Scoring of response based on suitability of dialogue-act and content similarity

Sota Matsumoto
Kyoto Institute of Technology, Japan
matsumoto@ii.kit.is.ac.jp

Masahiro Araki
Kyoto Institute of Technology, Japan
araki@kit.ac.jp

ABSTRACT

We present an approach to scoring candidate utterances in a large repository of short text conversation (STC) data to select those to be used as a suitable response to a newly given utterance. Candidate utterances are evaluated based on the suitability of a dialogue-act and the content similarity. The estimation of the suitability of a dialogue-act is implemented by learning the trend of a dialogue-act pair that frequently appears in the repository. Also, we calculated the content similarity between utterances by means of the cosine similarity of topic vectors using LDA and IDF. By multiplying these values, those candidates which are suitable in terms of function and content attain a high score. As a result of the experimental evaluation, for content similarity, it was found that increasing the weighting of the IDF produces a better accuracy.

Team Name

KIT15

Subtasks

STC (Japanese)

Keywords

short text conversation, response, dialogue-act, content similarity, LDA, IDF

1. INTRODUCTION

Several open-domain conversational systems have been developed over the past decades. However, the level of performance of these systems has not yet reached a level where they can be practically implemented. One of the major reasons for this is the lack of a large volume of real conversation data. Therefore, we consider a highly simplified version of the problem here: one round of conversation is formed by two short texts, with the former being an initial post from a user and the latter being a comment given by the computer. We refer to this as a short-text conversation (STC). Furthermore, rather than generating a suitable comment for a given post, we reuse suitable utterances as comments from the large repository of STC data. That is, we tackle a conversation between a human and a computer as a problem of information retrieval (IR) technology.

A previous work on selecting responses from the repository produced IR-Status [1] which returns a comment that is stored as a reply to an utterance in the repository that is similar to a given post, but because it does not check whether the candidate is suitable, a misdirected reply is often selected. Also, Inaba et al. proposed a method retrieving texts containing a specific topic word from the repository and scoring based on the importance of words in the text [2]. Thus, while it is possible to select the appropriate utterance along the topic word, since it doesn't consider other informations such as intention and function included in the utterance, it is insufficient for use in response. Therefore, to select candidates, we decided to score each utterance in the repository for two items, namely, dialogue functional relevance and content similarity. Thus, it is possible to choose candidates which are functionally reasonable in terms of conforming to a given utterance while being similar in content.

2. SCORING OF REPLY UTTERANCE

We evaluated each utterance in the repository based on two items, namely, functional relevance and content similarity, and by multiplying their assigned values, selected comments that satisfy both. We provide a means of formulating the problem in this paper to conform to the following formula and parameters.

$$Score(p, t_a) = ifs(p, t_a) * csim(p, t_a) \quad (1)$$

$Score$: Score of suitability of comment to initial post
ifs: Evaluator of interactive functional suitability
$csim$: Evaluator of content similarity
p: Given new post
t_a: Any utterance in the repository$(a = 1, 2, ..., n)$

3. METHOD

3.1 Interactive functional suitability

For a given post, to determine the interactive functional suitability of a comment for each utterance in the repository, we use a "dialogue-act" to indicate the type of utterance, such as "greeting", "question", or "supportive response". By this, and the likes of "it is likely to return the presentation as a response when a question is posted", it is possible to calculate the ease with which a dialogue-act can be returned when another one is posted. To estimate the appropriate dialogue-act which should be returned to the opponent of utterance, it is also possible to cope with formulaic responses [3]. Although it is general that dialogue-act is designed in accordance with the domain of the dialogue which takes place, it is tedious and expensive to manually analyze the kind of utterance that is present in the domain and to set the number of dialogue-acts matched

to the domain [4]. To overcome this problem, we used a method proposed by Crook et al. for automatically estimating those dialogue-acts that are matched to the domain by means of unsupervised clustering using the so-called "Chinese Restaurant Process" (CRP) [5].

