Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts

Feifan Liu, Deana Pennell, Fei Liu and Yang Liu
Computer Science Department
The University of Texas at Dallas
Richardson, TX 75080, USA
{ffliu,deana,feiliu,yangl}@hlt.utdallas.edu

Abstract

This paper explores several unsupervised approaches to automatic keyword extraction using meeting transcripts. In the TFIDF (term frequency, inverse document frequency) weighting framework, we incorporated part-of-speech (POS) information, word clustering, and sentence salience score. We also evaluated a graph-based approach that measures the importance of a word based on its connection with other sentences or words. The system performance is evaluated in different ways, including comparison to human annotated keywords using F-measure and a weighted score relative to the oracle system performance, as well as a novel alternative human evaluation. Our results have shown that the simple unsupervised TFIDF approach performs reasonably well, and the additional information from POS and sentence score helps keyword extraction. However, the graph method is less effective for this domain. Experiments were also performed using speech recognition output and we observed degradation and different patterns compared to human transcripts.

1 Introduction

Keywords in a document provide important information about the content of the document. They can help users search through information more efficiently or decide whether to read a document. They can also be used for a variety of language processing tasks such as text categorization and information retrieval. However, most documents do not provide keywords. This is especially true for spoken documents. Current speech recognition system performance has improved significantly, but there is no rich structural information such as topics and keywords in the transcriptions. Therefore, there is a need to automatically generate keywords for the large amount of written or spoken documents available now.

There have been many efforts toward keyword extraction for text domain. In contrast, there is less work on speech transcripts. In this paper we focus on one speech genre — the multiparty meeting domain. Meeting speech is significantly different from written text and most other speech data. For example, there are typically multiple participants in a meeting, the discussion is not well organized, and the speech is spontaneous and contains disfluencies and ill-formed sentences. It is thus questionable whether we can adopt approaches that have been shown before to perform well in written text for automatic keyword extraction in meeting transcripts. In this paper, we evaluate several different keyword extraction algorithms using the transcripts of the ICSI meeting corpus. Starting from the simple TFIDF baseline, we introduce knowledge sources based on POS filtering, word clustering, and sentence salience score. In addition, we also investigate a graph-based algorithm in order to leverage more global information and reinforcement from summary sentences. We used different performance measurements: comparing to human annotated keywords using individual F-measures and a weighted score relative to the oracle system performance, and conducting novel human evaluation. Experiments were conducted using both the human transcripts and the speech recognition (ASR) out-

Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pages 620–628,
Boulder, Colorado, June 2009.

put. Overall the TFIDF based framework seems to work well for this domain, and the additional knowledge sources help improve system performance. The graph-based approach yielded worse results, especially for the ASR condition, suggesting further investigation for this task.

2 Related Work

TFIDF weighting has been widely used for keyword or key phrase extraction. The idea is to identify words that appear frequently in a document, but do not occur frequently in the entire document collection. Much work has shown that TFIDF is very effective in extracting keywords for scientific journals, e.g., (Frank et al., 1999; Hulth, 2003; Kerner et al., 2005). However, we may not have a big background collection that matches the test domain for a reliable IDF estimate. (Matsuo and Ishizuka, 2004) proposed a co-occurrence distribution based method using a clustering strategy for extracting keywords for a single document without relying on a large corpus, and reported promising results.

Web information has also been used as an additional knowledge source for keyword extraction. (Turney, 2002) selected a set of keywords first and then determined whether to add another keyword hypothesis based on its PMI (point-wise mutual information) score to the current selected keywords. The preselected keywords can be generated using basic extraction algorithms such as TFIDF. It is important to ensure the quality of the first selection for the subsequent addition of keywords. Other researchers also used PMI scores between each pair of candidate keywords to select the top k% of words that have the highest average PMI scores as the final keywords (Inkpen and Desilets, 2004).

Keyword extraction has also been treated as a classification task and solved using supervised machine learning approaches (Frank et al., 1999; Turney, 2000; Kerner et al., 2005; Turney, 2002; Turney, 2003). In these approaches, the learning algorithm needs to learn to classify candidate words in the documents into positive or negative examples using a set of features. Useful features for this approach include TFIDF and its variations, position of a phrase, POS information, and relative length of a phrase (Turney, 2000). Some of these features may not work well for meeting transcripts. For example, the position of a phrase (measured by the number of words before its first appearance divided by the document length) is very useful for news article text, since keywords often appear early in the document (e.g., in the first paragraph). However, for the less well structured meeting domain (lack of title and paragraph), these kinds of features may not be indicative. A supervised approach to keyword extraction was used in (Liu et al., 2008). Even though the data set in that study is not very big, it seems that a supervised learning approach can achieve reasonable performance for this task.

