Vol.2019-UBI-63 No.1
Vol.2019-ASD-15 No.1
2019/8/26

Evaluating the Use of a Dialogue System For Activity Data Collection

Tittaya Mairittha[1,a)] Nattaya Mairittha[1,b)] Sozo Inoue[1,c)]

Abstract: The annotation of human activity is the essential process for human activity recognition system. The problem was that annotation systems require cumbersome equipment which deprives their use. This paper presents a method to collect training labels for human activity recognition by using a dialogue system. The experiments were carried to predict users' activity and provided valuable insights on how users interact with the system. The evaluation shows that the dialogue system can collect data efficiently by achieving the highest precision, recall, and F1 score is 0.78 and 0.75 and 0.76, respectively.

Keywords: data collection, dialogue system, activity recognition

1. Introduction

Human activity recognition using smartphone sensors has been studied in recent years [2]. It has been used in various products and research areas for many years such as human-computer interaction, healthcare, and assistive technology. To train machine learning algorithms for recognizing human activities, we need a labeled sequence of activities (i.e., the start and finish times of the events). Accuracy depends on how accurate labels (annotation) are collected. However, existing annotation systems have limited means to communicate and limitations. For example, using a mobile input, users incapable of self-labeling if their hands are not free, a privacy concern when using a video camera or voice recording for annotation.

A dialogue system that can converse with a human by using voice-based is called spoken dialog systems (SDS). There is a growing interest in a conversational user interface, as they can truly enable people to be mobile and hands-free such as Microsoft Cortana [7], Amazon Alexa [1], Google Assistant [4]. These devices become popular when designing deep learning-based dialogue systems, which are an attempt to converse with users in natural language. Interactive interfaces that permit a user to ask a natural language question and receive an answer, possibly in the context of a more extended multi-turn dialog. They also are much more practical for people who are multitasking or since screen fatigue is a concern — for example, nurses who need to record patient care when taking care of patients.

This paper presents a method to collect training labels for human activity recognition by using a dialogue system. The system uses speech data as input and processes speech to text, then detects intention and extracts meaning from this input. We set up a lab study with practical tasks for collecting activity data from the dialogue system and training model with that data. We also use the outcome of the user study to understand how users interact with dialogue.

2. Background

2.1 Dialogue system

Dialogue systems, conversational agents or chatbots are software programs that support conversational interaction between humans and machines in natural language [5]. It can be based on text-based or speech-based and can also be used on different devices. Typically, dialogue systems can classify into two categories: task-oriented dialogue system which is used in this paper; and non-task-oriented dialogue system or chatbot. The task-oriented dialogue system is designed for a particular task and set up to have short conversations [10, 16] such as booking flight tickets, talking to customer care service, and asking about the weather while non-task-oriented dialogue system or chatbot is designed for unstructured conversational as a conversation between human and human [18, 9]. The dialogue system requires an understanding of natural language in order to process user queries. The initial step after getting a user request in a dialogue system is to understand the intent correctly. Therefore, providing an appropriate response to the user. If the dialogue system fails to understand the meaning of the user ' s request, it may lead to giving an inappropriate response or no response.

[1] Kyushu Institute of Technology ,1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka, 804-8550, JAPAN
[†1] Presently with Kyushu Institute of Technology
[a)] fon@sozolab.jp
[b)] fah@sozolab.jp
[c)] sozo@sozolab.jp

2.2 Human activity annotation

Some of the previously published works focus on recognizing human activities from video recordings [17], which is suffering from a privacy concern and incredibly time-consuming process for the manual annotation of a large number of data [12]. Another way for collecting activity labels by using voice recording [15, 14], these require users to wear a headset or to put the microphone in the room, which is uncomfortable and a privacy concern as well. The voice records were affected by different environmental sounds that need to remove background noise from an audio file before applying the classification technique. While a common approach by inputting forms on a mobile app, the user incapable of self-labeling if their hands are not free. For example, in a nursing care service, nurses are not able to record activities while taking care of residents; they had to complete the record before or later that it may cause errors in forgetting to input activities accurately. We assume that if we can collect activity labels as human nature to chat, having such simplicity in interaction without any touching, and speed to capture things in real time by using a spoken dialogue system that can significantly reduce such annotation error and saving time needed for editing later on.