First, to estimate the dialogue-act for each utterance data in the repository and then generate an estimator for the dialogue-act in the domain, we perform clustering by applying the CRP. Based on the dialogue-act for each estimated utterance, we generate a lookup table WS for the weight from the relationship between the post-comment pairs in the repository. Each element in the table is calculated as follows:

$$W[i][j] = \frac{count(i,j)}{N} \quad (2)$$

Here, $count(i,j)$ is the number of pairs in the repository for which the dialogue-act of the post is i and that of the comment is j, and N is the total number of pairs in the repository. Thus, when the dialogue-act of the post is determined, it is possible to obtain the trend in the dialogue-act of the comments as a list of weightings. Therefore, when a post p is evaluated as $ifs(p,t_a)$, any utterance to get in the repository is: Therefore, when the post p is given the evaluation value $ifs(p,t_a)$ that any utterance t_a get in the repository is:

$$ifs(p,t_a) = W[dae(p)][dae(t_a)] \quad (3)$$

Here $dae(,)$ is an estimator of the dialogue-act learned from utterances in the repository. If the post is estimated for a dialogue-act other than those of domain $ifs(p,t_a)$, a uniform distribution is used.

3.2 Content similarity

To calculate the content similarity between a given post and an utterance in the repository, we used the cosine similarity of topic vectors generated for each utterance from the Latent Dirichlet Allocation (LDA) [6], which is one of the topic models. LDA can extract potential topics from an utterance which do not appear directly, but since it cannot handle any more than a finite number of previously learned topics, we use it in combination with the similarity determined from the inverse document frequency (IDF) to cope with any out-of-range topics covered by LDA. Although in order to calculate the visible similarity between documents it is common to use the cosine similarity of TF-IDF multiplied by the term frequency (TF) to IDF, because to target short texts, in this study, the number of words in each document is also small. Therefore, we considered that the value of TF is strongly influenced by the denominator (the total number of words in a document) than the molecule (the number of occurrences of the particular word in a document) in the similarity, and decided not using TF to prevent from appearing the large difference of the score by the number of words in document. When the post p is given the content similarity $csim(p,t_a)$, any utterance t_a retrieved from the repository is:

$$csim(p,t_a) = \alpha * lsim(p,t_a) + (1-\alpha) * isim(p,t_a) \quad (4)$$

Here $lsim(,)$ is a similarity calculator of LDA and $isim(,)$ is one of IDF. $\alpha (0 \leq \alpha \leq 1)$ is the parameter for adjusting the ratio occupied by LDA and IDF in the score.

4. EXPERIMENTS

4.1 Data

The STC data of the repository used in this study are utterances that were posted to the online social networking service Twitter during 2014 and which are selected at random according to the post-comment relationship. The number of pairs is 411,127. We refer to this as the "training data".

The data used for testing consisting of utterances that were posted to Twitter during 2015 and selected at random. The number of utterances was 202. We refer to this as the "testing data".

4.2 Experimental procedures

By performing the unsupervised clustering of dialogue-acts by applying the CRP to all of the utterances in the training data, we added the information of the dialogue-act to each utterance data and obtained an estimate of the dialogue-act in the training data domain. Here, we used the features of the bag-of-words for clustering. However, because low-frequency words could adversely affect the clustering, we selected only those words with a frequency of appearance in excess of 1,000 as features. The CRP hyperparameters α and β were set to 1, and 0.01, and Gibbs sampling was carried out 100 times. To refer to the evaluation value for the candidates obtained from the dialogue-act for a given utterance, we created a weight table based on a resulting combination of dialog-acts.

As the data for the training of the topic model for LDA, we used all of the articles of the free web encyclopedia Hatena Keyword[1] (until Feb 5th, 2016) and selected the bag-of-words of nouns as features. The number of dimensions of the topic vector was set to 300, and the other settings were the standard settings for LDAModel of the gensim[2] library for a topic model for Python. Using this, we generated topic vectors based on the bag-of-words for the noun for each utterance in the training data. In the same way as for the testing data, we calculated cosine similarities between generated topic vectors and the training data.

In IDF, we generated IDF vectors from the bag-of-words of nouns in utterances in the training data and the testing data and calculated this cosine similarity.

All of the morphological analysis was done by using MeCab[3].

Based on these evaluation values, we attempted four cases in which the parameter α of equation (4) was 0, 0.4, 0.5, and 1.0.

The resulting candidates were evaluated in terms of whether they are an appropriate response to a given utterance. Evaluations were carried out manually using ten annotators and for five candidates in descending order, starting from that having the top score in the testing data. Each candidate was labeled 0 (inappropriate), 1 (appropriate in some contexts), or 2 (appropriate) by multiple judges.

