Dialogue Act Classification in Reference Interview Using Convolutional Neural Network with Byte Pair Encoding

Seiya Kawano, Koichiro Yoshino, Yu Suzuki, and Satoshi Nakamura

Abstract Dialogue act classification is an important component of dialogue management, which captures the user's intention and chooses the appropriate response action. In this paper, we focus on the dialogue act classification in reference interviews to model the behaviors of librarians in the information seeking dialogues. Reference interviews sometimes include rare words and phrases. Therefore, the existing approaches that use words as units of input often do not work well here. We used the byte pair encoding compression algorithm to build a new vocabulary for the inputs of the classifier. By using this new unit as a feature of the convolutional neural network-based classifier, we improved the accuracy of the dialogue act classification while suppressing the size of vocabulary.

1 Introduction

Requests from users for information retrieval systems are often ambiguous, so this property makes it difficult to provide the exact information related to the real demand of a user in the information seeking process [16, 5]. It is known that clarifications such as "confirmation" or "asking for background information" help to find the requested information. A reference interview, a chat-style information-seeking dialogue at the reference service in a library, is an example of information seeking dialogue with these clarification action. Conducting this kind of reference interview in advance improves the accuracy of information provision in the reference service[7, 6]. We focused on reference interview for modeling the librarians' behavior to create a system that can provide information through interactions, even if the intention of the user at the first utterance is ambiguous.

Seiya Kawano, Koihciro Yoshino, Yu Suzuki, Satoshi Nakamura
Graduate School of Information Science, Nara Institute of Science and Technology, Takayama-cho, Ikoma, Nara, 6300192, e-mail: kawano.seiya.kj0@is.naist.jp

To model the response strategy, we focused on the task of dialogue act classification in the reference interview by using Inoue's dialogue act tag set [3]. We constructed classifiers with convolutional neural network (CNN), known as the state-of-the-art classifier using statistical methods [11]. To model the classifier with neural networks, we need enough training data with labels, which is difficult to obtain because the number of labeled dialogue data is limited. A trained model sometimes does not work well due to the lack of training data for rare and unusual words.

Subword approach is known as for reducing this problem [4]. On the other hand, information-seeking dialogues include phrases such as "May I help you?" and "Hold on, please". Using word units for this kind of expressions wastes the feature space and decrease the accuracy of classification. We implemented byte pair encoding (BPE) compression algorithm for effective use of the CNN feature space. Our investigation confirmed that the BPE-based features improve the accuracy of the dialogue act classification while suppressing the size of the vocabulary to be used.

2 Reference Interview in Libraries

Librarians in libraries provide documents that may contain answers to the user's questions. They try to clarify the information requirements of the user through a reference interview by asking about the subject, background, purpose, and motivation [8]. This kind of dialogue attracts high attention in the field of dialogue system research and is known as "information navigation" [16].

2.1 Corpus

We use chat logs of a virtual reference service QuestionPoint as the corpus [7, 6]. This corpus consists of 700 sessions of 12,634 utterances. Meta-data labels of the participants, dates, and times are given for each utterance. Personal information (user's name, email address, etc.) are anonymized.

Table 1 shows an example dialogue of the virtual reference interview. In this example, the librarian clarifies users background information (examination at school the next day, the grade of the user, search histories, etc.).

2.2 Dialogue Act in Reference Interview

In dialog systems, it is impractical to define comprehensive behaviors of the system by rules. Recent works tackle this problem with data-driven approaches, which learn behaviors of the system from dialogue corpora with statistical methods such as reinforcement learning [17, 15]. However, a data-driven approach requires very large-

Table 1 Example of virtual reference interview.

