A Supervised Framework for Keyword Extraction From Meeting Transcripts

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

***Abstract*—This paper presents a supervised framework for extracting keywords from meeting transcripts, a genre that is significantly different from written text or other speech domains such as broadcast news. In addition to the traditional frequency- or position-based clues, we investigate a variety of novel features, including linguistically motivated term specificity features, decision-making sentence-related features, prosodic prominence scores, as well as a group of features derived from summary sentences. To generate better system summaries, we propose a feedback loop mechanism under a supervised framework to leverage the relationship between keywords and summary sentences. Experiments are performed on the ICSI meeting corpus using both human transcripts and automatic speech recognition (ASR) outputs. Results have shown that our proposed supervised framework is able to outperform both unsupervised term frequency $\times$ inverse document frequency (TF-IDF) weighting and a supervised keyphrase extraction system which is known for its satisfying performance on written text. We conduct extensive analysis to demonstrate the effectiveness of the newly proposed features and the feedback mechanism used to generate summaries. Furthermore, we show promising results using n-best recognition output to address the problems of recognition errors.**

***Index Terms*—Automatic speech recognition (ASR), keyword extraction, meeting transcripts, summarization, supervised learning, term frequency $\times$ inverse document frequency (TF-IDF).**

I. Introduction

With the fast development of recording and storage techniques, large amounts of audio/video recordings are becoming readily available nowadays. Unlike news broadcast, a lot of the newly emerged recordings exhibit the "casual" nature of speech, that is, they are often naturally occurring conversations and cover a wide range of topics, such as meeting recordings, class lectures, interviews, voice mails, or even homemade audio podcasts. It is well known that browsing through these huge amounts of spoken audio documents can be a challenging task because of the temporal nature of sound. Providing keywords for these audio files thus allows users to quickly grab the gist of the lengthy recordings and helps information access effectively.

Manuscript received March 13, 2009; revised December 15, 2009; accepted January 18, 2010. Date of publication June 07, 2010; date of current version December 03, 2010. This work was supported by the National Science Foundation (NSF) under awards IIS-0714132 and IIS-0845484. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ruhi Sarikaya.

F. Liu and Y. Liu are with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: feiliu@hlt.utdallas.edu; yangl@hlt.utdallas.edu).

F. Liu is with Health Care Informatics, University of Wisconsin-Milwaukee, Milwaukee, WI 53211 USA (e-mail: liuf@uwm.edu).

Digital Object Identifier 10.1109/TASL.2010.2052119

Keywords for spoken documents can also be used for a variety of spoken language processing tasks. For example, [1] built a spoken document retrieval system by extracting keywords from both the transcripts and speech queries. [2] proposed a real time system to continually retrieve documents that are most related to the keywords being mentioned in a meeting, to remind the participants of similar discussions in past meetings and help the decision-making process. [3] developed a Japanese speech understanding system by exploiting the confidence measures, dependencies, and relationship among important words. Moreover, keywords are also known to benefit language modeling in speech recognition [4], [5], as well as speech summarization task [6].

Most of the related work for keyword extraction has been performed on the written text domain, often based on the following four clues: 1) frequency, 2) word association, 3) sentence/document structure or position, and 4) linguistic knowledge. Term frequency $\times$ inverse document frequency (TF-IDF) weighting, as one of the simple yet robust frequency-based strategies, has been shown to be very effective for selecting important words for various text domains [7]–[12]. It is based on the assumption that keywords should appear frequently in a specific document, but do not occur frequently in the entire document collection. Other frequency-related approaches use residual IDF, variance of term frequency, gain, burstness, and so on [13]. The word association based approaches assume that important words tend to co-occur within a domain. Various knowledge resources have been leveraged to measure the association between word pairs, including the encyclopedia based lexical resources or a domain-specific thesaurus [14]–[16]. Web-based resources, such as Wikipedia or search engines, have also been leveraged recently to determine the word associations [11], [17], [18]. Corpus-based methods, such as latent semantic indexing (LSI), mutual information (MI), and Chi-square statistic are also popular for computing word co-occurrence probabilities [19], [20]. Position-based approaches generally assume keywords are more likely to appear in special positions of the document, such as title, headline, first paragraph, or important sentences [12], [16], [21]–[23]. In particular, [16] and [22] attempted to use an iterative reinforcement approach to do keyword extraction and summarization simultaneously, on the assumption that "important" sentences usually contain keywords, and keywords are usually seen in "important" sentences. Linguistic clues also play a significant role in locating important words, such as part-of-speech (POS) tag patterns [14], [24]. These information sources have been widely used in both unsupervised and supervised methods in previous studies. Even though unsupervised approaches have the advantage of domain independence and requiring no training

data, a supervised framework can often better combine multiple knowledge sources and achieve strong discriminative power [7], [12], [20], [21], [25], [26].

Compared to the text domain, there have been very limited studies on speech data. [15] compared two lexical resources, WordNet and EDR electronic dictionary, for extracting keywords from multiparty meeting corpus. They showed that leveraging semantic resources can yield significant performance improvement compared to the approach based on the relative frequency ratio (similar to IDF). [17] attempted to eliminate mistranscribed keyphrases based on semantic coherence (measured using mutual information). They showed some positive results, however, sometimes correct keyphrases are also removed. [25] evaluated the performance of the tool "Extractor" on broadcast news transcripts with various quality. They evaluated keywords extracted from automatic speech recognition (ASR) output and compared them with those generated from reference transcripts.

This paper investigates keyword extraction on the meeting domain. Meeting transcripts generally differ from written text significantly. The following lists a few differences that may have a negative impact on a keyword extraction system.

1) Lexical density, measured using the percentage of content words, is low for meeting transcripts. According to [27], two content words per clause are quite typical in unplanned spoken text; in contrast, written text can often have around four to six content words (or even more) per clause. Fewer content words pose a main problem to the traditional frequency-based keyword extraction system.
2) Meeting transcripts lack structure information, such as title, paragraph, topic or sentence boundaries.
3) Sentences in meetings are often poorly structured. There are many incomplete sentences, interruptions, and disfluencies.
4) Multiple participants in meetings introduce new challenges. For example, different people have different speaking styles and word usage, participants also have different roles in topic discussions, and each participant can begin a new topic when starting his/her turn. These phenomena do not exist in most text domains where a document is generally written by one person (note there are text domains that have multiple authors, such as forum data). Therefore for multiparty meetings, keyword extraction algorithms may need to be speaker dependent to account for different speakers.
5) High ASR error rate implies that reference keywords may be incorrectly recognized. In addition, word errors significantly degrade many syntactic and semantic analysis techniques, such as POS tagging and parsing, thus making it more difficult for deeper understanding of the data.

