Recognizing Emotions in Dialogues with Acoustic and Lexical Features

Leimin Tian, Johanna D. Moore, Catherine Lai

School of Informatics, the University of Edinburgh
Informatics Forum, 10 Crichton Street, Edinburgh, UK, EH8 9AB
s1219694@sms.ed.ac.uk, j.moore@ed.ac.uk, clai@inf.ed.ac.uk

Abstract

Current research in emotion recognition focuses on identifying better feature representations and recognition models. The goal of this project is to improve on current automatic emotion recognition performance by identifying more predictive knowledge-driven features, and by building a hierarchical contextual model that combines state-of-the-art statistical and knowledge-driven features at different layers. Our model will have the potential to improve the quality of emotional interactions in current dialogue systems.

To improve on current approaches, we propose novel disfluency and non-verbal vocalisation (DIS-NV) based features, and show that they are highly predictive for recognizing emotions in spontaneous dialogues. We also propose an enhanced Long Short-Term Memory Recurrent Neural Network (LSTM) model that combines the DIS-NV features and other acoustic and lexical features at different layers.

Index Terms: emotion recognition, disfluency, non-verbal vocalisation, LLD, LSTM, human-computer interaction

1. Introduction

Research in cognitive science has shown that emotions are vital in human cognition and communication processes [1]. Thus, automatic recognition of emotion is crucial for advancing technologies related to human-computer interaction. For example, in affective game design, Non-Player Characters that are aware of the emotional states of the player and can generate emotional reactions are shown to keep players engaged and improve the gaming experience [2].

Our overall goal is to develop a predictive and robust emotion recognizer that can be applied to improve the interaction quality of current dialogue systems. State-of-the-art approaches for emotion recognition focus on identifying better feature representations and developing more predictive classification models. How to identify predictive features and models is still an open problem. In our current work, we proposed novel knowledge-driven features and multimodal contextual hierarchical models motivated by psychological studies that can improve on state-of-the-art of emotion recognition. We also conducted emotion recognition tasks on both spontaneous and acted dialogues to gain a better understanding of the data aspects that constrain the effectiveness of features and models.

Features used in emotion recognition can be extracted from various modalities (e.g., audio, visual, and lexical). Our work focuses on features describing the acoustic and lexical characteristics of the dialogues. Previous studies identified statistical Low-Level Descriptor (LLD) based acoustic features as a highly predictive type of feature for emotion recognition. These LLD features have been widely used in state-of-the-art models. However, the LLD features describe data at the frame-level from the signal perspective. Psycholinguistic studies have suggested characteristics at the utterance-level being useful clues in human emotion recognition [3]. Thus, in our previous work [4], we proposed features describing the occurrence of disfluencies and non-verbal vocalisations (DIS-NVs) in utterances, and our results showed that they were the most predictive type of feature for recognizing emotions in spontaneous dialogues. Our DIS-NV features obtained state-of-the-art performance, especially for recognizing the Expectancy dimension of emotion which relates to the uncertainty of the speaker.

Most widely used classification or regression algorithms have been applied to build emotion recognition models. There have also been studies on feature engineering for emotion recognition. Although many different algorithms exist and it is important to choose the appropriate one for a specific task, previous work has suggested that the predictiveness of features may have greater influence on performance, and there may not be significant differences between the performance of different machine learning algorithms when using the same feature set under similar circumstances [5]. In recent years, however, the development of deep learning models has yielded significant performance improvements in machine learning tasks, especially in computer vision and speech recognition [6]. The hierarchical structure of deep learning models enables the models to learn better feature representations automatically. Improved performance is also obtained in emotion recognition tasks using deep learning models over conventional machine learning algorithms. For example, deep neural networks obtained the best reported results in detecting Valence values and level of conflict [7]. The Long Short-Term Memory Recurrent Neural Network (LSTM) model has been widely used in emotion recognition because of its ability to model long-range contexts.

Combining multiple modalities generally improves emotion recognition performance. However, the improvement is often limited [8]. One reason may be that most multimodal models combine different modalities at the same level. However, different modalities may describe data at different levels. For example, the LLD features describe data at the frame-level, while the DIS-NV features describe data at the utterance-level. Combining these features at different levels may improve the benefits gained by modality fusion. Thus, we propose an enhanced LSTM model which will use different types of features at different layers of the network structure, and will study its performance compared to combining modalities at the same layer.

