Multi-Task Deep Learning for User Intention Understanding in Speech Interaction Systems

Yishuang Ning,[1,2] **Jia Jia,**[1,2] **Zhiyong Wu,**[1,2,3,*] **Runnan Li,**[1,2]
Yongsheng An,[1] **Yanfeng Wang,**[4] **Helen Meng**[3]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Tsinghua National Laboratory for Information Science and Technology (TNList)
[2]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[3]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, China
[4]Beijing Sougou Science and Technology Development Co., Ltd.
{ningys13,lirn15}@mails.tsinghua.edu.cn, jjia@mail.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn

Abstract

Speech interaction systems have been gaining popularity in recent years. The main purpose of these systems is to generate more satisfactory responses according to users' speech utterances, in which the most critical problem is to analyze user intention. Researches show that user intention conveyed through speech is not only expressed by content, but also closely related with users' speaking manners (e.g. with or without acoustic emphasis). How to incorporate these heterogeneous attributes to infer user intention remains an open problem. In this paper, we define **Intention Prominence (IP)** as the semantic combination of focus by text and emphasis by speech, and propose a multi-task deep learning framework to predict IP. Specifically, we first use long short-term memory (LSTM) which is capable of modeling long short-term contextual dependencies to detect focus and emphasis, and incorporate the tasks for focus and emphasis detection with multi-task learning (MTL) to reinforce the performance of each other. We then employ Bayesian network (BN) to incorporate multimodal features (focus, emphasis, and location reflecting users' dialect conventions) to predict IP based on feature correlations. Experiments on a data set of 135,566 utterances collected from real-world Sogou Voice Assistant illustrate that our method can outperform the comparison methods over 6.9-24.5% in terms of F1-measure. Moreover, a real practice in the Sogou Voice Assistant indicates that our method can improve the performance on user intention understanding by 7%.

Introduction

Recently speech interaction systems such as Apple Siri[1], Google Now[2] and Sogou Voice Assistant[3] have become widespread in all segments of society. Statistics from Microsoft indicate that more than 70% of the subscribers use these systems more than once a week[4]. In these systems, the key point to determine quality of service (QoS) and user experience is to understand user intention accurately and provide more satisfactory responses (Chen 2004). At present, these systems mainly generate responses by text-based natural language processing (NLP) (Bellegarda 2013). The focus determined by keywords of text is among the main features in most studies to understand user intention (Duan et al. 2008). However, the utterance with the same texts may carry different user intentions. As the example shows in Figure 1, two users' speech utterances have different acoustic emphasis but have the same texts. For case 1, the user emphasizes the word 'iPhone 7', the corresponding intention is 'I want to buy iPhone 7 at Xidan, but stores at other places are OK if iPhone 7 is not on sale at Xidan'. For case 2, the user expresses the intention 'I want to buy iPhone 7 at Xidan, but other mobile phones are OK if iPhone 7 is not on sale' by emphasizing the word 'Xidan'. For speech interaction systems, the incorporation of acoustic emphasis can provide more accurate understanding of the user's underlying intention, thus providing more related responses. This semantic supplementation phenomenon expressed by emphasis has also been reported in (Chakravartty 2001), which inspires us to consider whether we can combine focus with emphasis to improve the accuracy of user intention understanding.

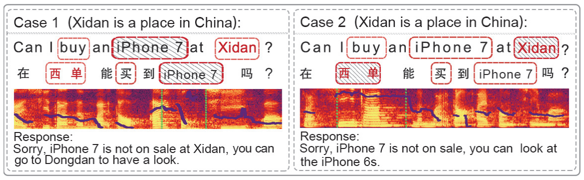
Motivated by this, in this paper, we define **Intention Prominence (IP)** as the semantic combination of focus by text and emphasis by speech that determines the underlying intention within an utterance. However, integrating focus and emphasis is a non-trivial task. To incorporate these heterogeneous attributes for user intention understanding, we propose a multi-task deep learning framework to predict IP from speech, as illustrated in Figure 2. In particular, we use long short-term memory (LSTM) to model the textual and acoustic information for focus and emphasis detection, respectively. LSTM has achieved state-of-the-art re-

[1] http://www.apple.com/ios/siri

[2] http://www.google.com/landing/now

[3] http://yy.sogou.com

[4] http://www.wtoutiao.com/p/ob5qG4.html



Case 1 (Xidan is a place in China):

Can I buy an iPhone 7 at Xidan ?

在 西 单 能 买 到 iPhone 7 吗 ?

Response:
Sorry, iPhone 7 is not on sale at Xidan, you can go to Dongdan to have a look.

Case 2 (Xidan is a place in China):

Can I buy an iPhone 7 at Xidan ?

