Benben: A Chinese Intelligent Conversational Robot

Wei-Nan Zhang, Ting Liu, Bing Qin, Yu Zhang, Wanxiang Che,
Yanyan Zhao, Xiao Ding
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology
{wnzhang,tliu,qinb,zhangyu,wxche,yyzhao,xding}@ir.hit.edu.cn

Abstract

Recently, conversational robots are widely used in mobile terminals as the virtual assistant or companion. The goals of prevalent conversational robots mainly focus on four categories, namely chit-chat, task completion, question answering and recommendation. In this paper, we present a Chinese intelligent conversational robot, Benben, which is designed to achieve these goals in a unified architecture. Moreover, it also has some featured functions such as diet map, implicit feedback based conversation, interactive machine reading, news recommendation, etc. Since the release of Benben at June 6, 2016, there are 2,505 users (till Feb 22, 2017) and 11,107 complete human-robot conversations, which totally contain 198,998 single turn conversation pairs.

1 Introduction

The research of conversational robot can be traced back to 1950s, when Alan M. Turing presented the Turing test to answer the proposed question "*Can machine think?*" (Turing, 1950). It then becomes an interesting and challenging research in artificial intelligence. Conversational robot can be applied to many scenarios of human-computer interaction, such as question answering (Crutzen et al., 2011), negotiation (Rosenfeld et al., 2014), e-commence (Goes et al., 2011), tutoring (Pilato et al., 2005), etc. Recently, with the widespread of mobile terminals, it is also applied as the virtual assistant, such as Apple Siri[1], Microsoft Cortana[2], Facebook Messenger[3], Google Assistant[4], etc., to make users acquire information and services with the terminals more conveniently.

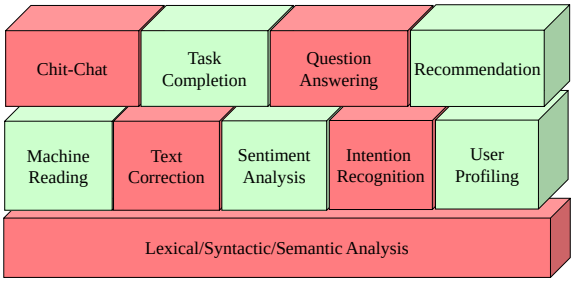

Figure 1: The technical structure of Benben.

The goals of the prevalent conversational robots can be grouped into four categories. First is chit-chat which is usually designed for responding greeting, emotional and entertainment messages. Second is task completion aiming to assist users to complete some specific tasks, such as restaurant and hotel reservation, flight inquiry, tourist guide, web search, etc. Third is question answering that is to satisfy the need of information and knowledge acquisition. Fourth is recommendation that can actively recommend personalized content through the user interest profiling and conversation history. Despite the success of the existing conversational robots, they tend to only focus on one or several goals and hardly achieve all of them in a unified framework.

In this paper, we present a Chinese intelligent conversational robot, Benben, which is based on massive Natural Language Processing (NLP) techniques and takes all of the goals in design.

Figure 1 shows the technical structure of Benben. The bottom layer contains the basic techniques of NLP, such as Chinese word segmentation, part-of-speech tagging, word sense disambiguation, named entity recognition, dependency parsing, semantic role labelling and semantic de-

[1] https://en.wikipedia.org/wiki/Siri
[2] https://en.wikipedia.org/wiki/Cortana_(software)
[3] https://www.messenger.com/
[4] https://assistant.google.com/

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 13–18
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-4003

