Incorporating Lexical and Prosodic Information at Different Levels for Meeting Summarization

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of Interspeech 2014

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Incorporating Lexical and Prosodic Information at Different Levels for Meeting Summarization

Catherine Lai, Steve Renals

Centre for Speech Technology Research
University of Edinburgh, Edinburgh, UK
clai@inf.ed.ac.uk, srenals@ed.ac.uk

Abstract

This paper investigates how prosodic features can be used to augment lexical features for meeting summarization. Automatic detection of summary-worthy content using non-lexical features, like prosody, has generally focused on features calculated over dialogue acts. However, a salient role of prosody is to distinguish important words within utterances. To examine whether including more fine grained prosodic information can help extractive summarization, we perform experiments incorporating lexical and prosodic features at different levels. For ICSI and AMI meeting corpora, we find that combining prosodic and lexical features at a lower level has better AUROC performance than adding in prosodic features derived over dialogue acts. ROUGE F-scores also show the same pattern for the ICSI data. However, the differences are less clear for the AMI data where the range of scores is much more compressed. In order to understand the relationship between the generated summaries and differences in standard measures, we look at the distribution of extracted content over meeting as well as summary redundancy. We find that summaries based on dialogue act level prosody better reflect the amount of human annotated summary content in meeting segments, while summaries derived from prosodically augmented lexical features exhibit less redundancy.

Index Terms: meeting summarization, prosody, dialogue.

1. Introduction

Automatic analysis of meetings has become an increasingly important task as more and more meetings are held online, recorded and archived. For example, automatic meeting summarization is a potentially useful tool for browsing and analysing dialogues for post-meeting tasks such as decision audits [1]. Methods for this have generally focused on lexical content. However, information in spoken dialogue can be characterized many other ways. For example, several studies have shown that automatic extractive summarization is possible using only prosodic features [2, 3, 4, 5]. Performance on this task is usually evaluated based on dialogue act or n-gram matching with gold standards (e.g. ROUGE [6]). Beyond this, however, we would like to understand how different aspects of prosody relate to what goes into meeting summaries. To do this, we need an understanding of where prosodic features can be incorporated into models and how this affects the generated summaries.

Current approaches generally use prosodic features calculated over dialogue acts (DAs) and the contribution of prosody is usually treated as independent to that of lexical content. Corpus analyses have found that extracted dialogue acts are generally produced with overall ‘bigger’ prosody, e.g. higher mean and maximum pitch and energy [3]. However, using DA level prosodic features washes out the use of prosody in marking out specific words as being important from an information structure point of view [7, 8]. It is possible that word level prosodic measurements could be used as lexical features in a similar way to term-frequency measures, such as tf.idf, which are commonly used to measure the importance of specific words relative to overall meeting content [9]. In fact, combining word prosody with tf.idf scores has been shown to help keyword extraction from voicemail messages [10], punctuation annotation [11] and topic tracking [12] in broadcast news. As such, we would like to know how these two types of information interact and how to combine them to improve summarization and dialogue understanding in general.

In this paper, we investigate whether augmenting lexical features with prosodic information improves extractive summarization performance in meetings. Our hypothesis is that integrating prosodic information at the word level will improve extractive summarization performance over plain lexical features like tf.idf. At the first stage, we combine lexical and prosodic features using an MLP to predict whether an isolated word belongs to an Extracted Dialogue Act (EDA). The probabilities generated, our augmented lexical features, are fed into the higher level DA extraction task. We compare the performance of our augmented features with DA level combinations of term-frequency and prosodic features, evaluating the resulting summarizers with commonly used retrieval measures, Area Under Receiver Operating Characteristic (AUROC) and ROUGE [6]. While feature analysis in summarization generally focuses on improvements of precision/recall based scores, these measures don’t tell us much about how summarizers based on different feature sets vary in what they selected as a region of interest. To start to tease some of these issues out we also look at the distribution of extracted DAs in the meeting timeline and also levels of redundancy of the different summaries.

