Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization

Sameer Maskey, Julia Hirschberg

Department of Computer Science
Columbia University, New York, NY
{smaskey, julia}@cs.columbia.edu

Abstract

We present results of an empirical study of the usefulness of different types of features in selecting extractive summaries of news broadcasts for our Broadcast News Summarization System. We evaluate lexical, prosodic, structural and discourse features as predictors of those news segments which should be included in a summary. We show that a summarization system that uses a combination of these feature sets produces the most accurate summaries, and that a combination of acoustic/prosodic and structural features are enough to build a 'good' summarizer when speech transcription is not available.

1. Introduction

Most text-based summarization systems rely upon lexical, syntactic, and positional information in determining which segments to include in a summary. News broadcasts contain additional sources of information that text news stories typically do not, including the broadcast structure and acoustic and prosodic information. While some proposed speech summarization systems have investigated subsets of these text-based and speech-based features [4, 5], there are many new features to consider. And, to date, no study has examined the relative contribution of different feature classes — lexical, structural, prosodic and discourse — as predictors for extractive summarization. In this paper, we propose new types of features in some categories and compare and contrast the utility of the different feature types for the summarization of Broadcast News stories.

In Section 2 we describe the corpus we use to train and test extractive summarization of news stories on. We describe the features we use and the types of machine learning algorithms we we have investigated for our classification task in Section 3. We then describe our evaluation experiments in Section 4 and present our results. In Section 5 we present our conclusions and discuss future work.

2. The Corpus and Annotations

Our summarization system [11] currently operates on Broadcast News shows from the TDT-2 corpus. It takes audio files and manual or automatic transcripts as input and presents an outline of the news broadcast in a GUI interface, which allows users to search the newscasts by content (transcriptions are time-aligned to the audio) and to access summaries of news stories.

To build this system, we use 216 stories from 20 CNN shows from the TDT-2 [17] corpus. This includes 10 hours of audio data. We used manual transcripts, Dragon ASR transcripts and audio files of each show for training and test. Extractive summaries were generated for each story in the show where stories were annotated by the same annotator. Manual transcriptions of 96 shows were further annotated with named entities, openings and closings, headlines, interviews and sound-bytes, following a labeling manual which followed ACE conventions for named entities. We make use also of sentence, turn and topic segmentation available in TDT-2.

Extractive summarization consists of extracting segments from original text or audio 'documents' and combining these to create a human readable/audible/viewable summary [9]. Although non-extractive summarization systems are a viable approach in the text domain, generative summaries of spoken documents must be produced in text or text-to-speech synthesis, in either case losing the non-lexical information in the original. Hence, in our work on spoken document summarization, we have chosen to extract summaries from sentence segments identified in the newscast. We view extractive summarization as a binary classification problem in which we determine whether a segment should be included in the summary or not. Below we describe the features we use to classify sentences to be included in a summary and the method we use to identify such sentences.

3. Features and Method

In this section, we describe the feature classes we use to predict sentences to be extracted and our method for selecting them, including lexical, structural, prosodic and discourse features.

3.1. Lexical Features

The lexical features we experimented with include counts of **person names** (NEI), **organization names** (NEII), and **place names** (NEIII) for each sentence. We also included the **total number of named entities** in a sentence as a feature, in addition to the **number of words in the current, previous and following sentence**.

Some of these features like named entities have previously been tested in other summarization systems [15, 4]. One of our findings is the importance of named entity features. Unlike text news, in broadcasts, multiple stories are presented in one broadcast, with each story containing its own distinctive named entities. While these named entities may not be repeated frequently over the broadcast, they are important clues to the selection of summary segments within a story. For example, a sentence containing many named entities in the introduction of a story by a news anchor often represents an overview of the story to be presented and, thus, is often included in a summary.

Our feature selection algorithm selects total number of NEs and number of words in the sentence as particularly useful features for predicting sentences to be included in a summary. For our current purposes, we have assumed that we can obtain accurate named entity labels from systems such as BBN's IdentiFinder$^{\text{TM}}$ [13].