Another line of research for keyword extraction has adopted graph-based methods similar to Google's PageRank algorithm (Brin and Page, 1998). In particular, (Wan et al., 2007) attempted to use a 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. We also find that this assumption also holds using statistics obtained from the meeting corpus used in this study. Graph-based methods have not been used in a genre like the meeting domain; therefore, it remains to be seen whether these approaches can be applied to meetings.

Not many studies have been performed on speech transcripts for keyword extraction. The most relevant work to our study is (Plas et al., 2004), where the task is keyword extraction in the 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). There is also some work using keywords for other speech processing tasks, e.g., (Munteanu et al., 2007; Bulyko et al., 2007; Wu et al., 2007; Desilets et al., 2002; Rogina, 2002). (Wu et al., 2007) showed that keyword extraction combined with semantic verification can be used to improve speech retrieval performance on broadcast news data. In (Rogina, 2002), keywords were extracted from lecture slides, and then used as queries to retrieve relevant web documents, resulting in an improved language model and better speech recognition performance of lectures. There are many differences between written text and speech — meetings in particular. Thus our goal in this paper is to investi-

gate whether we can successfully apply some existing techniques, as well as propose new approaches to extract keywords for the meeting domain. The aim of this study is to set up some starting points for research in this area.

3 Data

We used the meetings from the ICSI meeting data (Janin et al., 2003), which are recordings of naturally occurring meetings. All the meetings have been transcribed and annotated with dialog acts (DA) (Shriberg et al., 2004), topics, and extractive summaries (Murray et al., 2005). The ASR output for this corpus is obtained from a state-of-the-art SRI conversational telephone speech system (Zhu et al., 2005), with a word error rate of about 38.2% on the entire corpus. We align the human transcripts and ASR output, then map the human annotated DA boundaries and topic boundaries to the ASR words, such that we have human annotation of these information for the ASR output.

We recruited three Computer Science undergraduate students to annotate keywords for each topic segment, using 27 selected ICSI meetings.[1] Up to five indicative key words or phrases were annotated for each topic. In total, we have 208 topics annotated with keywords. The average length of the topics (measured using the number of dialog acts) among all the meetings is 172.5, with a high standard deviation of 236.8. We used six meetings as our development set (the same six meetings as the test set in (Murray et al., 2005)) to optimize our keyword extraction methods, and the remaining 21 meetings for final testing in Section 5.

One example of the annotated keywords for a topic segment is:

- **Annotator I**: analysis, constraints, template matcher;
- **Annotator II**: syntactic analysis, parser, pattern matcher, finite-state transducers;
- **Annotator III**: lexicon, set processing, chunk parser.

Note that these meetings are research discussions, and that the annotators may not be very familiar with the topics discussed and often had trouble deciding the important sentences or keywords. In addition, limiting the number of keywords that an annotator can select for a topic also created some difficulty. Sometimes there are more possible keywords and the annotators felt it is hard to decide which five are the most topic indicative. Among the three annotators, we notice that in general the quality of annotator I is the poorest. This is based on the authors' judgment, and is also confirmed later by an independent human evaluation (in Section 6).

For a better understanding of the gold standard used in this study and the task itself, we thoroughly analyzed the human annotation consistency. We removed the topics labeled with "chitchat" by at least one annotator, and also the digit recording part in the ICSI data, and used the remaining 140 topic segments. We calculated the percentage of keywords agreed upon by different annotators for each topic, as well as the average for all the meetings. All of the consistency analysis is performed based on words. Figure 1 illustrates the annotation consistency over different meetings and topics. The average consistency rate across topics is 22.76% and 5.97% among any two and all three annotators respectively. This suggests that people do not have a high agreement on keywords for a given document. We also notice that the two person agreement is up to 40% for several meetings and 80% for several individual topics, and the agreement among all three annotators reaches 20% and 40% for some meetings or topics. This implies that the consistency depends on topics (e.g., the difficulty or ambiguity of a topic itself, the annotators' knowledge of that topic). Further studies are needed for the possible factors affecting human agreement. We are currently creating more annotations for this data set for better agreement measure and also high quality annotation.

