Event Data Collection for Recent Personal Questions

Masahiro Mizukami, Hiroaki Sugiyama, Hiromi Narimatsu
NTT Communication Science Laboratories
{mizukami.masahiro, sugiyama.hiroaki, narimatsu.hiromi}@lab.ntt.co.jp

Abstract

In human-human conversation, people frequently ask questions about a person with whom to talk. Since such questions also asked in human-agent conversations, previous research developed a Person DataBase (PDB), which consists of question-answer pairs evoked by a pre-defined persona to answer user's questions. PDB contains static information including name, favorites, and experiences. Therefore, PDB cannot answer questions about events that occurred after it was built. It means that this approach does not focus on answering questions about more recent things (*recent personal questions*), e.g., *Have you seen any movies lately?* In contrast, since recent questions are frequently asked in a casual conversation, conversational agents are required to answer recent questions for maintaining a conversation. In this paper, we collect event data that consist of a large number of experiences and behaviors in daily lives, which enables to answer recent questions. We analyze them and show that our data is effective for answering recent questions.

1 Introduction

Questions about a conversational partner are called "personal question," which are an essential factor for expressing interest in conversational partners. Such questions frequently occur in casual human-human conversations. Nishimura *et al.* showed that such questions occurred in both human-human and human-agent conversations [Nisimura *et al.*, 2003]. Adequately answering them is an essential factor in the development of conversational agents [Sugiyama *et al.*, 2017].

To answer personal questions, previous works developed Person DataBase (PDB), which consists of question-answer pairs evoked by a pre-defined persona [Batacharia *et al.*, 1999; Sugiyama *et al.*, 2014]. Although their approach covers a wide variety of personal questions, developing a high-quality PDB is too expensive. The cost problem makes it difficult to update constantly; consequently, PDB usually contains only static information that rarely changes over time. Therefore, conversational agents using PDB cannot answer questions about recent events such as *What did you have for dinner yesterday?*. Also, it is easy to imagine that immutable responses to recent personal questions make conversational agents unnatural; therefore, conversational agents have to spend different days that like people spend different days, and it is more natural to return different answers to recent personal questions. To solve this problem, preparing other kinds of data which expresses recent experiences helps conversational agents to answer such questions about recent things (*recent personal questions*).

One simple idea is to collect data that express such recent events as a diary that is updated by the user. Previous work on response generation leveraged diaries or microblogs as a corpus that includes people's recent personal information [Li *et al.*, 2016]. Even though this approach seems reasonable, handcrafted-data-driven approach such as PDB has practical advantages in controllability and reliability. In this paper, we collected event data from participants who take part in short- and long-term periods. This collected data is hand-crafted, high-quality and easy to update (adding new day's data). We clarified the potential of event data to answer questions about recent behaviors/experiences in casual conversations through analysis.

2 Related Works

As mentioned in the introduction, PDB is the most closely related research to answer user questions. Batacharia *et al*. developed PDB about Catherine, a 26-year-old female living in New York City [Batacharia *et al.*, 1999]. To cover more questions and with different personas, Sugiyama *et al.,* developed a PDB with six personalities such as a 20-year-old female, a 50-year-old male, and robots [Sugiyama *et al.*, 2014]. Both PDBs contain only static information; therefore, they cannot answer recent questions. If we want to answer recent questions by PDB, we have to update PDB's contents constantly; however, updating PDBs constantly causes too expensive costs. The difficulty of updating PDB is the relationship of questions and answers (QA); for example, when the content of a base QA changes (e.g., Question:*Do you have any pets?*, Answer:*Yes, I have a dog.* change to new Answer:*No, I don't have.*), related contents of QAs should be changed depending on a changed content of base QA (e.g., Question:*Do you have a dog?*, Answer:*Yes, I have a dog.* should be changed to new Answer:*No, I don't have.*). PDB has many complicated relations of QAs, it makes updating

