Multi-modal User Intent Classification Under the Scenario of Smart Factory (Student Abstract)

**Yu-Ching Chiu,[1] Bo-Hao Chang, [1] Tzu-Yu Chen, [1] Cheng-Fu Yang, [2] Nanyi Bi, [3]
Richard Tzong-Han Tsai, [1*] Hung-yi Lee, [2] Jane Yung-jen Hsu [4]**

[1] Dept. of Computer Science and Information Engineering, National Central University, Taiwan
[2] Dept. of Electrical Engineering, National Taiwan University, Taiwan
[3] IoX Center, National Taiwan University, Taiwan
[4] Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan
crystalchiu@g.ncu.edu.tw, jack830808@gmail.com, elisachen1221@g.ncu.edu.tw, b05901082@ntu.edu.tw,
nb333@cornell.edu, thtsai@g.ncu.edu.tw, hungyilee@ntu.edu.tw, yjhsu@csie.ntu.edu.tw

Abstract

Question-answering systems are becoming increasingly popular in Natural Language Processing, especially when applied in smart factory settings. A common practice in designing those systems is through intent classification. However, in a multiple-stage task commonly seen in those settings, relying solely on intent classification may lead to erroneous answers, as questions rising from different work stages may share the same intent but have different contexts and therefore require different answers. To address this problem, we designed an interactive dialogue system that utilizes contextual information to assist intent classification in a multiple-stage task. Specifically, our system incorporates user's utterances with real-time video feed to better situate users' questions and analyze their intent.

Introduction

Question answering (QA) systems have been widely applied in various settings, especially in smart factories, where users have access to pre-built systems of expertise and are able to ask questions based on their immediate needs.

One popular way in which a QA system works is to classify users' different intents and match them with a pre-designed classification of expertise in the knowledge base (e.g., (Todhunter, Sovpel, and Pastanohau 2014)). If a match is achieved, the user is able to get an appropriate response. Otherwise, the user may be asked to rephrase the question to provide further classification (Rinaldi et al. 2003). In other words, the performance of QA systems largely depend on how successful they are able to identify users' needs in the first place.

However, user intent classification can be challenging in at least two ways. Firstly, it may be difficult for the users to pinpoint their real needs, either because they have few clues about the context or because they have problems with verbalizing their questions. Secondly, even if the users are able to describe what they need, in a relatively more complex task, where multiple stages may exist, their intent may be misunderstood by the system and is confused with ones that share similar features or needs in other stages, leading to irrelevant answers or useless information.

Either way, user intent classification can be overwhelmingly demanding for text-only models. Our solution to issues like these is to add in additional modalities to gather more contextual information to help the QA system enhance the intent detection as well as classification. For a more detailed illustration, please see our demo here[1].

Data Collection

To model an environment where user intent may be misunderstood by the QA system, thus creating challenges for its classification process, we designed a multi-step task of assembling a Meccanoid Robot and collected data on intent/question as well as visual information on our own.

Pre-Data Collection: Wizard-of-Oz Pilot Study To investigate problems that a user may encounter in the task, we conducted a Wizard-of-Oz experiment that simulated our system, where we had a research assistant, instead of the actual QA system, help the participant with the task by answering their questions. During the assembling process, the participant and our research assistant sat in different rooms and could only communicate with each other via Skype audio chat. We manually clustered collected user questions into 21 scenarios by observation. For each scenario, we picked the most representative question and then added it to our core list of questions.

User Question Collection To collect training data for the conversational model, we employed a Human Intelligence Task (HIT) on Amazon Mechanical Turk that asked turkers to review clips from the assembly process and the core questions that were asked and had them ask similar questions. Eventually, 3,769 variations of the original 21 core questions were collected in total.

Visual Data Collection For training data of object recognition model, we took pictures of different angles for all robot components. This is a widely adopted technique called data augmentation. In total, we obtained 447 pictures of different angles for all robot components.

[*] Corresponding author

[1] Demonstration Video - https://youtu.be/IpHWPpzxLpE

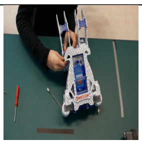
Class	1
Core question	Is there a right direction to lock the screws?
User question	Which way do I put the screws for locking it?
Video	

Table 1: Example data instance in our test set

Test Set For each of the 21 question intent cores, we have video recordings of the user asking the question. We then select 798 question variations as the test set. Each question in test set is paired with the video clip of its core question. Table 1 shows an example of our test set.

Model

Visual features combined with language modeling have shown great performance in question answering on images.

In this paper, we incorporate textual and visual information into our system. It should be noted that since our system needs to communicate with user in an interactive manner, real-time object recognition methods such as SSD (Liu et al. 2016) or YOLO (Redmon et al. 2016) are desired. However, for efficacy purposes, we used YOLOv3 Network (Redmon and Farhadi 2017).

Figure 1 depicts our model architecture. First, the question utterance from the user is converted to text by ASR and transformed into a representation vector by Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018). Meanwhile, the video frame of the user's current state is processed by our Visual State Discriminative Model (V-SDM) to generate an object vector which indicates the presence of the robot's components. With the integration of the textual and visual information that feeds into an MLP classifier, our system can provide the corresponding response to the user based on the predicted intent.

Visual State Discriminative Model

Visual Stage Discriminative Model (V-SDM) analyzes a video clip of the robot assembling process and outputs an object vector that indicates the components on the workbench. For each frame, V-SDM generates bounding boxes by YOLOv3 network (Redmon and Farhadi 2017). Each box contains probabilities of all robot's components which are head, neck, body, arms, feet, legs of the robot. The component with the highest probability is the predicted object. We denote the highest probability as the max probability of the box. Boxes with max probability lower than a threshold will be discarded. The probability vectors of all bounding boxes are treated as the input of the last softmax layer to generate the probability vector for current frame.

By accumulating the probability vectors of all the frames in this clip as the input, V-SDM can generate the video's probability vector. For each dimension of the probability vector, if the value is above the threshold, the same dimension of the object vector is set to 1, otherwise, the value is set to 0. In this case, the object vector represents the presence of components on the workbench.

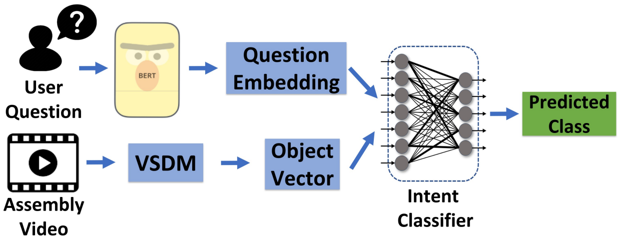


Figure 1: Multimodal Architecture

Conclusion

In this paper, we designed an interactive dialogue system by integrating the textual and visual modalities. To evaluate the performance, we will take the text-only MLP model with/without BERT embedding as our baseline models to compare the accuracy with our proposed model. With the proposed text+visual model, our goal is to mitigate the errors of the text-only model and improve the performance of user intent classification for providing better responses to user.

Acknowledgments

This research was supported in part by Intel Corporation.

References

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271.

Rinaldi, F.; Dowdall, J.; Kaljurand, K.; Hess, M.; and Mollá, D. 2003. Exploiting paraphrases in a question answering system. In *Proceedings of the second international workshop on Paraphrasing*, 25–32.

Todhunter, J.; Sovpel, I.; and Pastanohau, D. 2014. Question-answering system and method based on semantic labeling of text documents and user questions. US Patent 8,666,730.