Level of Interest Sensing in Spoken Dialog Using Multi-level Fusion of Acoustic and Lexical Evidence

Je Hun Jeon, Rui Xia, Yang Liu

Department of Computer Science
The University of Texas at Dallas, Richardson, TX, USA
{jhjeon,rx,yangl}@hlt.utdallas.edu

Abstract

Accurately sensing a user’s interest in spoken dialog plays a significant role in many applications, such as tutoring systems and customer service systems. In addition to the widely used acoustic evidence, we introduce different lexical features for interest level prediction and evaluate the impact of automatic speech recognition (ASR) on the effectiveness of lexical information. In order to capture contextual information, we combine the system’s hypothesis for the previous turn with the current one. Our final system uses a multi-level fusion method for this task. Each fusion step uses different information such as acoustic and lexical cues, contextual information, or hypotheses from different classifiers. Our experiments show that various combinations improve system performance. In particular, we found that even though the word error rate is quite high, there is still performance gain by incorporating lexical information obtained from ASR output.

Index Terms: Multi-level fusion, level of interest, affect sensing

1. Introduction

Information on interest or disinterest of users in spoken dialog systems has received a considerable amount of research attention. A major motivation comes from the desire to develop human machine interfaces that are more adaptive and responsive to a user’s behavior. Accurate recognition of the user’s interest plays a significant role in many applications, such as tutoring systems and customer service systems. However, this is a very challenging task because of various factors, such as the difficulty of proper definitions for affective states, speaker variation to express interest, poor speech recognition performance, task dependency, and lack of corpora.

There has been a lot of work on recognition of affective or emotional states, which is related to level of interest recognition in this study. Previous work has used various information sources, such as acoustic evidence from speech [1, 2, 3], vision information [4], linguistic analysis [5], and fusion of audio-visual cues [6, 7] and audio-linguistic evidence [8]. As shown in many studies, multi-modal processing is superior to relying on any single modality. In previous work, most linguistic analysis has been performed on human transcripts and not on actual ASR output. It is expected that high recognition errors for emotional speech will hurt affect/emotion recognition using lexical information.

This paper focuses on recognizing the level of user’s interest using acoustic and lexical cues as well as contextual information from spoken language. We propose using a multi-level fusion method to combine different information sources at different stages. For lexical information, we compare different term weighting methods and evaluate the impact of ASR word errors. Our experimental results show that combining contextual information at the decision level is useful and integrating lexical information achieves significant performance improvement, even though the word error rate is quite high.

2. Task and Our Approach

The task of this paper is to determine level of interest (LoI) in ordinal representation, and thus regression is used. This is one of the subchallenges in Interspeech 2010 Paralinguistic challenge [9]. The data used for this task is the Audiovisual Interest Corpus recorded at the Technische Universitat Munchen (“TUM AVIC”) [9]. Since details of the acoustic features are provided in [9], we only describe the lexical information and the approaches we use for this task.

2.1. Lexical Evidence

Since a listener provides feedback in dialog, we expect that the interest level is dependent on not only how the person says something (represented by acoustic features), but also what the person said. Because syntactic and semantic understanding techniques are not mature and they are also significantly impacted by word errors in ASR output, in this study we mainly use lexical information, i.e., word usage in an utterance.

The primary features we use are represented by a bag-of-word model. The pause, non-linguistic vocalizations (such as breathing, consent, hesitation, laughter, and coughing), and other human noise are also considered to be a type of word. We compared using different dictionary sizes based on different frequency cut off and found that a cut off threshold of 1 performed best; that is, we kept words that appear more than once in the training corpus and the others are collapsed into one word. We use different weighting methods for these words.

- Term frequency (TF). This is just the frequency of the word in a turn.
- Subjectivity information of the word (SUBJ). We use the subjectivity dictionary [10] to assign a value of 1 to strong subjective words in the dictionary, and 0.5 to weak subjective words.
- LoI information of the word (LoI). We first estimate the LoI value of each word from the training set. The LoI value of a word is the averaged LoI value of all the utterances that contain this word.

