Pattern Recognition Letters 51 (2015) 107–111

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

New feature weighting approaches for speech-act classification☆

Youngjoong Ko*

Department of Computer Engineering, Dong-A University, 840 Hadan 2-dong, Saha-gu, Busan 604-714, Republic of Korea

ARTICLE INFO

Article history:
Received 12 December 2013
Available online 10 October 2014

Keywords:
Natural language processing
Discourse analysis
Speech-act classification
Feature weighting scheme

ABSTRACT

Speech-act classification is essential to generation and understanding of utterances within a natural language dialogue system since the speech-act of an utterance is closely tied to a user intention. The binary feature weighting scheme has mainly been used for speech-act classification because traditional feature weighting schemes such as *tf.idf* are not effective in speech-act classification due to the short length of utterances. This paper studies two effective feature weighting schemes using the category distributions of features: (1) the first one exploits the entropy of whole category distributions and (2) the second one the log-odds ratio of positive and negative category distributions. As a result, the proposed schemes show significant improvement on SVM and *k*-NN classifiers in our experiments.

1. Introduction

A dialogue system is a software program that enables a user to interact with a computer using a natural language [5] . Since an essential task of the dialogue system is to understand what the user says, it must be able to determine the user's intention indicated in the user's utterance. A speech-act is a linguistic action and implies the user's intention. Therefore, the dialogue system must identify the speech-act of user's utterance. Although researchers have developed many techniques for the speech-act classification, they have mainly used the binary feature weighting scheme because it is simpler but more effective than other schemes such as *tf* (traditional term frequency), *idf* (inverse document frequency) and *tf.idf* [11,13,14]. An utterance is usually much shorter than a document, and it means that the utterance has only the small number of features. For example, as two major factors of traditional *tf.idf*, *tf* is the number of term occurrence in a document and *df* (document frequency) is the number of documents that a term occurs in a collection. In particular, since *tf* rarely becomes more than 2 in an utterance due to the short length of the utterance, terms with more than 2 frequencies make the distribution of term weights biased and it causes the poor performance of speech-act classification.

This paper explores to find more effective feature weighting schemes for the cases of classification with the small number of features such as speech-act classification, and proposes two weighting schemes that are based on feature distributions through categories. (1) One feature weighting scheme applies the entropy concept to estimate the feature importance using all category distributions of each feature, and (2) the other scheme utilizes the ratio of positive and negative category distributions to estimate the feature importance of each category. In the experiments, these weighting schemes achieved better performances than the binary feature weighting, *tf*, *idf* and *tf.idf* schemes.

The remainder of the paper is organized as follows. Section 2 describes the related work. Section 3 explains two proposed feature weighting schemes in detail and Section 4 is devoted to the analysis of our experimental results. The final section states the conclusions and future work.

2. Related work

Some previous studies on Korean speech-act classification have been based on rules extracted from a tagged dialogue corpus [3,8], while others have been based on statistical models learned from a tagged dialogue corpus [1,5, 6, 9].

The initial speech-act classification studies used rules that are extracted from a tagged dialogue corpus such as linguistic rules and dialogue grammar. Lee [8] developed a two-step speech-act classification system using linguistic rules and dialogue flow diagrams; the first step in this model classifies surface speech acts, whereas the second step classifies deep speech acts. Choi et al. [1] proposed a statistical dialogue classification model that performs both speech-act classification and discourse structure analysis using the maximum entropy model (MEM). This model automatically acquired discourse knowledge from a discourse-tagged corpus to resolve ambiguities. Lee and Seo [9] classified speech acts by a bigram hidden Markov model (HMM). Kim et al. [6] presented a speech-act classification model to utilize contextual information by adjacency pairs and a discourse

☆ This paper has been recommended for acceptance by Crocco Marco.

* Tel.: +82 51 200 7782; fax: +82 51 200 7783.
E-mail address: youngjoong.ko@gmail.com, yjko@dau.ac.kr

http://dx.doi.org/10.1016/j.patrec.2014.08.014
0167-8655/

Table 1
Example dialogue annotated with speech-acts.

Speaker	Utterance	Speech-act
Clerk	안녕하세요. 서울호텔입니다. (Hello. This is Seoul Hotel.)	Introducing-oneself
User	가족이 4명인데요. (I have four people in my family.)	Inform
User	방을 하나 예약하려고요. (I want to reserve one room.)	Request
Clerk	성함이 어떻게 되세요? (What is your name?)	Ask-ref
User	내 이름은 홍길동입니다. (My name is Kildong Hong.)	Response

Table 2
Example of lexical features.

Input utterance	내 이름은 홍길동입니다. (My name is Gildong Hong.)
Morphological analysis (morpheme/POS tag[a])	나/*np* 의/*j* 이름/*ncn* 은/*j* 홍길동/*nq* 이/*jcp* ㅂ니다/*ef* ./*s*. (My/*np* name/*ncn* is/*jcp* Gildong Hong/*nq* ./*s*.)
Content words features	나/*np* 이름/*ncn* 홍길동/*nq* ./*s*. (My/*np* name/*ncn* Gildong Hong/*nq* ./*s*.)
POS-bigram features	*np-j j-ncn ncn-j j-nq nq-jcp jcp-ef ef-s*.