3. Methods

3.1 A dialogue-based annotation

Figure 1 shows dialogue annotation processes based on a frame-based dialogue system. The frame-based dialogue system is designed for a task-oriented dialogue system to get information from the user and complete the task more efficiently [13]. Here the problem is similar to form filling, which asks the user questions to fill the slots (i.e., entities) in a frame (i.e., intent) and repeats until all the questions have been asked. In the human activity annotation, we proposed to ask two things from the user; (1) activity type; (2) timestamp (i.e., the start and finish times of the activities). For example, consider the utterances "stop walking 5 minutes ago". The intent of this utterance is a record. The "stop" is classified as an *action* entity, the "walking" is classified as an *activity* entity, and the textual span of "5 minutes ago" is classified as a *timestamp* entity.

3.2 Annotation process

Firstly, the user is made to enter the activity type, and then the system will find the match activity type. If not found, the system will request users to enter again. If it is matched, the activity will be recorded, and the predefined response is returned as output to the user. For completing a task, users need to tell the system to stop the activity. In this process, if users do not a specific time, the timestamp will be denoted as the current time.

3.3 Dialogue design

An algorithm behind the dialogue system for making decisions is that matching to a user utterance based on supervised machine learning models so that data resource

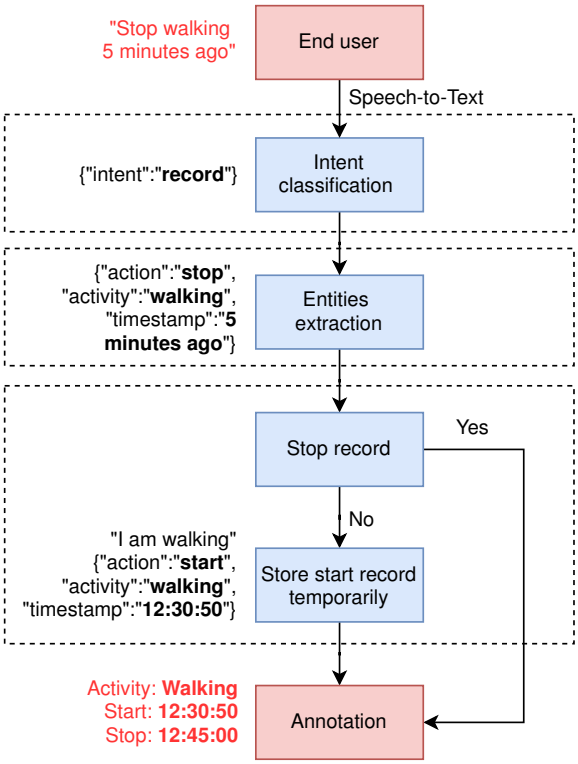

Fig. 1 Proposed Dialogue-Based Annotation.

is crucial in the development of effective intent classification model and modeling efforts in conversational. To collect relevant information and generate alternative replies when talking with users. We use a Machine-to-Machine (M2M) [8, 11] paradigm to collect training data. M2M is one of the most popular data collection approaches among intelligent virtual assistants on the market nowadays. The idea is to define a set of prompts for each intent and generate dialogue templates with each prompt, then paraphrasing to natural language by a human.

4. Experimental setup

Activity labels were collected through dialogue system application runs on Google Assistant [4] on a smartphone. We use Dialogflow [3] for building conversational. Dialogflow is a cloud-based NLU platform that provides a web interface to create bots which makes it easy to create initial bots. It also facilitates integration with Google Assistant that provides the functionalities of automatic speech recognition for converting speech to text on-device. Accelerometers were collected through FonLog [6] application in the background and uploaded to the cloud server by itself.

The system is tested with 7 participants, each participant was asked to carry the Android smartphone (Wiko Tommy 3 Plus) in the pants pockets and record activities while performing his/her routine for 2 days. The target activities include *walking*, *running*, *cycling*, *in a vehicle*, *sitting*, *standing*, *lying*, *downstairs*, *upstairs*, *take a train*, *carrying*, and *use a phone*. In total, we collected 243 annotations, where sitting and walking are the majority classes (see in Figure 2).

K-nearest neighbors (KNN), Decision tree (CART), Sup-

port Vector Machine (SVM), and Random forest (RF) were trained using mean, standard deviation, minimum, and maximum values in the interval as features. A window size is 1 minute with no overlapping. The dataset is split into 10 k-folds, each fold is then used once as a validation.