3.2. Prosodic/Acoustic Features

The intuition behind using prosodic/acoustic features for speech summarization is based on well-found research in speech prosody [6] that humans use intonational variation — expanded pitch range, phrasing or intonational prominence — to mark the importance of particular items in their speech. In Broadcast News, we note that a change in pitch, amplitude or speaking rate may signal differences in the relative importance of the speech segments produced by anchors and reporters — the professional speakers in our corpus. There is also considerable evidence that topic shift is marked by changes in pitch, intensity, speaking rate and duration of pause [7, 16], and new topics or stories in broadcast news are often introduced with content-laden sentences which, in turn, often are included in story summaries.

Prosodic/Acoustic features have been examined in research on speech summarization [5] and information extraction tasks [16]. Our acoustic feature-set includes features mentioned in [5, 4] as well as new acoustic features. It includes **speaking rate** (the ratio of voiced/total frames); **F0 minimum**, **maximum**, and **mean**; **F0 range** and **slope**; **minimum, maximum**, and **mean RMS energy** (minDB, maxDB, meanDB); **RMS slope** (slopeDB); **sentence duration** (timeLen = endtime - starttime). We extracted these features by automatically aligning the annotated manual transcripts or the ASR transcripts with the audio source. We then used Praat [3] to extract the features from the audio and experimented with both normalized and raw versions of each. Normalized features were produced by dividing each feature by the average of the feature values for each speaker, where speaker identify was determined from the Dragon speaker segmentation of the TDT-2 corpus. Normalized acoustic features performed better than raw values.

Our duration feature, 'sentence duration', represents the length in seconds of the sentence. Our motivation for including this features is twofold: Very short segments are not likely to contain important information. On the other hand, very long segments may not be useful to include in a summary, simply for concerns about providing over-long summaries. This length feature is can accommodate both types of information. We obtain sentence length by subtracting the end from the start time for each sentence.

Our feature selection algorithm finds that timeLen, minDB and maxDB are particular discriminatory, while pitch features are, curiously, among the least useful of the acoustic features.

3.3. Structural Features

Broadcast News programs exhibit similar structure, particularly broadcasts of the same show from the same news channel. Each usually begins with an anchor or anchors reporting the headlines, followed by the actual presentation of those stories by the anchor, reporters, and sometimes interviewees. Programs are usually concluded in the same conventional manner. We call the features which rely upon aspects of this patterning and from the overall structure of the broadcast *structural features* [11], comparable to [4]'s *style features*. We have previously shown that structural features are useful predictors of extractive summaries of Broadcast News [11].

The structural features we investigated for our study include **normalized /sentence position in turn**, **speaker type next-speaker type**, **previous-speaker type**, **speaker change**, **turn position in the show** and **sentence position in the show**. Only reporters' turns are so marked in the TDT-2 corpus, so our speaker type feature is binary, 'reporter or not'. This unfortunately conflates anchor turns with those of interviewees and soundbyte speakers.

3.4. Discourse Features

Some summarization systems [12] have included discourse features, such as [12]'s discourse trees, which models the rhetorical structure of a text to identify important segments for extraction. We have explored a different discourse feature, by computing a measure of 'givenness' in our stories. Following [14] we identify 'discourse given' information as information which has previously been evoked in a discourse, either by explicit reference or by indirect (in our case, stem) similarity to other mentioned items. Our intuition is that given information is less likely to be included in a summary, since it represents redundant information. Our *given/new* feature represents a very simple implementation of this intuition and proves to be a useful predictor of whether a sentence will be included in a summary. This feature is a score that ranges between -1 and 1 with a sentence containing only new information receiving a score of '1', and a sentence containing only 'given' information receiving '-1'. We calculate this score for each sentence by the following equation:

$$S(i) = \frac{n_i}{d} - \frac{s_i}{t-d} \quad (1)$$

Here, n_i is the number of 'new' noun stems in sentence i, d is the total number of unique noun stems in the story; s_i is the number of noun stems in the sentence i that have already been seen in the story; and t is the total number of noun stems in the story.

The intuition behind this feature is that, if a sentence contains more new noun stems, it is likely that more 'new information' is included in the sentence. The term n_i/d in the equation 1 takes account of this 'newness'. On the other hand, a very long sentence may have many new nouns but still include other references to items that have already been mentioned. In such cases, we would want to reduce the given-new score by the 'givenness' in the sentence; this givenness reduction is take into account by $\frac{s_i}{t-d}$. As we will show in Section 4, this simple measure improves our summarization F-measure. We have also experimented with variations on this scores but found 1 to yield the best performance.