Event name	Event reason	Event time	Event impressions
Played a mobile phone game	Habit before going to bed, To get daily bonus	0:00-4:00	Happy
Read a novel by a mobile phone	Habit before going to bed, To induce sleep	0:00-4:00	Fun, Sleepy
Got up	To prepare a lunch box	4:00-8:00	Sleepy, Tired
Went back to sleep	To rest before going to office	4:00-8:00	Sleepy
Got up	To go to office	8:00-12:00	Sleepy
Ate breakfast and made up	To go to office, Hungry	8:00-12:00	Delicious, Tired
Drove a car while listening to musics	To go to office, To motivate	8:00-12:00	Happy, Fun
Worked	I'm worker, To get a salary	10:00-12:00	Difficult
Ate lunch	Recess	12:00-16:00	Delicious
Listened to musics	To relax	12:00-16:00	Fun, Sleepy
Worked	I'm worker, To get a salary	12:00-16:00	Difficult
Worked overtime	To send mail	16:00-20:00	Tired
Drove a car	To go shopping	16:00-20:00	Sleepy
Went shopping	Shop received reservation products	20:00-24:00	Happy, Fun
Took a bath	To refresh oneself	20:00-24:00	Warm, Sleepy, Pleasant
Ate dinner	Prepared for me	20:00-24:00	Delicious
Did travel preparations	To go for a trip tomorrow	20:00-24:00	Tired, Pleasure
Looked for things	I have lost bought one	20:00-24:00	Sad, Laughing
Played a mobile phone game	Habit before going to bed	20:00-24:00	Happy, Sleepy
Went to bed	To prepare tomorrow	20:00-24:00	Sleepy

Table 1: Examples of collected event data

PDB difficult and expensive. A PDB's merit, which is found in handcrafted-data-driven methods, is the ability to generate answers based on facts and consistency from the data. Such handcrafted-data-driven approaches answer questions with consistent replies and without a lie. The consistency of the responses based on facts has the potential for improving the performance of conversational agents.

Although there are many studies on conversational agent's response generation [Ritter *et al.*, 2011; Inaba and Takahashi, 2016], few studies focus on the consistency depending on an agent's personality. Persona-based conversation models treat personality as speaker-embedding to increase the sentence quality [Li *et al.*, 2016]. This model is the state-of-the-art model to generate conversational agent's responses using an embedding vector that expresses agent's personality. This approach has potential to answer recent personal questions; however, it indicates two critical problems. One is that this approach cannot promise to answer without lies; this problem is strongly related with research of PDB. Hand-crafted database approach such as PDB can answer responses that reflected a right personality unless it gets wrong matching of questions. In contrast, neural network based approaches often answer questions with response sentences that do not exist in training data; since such models are optimized only for maximizing the naturalness of response sentences. Even though this approach has the potential to answer recent personal questions, it can offer no guarantee that the answers exist in training data.

Another problem is that this model does not consider the past consistency depended on the day, time, and past events. When we asked a question such as *What did you eat last night?* to conversational agents, this approach always replies the same response such as *I ate ramen.* This QA pair is natural when we check only this one pair; however, eating the same food every dinner is too unnatural in the daily life of conversational agents. Therefore, to establish a long-term conversation with human-agent and to make conversational agents more natural, we must solve this invariance problem of responses. Using information of date and time to train speaker-embedding vector, it may help to solve this problem. However, we can imagine easily that this model requires much training data which is insufficient with the amount we have now.

In this paper, we created event data for answering recent personal questions in casual conversations. This approach by the created event data is identical with PDB as the handcrafted-data-driven approach and is essential to verify answering based on facts, and we use it as the first step to develop a function that answers questions about recent experiences based on facts.

3 Data Collection

To answer recent personal questions that ask about recent experiences and behaviors, we collect the consistent data from humans as events that express experiences and behaviors. This event data has to be collected from participants with low-costs; because we need to update it constantly. Besides, we have to collect data from various participants because we do not know what kind of persona influences events.

We recruited 62 Japanese-speaking participants of roughly equal numbers of both genders whose ages ranged from 10s to 60s and collected daily experiences and behaviors as event data. They wrote down 20 events every day and at least two events every four hours. We collect event name, reasons, time, and impressions for each event; because these aspects are asked in casual conversation frequently. Participants were indicated not to write any descriptions including privacy. Such diary-like method to write down like a diary is low-cost compare than the PDB's collecting method. Specifically, we prepare an Excel file and ask participants to write four aspects such as name, reasons, time, and impressions,

for each column. The format of this Excel file is simple; one line is for one event, one sheet is for one day, one file is for one participant.

An event includes four aspects:

1. Event name: What is happened? What did you do?
2. Event reasons: Why did it happen? What did you do?
3. Event time: Selected from the following four-hour time blocks: 0:00-4:00, 4:00-8:00, 8:00-12:00, 12:00-16:00, 16:00-20:00, or 20:00-24:00.
4. Event impressions: How did you feel?