在 西 单 能 买 到 iPhone 7 吗 ?

Response:
Sorry, iPhone 7 is not on sale, you can look at the iPhone 6s.

Figure 1: Influence of emphasis on user intention understanding leading to different responses. In this figure, the words in blank and shadowed boxes are focus and emphasis, respectively, and the blue curve in the spectrogram plots the pitch contour of each utterance.

sults in various sequence processing tasks given its capability in leveraging long short-term contextual dependencies. Besides, our observation finds that focus and emphasis are consistent to a large degree, which indicates the two modalities may share some latent patterns. To leverage the shared patterns for modeling, we use multi-task learning (MTL) to reinforce the performance of focus and emphasis detection for each other. Furthermore, in human-mobile interaction scenarios, there are tremendous amounts of users. Due to the divergence of language habits, they may use different speaking ways to express intentions for the same thing. Motivated by this, we further incorporate location information that reflects users' dialect conventions with focus and emphasis. By virtue of Bayesian network (BN), it can take into account feature dependencies to improve the accuracy of IP prediction. Finally, we propose an unsupervised method using sparse auto-encoder (SAE)[5] to obtain the robust representations of input data before we fed the input to the LSTM hidden layers.

Experiments on a data set of 135,566 Mandarin utterances collected from real-world Sogou Voice Assistant illustrate that our proposed method can significantly improve the F1-measure for intention prominence prediction by 6.9-24.5% over the comparison methods. The feature contribution analysis also shows that besides focus by text, the emphasis by speech can achieve 2.8% improvement while the location can further enhance the performance by 3%.

Related Work

The use of textual information for user intention understanding has become a hot topic in natural language understanding (NLU) field. Existing methods mainly focus on classifying user intention into limited discrete categories. For example, (Deng et al. 2012) classified user intention into several semantic categories (Find Flight, Show Weather, etc.) and used kernel deep convex networks for spoken language understanding. (Shen et al. 2011) used sparse hidden-dynamics conditional random fields (SHDCRFs) to predict user intention based on users' dynamic actions. (Wang et al. 2015) proposed a graph-based semi-supervised approach to infer intent categories for tweets.

There are also some works focusing on inferring user intention from acoustic information. For instance, (Savino and Refice 2000) proposed to utilize acoustic cues such as F0 shape and duration to classify communicative intentions in dialog systems. (Matsubara et al. 2002) presented an example-based method for inferring speaker's intention. (Irie et al. 2004) used a decision tree learning method to understand speech intention based on a spoken dialog corpus.

However, there never exists methods to incorporate the two modalities. Different from various previous researches, our proposed method mainly dedicates to integrating the acoustic information with textual information for intention prominence prediction to generate more satisfactory responses in real-world speech interaction systems.

Problem Formulation

In this section, we first give several necessary definitions and then formulate the problem of predicting intention prominence within users' utterances. Given a set of utterances $\mathbf{U}$, for an utterance $\mathbf{u}_i \in \mathbf{U}$, we denote $\mathbf{u}_i = \{\mathbf{a}_i, \mathbf{t}_i, \mathbf{l}_i\}$ where $\mathbf{a}_i = \{a_i^1, a_i^2, ..., a_i^{N_i}\}$ and $\mathbf{t}_i = \{t_i^1, t_i^2, ..., t_i^{N_i}\}$ are the set of acoustic and textual features extracted at syllable level respectively, and N_i is the number of syllables of $\mathbf{u}_i$. $\mathbf{l}_i$ is user's location information by city (e.g. Beijing, Shanghai) infered from the GPS information. Specifically, for each syllable u_i^j of $\mathbf{u}_i$ ($1 \leq j \leq N_i$), we use a D_p dimensional vector $a_i^j = \langle a_{i1}^j, a_{i2}^j, ..., a_{iD_p}^j \rangle$ to indicate u_i^j's acoustic features (e.g. logarithmic F0, duration, energy and their statistic features), and a D_q dimensional vector $t_i^j = \langle t_{i1}^j, t_{i2}^j, ..., t_{iD_q}^j \rangle$ to indicate u_i^j's textual features (e.g. pronunciation features, prosodic features).

Definition 1. **Focus.** We adopt textual features to detect focus. Following previous works, we define focus of an utterance as a set of prosodic words or phrases that encode pragmatically and semantically salient information (Zhang et al. 2006). Generally, whether the current syllable of $\mathbf{u}_i$ is focus can be viewed as a binary classification problem. Therefore, we denote the focus categories as $\mathbf{F}_c$={*focal*,*neutral*}.