4. Experiments and Results

We compared the contribution of our lexical, acoustic, structural and discourse feature-sets to predicting sentences to be included in summaries using a number of different learning algorithms, test methods (cross-validation with resampling or held-out test sets), and feature selection algorithms. We found that the contribution of our various feature classes could best be examined using a Bayesian Network classifier with 10-fold cross validation (with resampling), on features selected by a procedure that combines subset evaluation with rank search and best-first search [18]. We measured performance by computing precision, recall, and F-measure. Results of these feature-set comparisons are shown in Figure 1.

We constructed a baseline for this task by concatenating the first 23% of sentences from each show, since our model summaries were, on average, 23% of the length of the source documents. Such a baseline is very strict for Broadcast News, since these stories are quite short, with average of 18.2 sentences in

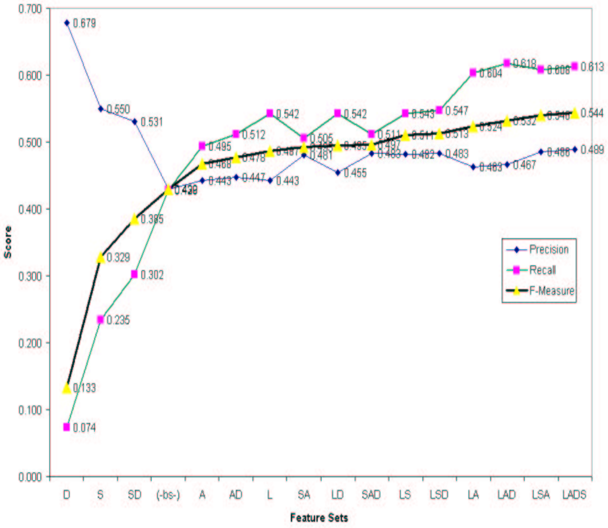

Figure 1: F-measure with 10 fold cross-validation

each story. Using this approach, our baseline F-measure is 0.43, recall is 0.43 and precision is 0.43.

From Figure 1 we can see that the best performing experiments combine all the feature-sets $L + A + S + D$. This gives an F-measure of 0.54, recall of 0.61, precision of 0.49 and an accuracy of 73.8% on our dataset with 10-fold cross validation. Our system thus has an F-measure which is 11% higher than the baseline. The F-measures for the individual feature-sets when tested alone are: discourse (0.13), structural (0.33), acoustic/prosodic (0.47) and lexical (0.49). So we see that the lexical and acoustic/prosodic feature-sets perform best alone, both surpassing the baseline. When we combine these two feature-sets, our F-measure is 0.52. Adding structural features improves performance to 0.54, and adding discourse features as well improves our F-measure to 0.544.

When we add only discourse features to the lexical and acoustic/prosodic feature-sets, performance is 0.53. However, when we look at the performance of structural features alone, compared to that of structural plus discourse features, we see that the F-measure improves from 0.33 to 0.39. Discourse features added to lexical improve the F-measure from 0.49 to 0.50, and, added to acoustic/prosodic features, improve the F-measure from 0.47 to 0.48. So, there appears to be more redundancy of our discourse features with acoustic/prosodic and lexical features than with the structural feature-set.

To look more specifically at which of the features in our feature-sets are most useful for predicting summary sentence selection, we performed feature selection on our entire set of features using a selection algorithm that computed individual predictive power of each feature and the redundancies between features. The five most useful features are shown in Table 1. Note that the best performing individual features include features from all four of our feature-sets with two from the lexical set and one from each of the others. Interestingly, this set of five was also selected as the optimal set of features by the feature selection algorithm. Our F-measure with just these features is 0.53 which is only 1% lower than the highest F-measure shown in Figure 1.

