Large-Scale Acquisition of Commonsense Knowledge via a Quiz Game on a Dialogue System

Naoki Otani[1] **Daisuke Kawahara**[1] **Sadao Kurohashi**[1]
Nobuhiro Kaji[2] **Manabu Sassano**[2]

[1]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[2]Yahoo Japan Corporation, Tokyo, Japan
otani.naoki.65v@st.kyoto-u.ac.jp {dk,kuro}@i.kyoto-u.ac.jp
{nkaji,msassano}@yahoo-corp.jp

Abstract

Commonsense knowledge is essential for fully understanding language in many situations. We acquire large-scale commonsense knowledge from humans using a game with a purpose (GWAP) developed on a smartphone spoken dialogue system. We transform the manual knowledge acquisition process into an enjoyable quiz game and have collected over 150,000 unique commonsense facts by gathering the data of more than 70,000 players over eight months. In this paper, we present a simple method for maintaining the quality of acquired knowledge and an empirical analysis of the knowledge acquisition process. To the best of our knowledge, this is the first work to collect large-scale knowledge via a GWAP on a widely-used spoken dialogue system.

1 Introduction

Large-scale knowledge is an essential resource in many natural language processing (NLP) applications. There have long been efforts devoted to collecting *commonsense knowledge*, *i.e.*, general knowledge that every person knows (Zang et al., 2013). We rely on such prior knowledge to understand languages. For example, consider the sentence "She went to get strawberries." A human might think she went to the refrigerator in the kitchen or a supermarket in the neighborhood. Computers, however, do not know that strawberries would be stored in refrigerators. This paper presents a methodology for acquiring large-scale commonsense knowledge from humans.

Early work on commonsense knowledge acquisition includes the Cyc project (Lenat, 1995), where a small group of human annotators organized resources. Manually curated resources are of high quality but require significant cost and time to build. Thus, several studies have automatically constructed knowledge bases on existing resources such as semi-structured or unstructured texts (for example, (Tandon et al., 2014)). However, commonsense knowledge is so clear for every person that it is often omitted in a text (Gordon and Van Durme, 2013). For instance, we rarely state in a text that strawberries are stored in refrigerators. Rather, we often talk about a major production region for strawberries. Therefore, manual effort is still required to build commonsense knowledge bases.

To reduce the cost of manual knowledge acquisition, some studies explored the use of crowdsourcing, a process that requests various tasks of non-expert workers on the Internet. The Open Mind Common Sense (OMCS) project (Liu and Singh, 2004; Speer and Havasi, 2012) recruited volunteers on the Internet and constructed *ConceptNet*, a large collection of commonsense knowledge such as (cake, *AtLocation*, supermarket). Whereas participants in the OMCS projects entered the commonsense knowledge in Web forms, some studies have transformed the knowledge acquisition process into a type of enjoyable game, called *games with a purpose* (GWAP) (von Ahn et al., 2006; Lieberman et al., 2007; Kuo et al., 2009; Nakahara, 2011; Herdağdelen and Barobni, 2012; Kuo and Hsu, 2011). The advantage of a GWAP is that it is more attractive to humans than the standard annotation processes and is able to collect accurate resources as a side effect of their enjoyment of the games.

Proceedings of the Open Knowledge Base and Question Answering (OKBQA) Workshop,
pages 11–20, Osaka, Japan, December 11 2016.

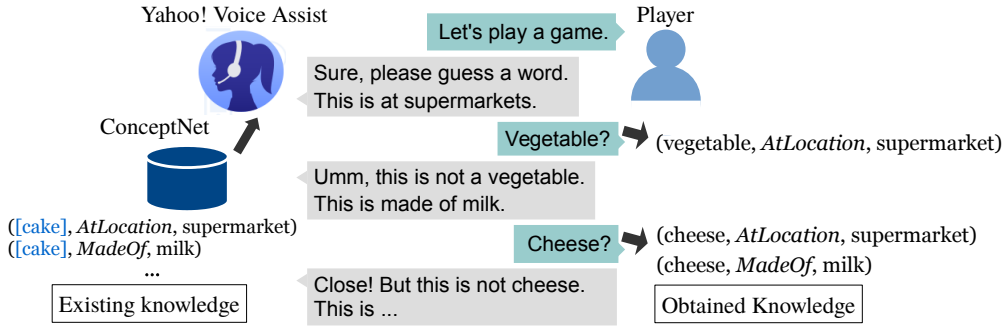

Figure 1: **Illustration of the quiz game.** A player is given clues about a certain word to be guessed. The clues are generated from existing knowledge in ConceptNet. We learn new knowledge from the player's guesses given the clues.