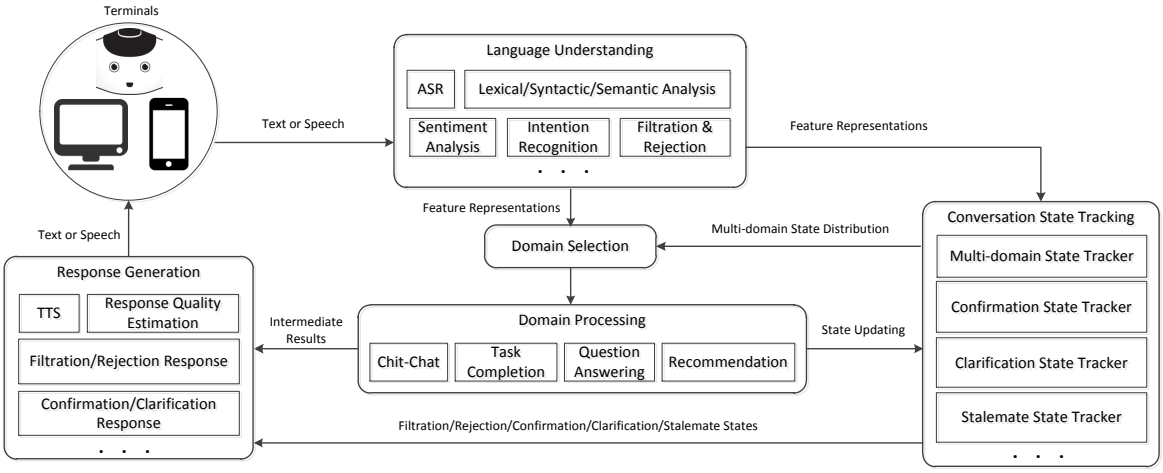

Figure 2: The simplified architecture of Benben.

pendency parsing, etc. These techniques are from the language technique platform (LTP)[5]. The middle layer includes the core techniques that are supported by the basic NLP techniques. The top layer is the four functionalities of Benben.

2 Architecture

In this section, we will introduce the architecture of Benben. Figure 2 shows the simplified architecture of Benben. It mainly consists of four components: 1) language understanding, 2) conversation state tracking, 3) domain selection and processing and 4) response generation. As can be seen, the architecture of Benben can be corresponded to the classic architecture of spoken dialogue systems (Young et al., 2013). Concretely, the natural language understanding, dialogue management and natural language generation in spoken dialogue systems are corresponding to the 1), 2) and 3), 4) components of the Benben architecture, respectively. We will next detail each component in the following sections.

2.1 Language Understanding

The user input can be either text or speech. Therefore, the first step is to understand both the speech transcription and text. In Benben, the LTP toolkit (Che et al., 2010) is utilized to the basic language processing, including Chinese word segmentation, part-of-speech tagging, word sense disambiguation, named entity recognition, dependency parsing, semantic role labelling and semantic dependency parsing. The results of these processing are finally taken as the lexical, syntactic and semantic features and transferred to different representations to the following processing steps.

We obtain the results of sentence-level sentiment analysis by using our proposed approach (Tang et al., 2015). The results are then used in two aspects. One is to directly generate consoling responses and the other is to take the sentiment as an implicit feedback of users to optimize the long-term goal of conversations.

We utilize the proposed weakly-supervised approach (Fu and Liu, 2013) to recognize the intention of users. User intention can be either used as clues for response generation or features for domain selection. For example, if a user says:"*I want to go to Beijing.*", he/she may want to book an airplane or train ticket or further reserve a hotel room in Beijing.

We also design a scheme to filter out the sentences that contain vulgar, obscene or sensitive words. A classifier is trained to automatically identify these sentences with manually collated lexicons. Meanwhile, the rejection scheme is also needed as Benben should cope with the inputs that are out of its responding scope.

2.2 Conversation State Tracking

After the language understanding step, an input sentence is transferred to several feature representations. These feature representations are then taken as the inputs of the conversation state tracking and domain selection. The conversation state

[5]http://www.ltp-cloud.com

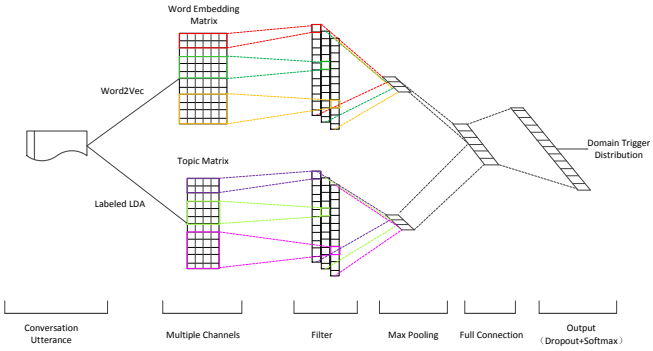

Figure 3: The framework of the proposed topic augmented convolutional neural network for domain selection.

