the goldstandard compression. If there is a tie, we choose the shorter one.

In total, 244 turkers participated in the first stage and 300 turkers performed the second stage annotation. Only 41 of them are the same as in the first stage. Since the annotation was performed by a large group of annotators, it is difficult to calculate the inter-annotator agreement. As an alternative, we present the turker agreement percentages in selecting the gold standard compression. We found that 16.12% of the goldstandard compressions are agreed by all of the 6 annotators; 21.07% are agreed by 5 annotators, 28.70% are agreed by 4 annotators; 27.99% are agreed by 3 annotators; 6.09% are agreed by 2 annotators, and 0.03% by 1 annotator. We refer to the human compressions obtained from the first and second stage as “multiple reference compressions” and “gold standard compressions” respectively in the rest of this study.

IV. Spoken Utterance Compression

We formulate the spoken utterance compression as a word deletion task, where the goal is to generate a compressed sentence that retains most of the important information while being as grammatical as possible. We explored both unsupervised and supervised approaches for this task. The unsupervised approach leverages the Integer Linear Programming (ILP) framework and a filler phrase detection module, with no human annotations required; while the supervised approach formulates the spoken utterance compression as a sequence labeling task and effectively integrates many speech-specific features under a Conditional Random Fields (CRF) model.

A. Unsupervised ILP Approach With Filler Phrase Detection

We first develop a filler phrase detection module to remove the filler words before applying the ILP compression approach. We define filler phrases (FPs) as the combination of two or more words, which could be discourse markers (e.g., I mean, you know), editing terms, as well as some terms that are commonly used by human but without critical meaning, such as, “for example,” “of course,” and “sort of.” Removing these fillers barely causes any information loss. We propose to use web information to automatically generate a list of filler phrases and filter them out in compression.

For each of the extractive summary sentences, we use it as a query to Google and examine the top N returned snippets (N is 400 in our experiments). The snippets may not contain all the words in a sentence query, but often contain the frequently occurring phrases. For example, when “of course” is in the query,

[1] Note that in this paper, we use “spoken utterance” and “sentence” interchangeably when there is no ambiguity. Both correspond to the dialogue act (DAs) segments in the human and ASR transcripts.

[2] http://www.mturk.com

it can be found with high frequency in the snippets. We collect all the phrases that appear in both the extracted summary sentences and the snippets with a frequency higher than three. Then we calculate the inverse sentence frequency (ISF) for these phrases using the entire ICSI meeting corpus. The ISF score of a phrase i is:

$$\mathrm{isf}_i = \frac{N}{N_i} \tag{1}$$

where N is the total number of sentences and N_i is the number of sentences containing this phrase. Phrases with low ISF scores mean that they appear in many occasions and are not domain- or topic-indicative. These are the filler phrases we want to remove to compress a sentence. The three phrases we found with the lowest ISF scores are “you know,” “i mean” and “i think,” consistent with our intuition.

We also noticed that not all the phrases with low ISF scores can be taken as FPs (“we are” would be a counter example). We therefore gave the ranked list of FPs (based on ISF values) to a human subject to select the proper ones. The human annotator crossed out the phrases that may not be removable for sentence compression, and also generated simple rules to shorten some phrases (such as turning “a little bit” into “a bit”). This resulted in 50 final FPs and about a hundred simplification rules. The FPs were filtered out from the spoken utterances before applying the ILP compression approach. Examples of the final FPs are: ‘you know,’ ‘and I think,’ ‘some of,’ ‘I mean,’ ‘so far,’ ‘it seems like,’ ‘more or less,’ ‘of course,’ ‘sort of,’ ‘so forth,’ ‘I guess,’ ‘for example.’

We employ the integer linear programming (ILP) approach in the same way as [31]. Given an utterance $S = w_1, w_2, \ldots, w_n$, the ILP approach forms a compression of this utterance by dropping words and preserving the word sequence that maximizes an objective function, defined as the sum of the significance scores of the consisting words and n-gram probabilities from a language model:

$$\begin{aligned} \max\ \lambda \cdot \sum_{i=1}^{n} y_i \cdot \mathrm{Sig}(w_i) \\ +(1-\lambda) \cdot \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} x_{ijk} \cdot P(w_k \mid w_i, w_j) \end{aligned} \tag{2}$$

where y_i and x_{ijk} are binary variables: $y_i = 1$ represents that word w_i is in the compressed sentence; $x_{ijk} = 1$ represents that the sequence w_i, w_j, w_k is in the compressed sentence. A trade-off parameter λ is used to balance the contribution from the significance scores for individual words and the language model scores. More details can be found in [31], [24]. We only used linear constraints defined on the variables, without any linguistic constraints.

We use the lp_solve toolkit.[3] The significance score for each word is its TF-IDF value, with the term frequency (TF) calculated on the meeting-level and the inverse document frequency (IDF) calculated using the entire ICSI meeting corpus. We trained a language model using the SRILM toolkit[4] on broadcast news data to generate the trigram probabilities. We leverage the balancing parameter λ to adjust the sentence compression ratio, which yields longer sentences when more weight was assigned to the word significance scores (fewer words were removed). This ILP-based approach is applied to the sentences after filler phrases (FPs) were filtered out. We refer to the output from this approach as “Unsupervised ILP.”

B. Supervised Sequence Labeling Approach

The supervised spoken utterance compression approach formulates the utterance compression as a sequence labeling task. We followed the experimental setup in [25], [39] and use the “BIO” labeling scheme, where “B” and “I” represent the beginning and inside of a word sequence to be preserved, “O” means a word is to be removed from the original utterance.

Given an original word sequence $X = (X_1, X_2, \ldots, X_n)$, the distribution of its corresponding label sequence $Y = (Y_1, Y_2, \ldots, Y_n)$ under the linear-chain Conditional Random Fields (CRF) model takes the following form:

$$\begin{aligned} p(Y \mid X) \propto \exp \sum_{k=1}^{n} \Bigg(\sum_j \lambda_j f_j(y_k, y_{k-1}, X) \\ + \sum_i \mu_i g_i(x_k, y_k, X) \Bigg) \end{aligned} \tag{3}$$

where f_j are the transition feature functions; g_i are the observation feature functions; λ_j and μ_i are the corresponding feature weights. We implemented a variety of features to effectively capture the situation where a word needs to be retained or removed from the original sentence. These features are listed below.

