Interactional Style Detection for Versatile Dialogue Response Using Prosodic and Semantic Features

Wei-Bin Liang, Chung-Hsien Wu, Chih-Hung Wang, Jhing-Fa Wang*

Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

{liangnet, chunghsienwu, sheleehom}@gmail.com, *wangjf@csie.ncku.edu.tw

Abstract
This work presents an approach to interactional style (IS) detection for versatile responses in spoken dialogue systems (SDSs). Since speakers generally express their intents in different styles, the responses of an SDS should be versatile instead of invariable, planned responses. Moreover, the IS of dialogue turns can be affected by dialogue topics and speakers’ emotional states. In this study, three base-level classifiers are employed for preliminary detection, including latent Dirichlet allocation for dialogue topic categorization, support vector machine for prosody-based emotional state identification and maximum entropy for semantic label-based emotional state identification. Finally, an artificial neural network is adopted for IS detection considering the scores estimated from the aforementioned classifiers. To evaluate the proposed approach, an SDS in a chatting domain was constructed for evaluation. The performance of IS detection can achieve 82.67%.

Index Terms: Interaction style, spoken dialogue system, versatile response, topic, emotional state

1. Introduction
Spoken dialogue systems have been applied to a wide range of domains, from simple goal-oriented applications such as airline travel information service [1], city tour guiding [2] and nursing home assistant [3] to complex conversational systems such as chatterbot- A.L.I.C.E. [4] and Jabberwacky [5]. Although significant efforts have been made in conversational SDSs, an invariable, planned response is generally employed to respond to the speakers’ speech acts [6]. However, speakers can express their intents in different styles. Thus, flexible, versatile interaction could be taken into account. For example, when a call center system detects an utterance produced in an angry and fire-eating style, the system can respond to the speaker gently or change to the customer service representative. Therefore, it is an interesting research issue that how the SDS can measure the expression level of the speaker while the speaker lost something. For choleric style, the listener can try to calm the speaker down while the speaker talks about politics. In an SDS, the interactional style is the listener’s perception of speech from the speaker. Though speech is the only material in an SDS but speech can be converted to word string and prosodic features. Furthermore, word string obtained from a speech recognizer could be used for semantic feature extraction and then passed to the state-of-the-art topic classification approach - Latent Dirichlet allocation (LDA) for dialogue topic categorization. Besides, prosodic information can express a speaker’s emotional state. For emotional state detection, two eminent base-level classifiers - support vector machine (SVM) and maximum entropy (MaxEnt) are utilized based on the prosody and semantic information, respectively. Finally, an artificial neural network (ANN) is adopted to estimate the complex non-linear dependencies on the three base-level classifiers for each interactional style.

The paper is organized as follows: Section 2 outlines the framework of this work and describes the detailed training phase for each component in the framework; Section 3 shows the data collection and detection results; Section 4 provides the conclusions.

2. Framework
Figure 1 illustrates the framework of interactional style detection based on three base-level classifications, including semantic label-based emotion recognition (SLER), prosodic information-based emotion recognition (PIER), and dialogue topic categorization (DTC). Because the three base-level classifiers process different types of input sources, all of the input sources should be converted to the corresponding feature spaces. Therefore, the training instances, including text and speech, are first used to train the semantic label-based emotion
recognition models (SERMs), prosodic information-based recognition models (PERMs), and dialogue topic categorization model (DTCM) using MaxEnt, SVM, and LDA, respectively. DTCM is employed to characterize the key words for topic categorization. SERMs and PERMs are utilized to model semantic labels and prosodic information for emotional state detection, respectively. Then, the same training instances are used again to estimate the confidence scores for each emotional state and each dialogue topic. To integrate the three base-level measures, an ANN is used to train the IS detection model. Finally, given a test instance, the confidence scores will be obtained from the three base-level models. The output node of the ANN with the highest score is used to output the detected IS. The detailed training procedure of each base-level classification will be described in the following sections.

2.1. Semantic Label-based Emotion Recognition

In a conversational system, speakers can express their emotional state in different ways. For example,

I am so glad I finished the annoying job

and

I finished the annoying job

represent the same emotion – Happy. However, the word “glad” does not appear in the latter sentence. To solve this problem, the SLRM [11] was employed for emotional state recognition based on textual information. First, the training data are divided into four categories – happy, angry, sad, and neutral. Then, for each emotional state, synonym mapping is employed to generate the synonyms of the recognized words using the Chinese knowledge base, HowNet [12]. Three categories of semantic labels (SL), including specific SLs, negative SLs, and disjunctive SLs, are extracted. Examples of the synonym set with the semantic labels are given in Table 1. Most of the specific SLs are verb and can express the speaker’s emotional state without emotional keyword such as “glad”. For two other categories of SLs, negative SLs are used to represent negative semantics. The atomic sentences before/after disjunctive SLs will be retained for further process.