4 Methods

Our task is to extract keywords for each of the topic segments in each meeting transcript. Therefore, by "document", we mean a topic segment in the remainder of this paper. Note that our task is different from keyword spotting, where a keyword is provided and the task is to spot it in the audio (along with its transcript).

The core part of keyword extraction is for the sys-

[1] We selected these 27 meetings because they have been used in previous work for topic segmentation and summarization (Galley et al., 2003; Murray et al., 2005).

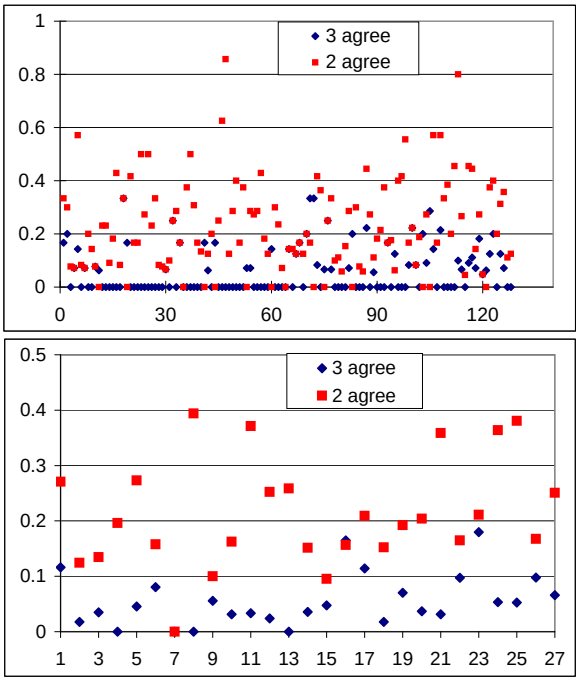

Figure 1: Human annotation consistency across different topics (upper graph) and meetings (lower graph). Y-axis is the percent of the keywords agreed upon by two or three annotators.

tem to assign an importance score to a word, and then pick the top ranked words as keywords. We compare different methods for weight calculation in this study, broadly divided into the following two categories: the TFIDF framework and the graph-based model. Both are unsupervised learning methods.[2] In all of the following approaches, when selecting the final keywords, we filter out any words appearing on the stopword list. These stopwords are generated based on the IDF values of the words using all the meeting data by treating each topic segment as a document. The top 250 words from this list (with the lowest IDF values) were used as stopwords. We generated two different stopword lists for human transcripts and ASR output respectively. In addition, in this paper we focus on performing keyword extraction at the single word level, therefore no key phrases are generated.

[2]Note that by unsupervised methods, we mean that no data annotated with keywords is needed. These methods do require the use of some data to generate information such as IDF, or possibly a development set to optimize some parameters or heuristic rules.

4.1 TFIDF Framework

(A) **Basic TFIDF weighting**

The term frequency (TF) for a word w_i in a document is the number of times the word occurs in the document. The IDF value is:

$$IDF_i = log(N/N_i)$$

where N_i denotes the number of the documents containing word w_i, and N is the total number of the documents in the collection. We also performed L_2 normalization for the IDF values when combining them with other scores.

(B) **Part of Speech (POS) filtering**

In addition to using a stopword list to remove words from consideration, we also leverage POS information to filter unlikely keywords. Our hypothesis is that verb, noun and adjective words are more likely to be keywords, so we restrict our selection to words with these POS tags only. We used the TnT POS tagger (Brants, 2000) trained from the Switchboard data to tag the meeting transcripts.

(C) **Integrating word clustering**

One weakness of the baseline TFIDF is that it counts the frequency for a particular word, without considering any words that are similar to it in terms of semantic meaning. In addition, when the document is short, the TF may not be a reliable indicator of the importance of the word. Our idea is therefore to account for the frequency of other similar words when calculating the TF of a word in the document. For this, we group all the words into clusters in an unsupervised fashion. If the total term frequency of all the words in one cluster is high, it is likely that this cluster contributes more to the current topic from a thematic point of view. Thus we want to assign higher weights to the words in this cluster.