Previous work on keyword extraction in meetings (e.g., [28] and [29]) has experimented with both unsupervised and supervised approaches. Unsupervised approaches utilize frequency and position clues and can achieve satisfying performance as a baseline; however, it is not straightforward to incorporate various knowledge sources in this method. In contrast, it is easier to integrate a rich feature set in the supervised framework. The concern of domain dependence of the supervised framework can be alleviated by using features that are not very domain specific, for example, without using n-gram word features. Therefore, we adopt the supervised direction in this work, and extend the previous study in the following ways. 1) Beyond the traditionally widely used features (e.g., TF-IDF, position information), we introduce additional rich features including term specificity information, decision-making sentence related features, speaker and prominence based features, and features extracted from system generated summaries. 2) We propose a feedback strategy to reinforce the impact of summary sentences on selecting effective keywords. We first use system generated keywords as query words and utilize a query-focused sentence retrieval approach to generate summaries; these summaries are in turn used to extract features to include in the feature set for keyword extraction. 3) We conduct analysis to evaluate feature effectiveness using different feature selection processes, and define various measurements to characterize the quality of summaries that can benefit the keyword extraction task. 4) We evaluate system performance using both human transcripts and different ASR output (1-best and n-best), and show promising improved keyword extraction results using n-best ASR output over 1-best hypotheses. 5) We compare the performance against the keyword extraction system "Kea," which is known for its satisfying performance on written text. Results show that our supervised feedback system outperforms the "Kea" system on all the test conditions.

The rest of this paper is organized as follows. Section II describes the meeting data we used in this paper. Section III presents our supervised framework for keyword extraction. Section IV shows the experimental results and analysis. We conclude the paper with a summary and future direction in Section V.

II. Data

We use the ICSI meeting corpus [30], 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) [31], topic boundaries, and extractive summaries [32].

We recruited two computer science students to annotate keywords and topic categories for each topic segment, using 26 meetings selected from the ICSI meeting corpus.[1] The annotators were asked to listen to the audio recordings, read human transcripts, and select up to five words/phrases that can best convey the main content of a topic segment. Other than that, no specific annotation instructions were given. On average, it took one to two times real time for annotators to select keywords for each meeting. For topic labeling, six predefined high-level categories were provided based on the structural function of each topic segment, including "On-topic Discussion," "Digits," "Chitchat," "Opening," "Closing," and "Agenda." In the experiments of this paper, we only use topic segments that are tagged

[1]We selected these 26 meetings because they have been used in other previous studies for topic segmentation and summarization [32], [33]. Compared to the data set used in [28], one meeting (Bro015) is dropped because it is much shorter than all the other meetings.

as "On-topic Discussion" by both annotators, since these represent the main content of the meeting conversations. In total, there are 134 such topic segments for the 26 meetings. 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 keywords with more than three words are very rare (0.52%).

In general, we found that annotators have quite high disagreement on keyword selection, similar to an observation in previous studies [34]. We use two metrics to quantify the inter-annotator agreement: Kappa coefficient [35] and consistency rate. Kappa measures the degree of inter-annotator agreement beyond the amount expected by chance. The Kappa score between the two annotators is 0.41 in our data. This moderate inter-annotator agreement seems reasonable considering the inherent difficulties due to the informal conversational style. Consistency rate, defined as the proportion of selected keywords that are agreed by the two annotators, is 26.71%. However, we also noticed that the consistency rate reaches 80% for several topic segments. This suggests that human annotation consistency may depend on different input. Some topic segments do not have a clear focus and the discussion in it is casual and lacks order; therefore, it is very difficult to decide what are the most representative keywords. In addition, some meetings contain technical discussions and are hard for annotators without the proper background to understand.

In this paper, we evaluate automatic keyword extraction performance using different testing transcripts. In addition to human transcripts, we use different ASR transcripts obtained from a state-of-the-art SRI recognizer [36], [37]. The first one is the final recognition output after rescoring. It has a moderate word error rate (WER) of about 36.2% on our corpus. For this condition, we obtained the DA and topic boundary information by aligning the human annotation to the ASR words. We will refer to this condition "ASR" when there is no ambiguity (compared to its general meaning of automatic speech recognition). The second is n-best hypotheses from the recognizer. We will use both the 1-best and additional candidates in our experiments. The WER for the 1-best is higher than the above ASR transcript, about 41.6% for the 26 meetings. For this condition, we used the speech recognizer's segments, which are typically pause based; therefore, each resulting transcript segment does not correspond to a DA. These different testing scenarios are used to examine the effect of recognition and sentence segmentation errors on keyword extraction.

In Table I, we show an excerpt of a human transcript and the corresponding recognition output. Because of space limits, the n-best output is not presented. Human annotated keywords for the corresponding topic segment are given at the bottom of the table, and also shown in bold in the transcripts. For human transcripts, speaker ID is provided at the beginning of each dialog act. This example clearly shows the difference between the meeting genre and written text. The sentences in meetings are often poorly structured or incomplete, and contain many informal expressions and disfluencies.

Table II shows some statistics about the corpus used in this study. These scores are generated and averaged over all the topic segments. The percentage of the covered keywords measures how many of the human annotated keywords appear in the recognition output. As mentioned before, since the segmentation for 1-best was pause-based, the average "sentence" length is generally longer for that condition than the others (human transcripts or the ASR output with DA boundaries aligned from the reference ones), and the variance is also much larger. We also notice that, although "1-best" output has higher WER than "ASR," the keyword coverage rate is about the same for the two conditions.

TABLE I
EXCERPT OF A MEETING SEGMENT WITH HUMAN TRANSCRIPTS AND ASR OUTPUT. HUMAN ANNOTATED KEYWORDS ARE SHOWN AT THE BOTTOM

human transcript:
mn015: we should distill them out
fe004: mm-hmm
mn015: and put them where the schemas are
mn015: if there are things that you know are intention-specific then we should put them up somewhere
fe004: so in general they'll be **bindings** across both **intentions** and the **actions**
mn015:
mn015: yep
fe004: so
mn015: that's wonderful
fe004: yeah
fe004: so it's it's general across all of these things
fe004: it's like i mean shastri would say you know **binding** is like an essential cognitive uh process

ASR transcript:
we should just tell them out
and put them where the scheme
i saw there's things that you know our **intention** specific they we should put them up somewhere
so in general be findings across both **intentions** in the **actions**
so
that's wonderful
yeah
so it's generates general across all these things
like i mean i think shuster was that you know finding is like jill cognitive uh process

1-best transcript:
we should just tell them out and put them where this came i saw there are things that you know our **intention** specific they worship put them up somewhere
uhhuh
so in general they'll be findings across both **intentions** in the **actions** too
yeah
yep
that's wonderful
yeah so it generates general across all the things it's like i mean i think shell she was that you know finding it's like jill cognitive process

Human annotated keywords:
Annotator 1: binding, intention, action
Annotator 2: binding, intentions, rad, parser input, domain object

III. SUPERVISED FRAMEWORK

Our task is to extract keywords for each of the topic segments in the meeting transcript. Therefore, we will use "topic segment" and "document" interchangeably in the following of the paper, to represent the individual processing unit. Note that our task is different from keyword spotting, where a keyword is provided and the task is to spot it from the audio (along with its

TABLE II
STATISTICS ABOUT THE 26-MEETING CORPUS. THE NUMBERS ARE GENERATED AND AVERAGED ACROSS ALL 134 TOPIC SEGMENTS. "S.D." STANDS FOR STANDARD DEVIATION

Statistics of corpus	Human	ASR	1-best
Avg. num of words	1,867	1,704	1,861
Avg. num of sentences	269	219	202
Avg. sentence length	6.94	7.77	9.20
S.D. of sentence length	7.91	7.55	12.48
Avg. num of annotated keywords	5.92	-	-
S.D. of annotated keywords	2.18	-	-
Percentage of covered keywords (%)	100	65.48	65.00
Avg. WER (%)	0.00	36.19	41.62

transcript). In this psper, we consider single words as keyword 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.