2. Experimental Setup

2.1. Data

The experiments described in the following were carried out on the ICSI [13] and AMI [14] meeting corpora. The ICSI corpus contains recordings of 75 naturally occurring meetings drawn from 8 different research groups (3-9 speaker per meeting). We use the scenario data from the AMI meetings corpus (140 meetings). Each of these meetings involved 4 speakers who worked on designing a remote control. Each group par-
participants in a series of 4 meetings focusing on different stages of the design process. In the following experiments, we used standard AMI development and test sets (20 meetings, 5 groups each). For the ICSI corpus, we used the same test set as [3] and a randomly selected development set (6 meetings each).

Human extractive summaries are available for all of the ICSI meetings and for 137 of the AMI meetings. There were 6 annotators involved in creating summaries, with 2 contributing to both corpora. Annotations were based on manually segmented dialogue acts and time aligned transcriptions. Annotators selected dialogue acts with the explicit goal of helping an external stakeholder (e.g., department head) understand what happened in the meeting. There was no upper limit on how many dialogue acts annotators could select for the extractive summary, although a rough guideline of 10% was given. Where possible extracted DAs (EDAs) were then linked to statements in human authored abstractive summaries that they supported. As in [3], we focus on detecting only linked EDAs as they are more likely to be actually important for understanding what happened in the meeting.

2.2. Features

2.2.1. Prosodic Features

F0 and intensity data was extracted using Praat at 10ms intervals. For F0, parameter settings were automatically determined using the method described in [15]. To reduce pitch tracking errors we calculate these parameters over spurs: segments separated by at least 500ms silence based on word alignments [16]. Missing F0 values were obtained through linear interpolation after octave jump removal. The F0 values were speaker normalized into semitones relative to speaker mean F0 value (Hz) for that conversation. Intensity measurements were normalized by subtracting the speaker mean for the conversation. We calculate prosody statistics — mean, standard deviation, maximum and minimum over F0 and intensity — over words and DAs. Utterances often exhibit natural downdrift which means prosodic gestures that are later in an utterance are produced with lower F0. To correct for this within spurs, we first subtract the predicted values from linear regression from observed values before calculating aggregate features when the slope of the spur is negative. We include the slope as a feature at the DA level along with the other features (DA-pros).

2.2.2. Term-Frequency Based Lexical Features

For each of the words in meeting, we calculate tf.idf and su.idf as described in [9] with standard stopwords set to zero. To represent a range of speech genres, the inverse document frequency component of these measures was calculated over the combined AMI, ICSI and TDT2 corpora. The su.idf metric takes speaker term frequency into account and was shown to have good performance in [9]. We also calculate two Pointwise Mutual Information (PMI) features which separately measure the association between words and EDA/non-EDA status in information theoretic terms [17]. PMI values are calculated over the training set only and words that do not appear in the training set are set to zero. We are primarily interested in the relationship between prosody and tf.idf, as this metric is widely used across spoken language processing tasks other than summarization. It is also applicable in single speaker situations (unlike su.idf) and does not require labelled data to estimate (unlike PMI). All values are derived after applying the Porter stemmer. For DA level prediction, we sum individual term-frequency values over words in a

<table>
<thead>
<tr>
<th>Feature</th>
<th>ICSI</th>
<th>AMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>tf.idf</td>
<td>0.509</td>
<td>0.558</td>
</tr>
<tr>
<td>su.idf</td>
<td>0.513</td>
<td>0.517</td>
</tr>
<tr>
<td>pmi</td>
<td>0.567</td>
<td>0.669</td>
</tr>
<tr>
<td>tf.idf.pros</td>
<td>0.532</td>
<td>0.601</td>
</tr>
<tr>
<td>su.idf.pros</td>
<td>0.546</td>
<td>0.599</td>
</tr>
<tr>
<td>pmi.pros</td>
<td>0.562</td>
<td>0.673</td>
</tr>
<tr>
<td>pros</td>
<td>0.532</td>
<td>0.602</td>
</tr>
<tr>
<td>tsp</td>
<td>0.569</td>
<td>0.668</td>
</tr>
<tr>
<td>tsp.pros</td>
<td>0.563</td>
<td>0.672</td>
</tr>
</tbody>
</table>

Table 1: Development set AUROC for word level EDA detection.

given DA (e.g. DA-tf.idf). This approach has been shown to be useful in a number of studies [9, 18]. We refer to the model including all three summed features as DA-tsp.