- **Word n-grams**
  The word n-gram features capture the identity of the current word; the two words before and after the current word; as well as all the bigrams and trigrams that can be formed by adjacent tokens and the current word.
- **Part-of-speech (POS) n-grams**
  Similar to the word n-grams, the POS n-gram features include the POS tags and the tag combinations that correspond to the unigram, bigram, and trigram word features. We use the TnT part-of-speech tagger [40] trained from Switchboard data for POS tagging.
- **Word position features**
  These include the absolute and relative position of the current word within the utterance. For the relative position, we use the following formula to decide the corresponding bin index: $(\mathrm{pos} + 1) * 10/\mathrm{sent\_len}$, where pos is the word's position in the sentence, indexed from 0, and sent_len is the length of the sentence. We take the integer part of this value and use the resulting number as the feature (11 bins: from 0 to 10). We also developed conjunction features to accurately pinpoint the word position, including the concatenation of the current word/POS tag with the absolute word position.

[3] http://sourceforge.net/projects/lpsolve/files/lpsolve/5.5.2.0/

[4] http://www.speech.sri.com/projects/srilm/

- **Shallow syntactic features**
  We hypothesize whether a word is removed or not relates to its position in the syntactic parsing tree. For example, words in the prepositional phrase (PP) or adverb phrase (ADVP) may be dropped more frequently. We use the Charniak-Johnson's reranking parser[5] to generate the sentence-level syntactic parse tree for each utterance. An example parse tree was shown in Fig. 1.
  We derive three types of features from the syntactic parse tree: (1) the phrase tag, which is the second-to-last syntactic tag along the path from the root to the word. It denotes whether the current word is included in the NP, VP, PP, ADVP, etc. We also include the phrase tag n-gram features as defined similarly for the word and POS n-gram features; (2) length of the path starting from the root node to the leaf node (POS tag); (3) length of the current path divided by the longest path in the parse tree. The last two features specify the absolute and relative depth of the current word in the parse tree, with the relative depth discretized into 11 bins as described above for the relative word position feature.
- **Deep syntactic features**
  We extend the shallow syntactic features with a set of conjunction features that capture the deep syntactic structures, including: (1) conjunction of the phrase tag with its relative depth in the tree; (2) conjunction of the phrase tag with the POS tag; (3) conjunction of the phrase tag with its parent tag along the tree path; (4) conjunction of the phrase tag, its parent tag, and the absolute word position; (5) conjunction of the phrase tag, POS tag, and the absolute word position. The last two features aim to differentiate the conjunction features (e.g., "VP_ADVP") that were found at different positions of the utterance.
  In addition, we define two "shared parent" features. One represents the lowest parent node that subsumes the current word and its previous word, (see the "VP" tag in Fig. 1), concatenated with the relative depth of this node in the tree. We use another such feature for the following word. These two features capture the relationship between the adjacent words, whether they share a local subtree or contribute to the global tree structure. We will show that these deep syntactic features are very effective in compressing the spoken utterances.
- **Distance to the same word**
  These features include the word distance from the current word to its previous/next same word, as well as the distance features of its previous/next word. They are designed to capture some disfluencies such as repetitions and revisions, where a word or phrase is repeated in the utterance.
- **TF-IDF scores**
  The TF-IDF scores capture the word significance. They are calculated in the same way as in the ILP approach.
- **LM probabilities**
  The LM probability for a word w_0 is defined as $\log(P(w_0 \mid w_{-1})) + \log(P(w_1 \mid w_0))$, which is the sum of the bigram probabilities of the current word given its previous word, as well as the next word given the current word. The LM probabilities capture the adhesion strength of the adjacent words. We use the same model trained from broadcast news data as in the ILP approach.
- **Transition features**
  The transition features include the combination of the current output label (y_i), the previous output label (y_{i-1}), with the current word/POS tag, or the previous word/POS tag, or the "shared parent" features with the previous/next word. The transition features capture the label transition probabilities with respect to the discriminative features.

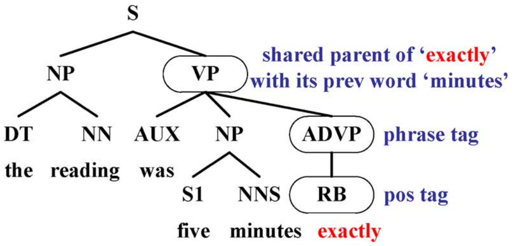


Fig. 1. An example parse tree showing the POS tag, phrase tag, and the shared parent tag of "exactly" with its previous word.

The above features integrate the ones that have been shown to be useful in previous studies [25], [39]. In addition, we proposed novel features that capture the deep syntactic structure of the spoken utterances. We expect the deep syntactic features to promote the compressed utterances that are also grammatically correct.

V. Speech Summarization With Compression

Our goal is to generate speech summaries which are concise representations of the speech recordings. To achieve this goal, we propose to couple sentence compression and summarization. In this work, we choose to use the maximum marginal relevance (MMR) framework for summarization due to its simplicity and verified competency in speech summarization. We expect this is a good starting point for coupling the spoken utterance compression and summarization systems. For each sentence S_i, its MMR score $\mathrm{MMR}(S_i)$ is the linear combination of its similarity to the original document (or a user query), $\mathrm{Sim}_1(S_i, D)$, and the similarity to the current selected summary sentences, $\mathrm{Sim}_2(S_i, \mathrm{Summ})$, as shown below:

$$\mathrm{MMR}(S_i) = \lambda \times \mathrm{Sim}_1(S_i, D) - (1-\lambda) \times \mathrm{Sim}_2(S_i, \mathrm{Summ}) \quad (4)$$

where λ is the balancing factor between the two components. We use cosine similarity under the vector space model for the similarity between two text segments (S_i and S_j):

$$\mathrm{Sim}(S_i, S_j) = \frac{\sum_k w_{i,k} \times w_{j,k}}{\sqrt{\sum_k w_{i,k}^2} \times \sqrt{\sum_k w_{j,k}^2}} \quad (5)$$

The term weight of word $w_{i,k}$ is determined by TF × IDF, where TF is its term frequency in the text segment S_i, and IDF

[5] https://github.com/BLLIP/bllip-parser

is the inverse document frequency generated from a large background corpus. The MMR approach iteratively selects the summary sentences until the given length limit is reached.