Under the supervised framework, each candidate word is represented by a feature vector. We only consider lemmatized content words (i.e., noun, verb, adjective, and adverb) as candidate words for training and testing. We use a maximum entropy classifier to assign the posterior probability of a word being a keyword. Section III-A provides details of the features we use in the supervised classification framework. Section III-B explains the feedback strategy we propose to reinforce the effect of important sentences on the task of keyword extraction.

A. Features

We first describe several widely used features in prior work for keyword extraction, and then list the features we introduce in this study.

- Commonly used features.
  - — We use three frequency related features: TF (term frequency), IDF (inverse document frequency), and the product of them, TF-IDF. These effectively identify words that appear frequently in a document, but do not occur frequently in the entire document collection.
  - — Position features are used to represent the first occurrence of a candidate word. We compute this on a sentence or word basis, named "dist-sent" and "dist-word," respectively.
  - — Sentence length and salience score features are extracted from the sentences containing the word. For a sentence, its length is represented by the number of word tokens it contains, and its salience score is calculated based on its cosine similarity with the entire meeting under the vector space model. If a candidate word appears in several sentences, we use the length of the longest sentence and the highest sentence salience score among those sentences.[2]
- Term specificity features.
  Term specificity is generally defined as "the extent to which the word's referent can be touched or felt" [38]. Terms with high specificity usually carry more semantic content. As an example, "chipmunk" is considered a more specific term than "animal." Intuitively, these words with high specificity should be weighted more heavily in tasks such as information retrieval, since these words play a significant role in characterizing the document [13]. By contrast, stopwords contribute the least amount of information content, and should be removed or downweighted. Based on this motivation, we introduce two types of features for term specificity.
  The first feature is the number of senses a word has. Since specific words generally have more precise meaning, it is natural that they have fewer senses. This negative correlation between term specificity and the number of word senses has been confirmed in [13]. In this paper, we find the number of word senses from WordNet (version 1.7.1).
  Another feature we use to represent term specificity is based on stopwords. Since a word with low IDF score means that it occurs in many documents and is not topic indicative, we create stopword lists consisting of words with the lowest IDF values. Three binary features are defined: "sw-200," "sw-300," and "sw-500" to denote whether a candidate word is in the corresponding list with 200, 300, and 500 stopwords, respectively.
- Decision-making (DM) sentence features.
  We notice that some human selected keywords are likely to appear in decision making (DM) sentences. Similar findings have been confirmed in [39] and [40]. According to [39], decision making conversations are more likely to contain the indicative word *we*, such as "we should," "we will"; thus, we extract those sentences with a structure of "we + any verb + any noun" and consider them as an approximated collection of decision-making sentences. A binary feature "DM-in" is defined to indicate whether the target word has appeared in the DM sentences. In addition, we found that some adverbs (such as "actually," "basically," "especially") are also commonly used in conversations to emphasize a sentence or clarify a point; therefore, we use a binary feature "DM-adv" to indicate the co-occurrence of a target word and some predefined adverbs that are collected from the entire meeting corpus (after removing some frequent adverbs since they have very weak discriminative power).
- Summary features.
  Given a summary, we extract four features from it: a binary feature indicating whether a candidate word has appeared in the summary (summary-in); the frequency of the word in the summary (summary-tf); the normalized frequency by its total number of occurrences (tf-norm); as well as the ratio of its occurrence in summary sentences and non-summary sentences (tf-ratio). We notice that these features favor different summary lengths in order to have more discriminative power, for example, shorter summaries by the first feature (summary-in) and longer summaries by the other three features; therefore, we use summaries of different length for these features. We used system summaries that contain about 20%, 30%, and 40% of the total word tokens. These percentages are selected

[2]We also experimented using the average length and salience score of the sentences, but these did not yield better performance.

empirically with the expectation that they can cover most of the salient sentences in the document and provide discriminative information for keyword extraction. In addition, these are in accordance with the typical compression rate range used in summarization [41].
One simple method to generate a system summary is to select the longest sentences in the document until reaching the predefined compression ratio. We name this method "TopLen" and use it as a baseline. This approach has been shown to be a very strong baseline for speech summarization [42]. In Section III-B, we will investigate other query-focused summary generation approaches.

- Speech-related features.
  Even though our task is to extract keywords from transcripts, we do have access to the speech signal and speaker information (the corpus we used contains recordings from separate channels for different speakers). We use two features specific to conversational speech.
  The first one is related to speakers. In meeting conversations, important words may be said by many participants. We therefore define the feature "spkr-num" as the number of speakers who have mentioned the candidate word.
  The second feature is based on prosody (how a word is said). We expect that words with high pitch accent (i.e., prominence) are more likely to be keywords. In this paper, a word-level pitch accent detection module trained using read speech [43] is employed to generate the pitch accent likelihood score for each candidate word. For a word that appears more than once, we take the average of the likelihood scores associated with each appearance. We refer to this feature as "prominence."
- POS features.
  For a candidate word w with POS tag t_i, a feature "$pos(t_i)$" is defined as the relative frequency of this word with tag t_i within the document. In addition, we hypothesize that the POS of words around a candidate word may be useful cues for determining keywords. Therefore, we include similar tag frequency features for the previous and the following word, denoted as "$pos\_before(t_i)$" and "$pos\_after(t_i)$."

B. Single-Loop Feedback Strategy

The relationship between summarization and keyword extraction has recently been investigated in written text domain [16], [22], [44]. However, the graph-based mutual reinforcement approach seems to under-perform on the meeting transcripts, as shown in [29]. This is partly due to the informal conversational style. In this paper, we propose a single-loop feedback strategy to generate more keyword-related sentences while maintaining the supervised framework. We first use the supervised method with only a few base features to generate keywords, then employ the query-focused sentence retrieval approach [45]–[47] to generate summaries. These summaries are further used to extract summary-related features described above. This procedure is illustrated in Fig. 1.