We used the SRILM toolkit (Stolcke, 2002) for automatic word clustering over the entire document collection. It minimizes the perplexity of the induced class-based n-gram language model compared to the original word-based model. Using the clusters, we then adjust the TF weighting by integrating with the cluster term frequency (CTF):

$$TF\_CTF(w_i) = TF(w_i) * \alpha^{(\sum_{w_l \in C_i, w_l \neq w_i} freq(w_l))}$$

where the last summation component means the total term frequency of all the other words in this document that belong to the same cluster C_i as the current

word w_i. We set parameter α to be slightly larger than 1. We did not include stopwords when adding the term frequencies for the words in a cluster.

(D) **Combining with sentence salience score**

Intuitively, the words in an important sentence should be assigned a high weight for keyword extraction. In order to leverage the sentence information, we adjust a word's weight by the salience scores of the sentences containing that word. The sentence score is calculated based on its cosine similarity to the entire meeting. This score is often used in extractive summarization to select summary sentences (Radev et al., 2001). The cosine similarity between two vectors, D_1 and D_2, is defined as:

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

where t_i is the term weight for a word w_i, for which we use the TFIDF value.

4.2 Graph-based Methods

For the graph-based approach, we adopt the iterative reinforcement approach from (Wan et al., 2007) in the hope of leveraging sentence information for keyword extraction. This algorithm is based on the assumption that important sentences/words are connected to other important sentences/words.

Four graphs are created: one graph in which sentences are connected to other sentences (S-S graph), one in which words are connected to other words (W-W graph), and two graphs connecting words to sentences with uni-directional edges (W-S and S-W graphs). Stopwords are removed before the creation of the graphs so they will be ineligible to be keywords.

The final weight for a word node depends on its connection to other words (W-W graph) and other sentences (W-S graph); similarly, the weight for a sentence node is dependent on its connection to other sentences (S-S graph) and other words (S-W graph). That is,

$$u = \alpha U^T u + \beta \hat{W}^T v$$
$$v = \alpha V^T v + \beta W^T u$$

where u and v are the weight vectors for sentence and word nodes respectively, $U, V, W, \hat{W}$ represent the S-S, W-W, S-W, and W-S connections. α and β specify the contributions from the homogeneous and the heterogeneous nodes. The initial weight is a uniform one for the word and sentence vector. Then the iterative reinforcement algorithm is used until the node weight values converge (the difference between scores at two iterations is below 0.0001 for all nodes) or 5,000 iterations are reached.

We have explored various ways to assign weights to the edges in the graphs. Based on the results on the development set, we use the following setup in this paper:

- **W-W Graph:** We used a diagonal matrix for the graph connection, i.e., there is no connection among words. The self-loop values are the TFIDF values of the words. This is also equivalent to using an identity matrix for the word-word connection and TFIDF as the initial weight for each vertex in the graph. We investigated other strategies to assign a weight for the edge between two word nodes; however, so far the best result we obtained is using this diagonal matrix.

- **S-W and W-S Graphs:** The weight for an edge between a sentence and a word is the TF of the word in the sentence multiplied by the word's IDF value. These weights are initially added only to the S-W graph, as in (Wan et al., 2007); then that graph is normalized and transposed to create the W-S graph.

- **S-S Graph:** The sentence node uses a vector space model and is composed of the weights of those words connected to this sentence in the S-W graph. We then use cosine similarity between two sentence vectors.

Similar to the above TFIDF framework, we also use POS filtering for the graph-based approach. After the weights for all the words are determined, we select the top ranked words with the POS restriction.

5 Experimental Results: Automatic Evaluation

Using the approaches described above, we computed weights for the words and then picked the top five words as the keywords for a topic. We chose five keywords since this is the number of keywords that

human annotators used as a guideline, and it also yielded good performance in the development set. To evaluate system performance, in this section we use human annotated keywords as references, and compare the system output to them. The first metric we use is F-measure, which has been widely used for this task and other detection tasks. We compare the system output with respect to each human annotation, and calculate the maximum and the average F-scores. Note that our keyword evaluation is word-based. When human annotators choose key phrases (containing more than one word), we split them into words and measure the matching words. Therefore, when the system only generates five keywords, the upper bound of the recall rate may not be 100%. In (Liu et al., 2008), a lenient metric is used which accounts for some inflection of words. Since that is highly correlated with the results using exact word match, we report results based on strict matching in the following experiments.