We developed a quiz game shown in Figure 1 as a module in the widely-used Japanese smartphone app, *Yahoo! Voice Assist*[1] (hereafter Voice Assist), which is a Siri-like spoken dialogue system that has been downloaded to more than 2.5 million devices. Although it is usually hard to get GWAP participants over the long term, our game is able to reach greater numbers of users than previous studies.

The quiz game follows the same framework as the previous work by Nakahara (2011), in which players are given clues about a certain word. From the players' incorrect guesses, we can obtain knowledge about the clues. For example, a hint "this is made of milk" is given to a player. The expected answer is a cake, but we can learn that "cheese is made of milk" when the player answers "cheese" in response to the hint.

Spoken dialogue systems on smartphones such as Siri and Cortana have attracted many industrial and research interests in recent years. They are promising as platforms of knowledge acquisition from humans because they have a large number of users. Furthermore, acquired knowledge is useful for developing sophisticated dialogue systems and attracting more users. Consequently, we can collect more knowledge from more users. Additionally, enjoyable user interaction without any goals is important. Even in a task-oriented dialogue system, Jiang et al. (2015) report that about 20% of user logs are chat, and some studies report that games on a dialogue system improve user engagement with the system (Kobayashi et al., 2015; Sano et al., 2016).

As the obtained knowledge contains incorrect facts, we aggregate facts collected from multiple players to determine the correct facts. Such incorrect facts come from players' guesses that are not relevant to the hints and automatic speech recognition (ASR) errors, which is a characteristic of spoken dialogue systems. To address these problems, it is not sufficient to only consider the number of players who give the same facts, as previous studies have done (von Ahn et al., 2006; Kuo et al., 2009; Nakahara, 2011). We present a method to reduce the ASR errors and estimate the confidence scores of facts.

We released the game in December 2015 and have collected more than 150,000 unique facts over eight months. The number of unique players was 70,000 in total, which is much larger than those of previous studies, for example, 6,899 players over six months for the game by Kuo et al. (2009). We evaluated the quality of the collected knowledge in two ways. We first evaluated the scoring method of the facts using crowdsourced annotations, and then manually evaluated samples of the collected knowledge. The results show that our quiz game is an effective way to acquire large-scale commonsense knowledge.

Our contributions are summarized as follows.

1. We collected large-scale, high quality knowledge from a quiz game on a dialogue system that has many users. We will make the collected resources publicly available.
2. We present a method to reduce ASR errors and maintain the quality of acquired knowledge.

2 Related Work

Research in knowledge base construction has contributed to the success of many applications including question answering and information extraction. Some knowledge bases are manually constructed,

[1] `http://v-assist.yahoo.co.jp` (in Japanese)

and others are automatically constructed from texts and existing resources. Automatically constructed large-scale knowledge bases include Freebase (Bollacker et al., 2008), which was constructed from Wikipedia's texts and existing resources such as WordNet. Freebase has been successfully used in many NLP tasks.

Whereas most large-scale knowledge bases focus on relations between named entities, we focus on commonsense knowledge, which has a wider range. The Cyc Project (Lenat, 1995), an early work on commonsense knowledge acquisition, recruited a small group of annotators to construct a knowledge base.

Manually curated resources are accurate but expensive to build. Several studies have attempted to construct knowledge bases automatically. For example, Tandon et al. (2014) extracted knowledge from WordNet and Web texts. However, commonsense knowledge is likely to be omitted from texts because it is assumed that every person knows such knowledge (Gordon and Van Durme, 2013). Li et al. (2016) addressed this problem using knowledge base completion, in which existing knowledge is used to acquire more knowledge. However, their method needs some amount of existing knowledge as a seed. Thus, manual effort is still required.

Crowdsourcing, which is a process that requests various tasks of non-expert workers on the Internet, can be used to reduce the cost of the manual process. The OMCS project (Liu and Singh, 2004; Speer and Havasi, 2012) collected commonsense knowledge by recruiting many volunteers on the Internet. The resulting knowledge base is called ConceptNet,[2] which we use and extend in our study.

Some studies transform the manual acquisition process into an enjoyable game, called a GWAP, to motivate players to participate in knowledge acquisition. GWAPs are a form of crowdsourcing[3] and have been used for validating (Herdağdelen and Barobni, 2012; Vannella et al., 2014; Machida et al., 2016) and collecting (von Ahn et al., 2006; Lieberman et al., 2007; Kuo et al., 2009; Nakahara, 2011; Kuo and Hsu, 2011; Nakahara and Yamada, 2013) language resources.

