Improve the chatbot performance for the DB-CALL system using a hybrid method and a domain corpus

Jin-Xia Huang[1], Oh-Woog Kwon[2],
Kyung-Soon Lee[3], and Young-Kil Kim[4]

Abstract. This paper presents a chatbot for a Dialogue-Based Computer Assisted Language Learning (DB-CALL) system. The chatbot helps users learn language via free conversations. To improve the chatbot performance, this paper adopts a Neural Machine Translation (NMT) engine to combine with an existing search-based engine, and also extracts a small domain corpus for the topics of the DB-CALL system so that the chabot's responses could be more related to the conversation topics. As a result of user evaluations, the performance of the chatbot was improved by using hybrid methods, achieving performance comparable to existing systems. The automatically extracted domain corpus has little help or even declines the chatbot performance as an auxiliary module of the DB-CALL system.

Keywords: DB-CALL, chatbot.

1. Introduction

We have developed a DB-CALL system, GenieTutor, to help English language learners in Korea (Kwon, Kim, & Lee, 2016). Similar to other DB-CALL systems, GenieTutor asks questions on different topics according to given scenarios, and the learners answer questions to practise what they learned. In order to allow the user to communicate more freely with the system, we developed a search-based chatbot to assist GenieTutor. Chatbot normally indicates an open-domain dialogue system for chitchat, which deals with the out of topic conversations in GenieTutor. However,

1. Electronics and Telecommunications Research Institute, Daejeon, Korea; hgh@etri.re.kr
2. Electronics and Telecommunications Research Institute, Daejeon, Korea; ohwoog@etri.re.kr
3. Chonbuk National University, Jeonju, Korea; selfsolee@chonbuk.ac.kr
4. Electronics and Telecommunications Research Institute, Daejeon, Korea; kimyk@etri.re.kr

How to cite this article: Huang, J.-X., Kwon, O.-W., Lee, K.-S., & Kim, Y.-K. (2018). Improve the chatbot performance for the DB-CALL system using a hybrid method and a domain corpus. In P. Taalas, J. Jalkanen, L. Bradley & S. Thouësny (Eds), *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018* (pp. 100-105). Research-publishing.net. https://doi.org/10.14705/rpnet.2018.26.820

the student satisfaction on the free-talking was lower than our expectations (Huang, Lee, Kwon, & Kim, 2017).

This paper describes how we improved the chatbot performance. We first implemented the hybrid chatbot by introducing the NMT engine, combining it with the search-based engine, and then extracted the small domain corpus for the topics of the DB-CALL system to improve the chatbot's performance as an auxiliary system.

2. Hybrid chatbot based on search and NMT engines

Last year, we developed a chatbot using the search engine Indri (Strohman, Metzler, Turtle, & Croft, 2005). It retrieves similar examples from dialogue corpuses which contain 410 thousand dialogue examples. A dialogue example consists of two utterances: one query utterance and one system response, which is also called one turn in a dialogue. If there is no similar example, the chatbot outputs random utterances to the user (Huang et al., 2017).

This year, we introduced an NMT engine OpenNMT (Klein et al., 2017) to generate responses if the search engine fails to get a similar example. The corpus for the NMT engine contains 1.4 million dialogue examples, which are from MovieDic (Banchs, 2012), BNC corpus[5], and our own dialogue corpus which has been built in the last decades.

A user evaluation involving 20 English learners was performed. The learners are the users of the DB-CALL system. They are asked to talk freely with the chatbot for 60 turns, and we got 1,211 user utterances in total. After chatting with the chatbot, the users assign 0 to 2 points to each response from the chatbot: 2 means the response is acceptable and satisfactory, 1 means it is acceptable but too general, 0 means the response is wrong. We evaluated the system with a percentage of responses that gained 1 and 2 points, and called it the acceptance rate.

Table 1 shows that comparing with the search engine (the first column), the NMT engine (the second column) gains a higher acceptance rate (72.15%>60.55) but lower satisfaction (2 points: 18.32%<34.74%). It means these two engines complement each other, and so a hybrid approach can help improve performance.

5. http://www.natcorp.ox.ac.uk/

As a result, the acceptance rate of the hybrid engine is 68.29% (the third column), which is much higher than 52.78% of the previous year (Huang et al., 2017).

Table 1. Evaluation on the hybrid chatbot compared with Siri and Cleverbot

	Our chatbot			**Siri**	**Cleverbot**
Score	**Search engine**	**NMT engine**	**Hybrid engine**		
2	34.74%	18.32%	23.78%	22.96%	39.31%
1	25.81%	53.84%	44.51%	37.57%	17.01%
0	39.45%	27.85%	31.71%	39.47%	43.68%
Acceptance rate (>1)	60.55%	72.15%	68.29%	60.53%	56.32%

For comparative evaluation, the user utterances are also input to Siri and Cleverbot to get their responses. Siri is a task-oriented dialogue system which allows chitchat. Cleverbot is a chatbot which has been in online service for about 20 years and contains 150 million dialogue examples[6]. Table 1 shows that the accept rate of our hybrid chatbot is higher than both Siri and Cleverbot, which is quite encouraging considering the time and cost invested.