In addition to these bag-of-word features, we develop three other features: (a) length of the turn represented using the number of words (LEN); (b) the number of strong subjective words...
all the words in the sentence (Max LoI). We use support vector machine (SVM) as the lexical classifier with an RBF kernel since it performed better than other classifiers in our preliminary experiments. For lexical information, we want to investigate two questions: Does it contribute anything in addition to the acoustic evidence? If so, will word errors in ASR output affect performance? We will use both human transcripts and ASR output to answer these questions.

2.2. Approach

We explore a few methods to improve performance upon the baseline. For acoustic and lexical evidence, we tried to optimize their individual performance by using feature selection, different feature configuration, different classifiers as well as their combination.

Our final system makes a prediction based on multi-level combination of different information sources; acoustic, lexical, context, and different classifiers. The overall system structure is shown in Figure 1. There are several different information fusions in our system.

- Combination 1: Acoustic level. We combine two hypotheses from two classifiers using acoustic evidence. The base classifier (RandomSubspace with REPTree) and GaussianProcess with NormalizedPoly kernel in the Weka toolkit [11] are used since they achieve the best performance among several individual classifiers.

- Combination 2: Context information. We expect that there is no sudden change of interest level from the listener in two consecutive turns, therefore we use contextual information to help predict LoI for the current utterance. We combine the predicted LoI of the previous turn with the current turn to make the new hypothesis for the current turn.

- Combination 3: Integrate lexical information. The decisions from the acoustic classifiers are combined with the results using lexical information at the decision level.

For all the fusion steps, a weighted linear interpolation method is used. The weights are estimated using a held-out development set. It is worth pointing out that the lexical combination part (Combination 3) may be done with different acoustic systems, the base classifier, combination of the two classifiers, or after combination with acoustic context. In addition, some combination components in Figure 1 can be skipped, and the combination does not necessarily follow the order as in the figure. For example, we may switch the order of context combination and lexical fusion, but we found using context for acoustic information only is better. For the test set, since we do not have time information for testing utterances, we can not use contextual information, therefore the contextual combination stage is not used.

3. Experimental Results

Since we do not have test set reference labels, all the following experiments will be based on the training and development set. We divide the development set into two folds in order to estimate the combination weights for different fusion steps. The optimal weights from one fold are applied to the other for testing. We used correlation coefficient (CC) and mean linear error (MLE) as the performance measure.

3.1. Baseline

For the baseline, the RandomSubspace model using acoustic evidence [9] is used. We also evaluated several other classifiers, but did not find one that consistently outperformed this baseline. Table 1 shows the baseline results. We also include the results using the GaussianProcess classifier that performs reasonably well compared to the RandomSubspace model. We will see later that their combination also yields some gain.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>CC</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RandomSubspace (baseline)</td>
<td>0.604</td>
<td>0.118</td>
</tr>
<tr>
<td>GaussianProcess</td>
<td>0.593</td>
<td>0.120</td>
</tr>
</tbody>
</table>

Table 1: Results using acoustic features using the RandomSubspace and GaussianProcess classifiers.

3.2. Using Lexical Information

For the bag-of-word model, we evaluated different vocabulary sizes based on word frequency in the training set. The size of the vocabulary changes from 1030 when using all the words (i.e., cut off frequency is 0) to 533 when words appearing more than once are used, and to 351 with a cut off of 2. For the reduced vocabulary case, words that are not preserved (frequency lower than threshold) are collapsed into one word. The best performance is achieved with a cut off threshold of 1, so we use this to form the bag-of-word features. Table 2 shows the results using the features and the weighting methods described in Section 2.1. The “BASE” features are the bag-of-word features, 534 words in total. These results are based on human transcripts provided for the development set.

<table>
<thead>
<tr>
<th></th>
<th>TF</th>
<th>SUBJ</th>
<th>LoI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE</td>
<td>CC</td>
<td>0.457</td>
<td>0.459</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>0.132</td>
<td>0.135</td>
</tr>
<tr>
<td>+ LEN</td>
<td>CC</td>
<td>0.502</td>
<td>0.499</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>0.128</td>
<td>0.130</td>
</tr>
<tr>
<td>+ SUBJ_W</td>
<td>CC</td>
<td>0.502</td>
<td>0.496</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>0.128</td>
<td>0.130</td>
</tr>
<tr>
<td>+ Max LoI</td>
<td>CC</td>
<td>0.507</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>0.128</td>
<td>0.129</td>
</tr>
</tbody>
</table>

Table 2: Results using lexical information: evaluation of different features and weighting methods.