For the aspects of reasons and impressions, participants can write more than one sentence with a space between phrases. We define two groups for collecting data. One is the long-term group which takes data with many days from a few participants, and this facilitates the comparison between participants. Another one is the short-term group which takes data with few days from a lot of participants; this is necessary to collect various event data. Five participants wrote 20 events per day for 30 days (long-term group), and 57 participants wrote 20 events per day for seven days (short-term group); finally, we collected a total of 10,980 events. Table 1 shows examples of events collected from a participant who belongs to the short-term group. The example shows that we obtain a variety of events even if the only one participant wrote.

4 Data analysis

We analyze next two viewpoints to show that our collected data helps to answer recent personal questions that related to personality and date. First, the tendency of events was varying among participants; it shows that we have to reflect participant's characteristics to answer recent personal questions. Second, the tendency of events was varying according to a day of the week; it shows that we have to reflect a day of the week and update event data constantly.

To analyze the tendency of events, we categorized the collected event data since they have slightly different event names, with which we cannot count the occurrence of each event. For example, we wish to handle two events such as *Went to school'* and *Went to high school* as the same event. To collect such similar events as the same event, we perform the word-based hierarchical clustering using word2vec that trained from Wikipedia data.

Next, we highlight the difference between event's tendencies among participants and days. We calculate frequency distributions of events for each participant and each day, and compare a JS divergence of these frequency distributions. This comparison clarifies two relationships of event tendencies: *Distributions of event frequency depend on each participant* and *Participants have different distributions of event frequency depending on each day.*

4.1 Event clustering

We performed hierarchical clustering to find similar events in the collected data [Larsen and Aone, 1999]. This clustering is both analysis and a necessary procedure to compare events among participants or days by collecting clusters. Before clustering, we trained word2vec [Mikolov *et al.*, 2013] from Wikipedia articles; word2vec is useful to convert event names to a word embedding. In clustering, we tokenize event names by mecab [Kudo, 2006], and restore tokenized words to original forms. Next, we calculate vectors by adding together word2vec of each tokenized words, and cluster these calculated vectors with Ward's method [Szekely and Rizzo, 2005]. Figure 1 shows a dendrogram and a heatmap of each vector. To confirm the difference of clustering results by the number of clusters, we respectively show the hierarchical clustering results of ten clusters and 30 clusters. Table 2 and Table 3 are ten and 30 lists of events. Event names are the nearest event to the center of each collected cluster, and cluster sizes are

Cluster	Event name	Size
E1	Cleaned up (掃除をした)	2480
E2	Got up (起床した)	1238
E3	Drove a car (車を運転した)	565
E4	Took a meal (ご飯を食べた)	1478
E5	Drank drink (飲み物を飲んだ)	1095
E6	Watched TV (テレビを見た)	633
E7	Took a bath (お風呂に入った)	436
E8	Went to a toilet (トイレに行った)	911
E9	Ate lunch (昼食を摂った)	1356
E10	Went to bed (寝た)	788

Table 2: List of 10 cluster's representative events

Cluster	Event name	Size
E1	Looked SNS by PC (PC でＳＮＳを閲覧した)	117
E2	Played a game (ゲームをした)	232
E3	Read mails (メールをチェックした)	269
E4	Cooked dinner (夕食の支度をした)	581
E5	Worked (仕事をした)	260
E6	Did the laundry (洗濯をした)	1021
E7	Got up (起床した)	876
E8	Going to bed (就寝する)	146
E9	Worked (仕事した)	216
E10	Took the train (電車に乗った)	202
E11	Came back home by car (車で帰宅した)	131
E12	Drove a car (車を運転した)	232
E13	Ate breakfast (朝食を食べた)	992
E14	Took a meal (ご飯を食べる)	486
E15	Drank Coffee (珈琲を飲んだ)	518
E16	Washed dishes (食器を洗った)	577
E17	Watched a video (動画を見た)	233
E18	Watched TV (テレビを見た)	274
E19	Watching TV (テレビを見る)	126
E20	Took a bath (お風呂に入った)	436
E21	Went to a toilet (トイレに行った)	222
E22	Went shopping (買い物に行った)	689
E23	Read a newspaper (新聞を読んだ)	218
E24	Tidying up dishes (食器を片づける)	540
E25	Ate lunch (昼食を摂った)	138
E26	Talked with guests (来客と話した)	143
E27	Sent a child to sleep (子供を寝かせた)	317
E28	Woke up (起きた)	143
E29	Slept in bed (ベッドで寝た)	300
E30	Slept (寝た)	345