The upper part of the figure shows the supervised keyword extraction process described above, with data flow along the solid lines and arrows. This approach uses a set of base features, including TF, IDF, TF-IDF, part-of-speech, and stopwords features. The lower part of the figure illustrates the feedback loop using query-focused summaries to extract features for another round of keyword extraction. The top $h\%$ (we use 60%) of words with highest confidence scores from the first pass keyword extraction are used as input query words, and all the sentences are ranked using a query-focused approach [45], [46]. The score of a sentence s given a query q, $p(s|q)$, is calculated as the weighted sum of the sentence's similarity to the query and its similarities with the other sentences in the document, along with those sentences' salience scores. Sentences' salience scores are calculated in an iterative process. For the $(k+1)^{th}$ iteration

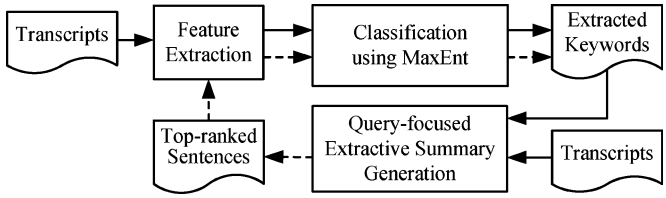

Fig. 1. Single-loop feedback strategy for keyword extraction.

$$p^{[k+1]}(s|q) = d\frac{sim(s,q)}{\sum_{z\in C} sim(z,q)} + (1-d)\sum_{v\in C}\frac{sim(s,v)}{\sum_{z\in C} sim(z,v)}p^{[k]}(v|q)$$

where $sim(\cdot,\cdot)$ is the similarity measure between two sentences or a sentence and a query, C is the set of all sentences in the document, and d is a trade-off parameter to weight the contribution from each component. A larger d means the sentence's relevance to the query is weighted higher. We set d as 0.6 in our experiments.

We employ the vector space model to calculate the sentence-sentence and sentence-query similarity, that is, each sentence and the query are represented as a vector of terms, then we calculate dot product between the two vectors. This approach favors longer sentences since previous research has shown that long sentences are very informative and can form appropriate summaries [42]. We consider two approaches to assign term weights when building the vectors for the sentences, and name the resulting summaries accordingly:

- TFIDF: product of a word's term frequency in the document and its inverse document frequency;
- CONF: confidence score from the supervised classifier in the first pass.

We further explored a simple reranking process to achieve maximum diversity in the top ranked sentences with respect to the contained query words. The basic idea is to push a sentence higher if it contains query words that are not seen before, and conversely lower if a sentence does not have any new query words. The reranking approach is shown in Algorithm 1. Its effect on diversifying the top-ranked sentences will be evaluated in Section IV-D.

Algorithm 1 Algorithm for reranking the summary sentences

Let S be the current ranked summary sentences
Let $N = |S|$ be the number of summary sentences in S
Let $R = \phi$ be the current ranked list of sentences
Let $Q = \phi$ be the set of query words
for i from 1 to N **do**
  if sentence s_i contains query words that are not in Q **then**
    add s_i to R and remove it from S
    add the new query words to Q
  end if
end for
Append S to R
Output R (reranked summary sentences)

IV. Experimental Results

A. Experimental Setup

For all the transcripts (human or recognition output), we used the TreeTagger [48] to lemmatize them, and the TnT part-of-speech tagger [49] trained from the Switchboard data to tag the transcripts. For the supervised framework, we use the maximum entropy classifier [50], and perform 9-fold cross validation on the 26-meeting corpus in all the experiments.

Two baseline systems are used in this study. First is an unsupervised system that uses only TF-IDF scores to rank all candidate words. Second is the state-of-the-art keyphrase extraction system "Kea"[3], which has been shown to perform satisfyingly on a variety of tasks. Kea uses Naive Bayes learning algorithm with four basic features: TF-IDF, first occurrence, length of the phrase, and node degree of the candidate phrase. We use the default setting of the Kea system without any controlled vocabulary.

Different performance metrics were employed in [29], including both automatic evaluation and human evaluation. Since in general human evaluation results are correlated with the automatic metrics, in this paper, we choose not to conduct human evaluation. For automatic evaluation, we use widely adopted precision/recall/F-score measurement. Given a set of system hypothesized unigrams and the corresponding human reference keywords, precision P is calculated as the number of matched unigrams, divided by the total number of system hypotheses; while recall R is the number of matched unigrams divided by the number of human reference words. The F-score is computed as the harmonic mean of precision and recall

$$F_1 = \frac{2 \times P \times R}{P + R}.$$

We use each annotation as a reference, and then compute the average score as the final result. Another automatic evaluation metric used in [29] is the weighted relative score that considers references from multiple annotators together and according to that, weights words differently. Because we use only two annotations in this paper, this evaluation is not substantially different from a simple average of the $P/R/F_1$ scores above; therefore, we only use one metric in this paper. 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.

[3]http://www.nzdl.org/Kea/

TABLE III
Keyword Extraction Results for Both Human Transcripts and ASR Output Using TF-IDF/Kea/Supervised Approaches. $\star$ and † Mean That the Improvement of the Supervised Approach Over the TF-IDF Method Is Statistical Significant at the Confidence Level of 95% and 90%, Respectively

		P(%)	R(%)	F(%)
Human	TF-IDF	35.33	32.50	33.78
	Kea	36.14	33.19	34.49
	Supervised-TopLen	39.02	35.38	37.01*
	Supervised-TFIDF	41.59	37.10	39.11*
	Supervised-CONF	**42.19**	**38.28**	**40.03***
ASR	TF-IDF	26.12	23.71	24.81
	Kea	27.95	25.34	26.50
	Supervised-TopLen	27.40	25.86	26.52
	Supervised-TFIDF	29.14	27.04	27.96†
	Supervised-CONF	**29.44**	**27.93**	**28.56**†

B. Results of Supervised Framework

Table III shows the results of our supervised keyword extraction approaches, in comparison with the two baseline systems, namely unsupervised TF-IDF weighting and Kea system. We performed significance test using McNemar test for the F-scores with respect to the TF-IDF baseline system. For each method, we output five unigrams in this experiment. For the supervised approach, we show results for three different settings according to the feedback module and summary generation approaches:

- **Supervised-TopLen:** Without feedback; summary-related features are extracted from "TopLen" summaries.
- **Supervised-TFIDF:** With feedback; summary-related features are extracted from "TFIDF" summaries.
- **Supervised-CONF:** With feedback; summary-related features are extracted from "CONF" summaries.

As can be seen from Table III, based on the F-score, the Kea system outperforms TF-IDF weighting by 0.71% and 1.65% (absolute) on human transcripts and ASR output, respectively. This indicates that the keyword extraction approaches developed in written text domain may not be directly applied well to the conversational speech data. By contrast, our supervised approach with "CONF" summaries outperforms TF-IDF weighting by 6.25% (abs.) on human transcripts, and 3.75% (abs.) on ASR output. The improvement is observed consistently across both recall and precision rates. When compared to the "Supervised-TopLen" system, which uses the same feature sets except the summary related features, the "Supervised-CONF" system yields an improvement of about 3% and 2%, respectively, on human and ASR transcripts. This shows that the summary features generated after the feedback loop are effective, boosting keyword extraction performance. More detailed analysis about using summaries for keyword extraction is provided in Section IV-D.