We can see that changing weighting methods (using TF, SUBJ, or averaged LoI) does not result in a significant performance difference. When adding the length of the turn (LEN) there is a large performance gain, consistent across different weighting methods. Adding the number of strong subjectivity words (SUBJ_W) and the maximum LoI value of words...
(Max LoI_W) achieves small performance gain. The overall performance of the TF and LoI weighting setups is similar: CC is slightly better for LoI weighting, but MLE is better for TF weighting. From our experiments using acoustic evidence, we found that reducing MLE seems more difficult than increasing CC, thus we decide to use all the features with TF weighting for bag-of-word features in our final system.

Next, we compare the performance change using ASR output vs. human transcripts. We use the best feature configuration based on the above result. For this experiment, Sphinx-3 was used to generate transcripts. We used TUM AVIC corpus for acoustic and language model training. We also added TIMIT corpus for acoustic and Fisher data for language model training to generate various transcripts with different word error rates (WER), resulting in WER of 34%, 51%, and 74%. For example, 34% WER is obtained using an acoustic model trained from all the training and development data as well as TIMIT data, and a language model trained from all the training and development data as well as Fisher corpus; 74% is obtained when using the training data for acoustic and language model training. To generate better ASR output, we randomly created errors in human transcripts, resulting in transcripts with 10% and 19% WER. In this process, we used a similar proportion of insertion, deletion, and substitution errors as in the real ASR results with higher WER.

![Figure 2: Results using different fusions. 'Base' is the baseline result using RandomSubspace for acoustic features; 'C 1' uses acoustic classifier combination; 'C 1+2' further adds the hypothesis from previous turn; 'C 1+2+3' combines the decision from the lexical classifier with that from 'C 1+2'; 'C 3' uses the combination of the baseline acoustic classifier and the lexical classifier.

![Figure 3: Results using the combination of the lexical and baseline acoustic classifier. Lexical information is extracted from transcripts with different WERs.

Table 3: Results using lexical features extracted using transcripts with different WERs.

<table>
<thead>
<tr>
<th>WER (%)</th>
<th>10%</th>
<th>19%</th>
<th>34%</th>
<th>51%</th>
<th>74%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>0.507</td>
<td>0.490</td>
<td>0.408</td>
<td>0.475</td>
<td>0.439</td>
</tr>
<tr>
<td>MLE</td>
<td>0.128</td>
<td>0.130</td>
<td>0.131</td>
<td>0.131</td>
<td>0.136</td>
</tr>
</tbody>
</table>

Table 3 shows the prediction results using lexical features extracted from the transcripts with various WERs. We can see that the major performance drop seems to occur when the WER is higher than 50%. For better ASR output, the performance is rather comparable to that of human transcription. Note that we use some artificial transcripts (WER of 10% and 19%) in this experiment. The errors in those transcripts are different from the real ASR errors, therefore there might be some artifacts in the results. Despite this, we expect the trend will hold using real ASR output, and will verify this in our future work.

3.3. Combination

Figure 2 shows the results using different levels of information fusion as described in Figure 1. Human transcripts are used for this experiment. We included the condition using the combination of the base acoustic classifier with the lexical classifier (C 3) to focus on the effect of lexical information. We can see that using more information achieves a higher performance gain. Combining contextual information improves CC but degrades MLE a bit (0.1775 vs 0.1776). Most of the performance gain comes from the addition of lexical information, as also demonstrated from the last result in the figure. When combining only lexical information with the baseline (without acoustic classifier combination or context information), there is slight performance degradation compared to using all the three fusion levels (‘C 1+2+3’).

We also evaluate the impact of WER on lexical information when it is combined with acoustic classifiers. Results are shown in Figure 3 using the same set of transcripts as used in Table 3. The two horizontal lines represent the baseline results of CC and MLE. For this experiment, we only used the baseline acoustic classifier combined with lexical information at the decision level. Not surprisingly, in general performance decreases as WER goes up. We also notice that the weighting value for the lexical model also decreased, from 0.38 for human transcripts (0 WER) to 0.16 for 74% WER. When the performance of the lexical classifier is better, its weight is higher in the system combination. When transcripts with WER higher than 50% are used, there is still some performance gain in CC, but MLE is slightly lower than baseline.