1.1. Research Questions

There are three important aspects for building an emotion recognition model: the data, the feature set, and the recognition model. There are two main approaches for collecting conversa-
tional emotional databases: by acting, or by recording spontaneous dialogues. There are two main types of features we can extract for most modalities: knowledge driven features describing cues identified in psychological studies of human emotion recognition, or statistical features describing properties of the data. For the model aspect, models may use information only from the current time, or they can include contextual information; From a structural view, flat models use the input feature representations directly, while hierarchical models are designed to learn better feature representations before performing classification or regression.

The key research question in emotion recognition is finding the features and models that will improve the recognition performance under the constraints of specific types of dialogue and size of data set. We work on this question by proposing the DIS-NV features and the enhanced LSTM model and experiment on different types of dialogues.

Most emotion recognition studies rely on intrinsic measures to evaluate different approaches (e.g., correlation coefficients or classification accuracies). This leads to another important question in emotion recognition: will performance improvements shown in intrinsic tests of emotion recognition models result in improvements in emotional interaction quality (e.g., higher engagement and satisfaction of the user), when the emotion recognition model is applied to a human-computer interaction system? We plan to work on this question in the future when we have an available dialogue system to apply our models to.

1.2. Hypotheses

• The DIS-NV based features will improve performance of existing emotion recognition models
• The enhanced LSTM model that combines features at different layers will have better performance than a LSTM model that combines features at the same layer
• An emotion recognition model using the proposed features and model structure will improve state-of-the-art emotion recognition performance

2. Background and Related Work

2.1. Psycholinguistic Studies

2.1.1. Emotion Theories

There are on-going debates in psychology on how to define, study, and explain emotions. Among all the theories, there are four major approaches that have influenced computational studies of emotions: the Darwinian, Jamesian, cognitive, and social constructivist perspectives [9]. Many current automatic emotion recognition studies follow the Darwinian emotion theory, which defines emotions as several primary and universal categories, such as Ekman’s Big-6 emotion categorization [10]. However, our work focuses on the cognitive emotion theory, which associates emotions with specific appraisals (assessment of stimulus that evoke changes in emotional states) and use a set of primitive appraisal components or dimensions to define emotions. This is because our goal is to build emotion recognition models that can be applied to the emotional interaction module of human-computer interaction systems, which are mostly developed with appraisal-based emotion models. In our work, we use four common emotional dimensions that have been identified as able to describe most of the everyday human emotions [11]: Arousal(A), Expectancy(E), Power(P), and Valence(V).

2.1.2. Human Emotion Recognition

The emotional state of a person during conversations tends not to change rapidly and thus depends on the context. Humans convey and perceive emotions through all communicative modalities. When recognizing emotions, human subjects are shown to have better performance when given information from multiple modalities [12]. These findings indicate that a contextual and multimodal model may have better emotion recognition performance. In our current work, we focus on emotion recognition from the audio and lexical modality, which includes acoustic information and contents of the speech.

Psycholinguistic studies have shown that prosodic cues and the lexical content of the speech are important in human emotion recognition (see [12] for a survey). However, disfluencies and non-verbal vocalisations (DIS-NVs) are also a common and interesting phenomena in natural speech. To the best of our knowledge, there have been no psycholinguistic studies showing direct relations between disfluencies and emotions. However, emotions can influence the neural mechanisms in the brain, and thus influence sensory processing and attention [13]. This in turn influences speech processing and production, which may result in disfluencies [14]. Current studies on human-human dialogues suggest that disfluency conveys information such as uncertainty [14], which relates to the Expectancy emotional dimension. Non-verbal vocalisations, especially laughter, have been identified as universal and basic cues in human emotion recognition [3]. Thus, we propose several DIS-NV features to study the predictiveness of DIS-NV for emotion recognition.

2.2. Automatic Emotion Recognition in Dialogues

2.2.1. Databases

Here we introduce the AVEC2012 database of spontaneous dialogues [15] and the IEMOCAP database of acted dialogues [16] as examples of state-of-the-art emotional databases. They are the most widely used databases of English dialogues annotated with dimensional emotions.

The Audio/Visual Emotion Challenge 2012 (AVEC2012) database [15] contains the Solid-SAL part of the SEMAINE corpus [17]. It includes approximately 8 hours of audio-visual recordings and manual transcripts of 24 subjects conversing with 4 on-screen characters with specific personalities role-played by human operators. Emotions in the AVEC2012 database were annotated at the word-level and the frame-level as real-value vectors in the Arousal-Expectancy-Power-Valence emotional space.