For the MMR system, we can use either the original transcript or the compressed ones as input. Furthermore, for the output, we can render either the original uncompressed (extractive) summaries or the compressed summaries. These different input and output configurations are described in the following.

- **1(a): Input original transcripts, output uncompressed (extractive) summaries**
  The uncompressed (extractive) summaries are generated by directly applying the MMR-based summarization approach to the original transcripts.
- **1(b): Input compressed transcripts, output uncompressed (extractive) summaries**
  We first apply the MMR-based summarization approach to the input compressed transcripts, then map the selected summary sentences to their uncompressed form to render the uncompressed summaries.
- **1(c): Input original transcripts, output compressed summaries**
  We first apply the MMR-based summarization approach to the original transcripts, then map the selected summary sentences to their compressed form to render the compressed summaries.
- **1(d): Input compressed transcripts, output compressed summaries**
  The compressed summaries are generated by directly applying the MMR-based summarization approach to the compressed transcripts.

The above settings 1(a)–1(d) use human transcripts as input. Similarly, we use 2(a)–2(d) to represent the settings using the ASR transcripts. In total, we evaluated eight different settings to render the compressed and uncompressed summaries: settings 1(a)–1(b), 2(a)–2(b) generate uncompressed (extractive) summaries; settings 1(c)–1(d), 2(c)–2(d) output condensed summaries.

VI. Experimental Results

In this section, we first evaluate the spoken utterance compression performance on both word- and sentence-level against the human compressions, including both the single goldstandard compression and the multiple human compressions obtained in Section III. In addition, we employ human annotators to judge the informativeness, grammaticality, and succinctness of the system and human compressions. We finally evaluate the compressed and uncompressed summaries generated using both human and ASR transcripts.

A. Compression Results

1) Experimental Setup: The utterance compression approaches are evaluated using the human-annotated summary sentences from the 26 meetings. For the unsupervised ILP-based approach, we directly apply it to all the summary sentences. For the supervised CRF-based approach, we performed 26-fold (leave-one-out) cross validation on the 26 meetings, and results were averaged across all folds. Similar to [39], in training the CRF system, we align the goldstandard compression with the original sentence in a consistent way, that is, the last appearance of the repeated words was always labeled as “preserved” while the earlier ones labeled as “deleted.” Before applying any compression or summarization approaches, we preprocess the human and ASR transcripts to remove the filled pauses and incomplete words and use the postprocessed transcripts for all experiments.

We compare the ILP and CRF systems using the same compression ratios, since it is unfair to compare the sentence compression systems with different compression ratios, as reported in [41]. Here we define the word compression ratio on the meeting level as the percentage of words that are retained after the compression. The final compression ratio is averaged across meetings. For the ILP system, we obtained the desired compression ratio by adjusting the balancing parameter λ between the word significance scores and language model scores. Giving more weight to the word significance scores will force the system to output more words. For the CRF system, we adjusted the system output based on the posterior probabilities of the labels. If the system output contains more words than expected, we dropped the words with low marginal probabilities of being retained until the desired compression ratio is met, and vice versa if the system output contains fewer words. Note that this word-level adjustment may be different from the optimal sentence-level result for the predefined compression ratio, but it is computationally much easier.

2) Automatic Evaluation: We evaluate the compressed spoken utterances against the human compressions using different metrics, including the word accuracy, sentence accuracy, lenient sentence accuracy, and the ROUGE scores.

- Word accuracy is defined as the percentage of words in the original sentence that both the system and goldstandard compressions agreed to keep or to remove. (i.e., treating compression as a word-level classification task).
- Sentence accuracy is calculated as the percentage of the system compressions that agree with the goldstandard compressions, as measured by the string match. In the multi-reference setting, we consider a match if the system compression matches any of the eight human reference compressions.
- Lenient sentence accuracy allows the system and human compressions to be “leniently” matched. We consider it a match if the length of the system compression is within one word distance of the human compression, and the length of their longest common subsequence is also within one word distance of the human compression. In the multi-reference setting, we consider it a lenient match if the system compression leniently matches any of the reference compressions.
- ROUGE-1 and ROUGE-2 [42] scores compare the system compression against one or multiple human compressions based on the unigram/bigram overlap. A higher ROUGE score means the system achieves better agreement with the human compressions. The ROUGE score does not consider the word sequence information, therefore is a slight lenient measure compared to the word accuracy.

TABLE II
SPOKEN UTTERANCE COMPRESSION USING UNSUPERVISED ILP AND SUPERVISED CRF APPROACHES. SYSTEM OUTPUT COMPARED AGAINST THE SINGLE GOLDSTANDARD COMPRESSION

System	Comp	ROUGE-1			ROUGE-2			Word	Sent	Lenient Sent
	Ratio	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	Accu.	Accu.	Accu.
Unsupervised ILP	70%	73.93	78.99	74.52	53.94	59.45	55.08	67.92	6.69	30.50
	75%	**74.05**	84.74	77.43	56.71	66.68	59.91	70.41	7.57	33.84
	80%	73.70	89.91	79.66	59.30	74.01	64.54	72.28	8.42	**34.67**
	85%	73.54	**95.06**	**81.73**	**61.46**	**81.31**	**68.82**	**74.20**	**9.98**	33.25
Supervised CRF	70%	**83.52**	91.15	85.83	**71.51**	78.62	73.51	**81.79**	**23.99**	**48.55**
	75%	81.43	94.30	**86.22**	69.91	81.83	**74.19**	81.57	22.68	46.92
	80%	79.02	96.76	85.86	67.91	84.32	74.02	80.42	20.70	42.74
	85%	76.17	**98.49**	84.79	65.09	**85.78**	72.84	78.28	15.97	36.93
Uncompressed	100%	**66.32**	**100.00**	**78.56**	**55.01**	**86.13**	**65.96**	**66.32**	**5.26**	**16.11**