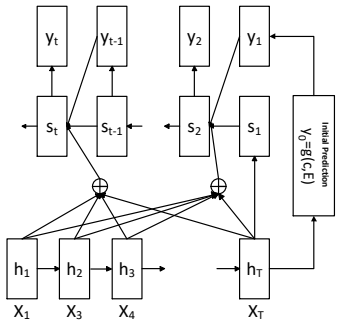

Figure 4: The framework of the proposed LTS model for response generation.

tracker records the historical content, the current domain and the historically triggered domains, the sequences of the states of confirmation, clarification, filtration, rejection, etc., and their combinations. Given the feature representations of an input, the multi-domain state tracker will produce a probability distribution of the states over multiple domains, which is then used to domain selection. The trackers of confirmation, clarification, filtration, rejection, etc., estimate their triggered probabilities, respectively. These probabilities are directly sent to the response generation as their current states or triggered confidences. It is worth noting that the conversations may come to a stalemate state, which indicates that the users are not interesting to the current conversation topic or they are unsatisfied to the responses generated by Benben. Once the stalemate state is detected, Benben will transfer the current conversation topic to another topic to sustain the conversation.

Meanwhile, as can be seen from Figure 2, there is an iterative interaction among conversation state tracking, domain selection and domain processing. The interactive loop denotes that the state tracking module provides the current state distribution of multiple domains for the domain selection. The triggered domains are then processing to update the conversation state as well as generate the intermediate results for response generation.

2.3 Domain Selection

The domain selection module is to trigger one or more domains to produce the intermediate results for response generation. It takes the feature representations from language understanding and the current multi-domain state distribution from conversation state tracking as inputs and estimates the triggered domain distribution using a convolutional neural network.

In Benben, we proposed a topic augmented convolutional neural network to integrate the continuous word representations and the discrete topic information into a unified framework for domain selection. Figure 3 shows the framework of the proposed topic augmented convolutional neural network for domain selection. The word embedding matrix and the topic matrix are obtained using the word2vec[6] and Labeled LDA (Ramage et al., 2009), respectively. The two representations of the input conversation utterance are combined in the full connection layer and output the domain triggered distribution. At last, the domains whose triggered probabilities are larger than a threshold are selected to execute the following domain processing step. Note that after the domain selection step, there may be one or more triggered domains. If there is no domain to be triggered, the conversation state is updated and then sent to the response generation module.

2.4 Domain Processing

Once a domain is selected, the corresponding processing step is triggered. We will next details the processing manners of the four domains.

Chit-Chat

The chit-chat processing consists of two components. First is the retrieval based model for generating chit-chat conversations. Here, we have indexed 3 million single turn post-response pairs collected from the online forum and microblog

[6]https://code.google.com/archive/p/word2vec/

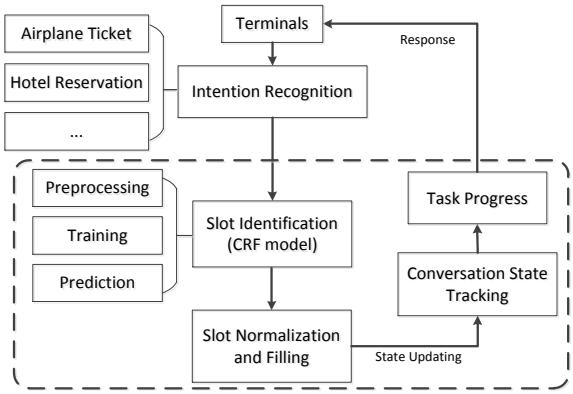

Figure 5: The process of task completion for each sub-domain.

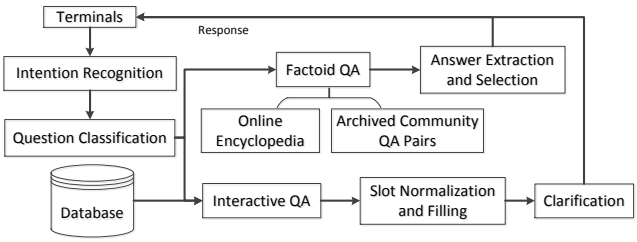

Figure 6: The process of the question answering in Benben.

conversations, using Lucene toolkit[7]. Second, the use of "<EOS>" to initialize the sequence to sequence (Seq2Seq) learning based response generation models usually leads to vague or non-committal responses, such as "*I don't know.*", "*Me too.*", etc. To address the problem, we used an optimized approach (Zhu et al., 2016), namely learning to start (LTS) model, to utilize a specific neural network to learn how to generate the first word of a response. Figure 4 is the framework of the proposed LTS model. 5 million post and comment pairs that are released by the short text conversation in NTCIR-12[8] are used to train the LTS model.