ID	Utterance
P1	here is a current in a metal wire due to the motion of electrons. sketch a possible path for the motion of a single electron in this wire, the direction of the electric field vector, and the direction of conventional current.
P2	you can just describe what they would look like
L3	Just a moment, please....
P4	Thanks
L5	Is this for a school assignment and if so what is your grade level?
P6	Im a junior in high school... its for a physics class... i have a test tomorrow and this stuff and Im still shakey on it
L7	What part of your physics books this question comes from: electricity?
P8	ya
L9	Let me check
L10	Hold on please
P11	ok
L12	I am still checking
L13	Hold on please
L14	http://www.swansontec.com/set.htm
L15	The source that I just sent has good graphics that shows the electic currents
L16	And the graphic is animated so you can see the movement
L17	Can you see the page?
P18	yes
L19	Let me check for more hold on please

scale datasets [16]. Using dialogue act is known to avoid this problem. Dialogue acts are defined as tags that indicate the intention of each utterance in dialogues [13].

In the dialogue acts of reference interviews defined by Inoue [3], librarians and users have two dialogue act categories to process the interview: 1) information transfer to request or provide information and 2) task management to assign or commit to tasks.

They also have two other dialogue act categories for smooth communication: 3) social relationship management to manage socio-emotional aspects of communication, and 4) communication management to manage physical aspects of communication. These four fundamental categories of dialogue acts are called dialogue act functions (DAF). They have detailed tags to model the behavior of participants, which is called dialogue act domain (DAD). The detail of the dialogue act definition are shown in Table 2.

3 Dialogue Act Classification for Reference Interview

A reference interview is an open-domain task, so the dialogues often contain out of vocabulary (OOV) words and low-frequency words. It is very difficult to train a good statistical model to classify utterances into dialogue act classes if there are

Table 2 Dialogue act tags in reference interview.

Dialogue Act Function (5 classes)	Dialogue Act Domain (19 classes)
Information Transfer	Information Problem
- Infomration Provision	Search Process
- Infomration Request	Information Object
	Feedback
	Other
Task Management	Librarian's Task
	User's Task
	Other
Social Relationship Management	Greeting
	Valediction
	Exclamation
	Apology
	Gratitude
	Downplay
	Closing Ritual
	Rapport Building
Communication Management	Channel Checking
	Pausing
	Feedback

many OOVs. Furthermore, the feature space will be wasted on some typical and frequent expressions if we use conventional word-based features. Such expressions can be compressed into one dimension of the feature vector. Therefore, we trained a domain-dependent tokenizer based on BPE, which is optimized with entropy, from the reference interview corpus to make efficient inputs for the dialogue act classifier.

3.1 Byte Pair Encoding

Byte pair encoding (BPE) is a simple form of data compression that recursively concatenates frequent consecutive symbols into one symbol to reduce the entropy [2, 12]. Symbols (vocabularies) defined in BPE for texts start from a set of characters. Thus, it can reduce the number of low-frequency words that often are OOVs in the test-set. On the other hand, BPE can create a long symbol if the set of characters is frequent. Although BPE was originally proposed in the field of data compression, Sennrich et al. [9] applied BPE to create a vocabulary for neural machine translation in order to reduce the number of OOVs. They also reported that reducing of the number of OOVs improved the bilingual evaluation understudy (BLEU) score of machine translation.

In this paper, BPE is regarded as a domain-dependent feature extractor and trained as a tokenizer to create a new unit. We need to give a size of vocabulary before the BPE training, which is decided by the number of initial symbols (= characters) and the number of merge operations of BPE. While Sennrich et al. [9] considered subwords in each word, we considered spaces as one token and trained the

tokenizer to extract not only subwords but also set phrases. We used SentencePiece[1] as the implementation of byte pair encoding.

3.2 Dialogue Act Classification Using Convolutional Neural Network

We used convolutional neural networks (CNN) [10] for dialog act classification. We tokenized each utterance into sequences of BPE units and made a matrix for each utterance as shown in Figure 1. Each unit was converted into a fixed length embedding vector. These vectors were placed as columns according to the sequence of the original units. We used 0-padding because the number of columns was set as the maximum number for units of one utterance. Our CNN consists of one convolution and global max-pooling layer, four hidden layers and a softmax output layer. Batch normalization was set for each layer, and RMSProp was used as the optimizer. The initial value of the word embedding vector was set randomly.

