Pattern Recognition Letters 101 (2018) 1–5

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances

Minkyoung Kim, Harksoo Kim*

Program of Computer and Communications Engineering, College of IT, Kangwon National University, 1 Gangwondaehak-gil, Chuncheon-si, Gangwon-do 24341, Republic of Korea

ARTICLE INFO

Article history:
Received 23 June 2017
Available online 6 November 2017

Keywords:
Integrated intention identification model
Speech act identification
Predicator identification
Sentiment identification
Partial error backpropagation

ABSTRACT

A dialogue system should capture speakers' intentions, which can be represented by combinations of speech acts, predicators, and sentiments. To identify these intentions from speakers' utterances, many studies have independently dealt with speech acts, predicators, and sentiments. However, these three elements composing speakers' intentions are tightly associated with each other. To resolve this problem, we propose a convolutional neural network model that simultaneously identifies speech acts, predicators, and sentiments. The proposed model has well-designed hidden layers for embedding informative abstractions appropriate for speech act identification, predicator identification, and sentiment identification. Nodes in the hidden layers are partially trained by three cycles of error backpropagation: training the nodes associated with speech act identification, predicator identification, and sentiment identification. In the experiments, the proposed model showed higher F1-scores than independent models: 6.8% higher in speech act identification, 6.2% higher in predicator identification, and 4.9% higher in sentiment identification. Based on the experimental results, we conclude that the proposed integration architecture and partial error backpropagation can help to increase the performance of intention identification.

1. Introduction

A dialog system should correctly understand speakers' utterances and should respond to their requests in their natural language. To realize the former, the dialogue system should identify the underlying intentions of the speakers' utterance [6]. The speakers' intentions in dialogues can be represented by combinations of speech acts and predicators (so-called main actions or concept sequences) [11]. In addition, the sentiments implicated in utterances can help to capture speakers' intentions. Table 1 shows some parts of a dialogue between a dialogue system and a human user.

In Table 1, we represent speaker's intention in a comma-separated triple format. The first element in the triplet is a speech act (*e.g.*, "inform," "ask-ref," "response," and "statement" in the example) that indicates a domain-independent intention associated with the conversational role of an utterance. The second element is a predicator (*e.g.*, "late," "be," "part," and "encourage" in the example) that captures a domain-dependent semantic focus associated with the main meaning of an utterance. The last element is a sentiment (*e.g.*, "none" and "sadness" in the example) that expresses speaker's attitude with respect to a dialogue topic. As shown in Table 1, a speech act and a predicator represent speaker's explicit intention, and a sentiment represents an implicit intention supplementing his/her explicit intention. As shown in Table 1, a current speech act is strongly dependent on previous speech acts. For example, the speech act "response" of the third utterance is affected by the previous speech act "ask-ref." If the previous speech act were not "ask-ref," it could be "inform." A predicator and a sentiment are less dependent on their contexts than a speech act is. On the other hand, they are affected by lexical meanings of the current utterance and are associated with each other. For example, the predicator "part" of the fourth utterance is determined by the word sense of the main verb phrase "was departed from." In addition, the predicator "part" of the fourth utterance aids the system to determine the sentiment "sadness," and the sentiment "sadness" of the fifth utterance aids the system to determine the predicator "encourage." In this paper, we propose an integrated model to simultaneously determine speakers' speech acts, predicators, and sentiments.

This paper is organized as follows. In Section 2, we review the previous work on intention analysis. In Section 3, we describe the integrated intention analysis model. In Section 4, we explain the experimental setup and report some experimental results. Finally, we draw conclusions in Section 5.

* Corresponding author.
E-mail address: nlpdrkim@kangwon.ac.kr (H. Kim).

https://doi.org/10.1016/j.patrec.2017.11.009
0167-8655/

Table 1
Example of a dialogue between a system and a user.

Speaker	Utterance	Intention
User	I was late in returning home yesterday.	(inform, late, none)
System	What time was it?	(ask-ref, be, none)
User	11 P.M.	(response, be, none)
User	In fact, I was parted from her.	(inform, part, sadness)
System	Come on.	(statement, encourage, sadness)

2. Previous work

Most recent studies on speech acts and predicators have been based on machine learning models. Stolcke et al. [16] proposed a speech act labeling model based on a hidden Markov model (HMM), in which acoustic features (prosodic features as well as lexical features) are used for dealing with speech inputs. Langley [8] proposed a speech act classifier and predicator classifier with memory-based learning (k-NN) to improve the performance of speech translation. Surendran and Levow [17] replaced the observation probabilities of an HMM with the class probabilities of a support vector machine (SVM) in order to reduce a sparse data problem. Kang et al. [4] proposed a multidomain model based on conditional random fields, in which input features are constructed according to application domains. Although many machine learning models based on various linguistic features have been proposed, the previous models have mainly dealt with speech act identification alone [4,14,16,17,19] or have separately dealt with speech act identification and predicator identification [8,9]. However, a pair consisting of a speech act and a predicator should be identified simultaneously to precisely capture speakers' intentions. Therefore, Lee et al. [10] proposed an integrated neural network model in which speech act identification results are used as inputs to predicator identification. To improve the performance of the integrated model, Seon et al. [13] proposed a mutual retraining method in which speech act identification results are repeatedly used as inputs to predicator identification while training, and vice versa. Although these integrated models showed that the integration architecture can help to increase performance, they did not consider speakers' sentiments as elements composing their intentions.