I earn a lot of money

belonging to happy emotion will be converted into the semantic labels “OBTAIN” and “MONEY” as earn is the synonym set with the semantic label of “OBTAIN” and “MONEY” being the synonym set with the semantic label of MONEY. If labels “OBTAIN” and “MONEY” form a frequent pair, this pair is an EAR. Finally, the MaxEnt model is employed to classify the emotional state based on the semantic label sequence, emotional association rules (EAR) selected based on a priori algorithm [13] are utilized for frequent trigger pair selection. As an illustrative example, the input text

Table 1. Some Examples of Semantic Labels

<table>
<thead>
<tr>
<th>Specific SLs</th>
<th>Synset in HowNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACHIEVE</td>
<td>VACHIEVE, FULFILL, END, FINISH</td>
</tr>
<tr>
<td>OBTAIN</td>
<td>OWN, OBTAIN, RECIEVE, EARN</td>
</tr>
<tr>
<td>LOSE</td>
<td>OWNNOT, LOSE, INDEBT</td>
</tr>
<tr>
<td>Negative SLs</td>
<td>Synset in HowNet</td>
</tr>
<tr>
<td>NEGATIVE</td>
<td>CANNOT, UNNECESSARY, NEVER</td>
</tr>
<tr>
<td>Disjunctive SLs</td>
<td>Synset in HowNet</td>
</tr>
<tr>
<td>Disj Extract</td>
<td>FINALLY, BUT</td>
</tr>
<tr>
<td>Disj Remove</td>
<td>OTHERWISE, ALTHOUGH</td>
</tr>
</tbody>
</table>

\[ Z(S) = \sum_{i=1}^{m} \exp \left( \lambda_k f_i(e_m, S) \right) \]  

(2)

Finally, given a word sequence, each semantic label-based emotion recognizer will output its corresponding confidence score.

2.2. Prosodic Information-based Emotion Recognition

In PIER, a front-end procedure extracts the prosody-related features that are commonly used in emotion recognition. Specially, the maximum, minimum, mean, and standard deviation of the pitch, energy, formant-related features, as well as speech rate are extracted to represent the acoustic expression of the utterance. These extracted values constitute a feature vector representation for an utterance, which is denoted by \( F \).

For each emotional state \( e_m \), a corresponding support vector machine \( V_m, m=1, \ldots, M \) is used to separate the feature vectors of the target emotional-state instances from the feature vectors of the non-target emotional-state instances. Note that the SVMs are trained by maximizing the margin \( \gamma(F) \) of the vectors in a training data set. Furthermore, we use the Platt’s formula [14]

\[ f_{svm}(e, F) = \frac{1}{1 + \exp(\alpha \gamma(F) + \beta)} \]  

(3)

to convert the binary SVM output to data likelihood, where the parameters \( \alpha \) and \( \beta \) are determined by maximizing the data likelihood of the training set. Finally, given an instance, each SVM-based emotion recognizer will output its corresponding confidence scores.
2.3. Dialogue Topic Categorization

Most of the approaches to topic detection focused on the techniques of information retrieval and projected the documents into a semantic space to acquire meaningful information. Although the performances of the proposed approaches in topic detection task are unstable, we employ the LDA [15] approach to realize dialogue topic categorization. Moreover, the topics for the dialogues in the daily-life domain are hard to annotate so that the latent topic of LDA can be used to estimate the appropriate number of topics. Recently, LDA becomes an attractive approach to topic detection. LDA extends the probabilistic latent semantic analysis, or simply PLSA, by treating the latent topic of each document as a random variable. Typically, LDA is a generative probabilistic model for documents in a text corpus. The documents are represented by the random latent topics, which are characterized by the distributions over words. The LDA parameters consist of \( \{ \alpha, \beta \} \) where \( \alpha = [\alpha_1, \alpha_2, ..., \alpha_c] \) denotes the Dirichlet parameters of \( C \) latent topic mixtures, and \( \beta \) is a matrix with multinomial entry \( \beta_{nc} = P(w_c | \theta) \). Using LDA, the probability of an \( N \)-word document \( w = [w_1, w_2, ..., w_N] \) is calculated by the following procedure. First, a topic mixture vector \( \theta \) is drawn from the Dirichlet distribution with parameter \( \alpha \). The corresponding topic sequence \( T = [T_1, T_2, ..., T_N] \) is generated based on the multinomial distribution with parameter \( \theta \) at document level. Each word \( w_n \) is generated by the distribution \( P(w_n | c_n, \beta) \). The joint probability of parameter \( \theta \), topic assignment \( c \) and document \( w \) is given by