Definition 2. **Emphasis.** We utilize acoustic features to detect emphasis. In previous researches, emphasis is represented as prominence of part of the words, phrases or even sentences. In (Tamburini 2003), according to the salient level of the current syllable, the emphasis of Chinese words can be divided into emphatic and neutral. Hence, the emphasis categories can be denoted as $\mathbf{E}_c$={*emphatic*,*neutral*}.

Definition 3. **Intention prominence.** In this paper, we define intention prominence as the combinaiton of focus and emphasis that carries the core information and specifies user intention within the speech. Actually, whether a syllable of $\mathbf{u}_i$ is intention prominence can be considered as a classification problem. Thus, we denote the intention prominence categories as $\mathbf{I}_c$={*prominent*,*neutral*}. In addition, we define the intention prominence of $\mathbf{u}_i$ as $\mathbf{D}_i$.

Problem. **Labeling each syllable of the utterances with intention prominence.** The proposed multi-task deep learning framework is implemented through 2 steps. 1) Learning a focus and emphasis detection model P: $\mathbf{U} = \{\mathbf{u}_i\}_i = \{\mathbf{a}_i, \mathbf{t}_i, \mathbf{l}_i\}_i \rightarrow \{P(f_c|t_i^j), P(e_c|a_i^j)\}_i^j$, where $P(f_c|t_i^j)$ and $P(e_c|a_i^j)$ are the focus and emphasis probability of u_i^j.

[5] http://ufldl.stanford.edu/wiki/index.php

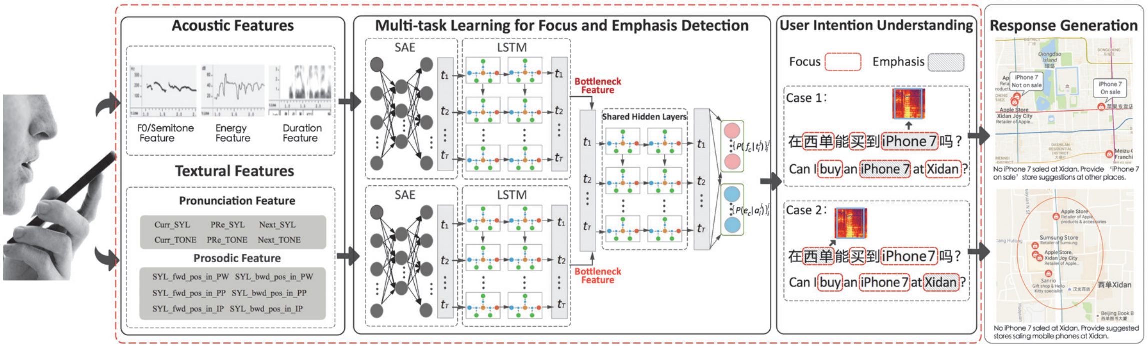

Acoustic Features
F0/Semitone Feature
Energy Feature
Duration Feature
Textural Features
Pronunciation Feature
Curr_SYL PRe_SYL Next_SYL
Curr_TONE PRe_TONE Next_TONE
Prosodic Feature
SYL_fwd_pos_in_PW SYL_bwd_pos_in_PW
SYL_fwd_pos_in_PP SYL_bwd_pos_in_PP
SYL_fwd_pos_in_IP SYL_bwd_pos_in_IP
Multi-task Learning for Focus and Emphasis Detection
SAE
LSTM
Bottleneck Feature
Shared Hidden Layers
t_1 t_2 t_T
$\{P(f_c|t_i^j)\}_i^j$
$\{P(e_c|a_i^j)\}_i^j$
User Intention Understanding
Focus
Emphasis
Case 1:
在西单能买到iPhone 7吗？
Can I buy an iPhone 7 at Xidan?
Case 2:
在西单能买到iPhone7吗？
Can I buy an iPhone 7 at Xidan?
Response Generation
iPhone 7 Not on sale
iPhone 7 On sale
Apple Store Xidan Joy City
No iPhone 7 saled at Xidan. Provide 'iPhone 7 on sale' store suggestions at other places.
Apple Store
Samsung Store
Apple Store, Xidan Joy City
Sanrio
西单Xidan
No iPhone 7 saled at Xidan. Provide suggested stores saling mobile phones at Xidan.

Figure 2: Overview of our method for user intention understanding which mainly focuses on the parts in red box.

2) Learning a intention prominence prediction model M: $\{P(f_c|t_i^j), P(e_c|a_i^j), \mathbf{l}_i\}_i^j \rightarrow \{\mathbf{D}_i\}_i$, where $f_c \in \mathbf{F}_c, e_c \in \mathbf{E}_c, \mathbf{D}_i^j \in \mathbf{I}_c$.