A word-guessing game was designed by von Ahn et al. (2006) to collect a large amount of knowledge within a short time and at a low cost. GWAPs have also been exploited to acquire knowledge in Chinese (Kuo et al., 2009; Kuo and Hsu, 2011) and Japanese (Nakahara, 2011; Nakahara and Yamada, 2013). The collected knowledge was registered in ConceptNet. Although it is generally hard to gather many players, our GWAP can reach many more users than previous studies because it is built on a running spoken dialogue system.

Spoken dialogue systems on smartphones have been attracting much industrial and research interest in recent years. Several studies report that enjoyable user interactions are beneficial for dialogue systems (Jiang et al., 2015; Kobayashi et al., 2015; Sano et al., 2016).

3 Rapid Knowledge Acquisition from Quiz Game

We use a quiz game on a spoken dialogue system to obtain large-scale, high-quality commonsense knowledge from many human players.

3.1 Japanese ConceptNet

Our knowledge acquisition method follows the scheme of ConceptNet (Speer and Havasi, 2012). In ConceptNet, knowledge is expressed as a triple of two concepts and a relation linking them, where a concept is a word or a short phrase, and a relation consists of about 30 relations such as *IsA*, *Causes*, or *Antonym*. We call a triple a *fact*, and the two concepts are called a *head* and *tail*, respectively.

Japanese ConceptNet has 95,468 facts in total.[4] We ignore facts obtained using the game by Nakahara (2011) with a weight of one (*i.e.*, only one player provided this fact) because these facts are likely to be inaccurate . The filtered ConceptNet has only 46,427 unique facts, and most of them are lexical knowledge (*e.g.*, *Antonym* and *DerivedFrom*). In contrast, Japanese WordNet,[5] for example, has 93,834

[2] `http://conceptnet5.media.mit.edu`

[3] For more information about crowdsourcing, readers can refer to (Law and von Ahn, 2011).

[4] From a snapshot taken on Sept. 10th, 2015 at `http://conceptnet5.media.mit.edu/downloads/20150910/`.

[5] `http://compling.hss.ntu.edu.sg/wnja/index.en.html`

words and many relations linking them. Thus, collecting more commonsense facts is essential to making them at least as useful in NLP tasks as WordNet is.

We only consider the filtered ConceptNet in the rest of this paper. Note that words in heads and tails are normalized into their representative forms (*e.g.,* { みかん mikan, ミカン mikan, 蜜柑 mikan} (orange) → みかん mikan (orange)) given by the morphological analyzer JUMAN++ (Morita et al., 2015).

3.2 Building a Quiz from using ConceptNet

We collect commonsense knowledge from many people. To motivate them, we transform the knowledge acquisition process into an enjoyable quiz game, where human players are given several hints about a certain word to be guessed, and we acquire knowledge from players' guesses. The hints are easily generated from existing knowledge in ConceptNet. Figure 1 shows examples. "This is at supermarkets" and "this is made of milk" are generated from the facts (cake, *AtLocation*, supermarket) and (cake *MadeOf*, milk), respectively. The word to be guessed (hereafter *keyword*) is "cake."

From the player's guesses, we can obtain knowledge about each relation and tail pair. For instance, we can learn that "cheese is made of milk" when the player answered "cheese" to the hint "this is made of milk." Note that we only hide the head of a fact and let players guess the word that fits with its relation and tail because this allows players to give diverse answers. For example, we use "X is made of milk" rather than "cake is made of X" to obtain many different responses.

If the player fails to guess the keyword, another hint is selected at random and given to the player. Our game gives up to five hints,[6] and 15 facts can be acquired from a player's guesses. We call the distance between a player's guess and a given hint the *hint distance*. The hint distances of (cheese, *MadeOf*, milk) and (cheese, *AtLocation*, supermarket) in Figure 1 are one and two, respectively.

The game is implemented as a part of the chat function of Voice Assist, which is the Japanese spoken dialogue system on smartphones, and is executed as follows:

1. A player utters a sentence such as "ゲームしよう" ("Let's play a game"), and the game session starts.
2. A keyword is selected randomly.
3. A hint about the keyword is drawn. The hint sentence is generated using predefined templates.
4. Given the hint, the player utters his/her guess.
5. If the guess matches the keyword, the game session ends and the system returns to the normal dialogue processing mode; otherwise the system returns to step 3 to add more hints to the quiz until the number of hints reaches its limit. If the number of hints reaches its limit, the system ends the game and returns to the normal dialogue processing mode.

To build a list of keywords and hints to be used in the game, we extracted the heads of facts in ConceptNet that have more than five facts with two or more different relations. Finally, the authors and developers of Voice Assist selected appropriate keywords and hints from the candidates. Note that we did not use lexical knowledge.