Table 1: Best Features for Predicting Summary Sentences

Rank	Type	Feature
1	A	Time Length in sec.
2	L	Num. of words
3	L	Tot Named Entities
4	S	Normalized SentPos
5	D	Given-New Score

To confirm that our results are unaffected by choice of classifier, we also computed ROC curves for the classifiers we tried. The area under the curve (AOC) of the ROC curve computes the ‘goodness’ of a classifier; the best classifier would have an AOC of 1. For the classifiers we examined, we obtained an AOC of 0.771 for Bayesian Networks, 0.647 for C4.5 Decision Trees, 0.643 for Ripper and 0.535 for Support Vector Machines. All results reported in Figure 1 are for a Bayesian Network classifier.

One conclusion we might draw from our results is that “the importance of **what** is said correlates with **how** it is said.” Intuitively, one might imagine that speakers change their amplitude and pitch when they believe their utterances are particularly important, to convey that importance to the hearer. If this is true, we would expect the sentences that our lexical features include in a summary to be the same as those predicted for inclusion by our acoustic/prosodic features. We computed the correlation coefficient between the predictions of these two feature-sets. The correlation of 0.74 supports our hypothesis.

Our findings also suggest it may be possible to do effective speech summarization without the use of transcription at all, whether manual (as employed here) or from speech recognition. Two of our feature-sets, acoustic/prosodic and structural, are independent of lexical transcription, except for sentence-level and speaker segmentation and classification, which have been shown to be automatically extractable using only acoustic/prosodic information [16, 1]. The accuracy of our acoustic/prosodic features alone ($F = 0.47$), and of our combined acoustic/prosodic and structural features ($F = 0.50$) compares favorably to that of our combined feature-sets ($F = 0.54$). So, even if transcription is unavailable, it seems possible to summarize broadcast news effectively, even when transcription is unavailable.

The results mentioned above assume an exact match of a predicted summary sentence to a labeled summary sentence. For summarization purposes, this measure is generally considered too strict, since a sentence classified incorrectly as a summary sentence may be very close in semantic content to another sentence which *was* included in the gold standard summary. Another metric standardly used in summary evaluation, which takes this synonymy into account, is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [10]. ROUGE measures overlap units between automatic and manual summaries. Units measured can be n-gram, word sequences or word pairs. For ROUGE-N, ROUGE-L, ROUGE-S and ROUGE-SU, then, N indicates (the size of) the n-grams computed, L, the longest common subsequence, and S and SU stand for skip bigram co-occurrence statistics with and without optional unigram counting. ROUGE-N is computed using the following equation.

$$ROUGE-N = \frac{\sum\limits_{S \in Ref.Sum} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in Ref.Sum} \sum\limits_{gram_n \in S} Count(gram_n)}$$

Figure 2 presents results of evaluating our feature-sets using the ROUGE metric, with $N = 1 - 4$ and all of the variants described above. The results shown in Figure 2 are similar

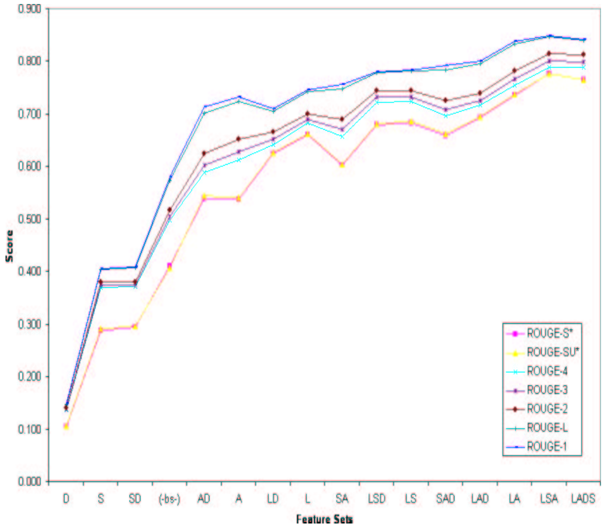

Figure 2: Evaluation using ROUGE metrics

to those shown in Figure 1 using the F-measure metric. However, the difference between the baseline and our best combined feature-set is even greater using ROUGE. We obtained our highest ROUGE score of 0.85 with ROUGE1 and ROUGE-L which is 27% higher than the baseline. If we take an average of the performance of different ROUGE systems, we get a mean score of 0.80, which is 30.3% above the baseline. This mean score may be a more reasonable measure than ROUGE1, since it includes the performance of versions which look for more than mere unigram overlap. Note that, the combined acoustic/prosodic and structural features alone obtain a ROUGE1 score of 0.76 and an average ROUGE score of 0.68.