Data Observation

The definition of intention prominence indicates that the performance of intention prominence prediction largely depends on the accuracy of focus and emphasis detection. How can we improve the accuracy of focus and emphasis detection? In this work, we conducted a data observation to reveal the implicit relation between focus and emphasis.

Data set. The raw data set includes 135,566 Mandarin utterances provided by Sogou Voice Assistant. Each utterance is assigned with its raw speech, the corresponding speech-to-text information with a word error rate (WER) of 5.5% and user's location (city only). We randomly chose 2,000 utterances, and invited 3 human labelers to mark the focus, emphasis and intention prominence of each utterance at syllable level. Specifically, focus is labeled from the key concept of the utterance while emphasis is labeled according to auditory perception (e.g. higher F0 or energy, longer duration, etc). Then intention prominence is labeled as the combination of focus and emphasis of the corresponding utterance. When the labelers had different opinions, they stopped and had a discussion until they reached an agreement.

Patterns. To investigate the relation between focus and emphasis, we first analyzed the distribution of focus and emphasis. Specifically, there are 1.2 foci in each utterance on average, and 57% of the utterances have emphasis.

We further concluded the relation patterns between focus and emphasis in the same utterance and summarized the following five typical patterns in Figure 3.

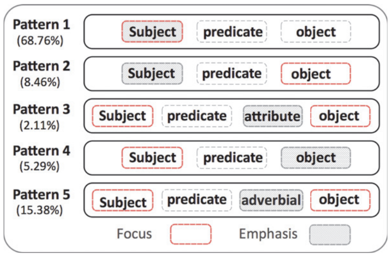


Figure 3: Relation patterns between focus and emphasis.

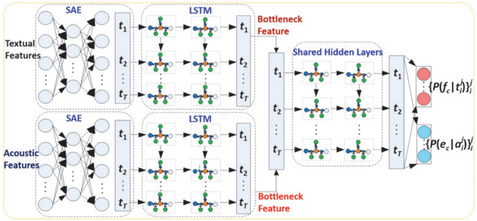


Figure 4: The framework of multi-task learning (MTL) for focus and emphasis detection.

Pattern 1. Emphasis and focus are consistent. The proportion of this pattern is 68.76%.

Pattern 2. The subject of the sentence will obtain emphasis while focus is on the object. The proportion of this pattern is 8.46%.

Pattern 3. The attribute before the object will obtain emphasis while focus is on the object. The proportion of this pattern is 2.11%.

Pattern 4. The object will obtain emphasis while focus is on the subject. The proportion of this pattern is 5.29%.

Pattern 5. The focus sensitive operators (the adverbs or modal verbs) will obtain emphasis while focus connected with them will not. The proportion of this pattern is 15.38%.

Summarization. From Figure 3, we can see that although focus and emphasis come from text and acoustic information respectively, the positions of them in utterances are consistent to a large degree. Besides, focus and emphasis usually present a certain modified relation in utterances. **These findings inspire us to introduce multi-task strategy to better capture the relations between focus and emphasis for intention prominence prediction.**

Our Approach

Based on the findings, we formulate the problem of intention prominence prediction in a multi-modal multi-task model. To compare with previous works, our technique contributions are in three aspects: 1) different from (Deng et al. 2012;

Wang et al. 2015) which classify user intention as discrete categories, we focus on combining focus with emphasis to generate more related responses in speech interaction systems, and propose a unified framework for user intention understanding. 2) different from (Ning et al. 2015), for the emphasis detection task, we utilize LSTM to model contextual dependencies which emphasis detection mainly relies on, and use bottleneck features to yield the compact representations; 3) to leverage the large-scale of unlabeled data, we adopt SAE to obtain robust representations of the input.

Multi-task Deep Learning for Focus and Emphasis Detection

Figure 4 shows the framework of multi-task deep learning for focus and emphasis detection. In this figure, SAE is used to generate the robust representations for input textual or acoustic features. LSTM is used to extract the bottleneck features which are closely related with focus and emphasis. Then the fusion of textual and acoustic bottleneck features are fed to shared LSTM hidden layers with MTL for focus and emphasis detection to derive the probability of focus and emphasis for each syllable.

Learning Robust Representations of Input. To leverage the large-scale of unlabeled data, we propose an unsupervised model to learn the robust representations of the input features with SAE. The advantage of SAE is that even when the number of hidden units is large (perhaps even greater than the dimension of input data), we can still discover a robust structure, by imposing other constraints on the network.