Task Completion

The domain of task completion also has sub-domains, such as restaurant and hotel reservation, airplane and train ticket booking, bus and metro line guidance, etc. For each sub-domain, the task completion process is shown in Figure 5. As can be seen, after recognizing the user intention, a conditional random field (CRF) model is utilized to identify the values, in the user input, to fill the semantic slots according to the characteristics of the sub-domain. For the same semantic slot, there may be different forms of values can be filled in. Therefore, we also proposed a value normalization scheme for the semantic slots. The conversation state is then updated after a slot has been filled. In a task progress, the task completion is an interactive process between terminals and users so that it needs a multi-turn conversation controller. In Benben, the multi-turn conversation is jointly controlled by the conversation state tracking and domain selection. The domain alternation is implemented by the confirmation and clarification state trackers as well as the response generation.

Question Answering

The question answering (QA) domain has two modes, namely factoid QA and interactive QA. After the intention recognition of user input, the question classification module routes the user questions into the factoid QA or interactive QA. For the factoid QA, we retrieve the candidate paragraphs, sentences, infobox messages from the online encyclopedia, and QA pairs from a large scale archived community QA repository, which are collected from a community QA website. As Benben currently processes the Chinese conversations, the Baidu Encyclopedia[9] and Baidu Zhidao[10] are selected as the online encyclopedia and the community QA website, respectively. The answer extraction module then extract the candidate answers from the retrieved paragraphs, sentences, infobox messages and QA pairs. The answer selection is to rank the candidate answers and the top 1 answer is sent to the user as a response. The interactive QA is similar to the task completion and they share the common processes of slot normalization, slot filling and clarification. An example interactive QA is the weather forecast and inquiry as the weather is related to the date and location. Figure 6 shows the process of the question answering in Benben.

Recommendation

The recommendation in Benben has two functions. The first is to satisfy the users' information need on specific content, such as news. The second is to break the stalemate in conversation.

[7] http://lucene.apache.org/core/
[8] http://ntcir12.noahlab.com.hk/stc.htm
[9] https://baike.baidu.com/
[10] https://zhidao.baidu.com/

Taking the news recommendation as an example, Benben can respond to the requirement of news reading in some specific topics, such as sports, fashion, movie, finance, etc. For example, users say: "*Show me the most recent news about movies.*" as well as they can also say "*Once more*" to see another movie news. Besides the querying mode, when a stalemate is detected during a conversation, Benben will recommend a recent news, according to the user profiling information, by a random alternation to the conversation topic transferring to break the stalemate. Note that the news recommendation is also in an interactive way, which means that Benben will ask the user whether he/she wants to read a news of a specific topic in an euphemism way.

2.5 Response Generation

As shown in Figure 2, the response generation takes the conversation states and the intermediate results as input to generate text or speech responses to users. The filtration, rejection, confirmation and clarification responses are generated by considering the corresponding states that obtained from the conversation state tracking. The transferred topic and recommendation responses are generated to break the stalemate in conversations. It is worth noting that there may be more than one triggered domains in domain selection and processing steps. Therefore, the intermediate results may contain multiple outputs from different domains. These outputs are actually the generated responses from the corresponding domains. However, in each turn of a conversation, there is only one response that should be responded to users. Hence, the response quality estimation module is proposed to generate a unique response to users. The quality estimation process considers the states, the output confidences and the response qualities of the domains. For example, if the triggered probability of QA is higher than other domains and the confidence of the generated answer is larger than a threshold, the answer is more likely to be a response to users. The module will also identify an answer type to check whether the generated answer is matched to the predicted type or not. For example, the expected answer type of the question "*When was the Titanic first on?*" is "*Date*". If the QA domain outputs a location or a human name, the answer type is mismatched so that the answer should not be a response to users.

3 Featured Functions of Benben

Diet Map based Conversation: The diet map is a database that contains the geographical distribution of diet in China. It is constructed by mining the location and diet pairs from a microblog service in China, named Sina Weibo. The diet map not only includes the related diet and location information, but also distinguishes the breakfast, lunch, dinner as well as the gender of users. These aspects can be seen as the slots in conversations. Based on the diet map, we develop a function for querying the location specific diet through chatting with Benben.

Implicit Feedback based Conversation: We find that users may express their emotion, opinion, sentiment, etc., on the inputs during the conversation process. These can be seen as the implicit feedback from users. We thus explore the implicit feedback in the conversation to optimize the long-term goal of conversation generation and model the implicit feedback as a reward shaping scheme towards a basic reward function in a reinforcement learning framework.