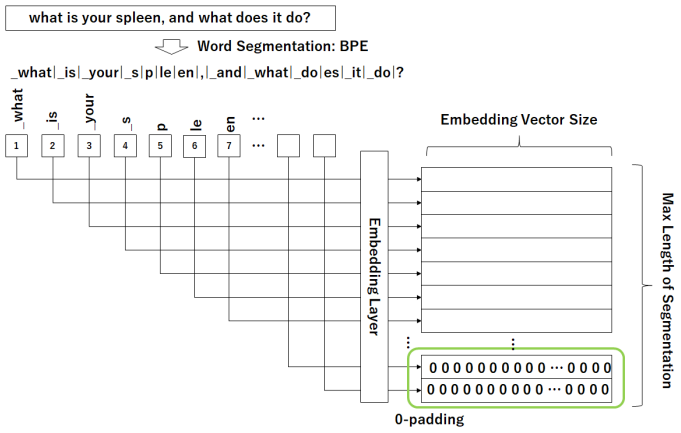

Fig. 1 The Input Generation to CNN.

[1] http://github.com/google/sentencepiece

4 Experimental Evaluation

4.1 Experimental Settings

In the Online Computer Library Center (OCLC) virtual reference interview dataset, 200 sessions are annotated with dialogue acts (200 labeled sessions) by Inoue [3] and the other 500 sessions do not have any annotations of dialogue acts (500 unlabeled sessions). In the dialogue act classification, we use 200 labeled session of 5,327 utterances with 10-fold cross-validation. The remaining 500 session of 7,307 utterances are used to learn the tokenizer and embed of the trained BPE units.

We examined the CNN-based classifier with the BPE units under several vocabulary settings. We prepared CNN based classifiers with word units and character units for comparison. The optimal parameters of CNN adopted for the experiment are listed in Table 3.

Table 3 Parameter settings for CNN.

Arguments	**Hyper-parameter**
Dimension of word embedding	50, 100
Max length of segmentation	300, 100, 50
Number of filters	128 256
Kernel size of convolution	2 3 5
Stride length of convolution	1

We also prepared classifiers based on multi-layer perceptron (MLP) and random forests (RF) algorithm by using features in previous works [18, 14]: 1) Bag-of-words (BoW), 2) Bag-of-bigrams (BoB), 3) Text segmentation length, 4) Speaker type (librarian of user), and 5) Message position in the dialogue.

4.2 Experimental Results & Discussions

Table 4 summarizes the results of the dialogue act classification in each setting. We targeted from 100 to 1000 as the vocabulary size of each BPE. However, the vocabulary is trained on difference dataset (500 unlabeled sessions), and some words were unseen in the training data (200 labeled sessions) of the classifier. Vocab is the size of the vocabulary, DAF and DAD are the accuracies of labeling for each category, OOV is the average number of OOVs in cross-validation. Maximum and Average Word Length means the maximum and the average of the units used in the classifier. As seen in Table 5, the BPE reduced the number of OOVs compared to the word-based methods, although there are fewer BPE units than words. In the following text segmentation example of BPE, if the vocabulary size is 100, tokens are similar to characters and only frequent words are tokenized. Frequently words (such