The objective of SAE is to learn a function $h_{W',b'} \approx x'$, so as to output $\hat{x}'$ that is similar to x' . Given the activation $a_j^{(2)}(x')$ of the hidden unit j when the network is given a specific input x', the objective function of SAE is denoted:

$$J_{sparse}(W',b') = J(W',b') + \beta \sum_{j=1}^{s_2} KL(\rho \| \hat{\rho}_j) \tag{1}$$

with

$$J(W',b') = \Big[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\big(\parallel h_{W',b'}(x'^{(i)}) - x'^{(i)} \parallel \big)\Big] + \frac{\lambda}{2} \parallel \xi \parallel^2 \tag{2}$$

where Eqn. (2) is the loss function of auto-encoder (AE) without sparsity constraints. In this equation, $\parallel \xi \parallel$ is the L2-norm of all weight matrices of AE, m is the number of training samples and λ is a weight decay parameter; The second term in Eqn. (1) is the Kullback-Leibler (KL) divergence between two Bernoulli random variables with mean ρ and $\hat{\rho}_j$ respectively, and ρ is a sparsity parameter (typically a small value close to 0, in our case, we set it to 0.05). s_2 is the number of hidden units, and β controls the weight of the sparsity penalty term. To train SAE, we use the back propagation (BP) algorithm to minimize equation (1). After training this model, we use the output of the hidden layer of SAE as the input of the LSTM.

Extracting Task Related Compact Features. For the same prosodic words, when they are in different contexts, the probabilities to become focus or emphasis are different. That means focus or emphasis is closely related with its contexts. However, traditional contextual models such as conditional random fields (CRFs) and window based methods can only model dependencies with limited and fixed time ranges. To address this problem, (Graves 2012) proposed LSTM to bridge long time lags and this approach has shown state-of-the-art performance in many classification tasks (Sutskever et al. 2015). Motivated by this, we adopt the LSTM framework to model textual and acoustic features for extracting bottleneck features related to focus and emphasis.

Given an input sequence $\mathbf{x} = (x_1, x_2, ..., x_T)$ ($\mathbf{x}$ could be the robust representations of textual or acoustic feature vectors), this model computes the hidden vector sequence $\mathbf{h} = (h_1, h_2, ..., h_T)$ and outputs vector sequence $\mathbf{y} = (y_1, y_2, ..., y_T)$ by iterating the following equations from $t = 1$ to T:

$$h_t = \phi(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{3}$$

$$y_t = W_{hy}h_t + b_y \tag{4}$$

where W_{xh}, W_{hh} and W_{hy} are the input-hidden, hidden-hidden and hidden-output weight matrices. b_h and b_y are the hidden and output bias vectors. $\phi(\cdot)$ is the activation function and can be implemented by LSTM block with equations in (Hochreiter and Schmidhuber 1997; Graves 2012).

To yield the compact representations of textual and acoustic information for each syllable, we use a bottleneck layer which has relatively smaller number of hidden units compared with the other hidden layers in the network.

Modeling Correlations with Multi-task Learning. Based on our findings in the previous section, we find that emphasis and focus are highly related with each other. This finding inspires us to incorporate focus and emphasis detection in the MTL framework. With MTL, the two tasks can share what they learn (the probability of focus and emphasis) to reinforce the detection performance of each other.

To learn the parameters of our model, we use the mini-batch-based adaptive gradient algorithm. In each iteration, a task t is selected randomly, and the model is updated according to the task-specific objective. This is actually to minimize the sum of two single-task objectives.

Intention Prominence Prediction

As described in the previous section, the task for intention prominence prediction can be viewed as a classification problem. The purpose is to derive the class labels given a set of feature variables $\mathbf{s} = \{P(f_c|t_i^j), P(e_c|a_i^j), \mathbf{l}_i\}_i^j$ for each syllable of the utterances where $P(f_c|t_i^j)$ and $P(e_c|a_i^j)$ are the probability of focus and emphasis for each syllable respectively, and $\mathbf{l}_i$ is the location information (i and j are the utterance and syllable identity respectively). Previous research indicates a more accurate modeling of the feature dependencies leads to improved classification accuracy (Friedman et al. 1997). This means the performance of a classifier may be improved if the learning procedure takes into account the correlations between features. Although LSTM can model the long short-term contextual dependencies, it cannot handle the feature correlations. As for this problem, Bayesian network (BN) is capable to learn the dependencies between different features. Therefore, we use this model for intention prominence prediction.

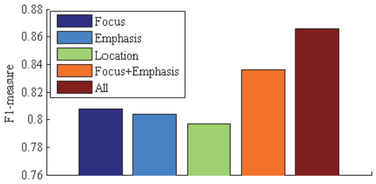


Figure 5: Comparison of results using different modalities.