The previous studies on sentiment classification can be divided into two groups: feature-focused methods [12,20] and learner-focused methods [2,5]. The feature-focused methods have mainly studied feature-weighting schemes based on various resources, such as sentiment dictionaries and sentiment snippets (*i.e.*, two or three sentences including sentiment words). The learner-focused methods have mainly studied how to apply various machine learning models to sentiment classification. To alleviate the feature engineering requirements for learner-focused methods, sentiment classification models based on neural networks using word embedding vectors as input features have been proposed [15]. Although there have been numerous studies on sentiment classification, most of the previous studies were focused on sentiment classification, not of dialogue utterances but short texts such as customers' reviews and blogs.

3. IIIM: integrated intention identification model based on neural networks

Given n utterances, $U_{1,n}$, in a dialogue D, let $S_{1,n}$, $P_{1,n}$, and $E_{1,n}$ denote n speech act tags, predicator tags, and sentiment tags in D, respectively. The integrated model can then be formally expressed

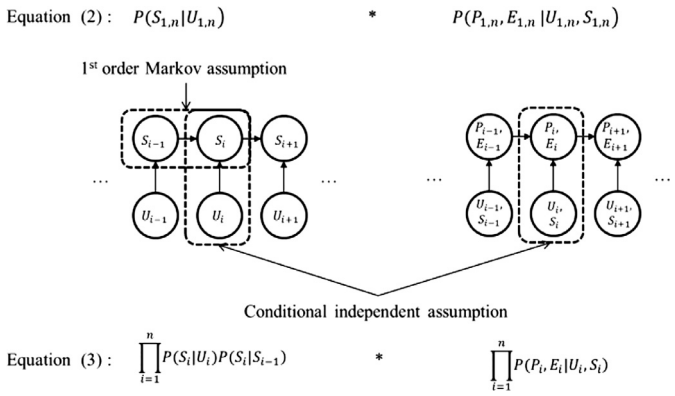

Fig. 1. Simplification of equation by the Markov assumption and independence assumption.

as the following equation:

$$IM(D) \overset{def}{=} \underset{S_{1,n},P_{1,n},E_{1,n}}{\arg\max} P(S_{1,n}, P_{1,n}, E_{1,n}|U_{1,n}) \quad (1)$$

According to the chain rule, Eq. (1) can be rewritten as follows:

$$IM(D) \overset{def}{=} \underset{S_{1,n},P_{1,n},E_{1,n}}{\arg\max} P(S_{1,n}|U_{1,n})P(P_{1,n}, E_{1,n}|U_{1,n}, S_{1,n}) \quad (2)$$

As shown in Eq. (2), the integrated model consists of two parts: the speech act identification model, $P(S_{1,n}|U_{1,n})$, and the predicator & sentiment identification model, $P(P_{1,n}, E_{1,n}|U_{1,n}, S_{1,n})$. To simplify the speech act identification model, we assume that a current speech act is dependent on the previous speech act (*i.e.*, a 1st order Markov assumption) because speech acts are strongly affected by their previous contexts. Then, we assume that a predicator and a sentiment are only dependent on their current observational information (*i.e.*, a conditional independent assumption) because predicators and sentiments are strongly affected by lexical meanings of current utterances. We also apply the conditional independent assumption to the speech act identification model. Fig. 1 depicts the process by which Eq. (2) is simplified into Eq. (3) according to these two assumptions.

$$IM(D) \overset{def}{=} \underset{S_{1,n},P_{1,n},E_{1,n}}{\arg\max} \prod_{i=1}^{n} \begin{Bmatrix} P(S_i|U_i)P(S_i|S_{i-1}) \\ P(P_i, E_i|U_i, S_i) \end{Bmatrix} \quad (3)$$

To obtain the sequence labels $S_{1,n}$, $P_{1,n}$, and $E_{1,n}$ that maximize Eq. (3), we propose an Integrated Intention Identification Model (IIIM) based on Convolutional Neural Networks (CNNs) [7], as shown in Fig. 2. In Fig. 2, W_i is a Word2Vec embedding vector with 50 dimensions of the ith words in the input utterance [3]. The embedding vectors are trained from a large balanced corpus called the 21st century Sejong project's POS-tagged corpus [18]. H_X represents a set of nodes fully connected with the output X. In other words, H_P means a set of nodes that is fully connected with the output vector representing the predicator P_i. Similarly, H_{XY} means a set of nodes fully connected with the outputs X and Y. In other words, H_{EP} means a set of nodes that is fully connected with two output vectors representing the sentiment E_i and the predicator P_i. In this paper, these sets of partially grouped nodes, such as H_{ES}, H_{SP}, H_{EP}, and H_{ESP}, are called shared nodes because they can contain weighting values associated with multiple outputs. While training, three types of utterance embedding vectors are generated by concatenating nodes in the hidden layer: the utterance embedding vector SE_i for speech act identification, the utterance embedding vector PE_i for predicator identification, and the utterance embedding vector EE_i for sentiment identification. To generate the embedding vectors, three cycles of partial error backpropagations