Table 4 Accuracies different dialogue act classifications

Methods	Vocab	DAF	DAD	OOV	Maximum Word Length	Average Word Length	Average Text Length
BPE-unit-level CNN	97	0.8601 *	0.7175	0.0	4	1.5	60.6
BPE-unit-level CNN	**197**	**0.8684 *** **	**0.7256 ***	0.0	10	2.5	46.35
BPE-unit-level CNN	295	0.8622 ***	0.7209 *	0.0	10	2.8	41.1
BPE-unit-level CNN	395	0.8620 **	0.7130	0.0	12	3.2	37.8
BPE-unit-level CNN	494	0.8570 **	0.7122	0.0	13	3.4	35.8
BPE-unit-level CNN	592	0.8585 **	0.7141	0.2	13	3.5	34.3
BPE-unit-level CNN	686	0.8585	0.7092	0.4	13	3.7	33.2
BPE-unit-level CNN	784	0.8556	0.7091	0.7	13	3.8	32.3
BPE-unit-level CNN	881	0.8536	0.7040	1.2	13	3.8	31.5
BPE-unit-level CNN	977	0.8541	0.7046	1.6	13	4.0	30.9
Character-level CNN	67	0.8538	0.7124	0.0	1	1.0	75.8
Word-level CNN	6091	0.8438	0.6937	333.9	80	6.9	16.5
Word-level LSTM	6091	0.8286	0.6745	333.9	80	6.9	16.5
MLP(BoW + BoB)	6091	0.8498	0.7119	333.9	80	6.9	16.5
MLP(All features)	6091	0.8515	0.7145	333.9	80	6.9	16.5
RF(BoW + BoB)	6091	0.8292	0.6790	333.9	80	6.9	16.5
RF(All features)	6091	0.8367	0.7008	333.9	80	6.9	16.5

paired t-test with MLP (All feature) :* p < 0.05** p < 0.01*** p < 0.001

as “how”, “help”, “information”, etc.) are tokenized according to the increasing vocabulary size.

- **Original utterance:** do you want information on pilot mountain or rock climbing? how can i help you?
- **Vocab size = 100:** _d o _you _w an t _i n f or m at i on _ on _p i l o t _m ou n t a in _ or _ r o c k _c l i m b ing ? _h o w _c an _i _h e l p _you ?
- **Vocab size = 500:** _do _you _w ant _information _on _p il ot _m ount ain _or _ ro c k _c li m b ing ? _how _can _i _help _you ?
- **Vocab size = 1000:** _do _you _want _information _on _p il ot _m ount ain _or _ ro ck _c li m b ing ? _how _can _i _help _you ?

In the dialogue act classification results, the accuracy of dialogue act classification was improved by the proposed BPE-unit-level CNN on each vocabulary size from 100 to 1000 compared to other models, even if they do not use additional information such as the role of the speaker or appearance position in dialogue. The word and Character-level CNN did not show better performance compared to the conventional method based on MLP.

In respects to DAF, the proposed methods estimated the dialogue acts with high accuracy. However, accuracies of DAD were not high enough, being only 0.7256 in the best condition. This can be improved by considering some additional information such as dialogue history. We also analyzed misclassified examples and found that there were some ambiguities caused by the annotation. Some utterances of par-

ticipants had several roles, but the original annotation scheme did not allow to annotate multiple dialogue acts to one utterance. Such ambiguity of annotations should be eliminated to improve annotation. Below are some examples of utterances that should have multiple dialogue acts.

- thank you so much. this looks great. can you find any reasons why tea would do this? (Social:Gratitude, Info:Problem)
- ”Name”, welcome to maryland askusnow! i’m looking at your question right now; it will be just a moment. (Social:Greeting, Task:Librarian, Comm:Pausing)

As a solution to the problem, we can introduce the ISO24617-2 dialogue act annotation scheme [1]. The scheme has general-purpose functions (GPF) for utterances that control the contents of dialogues, and domain-specific functions (DSF) that process the dialogues. This scheme allows to annotate multiple DSF tags for one utterance.

5 Conclusion

In this paper, we proposed a dialogue act classification model based on BPE tokenizer and CNN-based classifier in the reference interview. Experimental results show that the classification accuracy of the proposed model was significantly higher than that of any baseline model. Our proposed model efficiently built the input the classifier with BPE-based tokenizer. It performed better than the classifiers that use words and characters as input units. Our model performed well for the DAF category. However, improvement in the DAD category remains as a future challenge. In addition, it is necessary to investigate the effectiveness of our method in other dialogue domains, and compare the other approaches like lemmatization or word-CNN with pre-trained embedding model.