3.4. Discussion

3.4.1. Combination methods

In the experiments above, we used the linear combination of various input at different levels. We also investigated using a super-classifier approach that takes the predictions from various sources as input and learns a combination function. We use all the sources, hypotheses from different acoustic classifiers, context, and lexical classifier. We train an SVM with this feature set using one fold of the development set and apply it to the other fold. The best performance of this method is obtained from using only the predictions of the baseline acoustic and lexical model (result is 0.622 for CC and 0.115 for MLE for human transcripts). Compared with the best performance using different levels of linear combination (result is 0.627 CC and 0.117 MLE with all combination), this super-classifier approach reduces MLE a bit and has slightly worse CC.

3.4.2. Class distribution

For analysis, we want to know whether the distribution of the LoI values has an impact on system performance. Figure 4 shows the number of instances for different LoI levels in the training and development data, and for the hypotheses using different systems. We can see that most of the training data is located between 0.4 and 0.8. This skewed distribution may cause
a problem for model learning. As shown in the figure, the predicted values are more saturated in that range, even more than that in the training data. This is more noticeable for the lexical classifier than the acoustic one. Their combination lies in between. This kind of concentration reduces errors in that range, but causes more errors in regions with higher and lower LoI values. To reduce this data skew, we tried upsampling approaches to increase the number of instances in the regions with high and low values, but that did not improve performance. From the figure, we can see that the distribution of the development data is similar to the training data. If the distribution is different, we expect that there may be some performance degradation.

3.4.3. Feature issues

We tried to improve the performance at the feature level. For the acoustic features we performed various feature selection methods, but the reduced feature set did not improve performance. To take contextual information into consideration, we also used feature level combination where we added the features from the previous turn or the delta features to the current feature vector. There was no consistent gain from those feature level combinations, thus we did not use these features in the final system.

From the two folds of the development data, we notice that there is a big performance difference between them: 0.643 CC and 0.114 MLE in the first fold, and 0.567 and 0.122 for the second using the base acoustic classifier. The distribution of the LoI values is similar for the two folds. However, the speakers are different. The first fold includes 1 male and 2 females, and the second fold includes 2 males and 1 female. The big difference between them suggests that the current acoustic features may not be robust to represent speaker variation.

Regarding lexical features, we adopted frequency based selection on the manually transcribed training data. The chosen words are 533, which are less than 30% of all the words in the dictionary. This may result in low word coverage in testing, especially for the erroneous ASR transcripts. We need to develop better word selection methods to increase feature coverage and robustness for ASR condition.

3.4.4. Context information

Since the data contains only the turns from the listener and does not include the presenter’s speech, we cannot use information from the presenter. For the context information, we only use the listener’s previous turn. We expect that information from the presenter would be very useful. Furthermore, in some applications, presenter’s speech is made by a dialog system, and thus it will not suffer from erroneous transcription.

3.5. Results on Blind Test Data

For the blind test set, since context information is not available, we only used the combination from the two acoustic classifiers and the lexical classifier for the current utterance. We used the super-classifier (SVM) to combine these sources since on the development set we found this is more stable for MLE than combining them sequentially (first two acoustic classifiers, then with the lexical classifier). The baseline results (using the RandomSubspace acoustic classifier) are 0.421 and 0.146 for CC and MLE respectively [9]. Our results are 0.428 and 0.146. Our approach improved CC slightly but MLE was almost the same. From our experimental results on the development set, we think the low performance gain is mainly caused by poor ASR results on the test set, which might be about 70% WER. If we can reduce the WER, we expect more performance gain from using lexical information.

4. Conclusions

In this paper, we proposed using multi-level fusion for level of interest sensing. Each fusion incorporates different information from acoustic, lexical, or contextual evidence. The combination of lexical information achieves the most significant gain (when using human transcripts) upon the baseline system using acoustic features. We also investigated how ASR word errors affect the effectiveness of lexical information and showed that there is still some performance gain even when using erroneous ASR output. We plan to explore other features and models to improve system robustness in our future work.

5. References