TABLE IV
COMPARISON BETWEEN HUMAN-ANNOTATED KEYWORDS AND SYSTEM-GENERATED HYPOTHESIS FOR THE MEETING EXAMPLE SHOWN IN TABLE I. ALL THE KEYWORDS ARE LEMMATIZED

Annotator 1	action intention binding
Annotator 2	intention domain object rad binding parser input
System (Human)	action intention domain rad schemas
System (ASR)	action intention specific module general

In Table IV, we demonstrate keyword extraction performance for the example shown in Table I. The system-generated keywords are presented for both human transcripts and ASR output. These results are generated using the supervised system with "CONF" weighting for summary generation. For comparison, the lemmatized human annotated keywords are also provided. On human transcripts, our system successfully extracted "action," "intention," "domain," and "rad" as keywords, while on ASR output only "action" and "intention" were extracted. This is partly because as many as 34.52% of the human annotated keywords are missing in the ASR output due to the recognition errors. For example, "domain," "parser input," "binding" were recognized as "main," "part be towards," and "finding," respectively. This high rate of missing the reference keywords results in an upper bound of recall rate (around 61.5% in our data) — no matter what the keyword extraction approach is, these reference words will not be hypothesized in the system output. This example shows that ASR errors can significantly impact keyword extraction performance, resulting in misses of some important words, change of the weight of desired keywords, and maybe selection of incorrectly recognized words. We will make further investigation about the effect of n-best ASR output on keyword extraction in Section IV-E.

C. Analysis I: Feature Analysis

To measure the effectiveness of different features, we perform three feature selection processes.

- **Forward feature selection.** It begins with an empty set, and iteratively adds the feature that achieves the largest performance gain when combined with the current selected features.
- **Backward feature selection.** It starts with the entire feature set and removes one feature in each iteration to have the least performance degradation.
- **Dynamic programming (DP) based feature selection.** Similar to forward feature selection, this approach also starts with an empty set. Unlike forward and backward methods that only keep the best feature subset in each iteration, this approach maintains N best feature subsets from a set of N features in each iteration using a dynamic programming-based strategy [51].

In all three processes, the final feature set is the one that achieves the best performance during the iterations. The three approaches have their own merits, though none of them can guarantee the selection optimality. Forward and backward feature selection approaches adopt the greedy strategy but they are computationally more efficient; dynamic programming-based approach enlarges the searching space and results in feature subsets with more divergence.

TABLE V
FEATURE SELECTION RESULTS USING FORWARD (NOTED AS "FW"), BACKWARD (NOTED AS "BW"), AND DYNAMIC PROGRAMMING APPROACHES (NOTED AS "DP"). "X" INDICATES THE CORRESPONDING FEATURE IS SELECTED. "-" MEANS THE FEATURE IS NOT AVAILABLE. THE LAST ROW SHOWS THE F-SCORE USING THE SELECTED BEST FEATURE SUBSET

		Human			ASR		
Category	**Features**	Fw	Bw	Dp	Fw	Bw	Dp
Frequency	TF	X	X	X		X	X
	IDF		X	X	X		X
	TF-IDF	X	X	X	X	X	X
Position	dist-word	X	X	X	X		X
	dist-sent					X	X
Sentence	sent-len		X				
	sent-score	X	X	X	X	X	X
Term specificity	sense-num	X	X	X	X	X	X
	sw-200	X	X		X	X	X
	sw-300	X	X	X			
	sw-500			X			X
DM-related	DM-in		X	X	X	X	
	DM-adv	X	X	X			
Summary	summary-in	X	X	X	X	X	X
	summary-tf	X	X	X	X	X	X
	tf-norm	X	X	X	X	X	X
	tf-ratio	X	X	X	X	X	X
Speech	spkr-num			X		X	
	prominence	X	X		-	-	-
POS	pos	X		X	X	X	X
	pos-context		X	X		X	
F-score (%)		41.22	41.38	42.34	29.93	30.10	30.52

Since our annotated data is limited, we perform this feature analysis using all the data in the same cross-validation setup. When more data is available, we would like to test the selected feature set on a blind test set. From this experiment, we hope to exploit the best performed feature combinations on the conversational speech data for the keyword extraction task.

Results are shown in Table V. Not surprisingly, the traditional "TF-IDF," "pos," "sent-score" features all perform well on the speech data. In addition, the linguistically motivated term specificity feature, such as "sense-num," and the summary-related features also play a significant role in boosting the performance. By contrast, features "dist-sent" and "sent-len" are rarely selected in the final set. This could be because the spoken sentences are usually poorly structured and there is a large variation in length, as shown in Table II. The speech-related "prominence" feature also benefits the keyword extraction task on human transcripts. We did not include this feature for ASR condition, as the word alignment information was not available. It is also worth pointing out that since the prominence prosody model was trained using read speech, this may have limited its benefit on conversational speech due to the mismatched conditions. We will continue to investigate using prominence information in the future.

Overall, we observe improved performance after the feature selection processes, compared to using all the features. Among the three feature selection approaches, dynamic programming based approach achieves slightly higher performance; while there is no significant performance difference between the forward and backward feature selection processes. In addition, there are many common features in the final selected features from the three approaches. We also conducted an oracle experiment using human reference summaries [52] to generate summary-related features, and achieved an F-score of 43.30% and 43.18%, respectively, for forward and backward processes on human transcripts. This better performance over using

TABLE VI
KEYWORD EXTRACTION PERFORMANCE USING FEATURES FROM DIFFERENT SUMMARIES. ALSO SHOWN IS THE COVERAGE OF REFERENCE KEYWORDS IN DIFFERENT SUMMARIES, WITH LENGTH ABOUT 20%, 30%, AND 40% OF THE TOTAL WORD TOKENS. "NOR" MEANS WITHOUT THE RERANKING PROCESS

	Summary	Extraction F-measure %	Keyword Coverage %		
			Len. 20%	30%	40%
Human	TopLen	37.01	59.31	73.27	81.94
	TFIDF-NoR	37.63	68.46	77.21	83.95
	CONF-NoR	37.20	67.82	79.86	87.32
	TFIDF	**39.11**	**70.95**	**80.82**	**88.76**
	CONF	**40.03**	**71.67**	**83.87**	**91.33**
ASR	TopLen	26.52	34.03	42.86	49.60
	TFIDF-NoR	27.68	45.67	52.09	56.26
	CONF-NoR	26.86	44.70	51.52	56.58
	TFIDF	**27.96**	**47.59**	**53.21**	**58.51**
	CONF	**28.56**	**45.83**	**53.61**	**59.31**

system-generated summaries indicates that further improvements using better summaries are still possible for keyword extraction.

D. Analysis II: Impact of Summaries on Keyword Extraction

Although the relationship between keyword extraction and summarization has received some attention in recent years, limited research has been conducted on what kind of summary can boost keyword extraction performance. As shown in Table III, different summaries can result in very different keyword extraction performance. We have seen that the confidence score based "CONF" summaries perform better than the "TopLen" and "TFIDF" summaries on both human transcripts and ASR output. In this section, we will analyze the effect of different summaries, expecting this line of work can benefit further research.