Table 3: List of 30 cluster's representative events

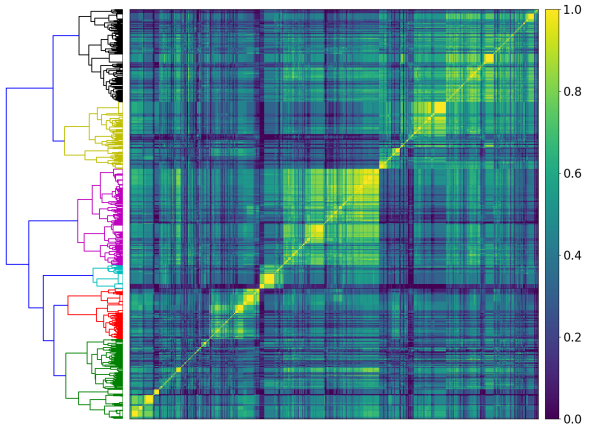

Figure 1: Result of hierarchical clustering for events

numbers of events included in each cluster.

From Table 2, we obtained common events which seem to happen to anyone such as *Got up*, *Took a meal*, *Drank drink*, *Took a bath*, *Ate lunch* and more. In contrast, from Table 3, we obtained detailed events which seem to happen to specific personas such as *Look SNS by PC*, *Played a game*, *Watched a video* and more. Such events which indicates participant's characteristic are important to highlight the difference between participants; therefore we use the 30 lists of events as clustering result to compare events between participants and days in following analysis sections.

Note that we defined size of clusters based on a few preliminary experiments. Proposing the clustering method that determines a size of clusters based on clusters variances or entropy has a potential to improve clustering performance; therefore, we will tackle defining a better model to handle event data in future work.

From Figure 1, we can find a few particularly bright clusters that include very similar events. In contrast, some clusters with less brightness include various events that are not so similar.

Since this hierarchical clustering successfully makes clusters, we can benefit by using clustering results for data analyses. This clustering method that considers word meaning as word2vec, could make clusters which gathered almost same meaning events.

4.2 Event analysis for each participant

First, we analyzed events among participants. To highlight the differences between participants, we calculate the distribution on clusters of every participant. To calculate it, we used the 30 clusters in Table 3. We compare these cluster distributions between each participants using JS divergence.

The averaged JS divergence of every participant was 0.39. The minimum JS divergence is 0.063, and the maximum JS divergence is 0.77, these scores were found among participants who are the short-term group. Averaged JS divergence is not close to 0; it means that distributions of event frequency are different on each participant.

		Participants				
	Cluster	P1	P2	P3	P4	P5
E1	Saw SNS by a PC	1	0	7	3	9
E2	Played a game	0	0	2	97	0
E3	Read mails	34	16	6	5	0
E4	Cooked dinner	73	22	53	5	8
E5	Worked	0	38	14	57	0
E6	Did the laundry	25	72	32	67	39
E7	Got up	63	16	50	96	13
E8	Going to bed	0	0	11	0	22
E9	Worked	9	3	32	8	14
E10	Took the train	6	15	0	0	7
E11	Came back home by a car	90	0	3	0	1
E12	Drove a car	0	18	28	4	0
E13	Ate breakfast	85	82	54	55	80
E14	Took a meal	2	6	20	2	62
E15	Drank Coffee	39	18	1	60	69
E16	Washed dishes	2	13	30	25	36
E17	Watched a video	7	23	3	5	0
E18	Watched TV	61	7	0	1	0
E19	Watching TV	0	0	29	0	2
E20	Took a bath	23	25	32	20	47
E21	Went to a toilet	1	0	20	0	0
E22	Went shopping	24	127	42	8	17
E23	Read a newspaper	44	5	1	4	13
E24	Tidying up dishes	7	6	54	38	19
E25	Ate lunch	0	0	18	6	2
E26	Talked with guests	2	3	13	1	3
E27	Sent a child to sleep	2	11	15	3	33
E28	Woke up	0	32	1	0	35
E29	Slept in bed	0	30	20	30	12
E30	Slept	0	12	9	0	57

Table 4: Cluster assignment of events in participant whom is only long-term group

To analyze details of event tendencies, we show counts of event cluster assignment in each participant who is a long-term group, in Table 4. Most participants have different distributions of events, but E4, E6, E7, E9(and E5), E13, E14, E15, E16, E20, E22, E23, E24, E26, and E27 occurred more than once in all participants of the long-term group. E4, E6, E16, E22, and E24 are clusters containing mainly housework such as *washing*, *cleaning*, *cooking*, *shopping* and more. E7, E13, E14, E15, and E20 are clusters that indicate physiological desires such as *eating, drinking*, *sleeping*, *taking a bath* and more. E9 (and E5) is the cluster that indicates mainly *working*, E23 indicates *reading*, and E26 indicate *talking*. The last E27 is a cluster included various events such as *child - rearing*, *one's hobby*, and *school life*. Such basic events that are related to living were observed in almost participants. In contrast, we obtained that events which relate to entertainment such as *Played a game* were observed in a specific participant such as P3.