We apply the above evaluation metrics to both the ILP and CRF systems. We use the human transcripts as input and compare the compressed sentences against the goldstandard compression or all of the eight reference compressions.[6]

Table II shows the spoken utterance compression results against the single goldstandard compressions, with word compression ratio ranging from 70% to 85%[7]. As a comparison, we also provide the results of using uncompressed sentences (filled pauses and incomplete words were removed). As can be seen from Table II, the unsupervised ILP approach achieves reasonable performance when using higher compression ratio, yielding 74.20% word accuracy when using 85% compression rate. On the contrary, the CRF approach maintains stable performance across different compression ratios, with best performance achieved at 70% compression ratio with word accuracy of 81.79% and sentence accuracy of 23.99%. When comparing to the goldstandard compressions, the sentence accuracy scores of both compression systems are low. This is because the sentence accuracy requires a strict match between the system and goldstandard compressions, while there can be multiple acceptable compressions other than the goldstandard. Note that the sentence accuracy of the uncompressed utterances (last row in Table II) indicates that among all the input spoken utterances (these are summary utterances), only 5.26% of them do not need any compression, while the vast majority of the spoken utterances contain redundant words.

Table III presents the system performance as compared to the multiple reference compressions. With the increasing of the word compression ratios, we also notice an increase in the ROUGE f-scores, sentence accuracies, and lenient sentence accuracies. This is observed for both the ILP and CRF approaches. This is because the average word compression ratio of the multi-reference is much higher than the goldstandard compression (74.84% v.s. 63%), therefore systems with higher word compression ratio tend to have better performance. Moreover, the sentence accuracy scores have increased dramatically as compared to evaluation against the single goldstandard (in Table II). For the CRF system with 85% compression ratio, the sentence accuracy score has raised from 15.97% to 55.11%, with a lenient sentence accuracy of 84.45%. This is very encouraging result, indicating the system compressions are of good quality and more than 80% of the system compressions are close to the human compressions.

TABLE III
SPOKEN UTTERANCE COMPRESSION USING UNSUPERVISED ILP AND SUPERVISED CRF APPROACHES. SYSTEM OUTPUT COMPARED AGAINST THE EIGHT HUMAN REFERENCE COMPRESSIONS

System	Comp Ratio	R-1 F (%)	R-2 F (%)	Sent Accu.	Lenient Sent Accu.
ILP system	70%	78.89	60.81	16.66	49.71
	75%	82.12	66.47	23.14	59.54
	80%	84.84	72.12	29.65	69.64
	85%	**87.37**	**77.56**	**40.46**	**78.53**
CRF system	70%	87.13	77.05	48.26	74.99
	75%	89.01	78.98	52.87	80.32
	80%	89.93	80.76	54.73	83.42
	85%	**90.12**	**81.63**	**55.11**	**84.45**
Goldstand.	63%	87.91	77.47	100.00	100.00
No comp.	100%	86.50	78.44	37.71	60.19

3) Human Evaluation: The automatic evaluation metrics count word matches but can hardly measure linguistic quality of the compressed sentences. We therefore employed the Amazon Mechanical Turk workers to manually judge the system and human compressions based on three criteria: (1) informativeness: the compressed sentence should contain most of the important information in the original sentence; (2) grammaticality: the compressed sentence being as grammatical as possible; (3) succinctness: the compressed sentence should contain the least redundant words. For each of the system or human compressions, we ask the human annotators to assign a score from 1 to 5, with 5 meaning the compressed sentence has the desired quality. We evaluated the ILP and CRF systems with 70% and 80% compression ratios, using 4 meetings with 646 sentences in total. Each sentence was rated by 7 annotators. The scores are averaged and presented in Table IV.

In general, the ILP system yields inferior performance compared to the CRF-based system. The latter is close to the goldstandard compression on the informativeness and grammaticality, although being slightly redundant with 80% compression ratio. One interesting finding is that, the human perceived "succinctness" is not just dependent on the sentence length. We notice that the ILP and CRF systems with both 80% compression ratio actually yield different "succinctness" scores (about 0.2 gap), and the ILP system with 80% compression ratio yields a

[6] We chose not to report the compression results on the ASR transcripts since we do not have goldstandard compressions available on the ASR output.

[7] The compression ratios were selected based on the previous research [25].

TABLE IV
HUMAN EVALUATION ON INFORMATIVENESS, GRAMMATICALITY, AND SUCCINCTNESS

System	Ratio	Informative	Grammatical	Succint
ILP	70%	3.68	3.48	3.41
	80%	3.97	3.72	3.47
CRF	70%	3.96	3.83	3.74
	80%	4.14	3.89	3.66
Goldstand.	63%	4.14	4.16	4.05

TABLE V
FEATURE EFFECTIVENESS EVALUATION. RESULTS COMPARED AGAINST THE GOLDSTANDARD COMPRESSION

Feature Category	R-1 F (%)	R-2 F (%)	Word Accu.	Sent Accu.
Word n-grams	82.69	70.05	78.96	18.80
+ POS n-grams	83.80	71.66	80.11	20.75
+ TFIDF features	82.47	69.85	78.65	18.00
+ Position of word/POS	**84.45**	**72.49**	**80.14**	**20.61**
+ Shallow synt feats	**84.40**	**72.15**	**80.28**	**20.97**
+ Deep synt feats	**84.25**	**72.13**	**80.18**	**21.02**
+ Dist to same word	**84.23**	**71.47**	**80.50**	**20.77**
+ LM probabilities	82.75	70.08	79.01	18.82
+ Transition feats	83.94	71.77	79.94	21.41
All features	**85.83**	**73.51**	**81.79**	**23.99**