3.3 Reducing ASR Errors

Player's utterances are likely to suffer from ASR errors because they are not accompanied by any context that is helpful for recognizing words. Table 1(a) provides examples. In Japanese, for example, "cheese" (チーズ chîzu) is sometimes recognized as "a map" (地図 chizu). To alleviate this problem, we automatically identify ASR errors and rewrite them into their correct forms.

We first identify ASR error pairs and the correct form based on pronunciations. We transcribe the collected words into *rōmaji*,[7] which represents their pronunciations, and calculate the string similarities between the transcribed words that were given in response to the same hint. We define the string similarity of transcribed strings X and Y as $1 - \frac{L(X,Y)}{\max\{|X|,|Y|\}}$, where L denotes the Levenshtein distance and $|\cdot|$ denotes the length of the string. The pair of words is considered to be identical if the similarity is higher

[6] We followed Nakahara (2011) and determined the maximum number of hints.

[7] We used KAKASI (`http://kakasi.namazu.org/index.html.en`) to convert a word into *rōmaji*.

Keyword	Recognized facts	Intended head
cake	(地図 chizu (map), *MadeOf*, milk)	チーズ chîzu (cheese)
orange	(佐藤 satô (Japanese family name), *HasProperty*, sweet)	砂糖 satô (sugar)
kitchen knife	(校長 kôchô (school principal), *UsedFor*, cut)	包丁 hôchô (kitchen knife)

(a) ASR errors: Intended heads were given by the authors.

Turn	Hint / Guess	Obtained facts
Hint (H) 1	This is herbivore.	
Guess (G) 1	Sheep.	(sheep, *HasProperty*, herbivore)
H2	This is yellow.	
G2	Tiger.	**(tiger, *HasProperty*, herbivore)**, (tiger, *HasProperty*, yellow)

(b) Hint distances: The hint distance of (tiger, *HasProperty*, herbivore) is two, and the others are one.

Table 1: **Illustration of inaccurate facts obtained from the quiz game.** For simplicity, only English translations are reported for some words.

than 0.7.[8] For example, チーズ (cheese) and 地図 (map) are identical because the their pronunciations, chîzu and chizu, are sufficiently similar.

Next, we rewrite the identical words to the correct words.We do not yet know which of the words is the correct form. The key to determining the correct form is that a player's guess will be semantically similar to the keyword because the player is attempting to answer the keyword in the game. We take cheese (チーズ chîzu) and a map (地図 chizu) given in response to the hint "this is made of milk," for example. Assume the keyword is "cake." We calculate the cosine similarities between the word vectors (Mikolov et al., 2013) of the guesses and the keyword, obtaining sim(cheese, cake) = 0.65, and sim(map, cake) = 0.13. Indeed, (cheese, *MadeOf*, milk) is correct, and (map, *MadeOf*, milk) is incorrect.

We assume that the word whose word vector is more similar to that of the keyword is correct. Thus, 地図 (map) is rewritten to チーズ (cheese), whose semantic similarity to the keyword, *i.e.*, sim(cheese, cake), is higher than sim(map, cake).

3.4 Aggregation of Acquired Knowledge

The quality of the acquired knowledge is not always good, and we must aggregate multiple facts obtained from the players to learn correct knowledge. To this end, we consider the following three aspects of the knowledge acquisition process from the quiz game.

1. Similarly to previous studies (von Ahn et al., 2006; Kuo and Hsu, 2011; Herdağdelen and Barobni, 2012; Nakahara, 2011), we assume that facts given by many players are likely to be correct. We use P_f to denote the set of players that answered the head of fact f. If P_f consists of many players, f obtains a high score.
2. A fact whose hint distance is large is less reliable than a fact whose hint distance is small because players focus on the last hint and tend to ignore earlier hints (see Table 1(b) for example). Thus, we weight frequency by hint distance (Section 3.2). The distance of fact f given by player p is denoted by $d(f, p)$. We weight fact f given by player p by $w_d(f, p) = g^{-d(f,p)}$, where g is a hyperparameter.
3. As explained in the previous section, a correct word is likely to be semantically similar to the keyword. Thus, we can also use the semantic similarity between a player's guess and keyword as prior knowledge during scoring. We define the weight of fact f as $w_s(f) = (\mathrm{sim}(f_{\text{head}}, f_{\text{keyword}}) + 1)/2$, where sim is the cosine similarity between word vectors, and f_{head} and f_{keyword} are the head and keyword of fact f, respectively.