These results show that participants have different event tendencies. This indicates that we should collect data which depends on each persona to answer recent personal questions.

	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
Mon.		.005	.005	.005	.006	**.011**	**.013**
Tue.	.005		.005	.007	.006	**.013**	**.019**
Wed.	.005	.005		.004	.004	**.008**	**.011**
Thu.	.005	.007	.004		.005	**.011**	**.017**
Fri.	.006	.006	.004	.005		**.013**	**.017**
Sat.	**.011**	**.013**	**.008**	**.011**	**.013**		.005
Sun.	**.013**	**.019**	**.011**	**.017**	**.017**	.005	

Table 5: JS divergences between days of the week (**bold** means the top 10 high scores)

4.3 Event statistics for each day

Second, we analyzed events among days. Like Section 4.2, we counted the clusters of every participant with 30 clusters in Table 3. We compare cluster counts of all participants in a total at a day of the week. We show JS divergences between days of the week in Table 5.

We focus on JS divergences between weekdays and weekends. These high JS divergences (We showed it as **bold** in Table 5) show the difference between weekdays and weekends; furthermore, the small JS divergence scores are concentrated in between weekdays and between weekends. This result shows that participants spend different life between weekdays and weekends. Such result that we can imagine easily lets us reconfirm the importance to answer depending on a day. Therefore, we need data which depend on each day to answer questions that ask about events.

5 Discussion

In this section, we discuss the potential to answer recent personal questions by our collected data. Our discussion follows the "comparison with the conversation corpus" in Sugiyama *et al*. [Sugiyama *et al.*, 2014], whose PDB covers 41.3% of questions in real conversations and explains why other questions were excluded. The top reason that questions were excluded is "limited by specific words, date, or time" such as *What did you eat for lunch today?* or *Where did you go this summer vacation?*, such questions are about 71.2% of a whole of excluded questions. We mainly focus on these excluded questions and show case studies which can answer by our collected data. Tackling to answer such questions helps to solve future works of the previous research.

First of all, we collect 286 questions that are the same as excluded questions by Sugiyama *et al.* [Sugiyama *et al.*, 2014], and extract 204 questions that were excluded by "limited by specific words, date, or time." In previous works, they said that these questions are difficult to maintain consistency with 5W1H answers in particular. We focus on these questions and find questions that can answer questions if we use event data. We show examples of such question which can answer based on an event in Table 6.

From Table 6, some questions which ask about speaker's recent behaviors can answer by our collected data. For example, we can answer a question such as *What did you eat for lunch today?*, an answer is *Yes, I ate a curry and rice* by using an event such as *Ate a curry and rice.* In this manner, we can make an answer utterance that based on an event matched with a question. These results show us the possibility to answer a part of questions that were unsolved future works of PDB with our collected data.

We can also answer questions that ask opinions. Such questions frequently occurred after disclosure or an answer that replied to first questions. To answer with opinions, we use aspects of event impressions. We show examples of questions which ask opinions about events in Table 7. Specifically, when a conversational agent say *I watched a movie.* as disclosure, and the user asks *Do you like it?* that asks conversational agent's opinion, a conversational agent can answer *It is fun.* by using an aspect of event impression from our collected data. In question-answering based on the conventional PDB, we cannot handle such kind of questions which continued the same topic as the previous turn. Answering questions about the details of the same one event, it shows the potential which improves the question-answering function to talk deeper.

However, we obtain some questions that we could not answer by our collected data; there are *Questions that ask about agent's past custom* and *Questions that ask about agent's future*. To answer questions that ask about agent's future, we have to prepare the other data such as plans made by agents. These plans may need the approach such as the belief-desire-intention model that is different from our event data. To answer questions that ask about agent's past custom, we need data which indicates habitual events and experiences. Such data seem closely related to our event data, because habitual events and experiences may be made by the accumulation of recent events. We clarify the relationship between past custom and events, and will propose a method that generates past custom based on accumulated recent events in future work.