3. Extracting the domain corpus for topic conversations

According to our experiments in the last year, the satisfaction on the chatbot as an auxiliary module of the DB-CALL system was much lower than that of the independent chatbot. We assumed that the performance could be raised if the chatbot responses could be more related to the given topics in DB-CALL systems (Huang et al., 2017). In this paper, we extracted a small domain corpus for the topics 'ordering food' and 'city tour' of the DB-CALL system to see if it helps.

To extract the domain corpus from the chatbot corpus, we firstly used the domain and topic labels of the examples. There are 156 thousand examples of domain labels like *study*, *business*, and *travel-meal*; and 39 thousand of them have more detailed topic labels like *reservation*, *cancel*, and *ordering*. Secondly, we extracted domain examples according to the domain weights they gain: the weight is directly proportional to the number of domain keywords in the example, and is

6. https://en.wikipedia.org/wiki/Cleverbot

inversely proportional to the example length. As the result, about 4.5 thousand and 2.8 thousand examples are extracted for 'ordering food' and 'city tour' topics, respectively.

The search engine corpus is separated into two parts. The search engine searches the small domain corpus before it searches the general corpus. The similarity threshold for in-domain search is lower than the one for general domain search, and the in-domain examples gain higher priorities when the similarities are the same with general domain examples.

The same 20 English learners were asked to have conversations with the DB-CALL system and finish the tasks like 'ordering food' or 'buying city tour tickets'. Free talking was allowed in the conversations. As results, we got 115 out of topic utterances which were replied to by the chatbot. The responses with and without the domain corpus were both produced for comparison.

Table 2 shows that the search engine gives responses to 59.13% of the user utterances without the domain corpus, and it is improved to 63.48% with the extracted domain corpus. The acceptance rate of the search engine is also improved from 30.43% to 32.17% with the domain corpus.

Table 2. Evaluation on the chatbot as an auxiliary module in the DB-CALL system

	Coverage (search engine)	**Acceptance rate (search engine)**	**Acceptance rate (hybrid engine)**
Without domain corpus	59.13%	30.43%	41.74%
With domain corpus	63.48%	32.17%	40.00%

However, the acceptance rate with the hybrid engine is rather declined from 41.74% to 40.00%. One reason is according to the hybrid approach – most of the similar examples tend to be matched whether they are in-domain examples or not. The coverage is more improved by less similar examples, which improves the acceptance rate of the search engine, but the opportunity decreases for the NMT engine to generate more acceptable responses. It causes the overall acceptance rate to drop in the hybrid engine.

The other reason is that a DB-CALL system is supposed to play different roles in different topics. For example, the system should act as a waiter in the 'ordering food' domain. Following, the chatbot response is considered wrong although it is an in-domain response but more with a role of accompanying guests. Therefore, role

information should be considered in addition to domain information in extracting a domain corpus.

System (DB-CALL): What would you like to order?

User: I'll have the sandwich the man is eating.

System (chatbot): I'll have that too.

4. Conclusion

This paper presented a chatbot which combined an NMT engine with a search engine, and the evaluation showed that the hybrid approach improved the chatbot performance. We also extracted the domain corpus for the out of topic conversations in the DB-CALL system. The evaluation showed that, unlike the search-based engine, the performance declined in the hybrid engine. A brief discussion was held, and it seems more in-depth research is required in the future to improve the performance of the chatbot as an auxiliary module of the DB-CALL system.

5. Acknowledgements

This work was supported by the ICT R&D program of MSIP/IITP [2015-0-00187, Core technology development of the spontaneous speech dialogue processing for the language learning], and Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [18ZS1140 Conversational AI Core Technology Research].

References

Banchs, R. E. (2012). MovieDic: a movie dialogue corpus for research and development. *Proceedings of ACL*, 203-207

Huang, J.-X., Lee, K.-S., Kwon, O.-W., & Kim, Y.-K. (2017). A chatbot for a dialogue-based second language learning system. In K. Borthwick, L. Bradley & S. Thouësny (Eds), *CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017* (pp. 151-156). Research-publishing.net. https://doi.org/10.14705/rpnet.2017.eurocall2017.705

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). OpenNMT: open-source toolkit for neural machine translation, *CoRR*, abs/1701.02810. http://arxiv.org/abs/1701.02810

Kwon, O.-W., Kim, Y.-K., & Lee, Y. (2016). Task-oriented spoken dialog system for second-language learning. In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 237-242). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.568

Strohman, T., Metzler, D., Turtle, H., & Croft, W. B., (2005). Indri: a language model-based search engine for complex queries. *Proceedings of the International Conference on Intelligence Analysis*.

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net

Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018
Edited by Peppi Taalas, Juha Jalkanen, Linda Bradley, and Sylvie Thouësny

Publication date: 2018/12/08

Typeset by Research-publishing.net
Cover theme by © 2018 Antti Myöhänen (antti.myohanen@gmail.com)
Cover layout by © 2018 Raphaël Savina (raphael@savina.net)
Drawings by © 2018 Linda Saukko-Rauta (linda@redanredan.fi)

ISBN13: 978-2-490057-22-1 (Ebook, PDF, colour)
ISBN13: 978-2-490057-23-8 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-21-4 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, UK: British Library.
Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: Décembre 2018.