By using the MTL framework, the learned probabilities of focus and emphasis have revealed the feature correlations. Besides, both of them are also correlated with geographic features. To derive the class label for each syllable, we use a feature vector $V = \{v_1, v_2, ..., v_T\}$. In the BN for our problem, each feature of V is a node and edges between pairs of nodes represent direct correlations between features. For example, an edge from v_i to v_j implies v_i is the parent of v_j in the BN, and the influence of v_j on the assessment of the class variable also depends on the value of v_i. Then the joint probability distribution (JPD) can be calculated as $P(V) = \Pi_{v \in V} p(v|pa(v))$ where $pa(v)$ is the parent of v. Finally, the structure can be learned with the BN search algorithm. In this paper, we use the tree augmented naive Bayes (TAN) search algorithm which attempts to maximize the Bayesian score metrics to learn the network structure.

To obtain the intention prominence of $\mathbf{u}_i$, we simply calculate $argmax_{\mathbf{D}_i^j} P(\mathbf{D}_i^j|\mathbf{s})(1 \leq j \leq N_i)$ using $P(V)$ with:

$$P(\mathbf{D}_i^j|\mathbf{s}) = P(V)/P(\mathbf{s}) \propto P(V) = \prod_{v \in V} p(v|pa(v)) \qquad (5)$$

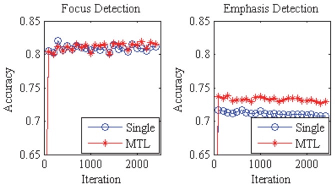


Figure 6: Comparison of results using single task and multi-task learning (MTL) for focus and emphasis detection.

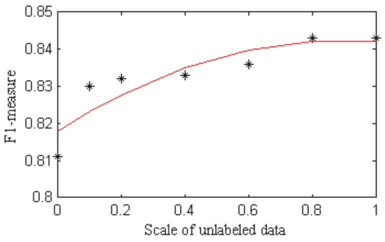


Figure 7: Influence of unlabeled data with different scales.

Experiments

In this section, we present extensive of experimental results to evaluate the effectiveness and efficiency of our method.

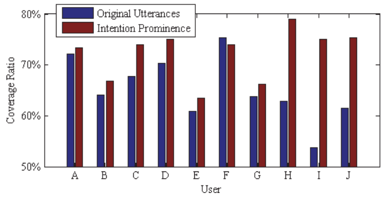


Figure 8: Differences of intention understanding among different people.

Experimental setup

Data set. We established a benchmark data set from a real-world speech interaction system, Sogou Voice Assistant, for predicting intention prominence. The raw data set includes 135,566 Mandarin utterances recorded in 2013.

Due to the massive scale of our data set, manually labeling the focus, emphasis and intention prominence is not only label-intensive, but also time-consuming. Hence, we randomly selected 2,000 utterances from the data set and labeled the focus, emphasis and intention prominence for each utterance (the detailed labeling method is introduced in the previous section). Beside the labeled data, the large scale of unlabeled data are also adopted to train SAEs to get the robust representations of the input.

Features. For focus detection, 48 dimensional textual features, including 6 dimensional pronunciation features and 42 dimensional prosodic features that are broadly adopted for Mandarin speech synthesis (Kang 2010) are used. The pronunciation features include the current, the previous, the next syllable and their tones. The prosodic units are from five levels: syllable, prosodic word, prosodic phrase, intonational phrase and sentence. These features can be divided into two types: one type of features are used to describe the forward or backward position of prosodic units (e.g. SYL_fwd_pos_in_PW is the forward position of the current syllable in the prosodic word), the other to denote the length of the subunits that the current, the previous, the next units contain (e.g. SYL_num_in_PW is the number of syllables in the current prosodic word).

As for emphasis detection, according to previous works, we find that emphasis usually has higher F0, longer duration and higher energy (Tamburini 2003). Besides, research shows the change of semitone is consistent with the distance of auditory perception (Zhao 2011). This indicates emphasis may be also closely related with semitone. Thus, we also choose semitone as one feature. Therefore, the features for emphasis detection are summarized as F0 related features (the mean, minimum, maximum and range of log F0), energy related features (mean, minimum and maximum values), duration as well as semitone feature.

Finally, we use focus by text and emphasis by speech together with location information to predict intention prominence of each utterance.

Comparison methods. To evaluate the effectiveness of our proposed method LSTM+BN, we compare the performance of focus and emphasis detection as well as intention prominence prediction with some well-known methods, in-

Table 1: Comparison of results using different models.