4.3 Results

4.3.1 *Clustering*

As a result of unsupervised clustering by CRP to STC

[1] http://d.hatena.ne.jp/keyword/
[2] https://radimrehurek.com/gensim/
[3] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

data repository, it was classified into 41 types of dialogue-acts. However, approximately half of the data converged at the one cluster, and several clusters containing it were those focused on case particles which don't affect the function of the conversation so much. Therefore, the learned weights also were strongly attracted to those, and basically the value of ifs becomes higher as the candidate is "text containing many case particle". As an exception, the case that there is a tendency that clusters of greeting type are relatively returned ones of the same type was observed.

4.3.2 Evaluation

For each pair consisting of a post in the testing data and a comment in the candidate utterances, by regarding those judged to be appropriate as being correct based on the evaluation value of each annotator, the accuracies of the candidate selection were calculated. In the following table, each value corresponds to the average of the accuracy of the four trial results in the four evaluation conditions. Case 2-1 is a case in which if only the evaluation value is 2 the result is correct, and thus only the best candidate is evaluated. Case 2-5 is the same as case 2-1 for a correct judgment, and the case when the top five candidates are evaluated. Cases 12-1 and 12-5 are those when, if the evaluation value is 1 or 2, the result is correct, and the same as in case 2-1 and case 2-5 in terms of the evaluated candidates, respectively.

α	case 2-1	case 2-5	case 12-1	case 12-5
0	0.2297	0.2050	0.5589	0.5380
0.4	0.1817	0.1743	0.4748	0.4535
0.5	0.1812	0.1660	0.4614	0.4317
1.0	0.0787	0.0787	0.2114	0.2130

Table 1: Averages of accuracies of each trial result for the four evaluation conditions

In Table 1, as the value of α falls, that is, as the context similarity constitutes a smaller proportion of LDA, the accuracy increases. Finally, the case in which LDA is not used showed the highest accuracy.

One of the factors to be considered as a reason for the LDA not providing a suitable level of accuracy is the lack of the number of dimensions of the topic vector. It can be said that it was not possible to achieve a correspondence to the vast number of topics in Twitter in 300 dimensions. Therefore, in response to a search including the LDA, we should be determining a topic model with a higher number of dimensions.

5. CONCLUSIONS

From a repository of pairs consisting of an initial post and corresponding comment, we searched for utterances which constitute a suitable comment to a given new post.

This paper has presented a scoring method whereby a higher weight is assigned to an utterance having a dialogue-act that is likely to be used in the reply to a given post by looking at the trend in the relationships of the dialogue-act pairs. By multiplying this value by the content similarity between a given post and a candidate utterance using LDA and IDF, we can select a comment which is suitable as a reply to a given post, and which is similar to it.

As a result, determining the content similarity by using only IDF produced a higher level of accuracy. In addition, the classification of the dialogue-act could only work to exclude extremely collapsed texts with the exception of some, such as greeting.

Future challenges involve the consideration of the part-of-speech of the words to be adopted as feature in clustering of dialogue-act, the verification of the efficacy of the functional interactive suitability by comparing this method with others and improving the accuracy by increasing the number of dimensions of the LDA model.

6. REFERENCES

[1] Alan Ritter, Colin Cherry and William B. Dolan. Data-Driven Response Generation in Social Media. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 583–593, 2011.

[2] Michimasa Inaba, Sayaka Kamizono, Kenichi Takahashi. Tsuittaa wo motiita Hitasukushikougatataiwashisutemu no tame no Hatsuwakouhobunkakutoku (Candidate Utterance Acquisition Method for Non-task-oriented Dialogue Systems from Twitter). Transactions of the Japanese Society for Artificial Intelligence, Vol. 29, No. 1, pages 21–31, 2014.

[3] Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami. Open-domain Utterance Generation for Conversational Dialogue Systems using Web-scale Dependency Structures. Proceedings of the SIGDIAL 2013 Conference, pages 334–338, 2013.

[4] Ryuichiro Higashinaka, Noriaki Kawamae, Kugatsu Sadamitsu, Yasuhiro Minami, Toyomi Meguro, Kohji Dohsaka, Hirohito Inagaki. Taiwakouisekkei no tame no Hatsuwakurasutaringu (Automatic Clustering of Utterances for a Dialogue Act Design). 63th SIG-SLUD, pages 37–42, 2011.

[5] Nigel Crook, Ramon Granell, and Stephen Pulman. Unsupervised Classification of Dialogue Acts using a Dirichlet Process Mixture Model. Proceedings of SIGDIAL, pages 341–348, 2009.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. in Journal of Machine Learning Research, pages 1107–1135, 2003.