TABLE VI
LEVERAGING MULTIPLE REFERENCE COMPRESSIONS IN THE CRF TRAINING STAGE

Multi. Reference n	Eval: Goldstand.			Eval: Multi-Ref		Tag Ratio (%)
	Word Accu.	Sent Accu.	Lenient Accu.	Sent Accu.	Lenient Accu.	
n = 4	80.70	21.51	46.48	44.61	71.30	79.65
n = 5	81.41	22.11	48.32	46.31	74.26	75.06
n = 6	**82.09**	**23.93**	**49.07**	**48.99**	**75.63**	69.80
n = 7	**82.19**	**24.30**	**49.16**	**50.30**	**76.48**	61.61
n = 8	81.59	22.29	47.06	46.12	74.29	46.25
Goldstand.	**81.79**	**23.99**	**48.55**	**48.26**	**74.99**	62.56

slightly better "succinctness" than the system with 70% compression ratio. The following shows two compressions generated from the ILP system using 80% and 70% compression ratios respectively: (1) "i wasn't sure whether wizard was the term for not a man" (80% compression); (2) "i wasn't whether wizard was the term for not a man" (70% compression). In the first compression, the sentence seems to be grammatically complete and succinct; while in the second compression, the word "sure" was dropped (probably because of its low significance score), and thus "i wasn't" became a dangling constituent that may be further removed. This suggests that the human perceived "succinctness" of a compressed sentence may be correlated with both the sentence length and "grammaticality." The ILP system with 70% compression ratio tends to generate sentences with poor grammar, and some dangling sentence constituents are often considered redundant.

4) Feature Effectiveness: For the CRF-based compression system, we analyzed the effectiveness of different feature categories. Descriptions of the features have been presented in Section IV.B. In Table V, we presented the results of feature effectiveness evaluation, as compared to the goldstandard human compression.[8] We use the word n-gram as the base feature category, and add each of the other feature categories separately. For the evaluations, we use the CRF system with a fixed compression ratio of 70%. We can see from Table V that the word n-gram is a robust feature category. It can achieve 78.96% word accuracy by itself. The TF-IDF scores and the LM probabilities are not very effective, yielding only marginal or no improvement upon the base features. Four other feature categories have been identified as more effective than others, they are (1) position of word/POS tag; (2) shallow syntactic features; (3) deep syntactic features; (4) distance to the same word features. The results show that the shallow and deep syntactic features we proposed can achieve satisfying performance on spoken utterance compression, while the distance to the same word features are effective since they capture the repetitions, revisions, etc. that are common in the spontaneous speech. Some of the effective feature categories are related to the speech-specific characteristics, indicating the traditional compression approaches need to be adapted to fit in the spoken utterance compression task.

5) Leveraging Multiple Reference Annotations: In this section, we investigate whether it is possible to leverage multiple human reference compressions in training the CRF system. The motivation is to use the majority vote to improve the labeling process, as compared to using the goldstandard compressions for labeling. We introduce a threshold n in producing the word labels during the CRF training stage. For example, when $n = 5$, we only label the word as "preserved" if at least 5 out of the 8 annotators chose to retain the word, otherwise, it is labeled as "removed." We generated different labeling schemes using varied threshold values. The results are presented in Table VI. In the last column, we also presented the percentage of tags in the data set that have been tagged as "preserved" for each of the n values. We compared the resulting compressed summaries against both the goldstandard compression as well as the multiple reference compressions. We observed that when n equals to 6 or 7, the CRF system can achieve similar or slight better performance than using the goldstandard compression for labeling. When n equals to 6 or 7, the tagging ratio is also similar to that of the goldstandard compression. In the future, we would like to investigate other approaches for utilizing the multiple human references in the compression system.

B. Summarization Results

1) Experimental Setup: We evaluate the summarization performance on the 6 test meetings from the ICSI corpus, using both human and ASR transcripts as system input. We use the ROUGE scores for evaluation, which has been widely used in other summarization tasks. Specifically, we use ROUGE-1 and ROUGE-2 evaluation metrics which measure the unigram and bigram overlap between the system summary and human reference summaries. Given its proved performance in previous sections, we use the CRF system with 70% word compression ratio and use the leave-one-out cross validation to generate the compressed transcripts.

[8] Similar results were observed when comparing the system against the multiple reference compressions.

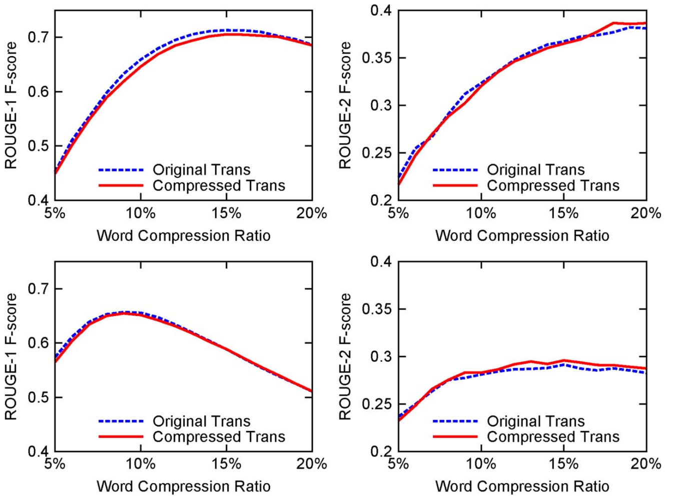


Fig. 2. Use original or compressed transcripts as input to the MMR system. Render either the uncompressed (extractive) summaries (upper two figures) or the compressed summaries (lower two figures).

2) *Impact of Using Original or Compressed Transcripts:* We first evaluate the impact of using either the original or the compressed transcripts as input to the MMR summarization system, and render both the uncompressed (extractive) and compressed summaries. These correspond to the settings 1(a)–1(d) as described in Section V. We used only human transcripts for this experiment. To evaluate the extractive summaries, we compare them to the three human annotated extractive summaries; to evaluate the system-generated compressed summaries, we compare them against the three compressed summaries formed by mapping the human annotated extractive summary sentences to the goldstandard compressions. Results are presented in Fig. 2, with summary length (word compression ratio) ranging from 5% to 20% of the total words of the original transcripts.

We found that in generating the compressed or uncompressed summaries, using different transcripts (original or compressed) as input to the MMR system only marginally affects the sentence selection. Similar findings were also reported in [25]: when generating the compressed summaries, using the original or compressed transcripts as input to the MMR system (corresponding to settings 1(c)–1(d)) only makes marginal difference, with the system achieving slightly better ROUGE-2 scores when using the compressed transcripts as input. Base on these findings, we choose to use the original transcripts as input to the MMR system in the rest experiments.

