Neurocomputing 168 (2015) 408–417

Contents lists available at ScienceDirect

Neurocomputing

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

Combining heterogeneous deep neural networks with conditional random fields for Chinese dialogue act recognition

Yucan Zhou [a], Qinghua Hu [a,*], Jie Liu [b], Yuan Jia [c]

[a] School of Computer Science and Technology, Tianjin University, Tianjin, China
[b] College of Computer and Control Engineering, Nankai University, Tianjin, China
[c] Institute of Linguistics Chinese Academic of Social Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 7 February 2015
Received in revised form
3 April 2015
Accepted 22 May 2015
Communicated by Jiayu Zhou
Available online 2 June 2015

Keywords:
Dialogue act recognition
Heterogeneous features
Deep learning
Conditional random fields

ABSTRACT

Dialogue act (DA) recognition is a fundamental step for computers to understand natural-language dialogues because it can reflect the intention of a speaker. However, it is difficult to adapt traditional machine learning models to the dialogue act recognition task due to the heterogeneous features, statistical dependence between the DA tags, and complex relationship between features and the DA tags. In this paper, we propose a new model which combines heterogeneous deep neural networks with conditional random fields (HDNN-CRF) to solve this problem. The proposed model has two main advantages. First, the heterogeneous deep neural networks (HDNN) model, which is extended from the deep neural networks (DNN), retains the powerful ability of representation learning and adds a new skill of dealing with heterogeneous features effectively. Second, the conditional random fields (CRF) can capture the statistical dependence between the DA tags which carries important information to determine the DA tag of the current utterance. To verify the effectiveness of the proposed model, we conduct several experiments on a Chinese corpus, called CASIA-CASSIL corpus. Ten kinds of features are extracted from the utterances. In the experiment, we give some quantitative analysis of these kinds of features. What's more, when comparing classification accuracies of the proposed model and some other models, the proposed model has achieved the best performance.

1. Introduction

Dialogue act (DA) falls into the category of shallow discourse structure [1]. It is derived from speech act. The speech act theory is first proposed by Austin [2], and then improved by Searle [3]. Speech act is the pragmatic force of an utterance, and characterizes the intention of the speaker which is critical to determine its pragmatic meaning. While in the dialogue system, speech act is evolved into DA. The DA tags not only include most of the tags in speech act, such as Statement, Imperative Sentence and Request [4]; but also relate to the conversational function [5]. For example, we may use different utterances to reject a request, but the DA tags of these utterances are the same.

DA recognition is very important for computers to understand natural-language dialogues. In the areas of human-to-computer dialogue, interactive information retrieval (IR) and interactive question answering, a key research is to recognize the intention of a speaker. But intention recognition is too complicated, and the DA can give some indications to the intention. So we can simplify the intention recognition to DA recognition.

However, it is still an ambitious goal to recognize the DA tag of an utterance automatically. There are four main challenges.

- *Heterogeneous features learning*:
  Several studies [6–12] have investigated different approaches to DA labeling. In their researches, three kinds of features are commonly used, i.e., DA-specific language models, DA-specific prosodic models and dialogue grammar. In DA-specific language models, the lexical feature is considered as an important feature. For the DA-specific prosodic models, most of the studies [13–16] are focused on the pitch contour, duration, energy and transformations thereof. Besides, our recent research results show that the accent of an utterance is also essential to determine the DA tags in a Chinese dialogue system [17]. While in the dialogue grammar, the position and context of an utterance are considered as important features. For example, it is usual that a dialogue does not start or end with a question.
  As we all know, the performance of machine learning methods heavily depends on the features on which they are applied [18]. So how to represent and combine these various kinds of

* Corresponding author.
E-mail addresses: zhouyucan@tju.edu.cn (Y. Zhou), huqinghua@tju.edu.cn (Q. Hu).

http://dx.doi.org/10.1016/j.neucom.2015.05.086
0925-2312/

features is a fundamental and imperative task for identifying the DA tag.