Models	Focus			Emphasis			Intention Prominence		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure	Precision	Recall	F1-measure
SVM	0.390	0.608	0.475	0.308	0.009	0.017	0.627	0.618	0.621
BN	0.704	0.760	0.731	0.462	0.272	0.343	0.797	0.789	0.791
CRF	0.724	0.755	0.739	0.457	0.036	0.066	0.769	0.754	0.761
LSTM	**0.763**	0.755	**0.759**	0.605	**0.568**	**0.575**	0.792	0.803	0.797
LSTM+BN	-	-	-	-	-	-	**0.868**	**0.865**	**0.866**

cluding SVM (Chang and Lin 2011), BN (Ning et al. 2015), CRF (Zang et al. 2014) and LSTM (Graves 2012).

Evaluation metrics. In all the experiments, we evaluate the performance in terms of Precision, Recall and F1-measure (Powers 2011). All the results reported in this paper were based on 5-fold cross validation.

Experimental Results

Performance. Table 1 lists the prediction performance on intention prominence by comparison methods. From the results, we can draw the following conclusions. 1) Using LSTM only can achieve good performance in terms of F1-measure compared with other machine learning methods such as SVM, BN and CRF, indicating the contextual dependencies are important; compared with CRF, LSTM can better leverage the contextual dependencies for modeling. 2) When LSTM is combined with BN, the performance has improved significantly (6.9-24.5% in terms of F1-measure), indicating the feature dependencies cannot be ignored as well. 3) The results by SVM are worse than other methods. One of the reason might be due to the poor capability of handling the imbalanced distribution between focus or emphasis and neutral segments while other methods are less affected by this factor. The experimental results demonstrate the contextual dependencies and feature dependencies are both important, and our model can better capture these dependencies to enhance the performance. Next we will further analyse the experimental results from the following two aspects.

1) Modality contribution analysis: To evaluate the contribution of different modalities (focus by text, emphasis by speech and location), we conduct a series of experiments with different combinations of modalities. Experimental results are illustrated in Figure 5. As can be seen that the performance of using focus or emphasis is higher than using location, indicating that when predicting intention prominence, focus by text or emphasis by speech may be more crucial than location. Besides, when we use the combination of focus and emphasis for prediction, it achieves much higher performance than use focus only, and emphasis can enhance the performance by 2.8%. Moreover, when the location factor is considered by combining information from all modalities, the performance is further improved by 3%. These results validate the necessity and effectiveness of taking focus, emphasis and location for consideration.

2) Multi-task learning (MTL) effect: Figure 6 shows the comparison results of using MTL and single task for focus and emphasis detection respectively. From this figure, we can see that when MTL is adopted, the performance of both tasks are improved. One of the main reasons might be the distribution of focus and emphasis in our data set. The relation patterns between focus and emphasis indicates that they are highly related with each other, which verifies the rationality of using MTL. Besides, more improvement can be found for emphasis detection than focus detection. The reason might be that emphasis is more dependent on focus. Where there is a focus, there is more likely to be an emphasis near it, but not vice-versa.

Table 2: Comparison of CPU time for different models.

Models	SVM	BN	CRF	LSTM+BN
CPU time (s)	1.803 $\times 10^4$	6.8	2.229 $\times 10^2$	1.7520 $\times 10^4$

Table 3: Experimental results of the top-10 coverage ratio of the original utterances and intention prominence.

	Coverage Ratio	CI
Original Utterances	65.25%	[0.607,0.697]
Intention Prominence	72.25%	[0.687,0.758]

Scalability. To verify the effectiveness of large-scale unlabeled data, we use different scale of unlabeled data for preprocessing. In Figure 7, as the scale of unlabeled data increase, the performance gets better gradually. When it exceeds 80% of the total amount (about 100,000 utterances), the performance reaches convergence. Thus, we conduct experiments on a data set containing 100,000 utterances.

Efficiency. Besides effectiveness, we also consider their efficiency performance. Table 2 lists the training time of each model. The results show that our model is still reasonable (about 5 hours) while it achieves the best performance.

Practicability

Since it is very difficult to find an objective way to evaluate the whole results, we would like to present an interesting real practice to validate the effectiveness of our method. To demonstrate that our model has prominent effect in helping understanding user intention, we carefully designed 32 original utterances, each of which has a different structure. Then we used our method to predict the intention prominence of each utterance. Both the original utterances and the corresponding intention prominence were provided to 10 subjects. The subjects were asked to search on the Sogou search engine with them, respectively, and gave a top-10 coverage ratio (Inkpen 2007) for each utterance. The experimental results are shown in Table 3, where the confidence intervals at confidence level 0.95 ($\alpha = 0.95$) are also given. It can be learned from the table that the accuracy of using intention prominence has been improved significantly, which demonstrates that users commonly agree the search responses obtained by using our method is much better than using original utterances directly.