3) *Impact of Using Human or ASR Transcripts:* We evaluate the impact of using human or ASR transcripts for generating both uncompressed (extractive) and compressed summaries. These correspond to the settings 1(a)/1(c) and 2(a)/2(c) as described in Section V. For this set of experiments, we use the original human/ASR transcripts as input to the MMR-based summarization system. Results are presented in Fig. 3. We observe that compared to using human transcripts, using the ASR transcripts results in some performance degradation in generating both the uncompressed (extractive) and compressed summaries, There is a larger performance gap when evaluated using the ROUGE-2 scores.

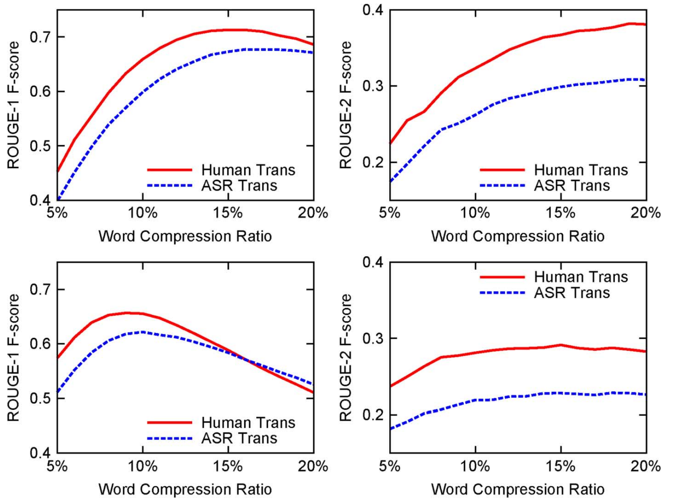


Fig. 3. Use human or ASR transcripts as input to the MMR system. Render either the uncompressed (extractive) summaries (upper two figures) or the compressed summaries (lower two figures).

TABLE VII
USE ORIGINAL HUMAN AND ASR TRANSCRIPTS AS INPUT, RENDER THE COMPRESSED AND UNCOMPRESSED SUMMARIES. RESULTS COMPARED TO THE REFERENCE COMPRESSED SUMMARIES

Eval	Summ Input	Summ Output	Compression Ratio 8%	9%	10%	11%	12%
R-1 F (%)	Human	Orig	59.77	**60.60**	60.44	60.02	59.14
		Comp	65.29	**65.67**	65.55	64.73	63.41
	ASR	Orig	56.06	57.21	57.82	**58.00**	57.83
		Comp	60.59	61.80	**62.17**	61.64	61.22
R-2 F (%)	Human	Orig	21.77	22.67	22.84	23.26	**23.74**
		Comp	27.53	27.75	28.12	28.43	**28.65**
	ASR	Orig	18.57	18.67	19.07	19.65	**19.86**
		Comp	20.68	21.33	21.92	21.94	**22.38**

In Table VII, we presented the ROUGE scores of comparing the system summaries against the reference compressed summaries, using both human and ASR transcripts as MMR input. When using the original utterances to render the uncompressed extractive summary, the system summaries contain many redundant words and result in inferior performance. When using the compressed utterances for summary rendering, more important contents can be included using the space saved by eliminating the redundant words. This yielded 5% absolute performance increase on both human and ASR transcripts as evaluated by the the ROUGE-1 scores. The gain on ROUGE-2 scores is 5% and 2.5% respectively when using the human and ASR transcripts.

Since our eventual goal is to investigate the possibilities of using the compressed speech summaries as a bridge to approach the abstractive summaries, we also compared the system-generated uncompressed and compressed summaries against the human abstracts. Results are presented in Fig. 4. In general, we found that the human abstracts are highly compact, and the best system performance was achieved when using 2% to 3% word compression ratio, as evaluated by the ROUGE-1 scores. We found using the compressed utterances achieves better performance on both human and ASR transcripts. But the gain is limited since the human abstracts only contain very few words in contrast to the lengthy meeting recordings.

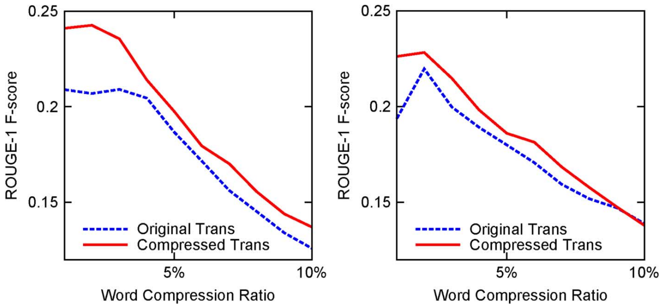


Fig. 4. Using human (left figure) or ASR transcripts (right figure) as input, render the uncompressed or compressed summaries. Results compared to the human abstracts and evaluated by ROUGE-1 F-scores.

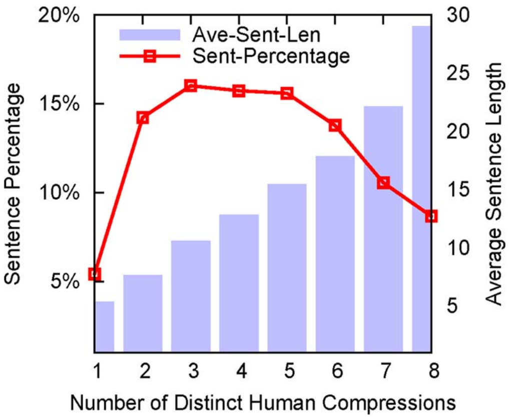


Fig. 5. Number of distinct human compressions, percentage of sentences with different number of compressions, and the average length of the sentences fall in each of the buckets.

VII. Discussion

We noticed that in many cases, there can be multiple ways of compressing a spoken utterance. Specifically, we observed that longer utterances may yield multiple compressions from the human annotators. In generating the human reference compressions using the Amazon Mechanical Turk, we employed 8 human annotators to compress each of the summary utterances. In Fig. 5, we present the number of distinct human compressions and the average length of the sentences within that bucket. We noticed that the longer sentences indeed tend to have more acceptable compressions, but the majority of the sentences yield 2, 3, 4 or 5 compressions, and fewer sentences have only one unique compression or too many compressions (7 or 8). This may be because some sentences are harder to compress than others, and different annotators may have their own preferences in performing the compressions. Moreover, it verifies that there can be multiple acceptable compressions for the majority of the spoken utterances.