The IEMOCAP database [16] contains approximately 12 hours of audio-visual recordings from 5 mixed gender pairs of actors. The recordings were manually transcribed. There are two types of dialogues in the IEMOCAP database: non-scripted and scripted dialogues. When collecting the non-scripted dialogues, the actors were instructed to act out emotionally intense scenarios. When collecting the scripted dialogues, the actors followed pre-written lines. Emotions were annotated at the utterance-level with a 1 to 5 integer score of the Arousal, Power, and Valence emotional dimensions.

2.2.2. Related Work

Related work on emotion recognition has shown that for the feature perspective, performance of different types of features vary for different types of dialogues, and combining multiple types of features will improve performance. For the model perspective, contextual and hierarchical models have better perfor-
mance than non-contextual and flat models.

Previous work on both the AVEC2012 and the IEMOCAP database has focused on LLD features for the acoustic model (e.g. [18]). However, there are results indicating that knowledge-driven features, such as global prosodic features, may also be highly predictive (e.g. [19]). Most of the recognition models on the AVEC2012 database used non-contextual models. However, models that included contextual information, in either the features extracted [4] or the recognition model used [20], have shown better performance in emotion recognition. To the best of our knowledge, the only previous work on the AVEC2012 database that applied deep learning models used the LSTM model to learn better feature representations, and then applied Support Vector Regression on the outputs of the LSTM models [21]. LSTM models were used directly for classification in previous work on the IEMOCAP database, and they obtained better performance than Hidden Markov Models (e.g. [18]).

3. Methodology

3.1. Features

Three types of acoustic and lexical features are studied. The utterance-level knowledge-driven features describe general emotional cues with a small feature set, and highlight the utterances that may be specifically interesting for emotion recognition. The frame-level statistical features give detailed information of all the data with a large feature set.

3.1.1. Knowledge-Driven Features

DIS-NV Features: We manually annotated three types of disfluencies (filled pauses, fillers, and stutters) and two types of non-verbal vocalisations (laughter and breath). Feature values are the ratios between durations of DIS-NV and the utterance duration. We have also tested other types of common DIS-NVs (speech repairs, silent pauses, turn-taking times, sigh, and prolongations). However, adding them to the DIS-NV feature set does not improve recognition performance, thus are omitted. Compared to the AVEC2012 database of spontaneous dialogues, DIS-NVs are less frequent in the IEMOCAP database of acted dialogues. This indicates fundamental differences between spontaneous and acted dialogues.

PMI Lexical Features: Point-wise Mutual Information (PMI) is a widely used measurement for the relation of words and emotions based on the frequency of a word having a class label. PMI based lexical features were the second most predictive features in previous work on the AVEC2012 database [4][20]. The lexical features we proposed are calculated as the total PMI values of all the words in an utterance for each binaryized emotional dimension.

3.1.2. Statistical Features

The LLD acoustic features were extracted using a frame-level sliding window. Functionals (e.g., mean) were applied to LLDs (e.g., MFCCs) and their corresponding delta coefficients. The OpenSMILE toolbox [22] was used to automatically extract these features from audio recordings. We choose the most widely used LLD feature set from previous work on each database as the reference set for experiments on this database.

3.2. Classification Models

The non-contextual and flat SVM models were built with the LibSVM tool [23]. We used the C-SVC approach with RBF kernel. All features were normalized to [-1,1] before classification. This is the setting widely used in previous work.

The contextual and hierarchical LSTM model is a recurrent neural network with a special structure called “the memory cell” that can model distant contexts. Our LSTM models were built with the PyBrain toolbox [24]. The LSTM model has more parameters that need to be trained compared to the SVM model.

4. Work Done So Far

Our previous results on the AVEC2012 database verified the predictiveness and robustness of the DIS-NV features on spontaneous dialogues. Our experiments have shown that the knowledge-driven DIS-NV features and PMI lexical features perform better than the statistical LLD features when recognizing emotions in spontaneous dialogues [4]. In contrast, the DIS-NV features and PMI lexical features are less predictive than the LLD features in previous experiments on acted dialogue, which may be because of the infrequency of DIS-NVs in acted dialogues compared to in spontaneous dialogues, and the use of scripts in part of the IEMOCAP database [25]. These findings verified that the performance of different types of features vary for different types of dialogues.