From analyses and case studies, we showed the potential of answering for recent personal questions that cannot be answered by the previous PDB. Our collected data helps to answer not only asking events but also asking opinions. However, we obtain some problems that remain about questions which ask about past custom and future. In future work, we tackle to answer questions that ask past custom such as habitual events using our event data. Furthermore, clarifying volumes and frequency to collect enough event data; these are *How many events do we need in one day?*, *How many times do we ask to write per one day?*, and *How many days do we ask to write events?*. Besides the data collection, to develop conversational agents that answer recent personal questions using event data, we have to propose a method that finds events which match with user's recent personal questions.

6 Conclusion

In this paper, we collect 10,980 events which express recent experiences and behaviors to help conversational agents answer questions about recent experiences. First of all, we analyze collected data to highlight the tendencies of events based on each participant and each weekday, and we show the necessity of our event data that make conversational agents more natural. Our analysis shows that event data reflect participant's characteristics and dependencies on weekdays, and we show two knowledge about tendencies of events. One, event tendencies are depending on each participant; we

Question	Answer	Event
Did you eat for lunch today?	Yes, I ate.	Ate a lunch
What did you eat for lunch today?	I ate a curry and rice.	Ate a curry and rice
Did you play video games lately?	Yes, I played video games.	Played video games
What did you play video games lately?	I played smartphone games.	Playing smartphone games
What kind of games did you play lately?	Smartphone games.	Playing smartphone games
Did you watch a movie?	Yes, I watch a movie.	Watched a movie
Where did you go out somewhere recently?	I went to the nearby French restaurant.	Went to the neighboring French restaurant
Where did you go out somewhere recently?	I went to the spa.	Going to the spa

Table 6: Examples of question which can answer based on an event

First question	First answer / Disclosure	Question	Opinions	Event	Event impressions
–	I watched a video	Do you like it?	It is fun.	Watched a video	Fun
–	I use PC for only presentation and playing games	Do you like it?	Yes, I love it.	Played a PC game	Fun / Tiresome
Did you go to the theater recently?	Yes, I watched "The Dark Night Rising" and "Library war"	How was it?	It is fun.	Watched a movie	Fun

Table 7: Examples of questions which ask impression or evaluation

should collect event data which depends on each conversational agent's persona. Another one, event tendencies are depending on each weekday; we should collect event data which depends on each day to make conversational agent's answers more natural.

In the discussion, we followed the previous works and obtained case studies that can answer by our collected event data. Our event data helps to answer recent personal questions such as *What did you eat for lunch today?* that asks about doing conversational agent's events; therefore, results show potential to achieve our first purpose that answers a part of questions that cannot be answered by the previous PDB. Furthermore, aspects of event impressions help to answer questions that ask opinions such as *Do you like it?*. This continuous question-answering shows the potential which improves the question-answering function to talk deeper.

In future work, we clarify volume and frequency to collect enough event data, and develop conversational agents that answer recent personal questions by collected event data.

References

[Batacharia *et al.*, 1999] B Batacharia, D Levy, R Catizone, A Krotov, and Y Wilks. Converse: a conversational companion. In *Machine conversations*, pages 205–215. Springer, 1999.

[Inaba and Takahashi, 2016] Michimasa Inaba and Kenichi Takahashi. Neural utterance ranking model for conversational dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 393–403, 2016.

[Kudo, 2006] Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. jp*, 2006.

[Larsen and Aone, 1999] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–22. ACM, 1999.

[Li *et al.*, 2016] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Nisimura *et al.*, 2003] Ryuhei Nisimura, Yohei Nishihara, Ryosuke Tsurumi, Akinobu Lee, Hiroshi Saruwatari, and Kiyohiro Shikano. Takemaru-kun: Speech-oriented information system for real world research platform. *Proceedings of International Workshop on Language Understanding and Agents for Real World Interaction*, 2003.

[Ritter *et al.*, 2011] Alan Ritter, Colin Cherry, and William B Dolan. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics, 2011.

[Sugiyama *et al.*, 2014] Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami. Large-scale collection and analysis of personal question-answer pairs for conversational agents. In *International Conference on Intelligent Virtual Agents*, pages 420–433. Springer, 2014.

[Sugiyama *et al.*, 2017] Hiroaki Sugiyama, Toyomi Meguro, and Ryuichiro Higashinaka. Evaluation of question-answering system about conversational agent ' s personality. In *Dialogues with Social Robots*, pages 183–194. Springer, 2017.

[Szekely and Rizzo, 2005] Gabor J Szekely and Maria L Rizzo. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of classification*, 22(2):151–183, 2005.