- *Various class variations and imbalanced distribution*:
  The number of DA tags is very large. From the specification of annotation [19], we can see three kinds of DA tag sets are commonly used, i.e., common tag set, interrupt tag set and specific tag set. There are 13 DA tags in the common tag set, 3 in the interrupt tag set and 36 in the specific tag set. And the specific tag set is a more detailed explanation of the common tag set. If we consider all the tag sets, there will be more than 200 DA tags. It is better to take all the tag sets into account because the more detailed the DA tag is, the better it can reflect the intention. However, a disappointing fact that it is very hard to collect and label the dialogues should not be ignored. So the spontaneous dialogues we can obtain are very precious and rare. When thinking about all the three DA tag sets, the number of samples belonging to a certain DA tag may be very small, making it difficult to apply machine learning techniques on this task.
  To simplify the task, most of the DA recognition works just take the common tag set and interrupt tag set into consideration. Even though, there are still 16 DA tags and the samples' distribution on these tags is heavily biased. In the CASIA-CASSIL corpus, 63% of the utterances belongs to the class of Statement, and just a few utterances fall into Abandoned, Interrupted, Exclamatory, etc. This phenomenon is consistent with our common sense that we use more statements in a dialogue. Fig. 1 presents the DA tags in the CASIA-CASSIL corpus and the samples' distribution. As for two certain DA tags, there are no utterances belonging to them, so there are only 14 DA tags. The numbers in the *x*-axis represent the DA tags of *Statement*, *Tag question*, *Y/N Question*, *Echo Question*, *Wh-Question*, *Imperative Sentence*, *Open-end Question*, *Rhetorical Question*, *Or Question*, *Position or negative question*, *Interrupted*, *Exclamatory Sentence*, *Or Clause After Y/N Question and No-question*.
- *Statistical dependence between the DA Tags*:
  As mentioned before, the context of an utterance is another important feature of the dialogue grammar. For example, if the DA tag of the previous utterance is *Yes or No Question*, then the DA tag of the current utterance is more likely to be *Statement*. So it is more reasonable to cast the DA recognition problem as a sequential labeling task. Classifiers for DA recognition should take these dependencies into consideration in order to improve their performance.
- *Complex relationship between the features and DA Tags*:
  It is a difficult task to recognize people's intention as it is usually not represented clearly by the utterances. People may use different utterances to express the same meaning. For example you can use the utterance that *I have promised to see a movie with my friends* or *it is going to rain and I don't want to get wet in the rain* to decline a dinner invitation. The movie and the weather seem to have no relationship with the dinner invitation, but they succeed in turning down the invitation with finesse. So even for human brain, the most powerful system, it is not so easy to identify people's intention. Although DA recognition is simplified from intention recognition, it is still a big challenge to adapt some traditional machine learning algorithms to model the complex relationship between the features and DA tags.

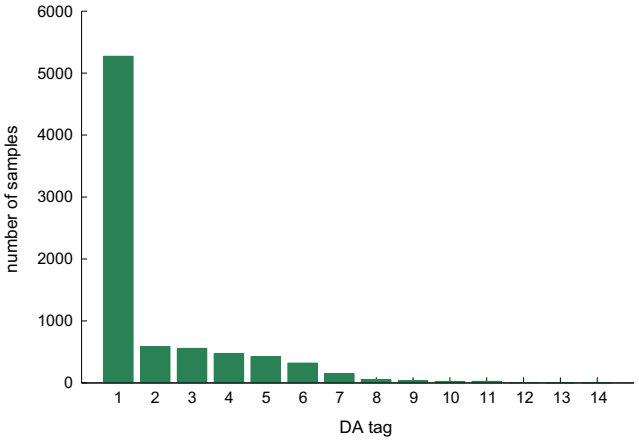

Fig. 1. DA tags and distribution of the samples in the CASIA-CASSIL corpus.

To address the issues mentioned above, a new sequential model based on deep learning and conditional random fields (CRF) is proposed. Deep learning is one of the hottest topics in the field of machine learning research in recent years and is famous for its powerful feature learning capability. And CRF is a very effective sequential learning model which has achieved good results in many sequential labeling tasks. With the combination of these two models, the proposed model can capture the statistical dependence between the dialogue act tags and overcome the shortcoming of weak feature learning ability of the standard CRF. Besides simple combination, here we introduce a variant of the classical deep neural networks (DNN). As for the DNN, it is unable to deal with heterogeneous information, so we adjust the DNN model to a new heterogeneous deep neural network (HDNN) to learn and combine high-level heterogeneous representations.

The rest of the paper is organized as follows. In Section 2, related works are introduced. Section 3 describes the task of DA recognition and discusses the proposed model in detail. Experimental results and analysis are shown in Section 4. Finally we draw some conclusions and talk about the future works in Section 5.

2. Related works

Various machine learning approaches have been introduced to automatically identify dialogue act tags in recent years. Roughly speaking, we can divide these researches into two parts. One considers the utterances of a dialogue to be separated, while the other models the utterances in a dialogue as a sequence.

Some shallow models consider the utterances in one dialogue are separated, so they recognize the DA tag of each utterance individually. One of the simplest techniques is *n*-grams. Louwerse use this method to model the correlation of the words and the DA tag [8]. The Bayes classifier is similar to *n*-grams except that any kind of features could be applied to it. Levin [20] and Grau [21] have applied the Bayes classifier with grammatical and bag-of-words features to DA recognition. Another model, decision tree, which can also deal with any kind of features and learn a set of well-understood rules, is adapted to this task by Mast in 1995 [22]. In 2004, Irie constructs a layered decision tree model to solve the problem of hierarchical intention recognition [23]. Other traditional models, such as Maximum entropy classifier [24,5,10] and Artificial Neural Networks [7] have also been introduced to DA recognition task. However, due to the complex dependence between the features and DA tags, the recognition performance is limited.

Some sequential models have also been used to solve the DA recognition problem. Unlike the learning methods described above, these sequential models aim to capture the dependency in the sequence. As the simplest sequential model, the *n*-gram method is based on the Markov assumption that the DA tag of the