2.2.3. Augmented Lexical Features

Our strategy for learning how to combine prosodic and term-frequency features at the word level was to cast the problem as classifying whether a word is in an EDA or not, using an MLP to learn feature combination weights. MLPs with one hidden layer were implemented using the theano toolkit [19] with the number of hidden states tuned on the development set. We use the probabilities determined by the top logistic regression layer as our augmented lexical features. We combined each of the three term-frequency features with all word level prosodic features \{tf.idf, su.idf, pmi\}.pros. We also looked at a prosody only model (pros), a model including all of the term-frequency features (word-tsp), and finally including all features (word-tsp.pros). We compare AUROC for augmented features with logistic regression classifiers using the single term-frequency features (Section 3.1).

2.3. EDA Detection and Evaluation

We use multilevel logistic regression to examine the efficacy of different feature sets for detecting EDAs. To account for differences between annotators and meeting types, as well as the unbalanced nature of the data, we include indicators for these at the group level. We model different annotators and meeting types as being drawn from normal distributions [20]. The meeting types were indicators for the 4 remote control design stages (AMI), and the 8 research groups (ICSI). For evaluation, we present AUROC and ROUGE-1 F-scores (i.e. unigram matching) calculated with DUC standard parameters [21] and a 15% word compression rate. Previous work has argued ROUGE-1 to be the most relevant version of ROUGE for meeting summarization [22]. We use gold standard annotations for calculating AUROC. Additional annotations were used to calculate ROUGE scores. The number of annotators for test set meetings was between 3-5 for the ICSI corpus and 2-3 for the AMI corpus.

3. Results

3.1. Word Level Prediction

Table 1 shows AUROC results for detecting words in summary linked EDAs on the development sets. The results indicate that augmenting tf.idf and su.idf with prosodic features improves classification performance, with a greater improvement seen for the AMI data. In fact, the prosody only based classifier performs better than both tf.idf and su.idf for both corpora. How-
and slope features only reduces AMI AUROC to 0.778, so the 
AMI: 0.73 vs 0.799). Excluding the new group level indicators 
are much better than reported in [3] (ICSI: 0.728 vs 0.811, 
as word prosody, it’s worth noting that DA prosody models 
augmented lexical features only helps for the ICSI data. 
However, adding DA prosody to the 
perform DA level combinations of the same features in AUROC 
monly provide better lexical coverage. 
3.2. DA Level Prediction 
Figure 1 shows the AUROC results when combining term fre-
quency features at the DA level (DA-lex), the word level (w-
lex), and using our prosodically augmented lexical features 
(w-lex.pros). We also look at the effect of adding DA level 
prosodic features (DA-lex). The results show that our aug-
mented lexical features (prosody or term-frequency based) out-
perform DA level combinations of the same features in AUROC 
(w-lex, w-lex.pros). The addition of DA prosody improves 
on bare lexical features. However, adding DA prosody to the 
augmented lexical features only helps for the ICSI data. 
8
Even though DA prosody does not appear to add as much 
as word prosody, it’s worth noting that DA prosody models 
am are much better than reported in [3] (ICSI: 0.728 vs 0.811, 
AMI: 0.73 vs 0.799). Excluding the new group level indicators and 
slope features only reduces AMI AUROC to 0.778, so the 