\[
p(\theta, T, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^{N} p(T_n | \theta) p(w_n | T_n, \beta)
\]

By integrating (4) over \( \theta \) and summing up the probabilities over \( c \), we obtain the marginal probability of document \( w \) by

\[
p(w | \alpha, \beta) = \int p(\theta | \alpha) \left( \sum_{c} \prod_{n=1}^{N} p(T_n | \theta) p(w_n | T_n, \beta) \right) d\theta
\]

Hence, the key words of each topic can be obtained using an unsupervised approach. Then, the naïve Bayes is employed to estimate the topic-level score. Finally, given a word sequence, \( N \) topic-level scores are obtained.

2.4. ANN-based Interactional Styles Detector

In this work, the multi-layer perceptron (MLP), an important class of ANN, is utilized to integrate the output of SLER, PIER, and DTCM for IS detection. The input layer has \( D = 2M + N \) nodes for the output values from SLER, PIER, and DTCM in which \( M \) denotes the number of emotional state and \( N \) represents the number of topics. There are \( J \) nodes in the output layer, each representing an interactional style. The number of nodes \( L \) in the hidden layer is varied according to different experimental setups.

Denoting the input values by

\[ I = I_1, ..., I_D, \]

the output value of the \( l \)-th hidden node is

\[ H_l = (1 + \exp \left( \sum_{a} \eta_a I_a \right))^{-1}, \quad l = 1, ..., L \]

Similarly, the output value of the \( j \)-th output node is

\[ O_j = (1 + \exp \sum_{l} \eta_l H_l)^{-1}, \quad j = 1, ..., J \]

Note that \( \eta_a \) is the weight of the link between the \( d \)-th input node and the \( l \)-th hidden node; \( \eta_l \) is the weight between the \( d \)-th hidden node and the \( j \)-th output node. These weights are trained by the well-known backpropagation algorithm. Although there are four output nodes for four ISs, only the IS with highest score will be selected as the detected IS.

3. Evaluations and Discussion

To evaluate the proposed method for interactional style detection for versatile response selection in a spoken dialogue system, we adopted the commonly-used Wizard-of-Oz approach [16] to collect the conversation corpus in the domain of daily life chatting. Berens' four interactional styles--incharge (IC), chart-the-course (CTC), get-things-going (GTG), and behind-the-scene (BTS) were used in this work. This corpus collected the speech utterances from seven male subjects in a lab environment using 16KHz sampling rate and 16-bit PCM wave format. Totally, 3,731 utterances were collected. In addition, each sentence was manually annotated by the subject who provided the utterances. In order to consider the difference between perceived and intended emotions, one annotator other than the speaker who provided the utterances also annotated the evaluation data. If different annotation occurred for the same utterance, they will discuss and determine the final annotation. For the annotation of interactional style, only the annotator other than the speakers will tag the interactional style because the interactional style is the listeners’ speech perception.

The software Praat [17] was utilized to extract the prosodic features. For PIER, each SVM for the \( m \)-th emotional state is trained to discriminate a specific emotion from the others using the open source-LIBSVM. For SERM, 15 specific SLs comprise totally 147 synonyms; one negative SL contains 47 Chinese negatives and two disjunctive SLs consist of 29 words. GibbsLDA++ [18] is employed for LDA training. An MLP with one hidden layer and four output nodes, each representing one interactional style, was constructed. Additionally, Mandarin speech recognition process was done using the HTK [19], which is commonly used in research and development of a speech recognition system. For speech features, 39 dimensions were used, including 12 dimensions of Mel-frequency cepstral coefficients (MFCCs), one dimension of log energy, and their delta and acceleration features. In total, the acoustic models are composed of 153 subsyllable and 37 particle models (e.g., EN, MA, OU) based on Hidden Markov Model (HMM) with 32 mixtures for each state. The average word accuracy of the ASR module is 86.1% with a lexicon of 554 words. \( K \)-fold (\( K = 5 \)) cross validation method was employed to evaluate the proposed approach; that is, 80% data were employed for model training and the remaining 20% data for testing in each iteration.