[a] The Korean POS tags in this example are as follows:
Noun: *np* (pronoun), *ncn* (common noun), *nq* (proper noun).
Particle: *j* (case particle), *jcp* (predicative case particle).
Ending: *ef* (final ending).
Symbol: *s*. (sentence closer).

stack, and Kang et al. [5] proposed to use a hierarchical structure to improve the speech-act classification. The both of these studies exploited SVM as a classifier, and they achieved better performances than other classification models. Feature weighting of their SVM classifiers is based on the binary weighting scheme because it is more effective than other feature weighting schemes.

3. Proposed feature weighting schemes for speech-act classification

Traditional feature weighting schemes of document classification are based on *tf*, *idf* and *tf.idf* [11,13,14]. They are calculated by using the unit of a document. That is, *tf* is the number of occurrences of a term in a document and *df* is the number of documents that a term occurs in a collection. In speech-act classification, the unit of an utterance has to be used instead of one of a document. Since the input of speech-act classification is only one utterance, its length is very short and it contains much smaller number of features than a document does. It causes a problem that *tf*, *idf* and *tf.idf* do not work well for speech-act classification.

This study explores new feature weighting schemes with category distributions to improve biased information from the short length of utterances; here, category means speech-act. The new schemes are expected to elevate the speech-act classification because they exploit the probabilistic distribution on a category that is larger than an utterance. Two different schemes are proposed in this study. One is based on entropy of probabilistic distributions of each feature on category and the other is on the distribution difference from a positive category to negative categories.

3.1. Examples of dialogue annotated with speech-acts and features extracted from an utterance

Table 1 shows an example dialogue between a clerk and a user in a hotel booking domain.

In general, speech-act analysis has exploited multiple knowledge sources in the form of lexical, syntactic, prosodic and contextual information [12]. These sources have typically been modeled using various stochastic models. Many previous studies on speech-act classification have applied syntactic patterns as intra-utterance features. Although a syntactic pattern can represent the syntactic and semantic features of utterances, previous studies have found that syntactic patterns from a conventional syntactic parser are incomplete owing to errors in the syntactic analysis and are dependent on time-consuming tasks and manually generated knowledge [9, 15]. To overcome this problem, a lexical feature extraction method is developed to use only the analysis results from a morphological analyzer so that our method becomes more robust to errors propagated from basic language analysis. A morphological analyzer generally creates fewer errors than a syntactic analyzer because the output of the morphological analyzer becomes the input of the syntactic one. We assume that content words and Part-Of-Speech (POS) tag sequences in an utterance can provide very effective information for detecting the speech act of that utterance. Based on this assumption, we extract informative features for speech-act analysis using only a morphological analyzer. Lexical features include content words annotated with POS tags and POS bigrams of all words in an utterance (see Table 2). Content words generally have noun, verb, adjective, adverb and symbol (punctuation/exclamation/question marks) POS tags. For example, the lexical features of the example utterance in Table 2 consist of four content words and seven POS bigrams. These features represent the linguistic function and meaning of an utterance. This lexical-based classification has demonstrated better and more robust performance than syntactic-based classification for speech-act analysis, because morphological analysis results have fewer errors than a syntactic parser in most cases [5,6]. Therefore, we also employed the same lexical features in this study.

Table 2 shows an example of lexical feature extraction using a morphological analyzer.

Eventually, the final feature set is composed of these lexical features and the speech-act of the previous utterance.

3.2. Estimation of feature probabilistic distribution for categories

The expected likelihood estimator [10,11] is used for the estimates of the probability of a feature in positive and negative categories as follows:

$$P(f_i|c_j) = \frac{N(f_i, c_j) + 0.5}{\sum_{t=1}^{|V|} N(f_t, c_j) + 0.5 \times |V|}, \tag{1}$$

$$P(f_i|\bar{c}_j) = \frac{N(f_i, \bar{c}_j) + 0.5}{\sum_{t=1}^{|V|} N(f_t, \bar{c}_j) + 0.5 \times |V|}, \tag{2}$$

where $N(f_i, c_j)$ is the count of the number of times that feature f_i occurs in category, c_j, $|V|$ is a vocabulary size and $\bar{c}_j$ is the negative categories of a positive category, c_j . Herein, 0.5 is a smoothing factor and it can be viewed as a linear interpolation between the maximum likelihood estimation and a uniform prior. This guarantees no zero probabilities yet retains the relative likelihoods for the frequently occurring values. In multiclass classification, a single text classifier is generally trained per a category to distinguish that category (positive category) from all other categories (negative categories). This strategy is called one-versus-all or one-versus-rest [2].

3.3. Entropy of probabilistic distributions

The first proposed feature weighting scheme is based on the Entropy value of Category Probabilities (ECP) for each feature. Since entropy is a measure of uncertainty in a feature for a collection, features with a high entropy value are regarded as bad features for speech-act classification. Therefore, the new feature weighting scheme is