Besides, we also investigated the differences of intention understanding among different users. From Figure 8, we can see that for the same 32 utterances, the coverage ratios of responses are quite different both for the original utterances and intention prominence, demonstrating different users have different understandings towards the intention conveyed by the same utterance. However, 9 out of the 10 users agree that the cases using intention prominence are more consistent with the real intention understandings. This indicates the efficiency of the proposed method.

Conclusions

To address the challenges in speech interaction systems, considering the influence of speech information on interaction, in this paper, we defined Intention Prominence (IP) as the semantic combination of focus by text and emphasis by speech, and proposed a multi-task deep learning framework which integrated the textual and acoustic information to predict IP. Experiments on a real-world data set validate the contribution of focus by text and emphasis by speech, and demonstrate the effectiveness of our method. In future, we will mine more speech information to improve the harmoniousness of speech interaction systems.

Acknowledgments

This work is supported by the National Key Research and Development Plan (2016YFB1001200), National Basic Research Program of China (2013CB329304), National Natural Science Foundation of China (NSFC) (61370023, 61375027, 61433018) and joint fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N_CUHK404/15). This work is partially supported by the major project of the National Social Science Foundation of China(13&ZD189). We also thank 2016IM010200 for the support.

References

Chen, F. 2004. Speech interaction system C how to increase its usability. In *Proceedings of the Fifth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2229-2232.

Bellegarda, J. 2013. Spoken language understanding for natural interaction: The Siri Experience. *Natural Interaction with Robots, Knowbots and Smartphones*, 3-14.

Duan, H.; Cao, Y.; Lin, C. and Yu, Y. 2008. Searching questions by identifying question topic and question focus. In *Proceedings of ACL*, 156-164.

Chakravartty, A. 2001. The semantic of model-theoretic of theories and scientific realism. *Synthese* 127(3): 325-345.

Zhong, X. and Luv, S. 2003. Sentence prominences function of disambiguity for Chinese homophone. *Journal of Psychology* 35(3): 333-339.

Deng, L.; Tur, G.; He, X. and Hakkani-Tur, D. 2012. Use of kernel deep convex networks and end-to-end learning for spoken language understanding. In *Spoken Language Technology Workshop (SLT)*, 210-215.

Shen, Y.; Yan, J. and Yan, S. 2011. Sparse hidden-dynamics conditional random fields for user intention understanding. In *Proceedings of the Twentieth International Conference on World Wide Web (WWW)*, 7-16. ACM.

Wang, J.; Cong, G.; Zhao, W. and Li, X. 2015. Mining User Intents in Twitter: A Semi-Supervised Approach to Inferring Intent Categories for Tweets. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 318-324.

Savino, M. and Refice, M. 2000. Acoustic cues for classifying communicative intentions in dialogue systems. In *International Workshop on Text, Speech and Dialogue*, 421-426.

Matsubara, S.; Kimura, S.; Kawaguchi, N.; Yamaguchi, Y. and Inagaki, Y. 2002. Example-based speech intention understanding and its application to in-car spoken dialogue system. In *Proceedings of International Conference on Computational Linguistics (ICCL)*, 1-7. ACL.

Irie, Y.; Matsubara, S.; Kawaguchi, N.; Yamaguchi, Y. and Inagaki, Y. 2004. Speech intention understanding based on decision tree learning. In *Proceedings of the Fifth Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Tamburini, F. 2003. Prosodic prominence detection in speech. In *Proceedings of ISSPA*, 385-388.

Ning, Y.; Wu, Z.; Lou, X.; Meng, H.; Jia, J. and Cai, L. 2015. Using Tilt for Automatic Emphasis Detection with Bayesian Networks. In *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Hochreiter, S. and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8): 1735-1780.

Graves, A. 2012. Supervised sequence labelling with recurrent neural networks. *Springer*.

Sutskever, I.; Vinyals, O. and Le, Q. 2015. Sequence to sequence learning with neural networks. In *Proceedings of Neural Information Processing Systems (NIPS)*, 1-9.

Kang, S. 2010. Prosodic modeling in HMM-based speech synthesis. Ph.D. diss. *Tsinghua University*.

Zhao, J.; Yuan, H.; Liu, J. and Xia, S. 2011. Automatic lexical stress detection using acoustic features for computer-assisted language learning. In *Proceedings of the Asia-Pacific Signal and Information Processing Association (APSIPA)*.

Chang, C., and Lin, C. 2011. LIBSVM: A library for support vector machines. ACM TIST 2:27:1-27:27.

Zang, X.; Wu, Z.; Meng, H. M.; Jia, J. and Cai, L. 2014. Using conditional random fields to predict focus word pair in spontaneous spoken English. In *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Powers, D. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* 2(1): 37-63.

Inkpen, D. 2007. Information Retrieval on the Internet.