In Table VIII, we show the most frequently dropped single words, part-of-speech tags, and their percentage in the total dropped words and POS tags. We also show the n-gram lengths and their percentages among all the dropped n-grams. These statistics are collected from the 26 meetings with 8 human compressions for each utterance. We see that the adverbs, prepositions, personal pronouns are among the most frequently dropped POS tags. We also notice that unigrams, bigrams, and trigrams together constitute 85.32% of the total dropped n-grams, which strongly support the CRF-based sequence labeling approach for spoken utterance compression. Note that in the preprocessing step, we removed the filled pauses "um/uh/eh," therefore these are not counted in the above statistics.

TABLE VIII
Most Frequently Dropped Words, Part-of-Speech Tags and Their Percentages in Total Dropped Words/POS Tags. N-Gram Lengths and Their Percentages in the Total Dropped n-Grams

Words	Per (%)	POS	Per (%)	N-gram	Per (%)
so	5.06	rb	16.27	1-gram	49.86
the	5.02	in	12.20	2-gram	25.06
i	4.95	prp	11.70	3-gram	10.40
and	4.94	dt	10.89	4-gram	5.88
that	4.17	cc	8.00	5-gram	3.11

TABLE IX
Most Frequently Dropped Bigrams, Trigrams, and Fourgrams by Human Annotators

Bigrams	**Trigrams**	**Fourgrams**
you know	a little bit	or something like that
sort of	so i think	and stuff like that
i mean	and so forth	i mean i guess
and then	more or less	and so i think
i think	so i mean	you know sort of
and so	and so on	on the other hand

TABLE X
System Compressed Summary Sentences for the Example Meeting Dialogue Segment Shown in Table I

System Compressed Summary	
423	there are a variety of ways of doing it
433	we could do something like a summary node of
444	a good to
446	necessarily try to think about what the decision variables are
450	the other trick which is not a technical trick it's a knowledge engineering trick is to make the each node sufficiently narrow that you don't get this combinatorics

In Table IX, we calculate the most frequently dropped bigrams, trigrams, and fourgrams by the human annotators. Many of these frequently dropped bigrams and trigrams can be effectively captured by our filler-phrase detection module. Actually, by only removing the filler phrases as done in Section IV.A, we can achieve 87.75% word compression ratio on the original human transcripts, which is an encouraging result.

In Table X, we show the system-compressed spoken utterances generated using the CRF system with 70% compression ratio for the example shown in Table I. We see that the system can successfully remove some redundancies, e.g., "so it's possible that," "some sort that." The system did not generate meaningful compression for utterance 444, since it contains some colloquial expressions such as "what i was gonna say," "at this point," which tend to be removed by the system. This example shows again there are a lot of challenges to the future work.

Finally it is worth pointing out that sentence compression is related to disfluency removal for spontaneous speech. Intuitively we expect that when humans compress a sentence, they will remove repetitions, revisions, and other disfluencies first, and then remove other words/phrases to further compress it. We performed some analysis to compare sentence compression with disfluency removal in [25] and showed that there is more compression than just disfluency removal, especially for long sentences. In this study we evaluated using sentence compression for summarization. There is also some previous work about the effect of disfluencies on summarization, such as [43], [44]. In our future work, we plan to investigate more how to more effectively use information about compression or disfluencies for summarization.

VIII. Conclusion

In this paper, we proposed to couple the spoken utterance compression and summarization systems for generating the compressed speech summaries. We investigated and compared the unsupervised ILP and the supervised CRF-based approach. We explored using multiple reference compressions to improve the labeling of the training process. The CRF-based system also integrated rich features, including the word and POS n-grams, position features, shallow and deep syntactic features, distance to the same words, etc. We evaluated the compression and summarization systems on both the human and ASR transcripts from the ICSI meeting corpus. Results show that the compressed summaries can incorporate more informative contents using the space created by eliminating the redundant words. In our future work, we would like to explore other speech-related features such as prosody to improve the CRF performance, as well as use these features to rerank the compression hypotheses in a two-step approach as used in [39]. In addition, we hope to recruit some graduate students to thoroughly analyze the compressed summaries, partly because the ICSI meetings are mainly scientific discussions and it is hard for the Mechanical Turk annotators to grasp the meeting discussions and give correct summary ratings. Finally, previous studies on joint selection and compression of the summary sentences mainly performed on the news domain [32], [34]. We would like to explore these possibilities on the speech transcripts.

Acknowledgment

We thank the reviewers for their insightful and enlightening comments.

References

[1] T. Cohn and M. Lapata, "Sentence compression as tree transduction," *J. Artif. Intell. Res.*, vol. 34, pp. 637–674, 2009.

[2] K. Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," *Comput. Linguist.*, vol. 28, no. 4, pp. 447–485, 2002.

[3] D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tür, "A global optimization framework for meeting summarization," in *Proc. ICASSP*, 2009, pp. 4769–4772.

[4] S. Xie, D. Hakkani-Tür, B. Favre, and Y. Liu, "Integrating prosodic features in extractive meeting summarization," in *Proc. ASRU*, 2009.

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

[6] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Proc. Eurospeech*, 2005.

[7] G. Murray, S. Renals, J. Carletta, and J. Moore, "Evaluating automatic summaries of meeting recordings," in *Proc. ACL MTSE Workshop*, 2005, pp. 39–52.

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

[9] J. Zhang, H. Y. Chan, P. Fung, and L. Cao, "A comparative study on speech summarization of broadcast news and lecture speech," in *Proc. Interspeech*, 2007.

[10] S. Xie, Y. Liu, and H. Lin, "Evaluating the effectiveness of features and sampling in extractive meeting summarization," in *Proc. IEEE Workshop Spoken Lang. Technol.*, 2008, pp. 157–160.

[11] S. Maskey and J. Hirschberg, "Summarizing speech without text using hidden Markov models," in *Proc. HLT/NAACL*, 2006.

[12] A. H. Buist, W. Kraaij, and S. Raaijmakers, "Automatic summarization of meeting data: A feasibility study," in *Proc. CLIN*, 2005.