First, we evaluate if the reranking procedure (see Section III-B) helps generate more informative summaries for keyword extraction. Table VI shows the performance of different supervised systems. "NoR" means the corresponding summary is generated via the feedback mechanism but without the reranking process. We can see that the supervised "CONF" system achieves the best performance with the reranking module, with an improvement of F-score from 37.20% to 40.03% on human transcripts, and 26.86% to 28.56% on the ASR output, compared to without using reranking. For "TFIDF," the improvement from reranking is not as large as in "CONF," especially for the ASR condition. In the same table, we also show the keyword coverage statistics of different summaries. After reranking, "CONF" summary contains more than 10% of the human annotated keywords compared to the "TopLen" summary. This holds for different summary lengths and on both human and ASR transcripts. The better keyword coverage in the summaries partly explains the better performance using the corresponding summaries for keyword extraction.

For further analysis, we evaluate the correlation between the summary quality and its impact on keyword extraction. In total, we define seven measurements. Four keyword-related statistics are calculated to measure the distribution of keywords in the summaries.

TABLE VII
CORRELATION BETWEEN KEYWORD EXTRACTION F-SCORES AND VARIOUS STATISTICS FOR SUMMARIES

	Human	ASR
ST-1	**0.7808**	**0.8315**
ST-2	**0.8742**	**0.8938**
ST-3	0.3999	0.4084
ST-4	-0.0116	-0.2821
ROUGE-1	-0.2320	-0.0049
ROUGE-2	0.5037	0.7796
ROUGE-SU4	0.2173	0.5611

1) Keyword Coverage (ST_1): percentage of reference keyword types that are covered in summary. This is the statistic shown before.
2) Normalized Keyword Frequency (ST_2): total frequency of reference keywords contained in the summary divided by the total frequency of reference keywords in the entire transcript.
3) Keyword Percentage (ST_3): total number of reference keywords contained in the summary divided by the total number of word tokens in the summary.
4) Sentence Percentage (ST_4): number of sentences that contain at least one reference keyword, divided by the total number of sentences in the summary.

Another three measurements come from the ROUGE scores [53], which calculate the word overlap between a system summary and the human reference summary. We use the reference summary that is described in [52]. In total, we generated three ROUGE scores:

1) ROUGE-1: unigram match;
2) ROUGE-2: bigram-based match;
3) ROUGE-SU4: skip-bigram plus unigram-based match.

For the three types of summaries ("TopLen," "TFIDF," and "CONF"), we used different length from 10% to 50% of the total word tokens, with 5% interval. The unsupervised TF-IDF weighting was used to perform keyword extraction using these summary sentences with different compression ratios. This approach is the same as that used for the entire document, except that only a subset of the original sentences are used as input in this experiment. We then compute the correlation (Spearman's rho) between the F-scores and the different measurements described above. The correlation results are shown in Table VII for both human transcripts and ASR output.

From the correlation results, we observe that Normalized Keyword Frequency (ST_2) is highly correlated with keyword extraction performance, for both human transcripts and ASR output. Keyword Coverage (ST_1) measure is the second most correlated one. This is also consistent with our findings in the feature selection results, where the summary-related features "summary-in," "summary-tf" and "tf-norm" features are all included in the final feature combinations. In contrast, among the three ROUGE scores, only ROUGE-2 shows reasonable correlation with the F-score on ASR output, while ROUGE-1 and ROUGE-SU4 scores have lower correlation with the keyword extraction performance, suggesting that the criteria used to optimize summarization and keyword extraction are different: ROUGE measures the matches of all the words in the

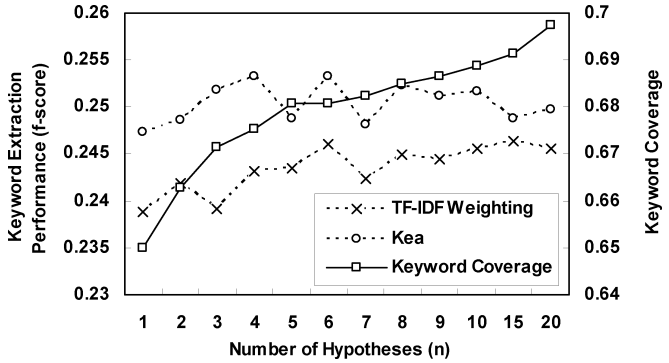

Fig. 2. Keyword extraction performance (left Y-axis) and reference keyword coverage (right Y-axis) using n-best output.

summaries, whereas for keyword extraction, it is more important to ensure a good coverage of keywords. Overall it seems that both ST_1 and ST_2 statistics can be used as reasonable indicators of whether the generated summaries are useful for the keyword extraction task.

For this analysis, we used unsupervised keyword extraction. We made this choice for different reasons. The unsupervised method is much simpler computationally and its performance is reasonably good. Also, the unsupervised approach can use different summary lengths and thus creates more data points for statistical analysis (e.g., nine results for each of the summary types, corresponding to compression ratios ranging from 10% to 50% with 5% interval). In contrast, the supervised approach uses the summaries with different length together to extract features and provides one result corresponding to one summary type. Even though the unsupervised performance is not as good as the supervised approach, the general trend is similar between the two methods in terms of the effectiveness of the summary sentences (we performed analysis using supervised keyword extraction and observed similar patterns). Note that this correlation study is only for analysis purpose. We cannot use these measurements during keyword extraction since the reference keywords are unknown beforehand.

E. Using N-Best Hypotheses for Keyword Extraction

As we have seen, there is a degradation when using ASR output for keyword extraction, mainly due to the word errors. In this section, we investigate if using rich ASR output, e.g., n-best hypotheses, can help improve performance. We use n-best hypotheses for $n = 1, 2, \ldots, 10, 15, 20$ in this experiment. Since our supervised framework requires the generation of different summaries, it is not straightforward to use it on n-best ($n > 1$) output (summarization using n-best is still an area that is understudied). Hence, we applied two keyword extraction systems, the unsupervised TF-IDF weighting and the Kea system. For both methods, we put n-best hypotheses together for each sentence, and then apply the algorithm as is to the expanded transcripts for each document. The keyword extraction results are shown in Fig. 2 (left Y-axis). For each n-best, we also calculate its coverage of reference keywords, shown in the same figure (right Y-axis).

We can see from the figure that the keyword coverage gradually improves as the number of hypotheses increases, compared to the coverage of 1-best result (65%). For both keyword extraction systems, there is a general trend of improved performance when n increases from 1 up to 6, even though there is some fluctuation as n changes. When n is 6, both TF-IDF weighting and Kea system achieve the best performance. After that, although the keyword coverage continues to increase, there is no improvement in keyword extraction performance. This preliminary study on n-best hypotheses shows that, more ASR hypotheses may contain more reference keywords and 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 [25]. Therefore, how to effectively utilize rich ASR output (n-best or lattices) is still a challenging problem for keyword extraction on speech transcripts. One possible way is to take into account the confidence measures of the recognition hypotheses. We will investigate this in our future study.

F. Discussions

In this section, we summarize a few findings regarding keyword extraction from speech transcripts based on our error analysis. We hope these different error categories will point out some future research directions.