The second metric we use is similar to Pyramid (Nenkova and Passonneau, 2004), which has been used for summarization evaluation. Instead of comparing the system output with each individual human annotation, the method creates a "pyramid" using all the human annotated keywords, and then compares system output to this pyramid. The pyramid consists of all the annotated keywords at different levels. Each keyword has a score based on how many annotators have selected this one. The higher the score, the higher up the keyword will be in the pyramid. Then we calculate an oracle score that a system can obtain when generating k keywords. This is done by selecting keywords in the decreasing order in terms of the pyramid levels until we obtain k keywords. Finally for the system hypothesized k keywords, we compute its score by adding the scores of the keywords that match those in the pyramid. The system's performance is measured using the relative performance of the system's pyramid scores divided by the oracle score.

Table 1 shows the results using human transcripts for different methods on the 21 test meetings (139 topic segments in total). For comparison, we also show results using the supervised approach as in (Liu et al., 2008), which is the average of the 21-fold cross validation. We only show the maximum F-measure with respect to individual annotations, since the average scores show similar trend. In addition, the weighted relative scores already accounts for the different annotation and human agreement.

Methods	F-measure	weighted relative score
TFIDF	0.267	0.368
+ POS	0.275	0.370
+ Clustering	0.277	0.367
+ Sent weight	0.290	0.404
Graph	0.258	0.364
Graph+POS	0.277	0.380
Supervised	0.312	0.401

Table 1: Keyword extraction results using human transcripts compared to human annotations.

We notice that for the TFIDF framework, adding POS information slightly helps the basic TFIDF method. In all the meetings, our statistics show that adding POS filtering removed 2.3% of human annotated keywords from the word candidates; therefore, this does not have a significant negative impact on the upper bound recall rate, but helps eliminate unlikely keyword candidates. Using word clustering does not yield a performance gain, most likely because of the clustering technique we used — it does clustering simply based on word co-occurrence and does not capture semantic similarity properly.

Combining the term weight with the sentence salience score improves performance, supporting the hypothesis that summary sentences and keywords can reinforce each other. In fact we performed an analysis of keywords and summaries using the following two statistics:

$$(1) \qquad k = \frac{P_{summary}(w_i)}{P_{topic}(w_i)}$$

where $P_{summary}(w_i)$ and $P_{topic}(w_i)$ represent the the normalized frequency of a keyword w_i in the summary and the entire topic respectively; and

$$(2) \qquad s = \frac{PS_{summary}}{PS_{topic}}$$

where $PS_{summary}$ represents the percentage of the sentences containing at least one keyword among all the sentences in the summary, and similarly PS_{topic} is measured using the entire topic segment. We found that the average k and s are around 3.42 and 6.33 respectively. This means that keywords are

more likely to occur in the summary compared to the rest of the topic, and the chance for a summary sentence to contain at least one keyword is much higher than for the other sentences in the topic.

For the graph-based methods, we notice that adding POS filtering also improves performance, similar to the TFIDF framework. However, the graph method does not perform as well as the TFIDF approach. Comparing with using TFIDF alone, the graph method (without using POS) yielded worse results. In addition to using the TFIDF for the word nodes, information from the sentences is used in the graph method since a word is linked to sentences containing this word. The global information in the S-S graph (connecting a sentence to other sentences in the document) is propagated to the word nodes. Unlike the study in (Wan et al., 2007), this information does not yield any gain. We did find that the graph approach performed better in the development set, but it seems that it does not generalize to this test set.

Compared to the supervised results, the TFIDF approach is worse in terms of the individual maximum F-measure, but achieves similar performance when using the weighted relative score. However, the unsupervised TFIDF approach is much simpler and does not require any annotated data for training. Therefore it may be easily applied to a new domain. Again note that these results used word-based selection. (Liu et al., 2008) investigated adding bigram key phrases, which we expect to be independent of these unigram-based approaches and adding bigram phrases will yield further performance gain for the unsupervised approach. Finally, we analyzed if the system's keyword extraction performance is correlated with human annotation disagreement using the unsupervised approach (TFIDF+POS+Sent_weight). The correlation (Spearman's ρ value) between the system's F-measure and the three-annotator consistency on the 27 meetings is 0.5049 (p=0.0072). This indicates that for the meetings with a high disagreement among human annotators, it is also challenging for the automatic systems.