[13] P. Fung, R. H. Y. Chan, and J. J. Zhang, "Rhetorical-state hidden Markov models for extractive speech summarization," in *Proc. ICASSP*, 2008, pp. 4957–4960.

[14] J. J. Zhang and P. Fung, "Learning deep rhetorical structure for extractive speech summarization," in *Proc. ICASSP*, 2010, pp. 5302–5305.

[15] J. J. Zhang, R. H. Y. Chan, and P. Fung, "Extractive speech summarization by active learning," in *Proc. IEEE Workshop Autom.. Speech Recognit. Understanding*, 2009, pp. 392–397.

[16] S. Xie, H. Lin, and Y. Liu, "Semi-supervised extractive speech summarization via co-training algorithm," in *Proc. INTERSPEECH*, 2010.

[17] S.-H. Lin and B. Chen, "A risk minimization framework for extractive speech summarization," in *Proc. ACL*, 2010.

[18] S.-H. Lin, Y.-T. Lo, Y.-M. Yeh, and B. Chen, "Hybrids of supervised and unsupervised models for extractive speech summarization," in *Proc. INTERSPEECH*, 2009.

[19] Y. Liu, S. Xie, and F. Liu, "Using n-best recognition output for extractive summarization and keyword extraction in meeting speech," in *Proc. ICASSP*, 2010, pp. 5310–5313.

[20] S. Xie and Y. Liu, "Using confusion networks for speech summarization," in *Proc. NAACL*, 2010.

[21] C. Hori and S. Furui, "A new approach to automatic speech summarization," *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 368–378, Sep. 2003.

[22] G. Murray, G. Carenini, and R. Ng, "Interpretation and transformation for abstracting conversations," in *Proc. NAACL*, 2010.

[23] G. Murray, G. Carenini, and R. Ng, "Generating and validating abstracts of meeting conversations: A user study," in *Proc. 6th Int. Natural Lang. Gener. Conf.*, 2010.

[24] F. Liu and Y. Liu, "From extractive to abstractive meeting summaries: Can it be done by sentence compression?," in *Proc. ACL*, 2009.

[25] F. Liu and Y. Liu, "Using spoken utterance compression for meeting summarization: A pilot study," in *Proc. IEEE Workshop Spoken Lang. Technol.*, 2010, pp. 37–42.

[26] K. Knight and D. Marcu, "Summarization beyond sentence extraction: A probabilistic approach to sentence compression," *Artif. Intell.*, vol. 139, pp. 91–107, 2002.

[27] J. Turner and E. Charniak, "Supervised and unsupervised learning for sentence compression," in *Proc. ACL*, 2005.

[28] M. Galley and K. McKeown, "Lexicalized Markov grammars for sentence compression," in *Proc. NAACL/HLT*, 2007.

[29] T. Nomoto, "Discriminative sentence compression with conditional random fields," *Inf. Process. Manage.*, vol. 43, pp. 1571–1587, 2007.

[30] T. Nomoto, "A generic sentence trimmer with CRFs," in *Proc. ACL*, 2008.

[31] J. Clarke and M. Lapata, "Global inference for sentence compression: An integer linear programming approach," *J. Artif. Intell. Res.*, vol. 31, pp. 273–381, 2008.

[32] D. M. Zajic, J. Lin, B. Dorr, and R. Schwartz, "Sentence compression as a component of a multi-document summarization system," in *Proc. DUC*, 2006.

[33] A. F. T. Martins and N. A. Smith, "Summarization with a joint model for sentence extraction and compression," in *Proc. Workshop Integer Linear Programming for Natural Lang. Process.*, 2009.

[34] T. Berg-Kirkpatrick, D. Gillick, and D. Klein, "Jointly learning to extract and compress," in *Proc. ACL/HLT*, 2011.

[35] 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, pp. 364–367.

[36] F. Liu, F. Liu, and Y. Liu, "A supervised framework for keyword extraction from meeting transcripts," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 538–548, Mar. 2011.

[37] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. SIGdial Workshop Discourse and Dialogue*, 2004, pp. 97–100.

[38] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciearena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lin, N. Tim, M. Ostendorf, K. Sönmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu, "Recent innovations in speech-to-text transcription at SRI-ICSI-UW," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1729–1744, Jul. 2006.

[39] D. Wang, X. Qian, and Y. Liu, "A two-step approach to sentence compression of spoken utterances," in *Proc. ACL*, 2012.

[40] T. Brants, "TnT—A statistical part-of-speech tagger," in *Proc. 6th Appl. NLP Conf.*, 2000, pp. 224–231.

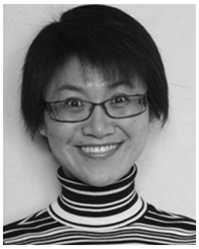
[41] C. Napoles, B. V. Durme, and C. Callison-Burch, "Evaluating sentence compression: Pitfalls and suggested remedies," in *Proc. ACL Workshop Monolingual Text-to-Text Generation*, 2011.

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

[43] Y. Liu, F. Liu, B. Li, and S. Xie, "Do disfluencies affect meeting summarization? A pilot study on the impact of disfluencies," in *Proc. MLMI*, 2007.

[44] X. Zhu and G. Penn, "Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization," in *Proc. HLT-NAACL*, 2006.

Fei Liu (M'11) received her B.S. degree in computer science and M.S. degree in computer science and engineering from Fudan University, Shanghai, China, in 2004 and 2007 respectively. She obtained her Ph.D. degree in computer science from the University of Texas at Dallas in 2011. She is a recipient of the Erik Jonsson Distinguished Research Scholarship from 2007 to 2011. Her research interests are in the areas of natural language processing, spoken language processing, machine learning, and social media text processing.



Yang Liu (M'05–SM'13) received B.S. and M.S. from Tsinghua University in China in 1997 and 2000 respectively, and a Ph.D. in electrical and computer engineering from Purdue University in 2004. She was a researcher at the International Computer Science Institute at Berkeley from 2002 to 2005. She is currently an Associate Professor at The University of Texas at Dallas. Her research interests are in the area of spoken language processing, natural language processing, and machine learning.