- **Frequency-based methods.** Most of the approaches, whether it is unsupervised TF-IDF weighting or the supervised approach as used in this study, rely heavily on a word's frequency in the document to determine its importance. However, human annotated keywords may not occur frequently in the transcripts. In fact, we found that about 35% of the human annotated keywords occur only once or twice, and about 57% of the annotated keywords occur less than five times in the corresponding meeting transcripts. For example, "hire" was selected as a keyword by one annotator for a topic segment, because the meeting participants discussed to hire a student worker. However, "hire" was only mentioned once since the conversation later shifted to how to put the student on the payroll. A more intelligent system needs to understand the semantic relatedness to identity the low-frequency keywords.
- **Human annotation agreement and evaluation issues.** As discussed in Section II, annotating keywords is considered difficult by our annotators, and the human annotation agreement is not high for meetings. This poses problems for both training classification models and evaluating system performance. Furthermore, since human annotators only mark a fixed number of keywords, a system may generate a keyword that is not on the list of human annotated keywords but still acceptable. Hence, the current evaluation metric may be too strict. Human evaluation was used in [29] where human judges were asked to reject unacceptable keywords. That resulted in higher keyword extraction scores. However, human evaluation is expensive and not always feasible for system development. How to define a more effective annotation and evaluation scheme for speech transcripts therefore remains to be an interesting topic.
- **ASR errors.** High WER explains some errors in keyword extraction. Compared to F-measure of 40.05% using

human transcripts, we obtain an F-score of 28.56% on ASR output (WER is 36.19%), and 26.22% on 1-best output (with WER of 41.62%), all using the "Supervised-CONF" system. The word errors cause serious problems especially when the reference keywords are incorrectly recognized. In addition, they have a negative impact on various features, such as POS tagging, parsing, semantic relatedness, and as a consequence degrade keyword extraction performance. Our preliminary investigation using n-best hypotheses has shown some promising results. Therefore, effectively using the n-best hypotheses or lattices may compensate for the high word errors and is an important direction for developing keyword extraction systems for spoken documents.

- **Unigram-based approaches.** The current methods we use are based on single words. As described in Section II, a large percent of the human annotated keywords are phrases. Since we only generate five unigram words, there will be missing keywords. In [28], a web resource-based approach was used to generate bigrams. In a preliminary study, we also noticed that some top ranked unigrams can be naturally combined into bigrams or trigrams. After this combination, we can then generate additional unigram keyword candidates. In this way, the system output contains both unigrams and phrases, with a total of more than five words. This yields a higher recall rate, without much loss of precision, and thus better F-measure. We will investigate along this direction and try to extract more robust keyphrases in our future work.

V. Conclusion

In this paper, we investigated a supervised framework for extracting keywords from the meeting transcripts. A variety of features are proposed beyond the traditional frequency- and position-based features, including term specificity features, decision-making sentences-related features, prosodic prominence score, and a set of features extracted from the system generated summaries. We also proposed a single-loop feedback strategy under the supervised framework to leverage the relationship between keyword extraction and summarization. We performed experiments using the ICSI meeting corpus for both human transcripts and different ASR outputs. Our experiments show that the supervised framework with confidence score-based summaries achieves the best performance, and our proposed method outperforms the unsupervised TF-IDF weighting and a state-of-the-art keyphrase extraction system on all testing conditions. We conducted extensive analysis and demonstrated the effectiveness of the newly proposed features and the feedback summary generation method. In addition, we defined some measurements to characterize the summary quality in order to understand which summaries benefit the keyword extraction task. We also evaluated using n-best hypotheses for keyword extraction and presented promising results. Our analysis showed different reasons for errors in the current keyword extraction systems that are due to the frequency-oriented approach, human annotation and evaluation issues, high ASR error rates, and unigram-based setup. These point out interesting directions for future studies.

Acknowledgment

The authors would like to thank the three anonymous reviewers of this paper for their insightful comments. They would also like to thank the University of Edinburgh for providing the annotation for the ICSI meeting corpus, J. H. Jeon for generating the word prominence information, and D. Pennell for useful discussions. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of NSF.

References

[1] C. H. Wu, C. L. Huang, C. S. Hsu, and K. M. Lee, "Speech retrieval using spoken keyword extraction and semantic verification," in *Proc. TENCON*, 2007, pp. 1–4.

[2] A. Popescu-Belis, E. Boertjes, J. Kilgour, P. Poller, S. Castronovo, T. Wilson, A. Jaimes, and J. Carletta, "AMIDA automatic content linking device: Just-in-time document retrieval in meetings," in *Proc. MLMI*, 2008, pp. 273–284.

[3] K. Shimada, T. Endo, and S. Minewaki, "Speech understanding based on keyword extraction and relations between words," *Comput. Intell.*, vol. 23, no. 1, pp. 45–60, 2007.

[4] I. Bulyko, M. Ostendorf, M. Siu, T. Ng, A. Stolcke, and O. Çetin, "Web resources for language modeling in conversational speech recognition," *ACM Trans. Speech Lang. Process.*, vol. 5, no. 1, pp. 1–25, 2007.

[5] C. Munteanu, G. Penn, and R. Baecker, "Web-based language modelling for automatic lecture transcription," in *Proc. Interspeech*, 2007, pp. 2353–2356.

[6] K. Riedhammer, B. Favre, and D. Hakkani-Tur, "A keyphrase based approach to interactive meeting summarization," in *Proc. IEEE Workshop Spoken Lang. Technol.*, 2008, pp. 153–156.

[7] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning, "Domain-specific keyphrase extraction," in *Proc. IJCAI*, 1999, pp. 668–673.

[8] I. H. Witten, G. W. Paynter, E. Fr, C. Gutwin, and C. G. Nevill-Manning, "KEA: Practical automatic keyphrase extraction," in *Proc. ACM Digital Libraries*, 1999, pp. 254–256.

[9] S. Jones and G. W. Paynter, "Automatic extraction of document keyphrases for use in digital libraries: Evaluation and applications," *J. Amer. Soc. Inf. Science Technol.*, vol. 53, no. 8, pp. 653–677, 2002.

[10] Y. HaCohen-Kerner, "Automatic extraction of keywords from abstracts," in *Proc. 7th Int. Conf. Knowledge-Based Intell. Inf. Eng. Syst.*, 2003, vol. 2773, pp. 843–849.

[11] P. Turney, "Coherent keyphrase extraction via web mining," in *Proc. IJCAI*, 2003, pp. 434–439.

[12] Y. HaCohen-Kerner, Z. Gross, and A. Masa, "Automatic extraction and learning of keyphrases from scientific articles," *Comput. Linguist. Intell. Text Process.*, pp. 657–669, 2005.

[13] K. Kireyev, "Semantic-based estimation of term informativeness," in *Proc. NAACL*, 2009, pp. 530–538.

[14] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in *Proc. EMNLP*, 2003, pp. 216–223.

[15] 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 *Proc. LREC*, 2004, pp. 2205–2208.

[16] X. Wan, J. Yang, and J. Xiao, "Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction," in *Proc. ACL*, 2007, pp. 552–559.

[17] D. Inkpen and A. Désilets, "Extracting semantically-coherent keyphrases from speech," *Canadian Acoust. Assoc.*, vol. 32, pp. 130–131, 2004.