Our current results comparing the DIS-NV and LLD features and the SVM and LSTM models on the AVEC2012 and IEMOCAP database are shown in Table 1. Numbers are weighted F-measures (%). “Mean” is the unweighted average of results on all emotion dimensions. “~S” is using the SVM model, “~L” is using the LSTM model. “DN” is using the DIS-NV features. “DN+LLD” represents concatenating the DIS-NV and the LLD features at the input layer of a LSTM model.

As we predicted, the complex structure of the hierarchical LSTM model constrains its predictiveness. Because of the small size of the DIS-NV feature set compared to the LLD feature set, the DN-L model has less parameters to train than the LLD-L model. Thus, the DN-LSTM models outperformed the DN-SVM models in both databases, while the LLD-L models are less predictive than the LLD-SVM models in both databases. Consistent with our previous results, the DIS-NV features are more predictive than the LLD features on the AVEC2012 database of spontaneous dialogues, but not on the IEMOCAP database of acted dialogues. The result that the DN+LLD model only outperformed both the DN-LSTM and the LLD-L model on the IEMOCAP database indicates that a better fusion strategy is needed.

Table 1: Current results with LSTM models.

<table>
<thead>
<tr>
<th>Models</th>
<th>A</th>
<th>E</th>
<th>P</th>
<th>V</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>DN-S</td>
<td>52.4</td>
<td>61.4</td>
<td>67.4</td>
<td>59.2</td>
<td>60.1</td>
</tr>
<tr>
<td>LLD-S</td>
<td>52.4</td>
<td>60.8</td>
<td>67.5</td>
<td>59.2</td>
<td>60.0</td>
</tr>
<tr>
<td>DN-L</td>
<td>54.1</td>
<td>65.8</td>
<td>68.3</td>
<td>60.1</td>
<td>62.0</td>
</tr>
<tr>
<td>LLD-L</td>
<td>52.4</td>
<td>60.7</td>
<td>66.1</td>
<td>58.1</td>
<td>59.3</td>
</tr>
<tr>
<td>DN+LLD</td>
<td>52.5</td>
<td>61.2</td>
<td>65.8</td>
<td>58.0</td>
<td>59.4</td>
</tr>
<tr>
<td>Cross-Validation Results on the IEMOCAP Database</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DN-S</td>
<td>36.3</td>
<td>#</td>
<td>40.7</td>
<td>32.8</td>
<td>36.6</td>
</tr>
<tr>
<td>LLD-S</td>
<td>65.2</td>
<td>#</td>
<td>53.8</td>
<td>53.5</td>
<td>57.5</td>
</tr>
<tr>
<td>DN-L</td>
<td>41.6</td>
<td>#</td>
<td>37.8</td>
<td>34.0</td>
<td>37.8</td>
</tr>
<tr>
<td>LLD-L</td>
<td>53.7</td>
<td>#</td>
<td>46.2</td>
<td>38.6</td>
<td>46.2</td>
</tr>
<tr>
<td>DN+LLD</td>
<td>53.9</td>
<td>#</td>
<td>51.6</td>
<td>39.5</td>
<td>48.3</td>
</tr>
</tbody>
</table>
5. Future Plans

In the remaining time of my PhD study, I plan to improve performance of the emotion recognition models by including global prosodic features describing duration, speaking rate, pitch, energy, amplitude, and spectral features of the utterances, as suggested by psycholinguistic studies [26]. I’ll also extract more robust PMI lexical features based on PMI values calculated from several available corpora with dimensional emotion annotations (e.g., [27]). We have also conducted pilot experiments on our enhanced LSTM model which uses the LLD features at the input layer and includes the DIS-NV features at a higher hidden layer. Results on the AVEC2012 databases have shown better performance gained compared to using the DIS-NV and LLD features at the input layer ($A = 53.4, E = 63.2$). In the next step, we will use the LLD features at the input layer, and the DIS-NV, the PMI lexical, and the global prosodic features at a higher layer in the network structure. The enhanced LSTM model will also be trained on a merged database of the AVEC2012 and IEMOCAP databases. If there is an available human-agent dialogue system, I will also apply the best performing model to this dialogue system and evaluate our model’s influence on the quality of interaction in the future.

6. References