In the error analysis, we found that some problems were caused by the annotation scheme. The lack of data was also a problem, therefore, as future work, we plan to improve the number and quality of the data. By using the classifier that we proposed, we will develop a dialogue manager that models the strategy of librarians in reference interviews to help find the exact information for the user, even if the requests of the users are ambiguous.

Acknowledgements This research and development work was supported by the JST PREST(JPMJPR165B) and JST CREST(JPMJCR1513).

References

1. Bunt, H., Alexandersson, J., Choe, J.W., Fang, A.C., Hasida, K., Petukhova, V., Popescu-Belis, A., Traum, D.R.: Iso 24617-2: A semantically-based standard for dialogue annotation. In: LREC, pp. 430–437 (2012)

2. Gage, P.: A new algorithm for data compression. The C Users Journal **12**(2), 23–38 (1994)
3. Inoue, K.: An investigation of digital reference interviews: A dialogue act approach. Ph.D. thesis, Syracuse University (2013)
4. Mikolov, T., Sutskever, I., Deoras, A., Le, H.S., Kombrink, S., Cernocky, J.: Subword language modeling with neural networks. preprint (http://www. fit. vutbr. cz/imikolov/rnnlm/char. pdf) (2012)
5. Misu, T., Kawahara, T.: Dialogue strategy to clarify users queries for document retrieval system with speech interface. Speech Communication **48**(9), 1137–1150 (2006)
6. Radford, M.L., Connaway, L.S.: Seeking synchronicity: Evaluating virtual reference services from user, non-user, and librarian perspectives. Proposal for a research project, submitted February **1**, 2005 (2005)
7. Radford, M.L., Connaway, L.S., Confer, P.A., Sabolcsi-Boros, S., Kwon, H.: " are we getting warmer?" query clarification in live chat virtual reference. Reference & User Services Quarterly pp. 259–279 (2011)
8. Ross, C.S., Radford, M.L., Nilsen, K.: Conducting the reference interview: a how-to-do-it-manual for librarians. Neal-Schuman Publishers, Inc. (2009)
9. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)
10. Severyn, A., Moschitti, A.: Unitn: Training deep convolutional neural network for twitter sentiment classification. In: SemEval@ NAACL-HLT, pp. 464–469 (2015)
11. Shi, H., Ushio, T., Endo, M., Yamagami, K., Horii, N.: A multichannel convolutional neural network for cross-language dialog state tracking. In: Spoken Language Technology Workshop (SLT), 2016 IEEE, pp. 559–564. IEEE (2016)
12. Shibata, Y., Kida, T., Fukamachi, S., Takeda, M., Shinohara, A., Shinohara, T., Arikawa, S.: Speeding up pattern matching by text compression. Algorithms and Complexity pp. 306–315 (2000)
13. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational linguistics **26**(3), 339–373 (2000)
14. Webb, N., Hepple, M., Wilks, Y.: Dialogue act classification based on intra-utterance features. In: Proceedings of the AAAI Workshop on Spoken Language Understanding, vol. 4, p. 5 (2005)
15. Williams, J.D., Young, S.: Partially observable markov decision processes for spoken dialog systems. Computer Speech & Language **21**(2), 393–422 (2007)
16. Yoshino, K., Kawahara, T.: Conversational system for information navigation based on pomdp with user focus tracking. Computer Speech & Language **34**(1), 275–291 (2015)
17. Young, S., Gašić, M., Thomson, B., Williams, J.D.: Pomdp-based statistical spoken dialog systems: A review. Proceedings of the IEEE **101**(5), 1160–1179 (2013)
18. Yu, B., Inoue, K.: An investigation of digital reference interviews: Dialogue act annotation with the hidden markov support vector machine. OCLC/ALISE research grant report published electronically by OCLC Research (2012)