Interactive Machine Reading: Given a document or a paragraph about a specific topic or event, Benben can continuously talk to users about the given content which is our proposed interactive machine reading function. Benben will first reads and understands the given material using the proposed approach (Cui et al., 2016) and users can ask several factoid questions according to the material content. Note that as these questions are context related, there are many anaphora and ellipsis phenomenons. We thus utilize the proposed approaches (Liu et al., 2016; Zhang et al., 2016) for anaphora and zero-anaphora resolution.

4 Implementation

There are three implementations of Benben. 1) First is the webpage version for PC or mobile phone(The link is http://iqa.8wss.com/dialogue-test). Users can open the link and type to chat with Benben in Chinese. 2) Second is the Nao robot version. We carry the Benben service in a cloud server and link a Nao robot to the service. Users thus can chat with Benben in speech. The ASR and TTS are implemented by calling the services from the voice cloud[11] of iFLYTEK[12]. Besides the conversation, the Nao robot version can

[11] http://www.voicecloud.cn/

[12] http://www.iflytek.com/en/index.html

Figure 7: The QR code of Benben in WeChat.

be also controlled by speech instructions of spoken language to execute some actions. Please see the video in Youtube[13] 3) Third, Benben is also carried in WeChat[14] App, which is the most convenient platform as it allows to chat with Benben in text and speech as well as images and emoticons. The quick response (QR) code is shown in Figure 7. Please scan the QR code using the WeChat App and chat with Benben.

5 Conclusion

In this paper, we present a Chinese conversational robot, Benben, which is designed to achieve the goals of chit-chat, task completion, question answering and recommendation in human-robot conversations. In the current version, Benben is implemented in three platforms, namely PC, mobile phone and Nao robot. In the future, we plan to apply it to other scenarios such as vehicle, home furnishing, toys, etc. Meanwhile, we plan to transfer Benben from Chinese version to English version.

Acknowledgments

The authors would like to thank all the anonymous reviewers for their insight reviews and the members of the conversational robot group of the research center for social computing and information retrieval, Harbin Institute of Technology. This paper is funded by 973 Program (No. 2014CB340503) and NSFC (No. 61502120, 61472105).

References

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: A chinese language technology platform. In *COLING*. pages 13–16.

Rik Crutzen, Gjaltjorn Y Peters, Sarah Dias Portugal, Erwin M Fisser, and J J J Grolleman. 2011. An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: An exploratory study. *Journal of Adolescent Health* 48(5):514–519.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2016. Attention-over-attention neural networks for reading comprehension .

B. Fu and T. Liu. 2013. Weakly-supervised consumption intent detection in microblogs. *Journal of Computational Information Systems* 9(6):2423–2431.

Paulo Goes, Noyan Ilk, Wei T. Yue, and J. Leon Zhao. 2011. Live-chat agent assignments to heterogeneous e-customers under imperfect classification. *ACM T-MIS* 2(4):1–15.

Ting Liu, Yiming Cui, Qingyu Yin, Shijin Wang, Weinan Zhang, and Guoping Hu. 2016. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. *CoRR* abs/1606.01603.

Giovanni Pilato, Giorgio Vassallo, Manuel Gentile, Agnese Augello, and Salvatore Gaglio. 2005. Lsa for intuitive chat agents tutoring system. pages 461–465.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*. pages 248–256.

Avi Rosenfeld, Inon Zuckerman, Erel Segalhalevi, Osnat Drein, and Sarit Kraus. 2014. Negochat: A chat-based negotiation agent. *Autonomous Agents and Multi-Agent Systems* .

Duyu Tang, Bing Qin, Furu Wei, Li Dong, Liu Ting, and Zhou Ming. 2015. A joint segmentation and classification framework for sentence level sentiment classification. *IEEE/ACM TASLP* 23(11):1750–1761.

Alan M Turing. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460.

S Young, M Gasic, B Thomson, and J. D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE* 101(5):1160–1179.

Weinan Zhang, Ting Liu, Qingyu Yin, and Yu Zhang. 2016. Neural recovery machine for chinese dropped pronoun. *CoRR* abs/1605.02134.

Qingfu Zhu, Weinan Zhang, Lianqiang Zhou, and Ting Liu. 2016. Learning to start for sequence to sequence architecture. http://arxiv.org/abs/1608.05554.

[13] https://youtu.be/wfPv9-I4q7s.

[14] https://en.wikipedia.org/wiki/WeChat