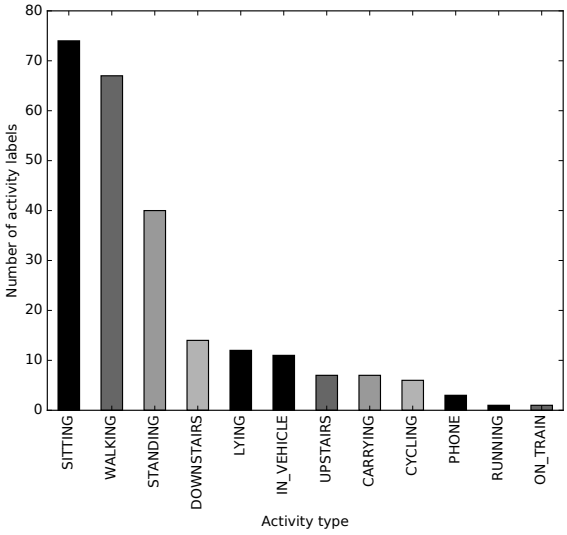

Fig. 2 A number of activity labels of each activity class.

5. Results

5.1 Activity recognition performance

To test the performance of the classification model, we use classification metrics to evaluate the accuracy in classifying activity labels following precision, recall, and f-measure. Table 1 reports results we obtained from the methods. Random forest achieved the highest scores, the precision was 0.75, the recall was 0.78, and the F-measure was 0.76 on average. However, other models do not seem to make much difference. We can see the average values still not good enough. We suspect that this is mainly because when users forget to log start or finish recording an activity, they do not have the ability to precisely recall an event at an absolute timestamp, as often approximate recording times were used such as "sitting 30 minutes ago".

Model	Recall	Precision	F-measure
KNN	0.7805	0.7374	0.7501
CART	0.7301	0.7229	0.7260
SVM	0.7805	0.7315	0.7234
RF	**0.7888**	**0.7558**	**0.7642**

Table 1 A comparison of different learning algorithms.

5.2 Intent classification performance

We then evaluate the quality of the dialogue system by measuring error rates. The system does not recognize user input on the average 9.9% of all utterances. When looking at errors into the conversation and the subsequent intentions invoked. We found that almost mistakes occur from speech recognition, and that lead to misinterpretation such as speakers pronunciation errors for non-native English speakers. We accept that it is usual for the development that these technologies get things wrong on occasion. However, speech recognition improvements will ensure you have satisfied users. The system has been useful in understanding the user's intention, and almost all annotations have been solved from a single utterance-response pair from a conversation. One of the reasons is that users generally used simple words (e.g., show, stop sitting, sitting), with 47.7% being just 2 words and 35.7% being a single word.

6. Conclusion

In this paper, we present a new approach to collect training labels for activity recognition through the use of the dialogue system based on a smartphone. We evaluate the performance of recognizing activity labels to show the feasibility of using the dialogue-based annotation. In our future work, we will evaluate the system with long-term data collection and more diverse samples. We will also find out the usage pattern and the implementation of another type of dialogue system.

References

[1] Amazon. *Amazon Alexa*, 2014.
[2] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3):33, 2014.
[3] Google. Dialogflow, 2010.
[4] Google. *Google Assistant*, 2016.
[5] Dan Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson London, 2014.
[6] Nattaya Mairittha, Tittaya Mairittha, and Sozo Inoue. A mobile app for nursing activity recognition. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 400–403. ACM, 2018.
[7] Microsoft. *Microsoft Cortana*, 2014.
[8] Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*, 2017.
[9] Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, Mirna Adriani, and Satoshi Nakamura. Developing non-goal dialog system based on examples of drama television. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 355–361. Springer, 2014.
[10] Antoine Raux and Maxine Eskenazi. Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges. In *InSTIL/ICALL Symposium 2004*, 2004.
[11] Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*,

2018.

[12] Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert. Temporal segmentation and activity classification from first-person sensing. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 17–24. IEEE, 2009.

[13] William K Thompson and Harry M Bliss. Frame goals for dialog system, February 2 2010. US Patent 7,657,434.

[14] Tim Van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse. Accurate activity recognition in a home setting. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 1–9. ACM, 2008.

[15] TLM Van Kasteren, Gwenn Englebienne, and Ben JA Kröse. Activity recognition using semi-markov models on real world smart home datasets. *Journal of ambient intelligence and smart environments*, 2(3):311–325, 2010.

[16] Joost Van Oijen, Willem Van Doesburg, and Frank Dignum. Goal-based communication using bdi agents as virtual humans in training: An ontology driven dialogue system. In *International Workshop on Agents for Games and Simulations*, pages 38–52. Springer, 2010.

[17] Jianxin Wu, Adebola Osuntogun, Tanzeem Choudhury, Matthai Philipose, and James M Rehg. A scalable approach to activity recognition based on object use. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.

[18] Zhou Yu, Alexandros Papangelis, and Alexander Rudnicky. Ticktock: A non-goal-oriented multimodal dialog system with engagement awareness. In *2015 AAAI Spring symposium series*, 2015.