[18] K. Coursey, R. Mihalcea, and W. Moen, "Automatic keyword extraction for learning object repositories," in *Proc. Conf. Amer. Soc. Inf. Sci. Technol.*, 2008.

[19] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *Int. J. Artif. Intell.*, vol. 13, no. 1, pp. 157–169, 2004.

[20] A. Csomai and R. Mihalcea, "Linguistically motivated features for enhanced back-of-the-book indexing," in *Proc. ACL*, 2008, pp. 932–940.

[21] P. D. Turney, "Learning algorithms for keyphrase extraction," *Inf. Retrieval*, vol. 2, no. 4, pp. 303–336, 2000.

[22] H. Zha, "Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering," in *Proc. SIGIR*, 2002, pp. 113–120.

[23] Y. HaCohen-Kerner, I. Stern, D. Korkus, and E. Fredj, "Automatic machine learning of keyphrase extraction from short html documents written in Hebrew," *Cybern. Syst.*, vol. 38, no. 1, pp. 1–21, 2007.
[24] E. D'Avanzo, B. Magnini, and A. Vallin, "Keyphrase extraction for summarization purposes: The LAKE system at DUC-2004," in *Proc. Document Understanding Conf.*, 2004.
[25] A. Désilets, B. D. Bruijn, and J. Martin, "Extracting keyphrases from spoken audio documents," *Inf. Retrieval Tech. for Speech Applicat.*, vol. 2273, pp. 36–50, 2002.
[26] A. Hulth, "Reducing false positives by expert combination in automatic keyword indexing," in *Proc. RANLP*, 2003, pp. 197–203.
[27] M. A. K. Halliday, "Some grammatical problems in scientific english," *Writing Science: Literacy and Discursive Power*, pp. 69–85, 1993.
[28] F. Liu, F. Liu, and Y. Liu, "Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion," in *Proc. IEEE Workshop Spoken Lang. Technol.*, 2008, pp. 181–184.
[29] F. Liu, D. Pennell, F. Liu, and Y. Liu, "Unsupervised approaches to automatic keyword extraction using meeting transcripts," in *Proc. HLT-NAACL*, 2009, pp. 620–628.
[30] 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 *Proc. ICASSP*, 2003, pp. 364–367.
[31] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. SIGdial Workshop Discourse Dialogue*, 2004, pp. 97–100.
[32] G. Murray, S. Renals, J. Carletta, and J. Moore, "Evaluating automatic summaries of meeting recordings," in *Proc. ACL 2005 MTSE Workshop*, 2005, pp. 39–52.
[33] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proc. ACL*, 2003, pp. 562–569.
[34] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Commun. ACM*, vol. 30, no. 11, pp. 964–971, 1987.
[35] J. Carletta, "Assessing agreement on classification tasks: The Kappa statistic," *Comput. Linguist.*, vol. 22, pp. 249–254, 1996.
[36] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciearena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lin, T. Ng, 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, Oct. 2006.
[37] A. Janin, A. Stolcke, J. Frankel, O. Çetin, K. Boakye, X. Anguera, and C. Wooters, "The ICSI-SRI spring 2006 meeting speech-to-text system," in *Proc. MLMI*, 2006, pp. 444–456.
[38] J. Reilly and J. Kean, "Formal distinctiveness of high- and low- imageability nouns: Analyses and theoretical implications," *Cognitive Sci.*, vol. 31, no. 1, pp. 157–168, 2007.
[39] P. Y. Hsueh and J. Moore, "What decisions have you made: Automatic decision detection in conversational speech," in *Proc. NAACL-HLT*, 2007, pp. 25–32.
[40] R. Fernández, M. Frampton, P. Ehlen, M. Purver, and S. Peters, "Modelling and detecting decisions in multi-party dialogue," in *Proc. SIGDial Workshop Discourse Dialogue*, 2008, pp. 156–163.
[41] I. Mani, *Advances in Automatic Text Summarization*, M. T. Maybury, Ed. Cambridge, MA: MIT Press, 1999.
[42] G. Penn and X. Zhu, "A critical reassessment of evaluation baselines for speech summarization," in *Proc. ACL*, 2008, pp. 470–478.
[43] J. H. Jeon and Y. Liu, "Automatic accent detection: Effect of base units and boundary information," in *Proc. Interspeech*, 2009, pp. 180–183.
[44] F. Wei, W. Li, Q. Lu, and Y. He, "Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization," in *Proc. SIGIR*, 2008, pp. 283–290.
[45] R. Mihalcea and P. Tarau, "Text Rank: Bringing order into texts," in *Proc. EMNLP*, 2004, pp. 404–411.
[46] J. Otterbacher, G. Erkan, and D. R. Radev, "Using random walks for question-focused sentence retrieval," in *Proc. HLT-EMNLP*, 2005, pp. 915–922.
[47] L. Zhao, L. Wu, and X. Huang, "Using query expansion in graph-based approach for query-focused multi-document summarization," *Inf. Process. Manage.*, vol. 45, pp. 35–41, 2009.
[48] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proc. Int. Conf. New Methods Lang. Process.*, 1994, pp. 44–49.
[49] T. Brants, "TnT – A statistical part-of-speech tagger," in *Proc. 6th Appl. NLP Conf.*, 2000, pp. 224–231.
[50] H. Daumé, III, "Notes on CG and LM-BFGS optimization of logistic regression," 2004 [Online]. Available: http://www.cs.utah.edu/~hal/docs/daume04cg-bfgs.pdf, http://www.cs.utah.edu/~hal/megam/
[51] R. S. Cheung and B. A. Eisenstein, "Feature selection via dynamic programming for text-independent speaker identification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, no. 5, pp. 397–403, Oct. 1978.
[52] F. Liu and Y. Liu, "What are meeting summaries? A n analysis of human extractive summaries in meeting corpus," in *Proc. SIGDial Workshop Discourse Dialogue*, 2008, pp. 80–83.
[53] C. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop Text Summarization Branches Out*, 2004, pp. 74–81.

Fei Liu received the B.S. degree in computer science and the M.S. degree in computer science and engineering from Fudan University, Shanghai, China, in 2004 and 2007, respectively. She is currently pursuing the Ph.D. degree at the University of Texas at Dallas, Richardson.

She is currently a Research Assistant in the Computer Science Department, University of Texas at Dallas. Her research interests are in the field of natural language processing, machine learning, information retrieval, and spoken document understanding.

Feifan Liu received the Ph.D. degree in pattern recognition and intelligent systems from the Chinese Academy of Sciences, Beijing, in 2006.

He was a Postdoctoral Researcher at the University of Texas at Dallas, Richardson, from 2006 to 2009, and is currently an Associate Scientist (Research Faculty) at the University of Wisconsin-Milwaukee. His main research interests are natural language processing and biomedical text mining.

Yang Liu (M'05) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1997 and 2000, respectively, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 2004.

She was a Researcher at the International Computer Science Institute, Berkeley, CA, from 2002 to 2005. She has been an Assistant Professor at the University of Texas at Dallas, Richardson, since 2005. Her research interests are in the areas of spoken language processing, natural language processing, and machine learning.