Table 2 shows the results using ASR output for various approaches. The performance measure is the same as used in Table 1. We find that in general, there is a performance degradation compared to using human transcripts, which is as expected. We found that only 59.74% of the human annotated keywords appear in ASR output, that is, the upper bound of recall is very low. The TFIDF approach still outperforms the graph method. Unlike on human transcripts, the addition of information sources in the TFIDF approach did not yield significant performance gain. A big difference from the human transcript condition is the use of sentence weighting — adding it degrades performance in ASR, in contrast to the improvement in human transcripts. This is possibly because the weighting of the sentences is poor when there are many recognition errors from content words. In addition, compared to the supervised results, the TFIDF method has similar maximum F-measure, but is slightly worse using the weighted score. Further research is needed for the ASR condition to investigate better modeling approaches.

Methods	F-measure	weighted relative score
TFIDF	0.191	0.257
+ POS	0.196	0.259
+ Clustering	0.196	0.259
+ Sent weigh	0.178	0.241
Graph	0.173	0.223
Graph+POS	0.183	0.233
Supervised	0.197	0.269

Table 2: Keyword extraction results using ASR output.

6 Experimental Results: Human Evaluation

Given the disagreement among human annotators, one question we need to answer is whether F-measure or even the weighted relative scores compared with human annotations are appropriate metrics to evaluate system-generated keywords. For example, precision measures among the system-generated keywords how many are correct. However, this does not measure if the unmatched system-generated keywords are bad or acceptable. We therefore performed a small scale human evaluation. We selected four topic segments from four different meetings, and gave output from different systems to five human subjects. The subjects ranged in age from 22 to 63, and all but one had only basic knowledge of computers. We first asked the eval-

uators to read the entire topic transcript, and then presented them with the system-generated keywords (randomly ordered by different systems). For comparison, the keywords annotated by our three human annotators were also included without revealing which sets of keywords were generated by a human and which by a computer. Because there was such disagreement between annotators regarding what made **good** keywords, we instead asked our evaluators to mark any words that were **definitely not** keywords. Systems that produced more of these rejected words (such as "basically" or "mmm-hm") are assumed to be worse than those containing fewer rejected words. We then measured the percentage of rejected keywords for each system/annotator. The results are shown in Table 3. Not surprisingly, the human annotations rank at the top. Overall, we find human evaluation results to be consistent with the automatic evaluation metrics in terms of the ranking of different systems.

Systems	Rejection rate
Annotator 2	8%
Annotator 3	19%
Annotator 1	25%
TFIDF + POS	28%
TFIDF	30%

Table 3: Human evaluation results: percentage of the rejected keywords by human evaluators for different systems/annotators.

Note this rejection rate is highly related to the recall/precision measure in the sense that it measures how many keywords are acceptable (or rejected) among the system generated ones. However, instead of comparing to a fixed set of human annotated keywords (e.g., five) and using that as a gold standard to compute recall/precision, in this evaluation, the human evaluator may have a larger set of acceptable keywords in their mind. We also measured the human evaluator agreement regarding the accepted or bad keywords. We found that the agreement on a bad keyword among five, four, and three human evaluator is 10.1%, 14.8%, and 10.1% respectively. This suggests that humans are more likely to agree on a bad keyword selection compared to agreement on the selected keywords, as discussed in Section 3 (even though the data sets in these two analysis are not the same). Another observation from the human evaluation is that sometimes a person rejects a keyword from one system output, but accepts that on the list from another system. We are not sure yet whether this is the inconsistency from human evaluators or whether the judgment is based on a word's occurrence with other provided keywords and thus some kind of semantic coherence. Further investigation on human evaluation is still needed.