5. Conclusion

In this paper, we have presented results of an empirical study comparing different types of features that may be useful for speech summarization. We have shown that a combination of lexical, acoustic/prosodic, structural, and discourse features performs best at classifying sentences to be included in a summary. With this combined feature-set we obtain an F-measure of 0.54 on exact matching of summary sentences and a ROUGE score of 0.84 from ROUGE1 and ROUGE-L evaluations. Our findings also suggest that accurate speech summarization is possible in the absence of transcription, since our acoustic/prosodic and structural features alone obtain an F-measure of 0.50 and a ROUGE score of 0.76.

6. Acknowledgments

The work reported here was supported in part by the Defense Advanced Research Projects Agency under TIDES grant NUU01-00-1-8919. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the views of the funding agencies. We thank Martin Jansche, Irina Likhtina, Aaron Roth, and Lauren Wilcox for useful discussions.

7. References

[1] Barras, C., Zhu, X., Meignier, S., Gauvain, J.L. Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain. “Improving Speaker Diarization” Proc. DARPA RT04, 2004.

[2] Bikel, M.D., Miller, S., Schwartz, R and Weisschedel, R. “An algorithm that learns what’s in a name”, Machine Learning, 34(1/2/3):211-231, 1999.

[3] Boersma, P. “Praat, a system for doing phonetics by computer”, Glot International 5:9/10, 341-345. 2001.

[4] Christensen, H., Kolluru, B., Gotoh, Y., Renals, S. “From text summarisation to style-specific summarisation for broadcast news”, ECIR, 2004.

[5] Inoue, A., Mikami, T., Yamashita, Y. “Improvement of Speech Summarization Using Prosodic Information”, Proc. Speech Prosody, 2004, Japan.

[6] Hirschberg J. “Communication and Prosody: Functional Aspects of Prosody”, Speech Communication, Vol 36, pp 31-43, 2002.

[7] Hirschberg, J., Nakatani, C. “A Prosodic Analysis of Discourse Segments Direction-Giving Monologues”, ACL 1996.

[8] Hori, C., Furui, S., Malkin, R., Yu, H., Waibel, A. “Automatic Speech Summarization Applied to English Broadcast News Speech,” ICASSP 2002.

[9] Kupiec, J., Pedersen, J., Chen, F. “A trainable document summarizer”, SIGIR 1995.

[10] Lin, Chin-Yew “ROUGE: A Package for Automatic Evaluation of Summaries”, Proc. Workshop on Text Summarization, ACL 2004, Barcelona.

[11] Maskey, S., Hirschberg, J., “Automatic Summarization of BroadcastNews using Structural Features”, Eurospeech 2003.

[12] Marcu, D. “Discourse trees are good indicators of importance in text” In Advances in Automatic Text Summarization, pages 123-136, 1999.

[13] Miller, D., Schwartz, R., Weischedel, R., Stone, R. “Name Entity Extraction from Broadcast News”, DARPA Broadcast News Workshop, 1999.

[14] Prince, E.F. “The ZPG letter: subjects, definiteness, and information-status.” In Thompson, S. and Mann, W., eds. Discourse description, pp 295-325. 1992.

[15] Schiffman, B., Nenkova, A., McKeown, K. “Experiments in multidocument summarization”, HLT 2002.

[16] Shriberg, E., Stolcke, A., Tur, D.H., Tur, G. “Prosody-Based Automatic Segmentation of Speech into Sentences and Topics”, Speech Communication, Vo. 32. pp 127-154 2000.

[17] Language Data Consortium “TDT-2 Corpus”, Univ. of Pennsylvania.

[18] Witten, I.H., E. Frank, L. Trigg, M. Hall, G. Holmes and S. J. Cunningham, “Weka: Practical machine learning tools and techniques with Java implementations,” in H. Kasabov and K. Ko, eds., ICONIP/ANZIIS/ANNES’99 International Workshop, Dunedin, 1999.

[19] Zechner, K. “Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains”, R and D in IR, 199-207, 2001.