majority of the gain seems likely due to improved feature ex-
traction. In fact, contra to previous analyses, with improved F0 
analysis, we find that EDAs actually have significantly lower 
mean pitch than non-EDAs on average, but keep expanded pitch 
range. We also tried including prosodic delta features, however 
they did not produce much of a change in performance (ICSI: 0.812, AMI: 0.780, ± 4 DAs) 
ICSI AUROC for augmented word-tf.idf.pros and word-
tsp perform better than the best performing (full) feature set in 
[3] (0.829, 0.837 vs 0.818). Similarly, the word-tsp combi-
nation performs better than the best previously reported for 
AMI (0.861 vs 0.855). The improvement is greater for the ICSI 
data where term-weight features were previously reported as 
less effective. The augmented features perform better than DA 
duration (ICSI: 0.813, AMI: 0.82) and length in words (ICSI: 
0.831, AMI: 0.845), however the differences are quite small. 
Note, although PMI was a dominant predictor for the word 
level classification task, DA-PMI performs worse than DA-tsp 
(ICSI: 0.699 vs 0.726, AMI: 0.798 vs 0.833), which in turn per-
forms worse than word-tsp. In fact, incorporating prosody into 
word-tsp basically gives the same performance (AMI: ICSI: w-
tsp = 0.862, 0.838, w-tsp.pros = 0.861, 0.837). So, including ex-
tra term frequency features is more useful at the word level, 
however prosodic features can bridge the gap in their absence. 
In summing over lexical features, we assume they weight 
words as being more noteworthy for summarization purposes. 
So, we can view our augmented lexical features improving the 
weighting over bare tf.idf or PMI for DA aggregation. Our aug-
mented features provide more information than the DA length in 
words (a uniform weighting). While summing over such term 
weights has generally proven dominant in utterance level re-
trieval statistics like AUROC [23, 3], they have been reported 
as less effective when looking at n-gram based ROUGE [4, 22]. 
Figure 2 shows ROUGE-1 results for the different features 
sets. The ROUGE-1 scores mirror AUROC results for the ICSI 
data, with scores for the augmented lexical features significantly 
higher than DA-tf.idf. However, the AMI data is less clear. Al-
though DA-tf.idf still has the lowest score, the best performance 
for this data set comes from DA-tsp (0.595) though the range of 
differences are all within bootstrap confidence intervals. Note 
that ROUGE scores are several points lower than for ICSI al-
though AMI AUROC is similar or higher. Given these differ-
ences in rankings, we would like to know how these differences 
manifest in actual generated summaries. The next sections look 
at some other ways to measure summary differences. 
3.3. Redundancy 
While ROUGE scores give us an indication of how much over-
lap there is with a human summary, it doesn’t tell us much about 
redundancy, although this often appears as a constraint in un-
supervised approaches [24, 22]. We measure summary redu-
dancy by holding out each DA and measuring its cosine distance 
to the rest of the summary and sum the distances [24]. Figure 3 
shows a similar pattern to what we saw for AUROC measures. 
Summaries based on augmented lexical features were all signif-
icantly less redundant than those based on bare lexical features 
with and without DA level prosody (Wilcoxon p < 0.01, Holm 
corrected). Note, ROUGE-1 recall scores were the same or bet-
ter for the augmented lexical features than DA-level prosodic 
features. So, while the feature based classification method 
doesn’t explicitly take redundancy into account, longer dialogue 
acts tend to contain more independent information and poten-
tially provide better lexical coverage.
Our experiments indicate that augmenting lexical features such as tf.idf can improve extractive summarization and that incorporating prosodic information at the word level provides performance that outpaces lexically augmented features in AUROC terms. We found that summaries derived from prosodically augmented lexical features exhibit less redundancy. While DA prosody generally performed worse in retrieval terms, it may provide important information for temporally locating larger regions of interest. Although the objective measures discussed above present different perspectives on meeting summaries, understanding what prosodic models select beyond ROUGE style content matching is important for improving meeting summarization by isolated DA extraction. Although our DA prosody models did not have as high AUROC or ROUGE-1 F-scores as the text based features, they may be good indicators of regions of interest for browsing purposes. Identifying such regions rather than DAs may also help incorporate temporal aspects of meetings into automatic abstractive summaries. Predicting the amount of summary content could be useful for analyzing other aspects of group communication such as meeting productivity. The proportional measure we presented for looking at the distribution of predicted EDAs is just a starting point and a more detailed analysis is still required. Similarly, while reducing summary redundancy is theoretically good for improving meeting coverage lexically, its utility in meeting browsing requires further user based testing.

5. Conclusion and Future Work

Our next steps will embed summaries into meeting browsing tasks in order to shed light on whether notions like redundancy or the temporal distribution EDAs affect user efficiency and satisfaction. We are also working on inclusion of features related to visual emphasis. Given differences between ICSI and AMI ROUGE-1 results, the effect of different meeting structures between corpora also requires further investigation. Similarly, since AUROC and ROUGE-1 rankings only matched for the ICSI data, so it would be interesting to see how applying compression techniques as in [26] change ROUGE scores. Another approach is to incorporate prosodic features in unsupervised summarization methods that more closely fit ROUGE’s objectives [22]. Similarly, it may be useful to look at keyword identification as an objective for the generation of augmented lexical features, an approach initially explored in [10].
6. References