7 Conclusions and Future Work

In this paper, we evaluated unsupervised keyword extraction performance for the meeting domain, a genre that is significantly different from most previous work. We compared several different approaches using the transcripts of the ICSI meeting corpus. Our results on the human transcripts show that the simple TFIDF based method is very competitive. Adding additional knowledge such as POS and sentence salience score helps improve performance. The graph-based approach performs less well in this task, possibly because of the lack of structure in this domain. We use different performance measurements, including F-measure with respect to individual human annotations and a weighted metric relative to the oracle system performance. We also performed a new human evaluation for this task and our results show consistency with the automatic measurement. In addition, experiments on the ASR output show performance degradation, but more importantly, different patterns in terms of the contributions of information sources compared to using human transcripts. Overall the unsupervised approaches are simple but effective; however, system performance compared to the human performance is still low, suggesting more work is needed for this domain.

For the future work, we plan to investigate different weighting algorithms for the graph-based approach. We also need a better way to decide the number of keywords to generate instead of using a fixed number. Furthermore, since there are multiple speakers in the meeting domain, we plan to incorporate speaker information in various approaches. More importantly, we will perform a more rigorous human evaluation, and also use extrinsic evaluation to see whether automatically generated keywords facilitate tasks such as information retrieval or meeting browsing.

Acknowledgments

This work is supported by NSF award IIS-0714132. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of NSF.

References

T. Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference*.

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30.

I. Bulyko, M. Ostendorf, M. Siu, T. Ng, A. Stolcke, and O. Cetin. 2007. Web resources for language modeling in conversational speech recognition. *ACM Transactions on Speech and Language Processing*, 5:1–25.

A. Desilets, B.D. Bruijn, and J. Martin. 2002. Extracting keyphrases from spoken audio documents. In *Information Retrieval Techniques for Speech Applications*, pages 339–342.

E. Frank, G.W. Paynter, I.H. Witten, C. Gutwin, and C.G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *Proceedings of IJCAI*, pages 688–673.

M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of ACL*.

A. Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of EMNLP*, pages 216–223.

D. Inkpen and A. Desilets. 2004. Extracting semantically-coherent keyphrases from speech. *Canadian Acoustics Association*, 32:130–131.

A. Janin, D. Baron, J. Edwards, D. Ellis, G . Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus. In *Proceedings of ICASSP*.

Y.H. Kerner, Z. Gross, and A. Masa. 2005. Automatic extraction and learning of keyphrases from scientific articles. In *Computational Linguistics and Intelligent Text Processing*, pages 657–669.

F. Liu, F. Liu, and Y. Liu. 2008. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In *Proceedings of IEEE SLT*.

Y. Matsuo and M. Ishizuka. 2004. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence*, 13(1):157–169.

C. Munteanu, G. Penn, and R. Baecker. 2007. Web-based language modeling for automatic lecture transcription. In *Proceedings of Interspeech*.

G. Murray, S. Renals, J. Carletta, and J. Moore. 2005. Evaluating automatic summaries of meeting recordings. In *Proceedings of ACL 2005 MTSE Workshop*, pages 33–40.

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarization: the pyramid method. In *Proceedings of HLT/NAACL*.

L. Plas, V. Pallotta, M. Rajman, and H. Ghorbel. 2004. Automatic keyword extraction from spoken text. a comparison of two lexical resources: the EDR and WordNet. In *Proceedings of the LREC*.

D. Radev, S. Blair-Goldensohn, and Z. Zhang. 2001. Experiments in single and multi-document summarization using MEAD. In *Proceedings of The First Document Understanding Conference*.

I. Rogina. 2002. Lecture and presentation tracking in an intelligent meeting room. In *Proceedings of ICMI*.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of SIGDial Workshop*, pages 97–100.

A. Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of ICSLP*, pages 901–904.

P.D. Turney. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2:303–336.

P.D. Turney. 2002. Mining the web for lexical knowledge to improve keyphrase extraction: Learning from labeled and unlabeled data. In *National Research Council, Institute for Information Technology, Technical Report ERB-1096*.

P.D. Turney. 2003. Coherent keyphrase extraction via web mining. In *Proceedings of IJCAI*, pages 434–439.

X. Wan, J. Yang, and J. Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of ACL*, pages 552–559.

C.H. Wu, C.L. Huang, C.S. Hsu, and K.M. Lee. 2007. Speech retrieval using spoken keyword extraction and semantic verification. In *Proceedings of IEEE Region 10 Conference*, pages 1–4.

Q. Zhu, A. Stolcke, B. Chen, and N. Morgan. 2005. Using MLP features in SRI's conversational speech recognition system. In *Proceedings of Interspeech*.