Table 2 shows the average recognition accuracy of PIER and SERM. Compared to SLER, the column captioned “Text” shows the performance of emotion recognition without the semantic labels and EARs. The result reveals that the SLER is beneficial to IS detection. Table 3 demonstrates the number of utterance of each interaction style in each emotional state. In the evaluation corpus, there is high correlation between interactional style and emotional state. Especially, the speakers are often in angry when they talked about the political topic.

Table 4 shows the performance of topic detection with different number of latent topics. According to the results, eight latent topics can achieve the best performance of topic detection. Table 5 illustrates the example words of some latent topics in eight topics (\( N = 8 \)). One can observe that the topics can be defined by these keywords. For example, the columns captioned as “Topic 1” and “Topic 4” denote the dialogue topics “Politics” and “Food”, respectively. This result shows LDA is useful to this work. Finally, the evaluation of ANN
Table 2. Average recognition accuracy (ARA) of emotion recognition

<table>
<thead>
<tr>
<th>Approaches</th>
<th>PIER</th>
<th>SLER</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARA(%)</td>
<td>78.16</td>
<td>80.92</td>
<td>60.74</td>
</tr>
</tbody>
</table>

Table 3. Number of utterance of each IS in each emotional state.

<table>
<thead>
<tr>
<th></th>
<th>IC</th>
<th>CTC</th>
<th>GTG</th>
<th>BTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>795</td>
<td>63</td>
<td>23</td>
<td>7</td>
</tr>
<tr>
<td>Happy</td>
<td>31</td>
<td>3</td>
<td>21</td>
<td>371</td>
</tr>
<tr>
<td>Neutral</td>
<td>126</td>
<td>132</td>
<td>1169</td>
<td>679</td>
</tr>
<tr>
<td>Sad</td>
<td>0</td>
<td>490</td>
<td>16</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 4. Recognition performance of topic detection with different number of latent topics

<table>
<thead>
<tr>
<th>#(Topic)</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Rate(%)</td>
<td>78.97</td>
<td>79.25</td>
<td>81.23</td>
<td>76.82</td>
</tr>
</tbody>
</table>

Table 5. Example words of some latent topics in eight topics ($N=8$).

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politician</td>
<td>Everyone</td>
<td>Now</td>
<td>McDonald</td>
</tr>
<tr>
<td>President</td>
<td>Smart</td>
<td>Today</td>
<td>KFC</td>
</tr>
<tr>
<td>Lawmaker</td>
<td>Problem</td>
<td>Weekend</td>
<td>Buffet</td>
</tr>
<tr>
<td>Budget</td>
<td>Gossip</td>
<td>Male</td>
<td>Steak</td>
</tr>
<tr>
<td>Weapon</td>
<td>Know</td>
<td>Female</td>
<td>Vegetable</td>
</tr>
</tbody>
</table>

Figure 2: MLP-based interactional style detection with different number of hidden nodes.

Based IS detection with various numbers of hidden nodes is shown in Figure 2. The best one can achieve 82.67% detection accuracy with eight topics and twenty hidden nodes. However, the results with fifteen hidden nodes can obtain acceptable performance for different number of latent topics.

4. Conclusions

This work presents an approach to interactional style detection using prosodic features and textual information. In order to deal with the synonym problem, four categories of semantic labels are extracted from HowNet and conversational data. Then, MaxEnt is employed to model the relationship between emotional states and the semantic label sequence derived from the word sequence. For prosodic information, SVM is utilized to train a prosodic information-based emotion recognizer. Besides emotion, words are also the key role in IS detection. Hence, LDA is employed to obtain the keywords in the latent topics. Finally, an artificial neural network is adopted for IS detection considering the scores estimated from the above classifiers. In order to evaluate this work, more than 1800 sentences in the domain of daily life chatting were collected. Finally, the proposed approach can achieve 82.67% recognition accuracy with twenty hidden nodes used in the MLP-based IS detector. Moreover, the prosodic and semantic information based emotion recognizer can achieve 80.92% and 78.16% accuracy, respectively. Eight topics could obtain the best topic detection result. In our future work, although the proposed approach can work well on IS detection, more training data are needed for fair evaluation.

5. References