Our goal is to give a high score to correct facts (e.g., (cheese, *MadeOf*, milk)) and a low score to incorrect facts (e.g., (noodle, *MadeOf*, cheese)). Combining the ideas above, we define the score of fact f as

$$\sum_{p \in P_f} w_d(f, p) \times w_s(f).$$

[8] We empirically determined the threshold using a small set of word pairs.

Unique players	Games	Utterances	Unique facts
74,375	206,305	588,189	155,683

Table 2: **Data collected by the quiz game from December 2015 to August 2016.**

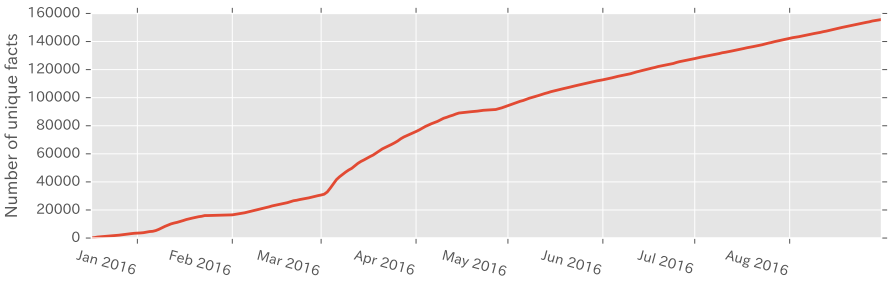

Figure 2: **Amount of newly acquired knowledge.** 98% of the facts did not exist in ConceptNet.

4 Empirical Evaluation

4.1 Knowledge Acquisition Speed

We created quizzes by following the procedure in Section 3.2 and released the game to a subset of the Voice Assist users in December 2015 and all users in March 2016. Table 2 shows the statistics of the collected data. At its peak, 1,300 unique players participated per day. These numbers are much larger than those of previous studies (for instance, 6,899 players over six months for the pet game (Kuo et al., 2009)).

Figure 2 shows the amount of newly acquired knowledge (*i.e.*, a cumulative sum of the number of unique facts). By March 2016, we had collected over 30,000 unique facts using only 92 hints. Because we increased the number of hints to 181 and published the game to all users in March 2016, the acquisition speed was accelerated. We obtained over 60,000 unique facts in March and 20,000 unique facts per month after that. There are 805 facts in the Japanese ConceptNet that have the same (*relation*, tail) as the hints added in March 2016. Surprisingly, 70% of them were obtained within the first nine days of our quiz game.

We cleaned the collected logs before this analysis because they contained meaningless utterances. **(1) Time out:** We discarded the rest of the utterances if the utterance interval exceeded one minute. **(2) Activation of Voice Assist functions:** If players uttered a command for one of the other functions such as calling, weather information or navigation, we considered it to be the last utterance of the session and discarded the rest of the utterances. **(3) Trivial utterances:** Utterances matching trivial patterns defined by the authors were discarded. **(4) Part-of-speech (POS):** Heads that do not meet a constraint on the POS of the relation type were filtered out from the extracted facts, where we used JUMAN++ (Morita et al., 2015) for morphological analysis. The constraints can be found in Speer and Havasi (2012).

4.2 Evaluation Using Crowdsourcing

We use crowdsourced judgments as the gold standard for evaluating the collected knowledge. We recruited crowd workers on Yahoo! Crowdsourcing[9] and evaluated 6,669 facts that were collected from multiple players. The facts were sampled from the data collected up to the end of February 2016. The workers answered whether a given fact was correct or not. We requested the judgments of five workers for each fact and aggregated them using the multi-class minimax entropy algorithm (Zhou et al., 2014).

The facts that were labeled as true consisted of 54% of all the facts. This is lower than those of previous studies because our game suffered from ASR errors. To obtain correct knowledge from such noisy collected facts, we needed to aggregate them and estimate confidence scores for each fact.

In this analysis, we validate the performance of the scoring method explained in Section 3.4 in terms of ROC-AUC, performing 3-fold cross validation on the evaluation set. To calculate a weight based on hint distances, we determined g by doing a grid search over $\{2, 4, 8, 16\}$, searching for the values that

[9] `http://crowdsourcing.yahoo.co.jp/`

	baseline		+ hint distance		+ semantic similarity		+ both	
Reducing ASR errors		✓		✓		✓		✓
	0.695	0.707	0.752	0.769	0.709	0.718	0.764	**0.777**

Table 3: **ROC-AUC of estimated confidence scores.** Note that the number of evaluated facts is different before and after ASR error reduction using the method explained in Section 3.3 (6,669 and 5,669 facts, respectively).

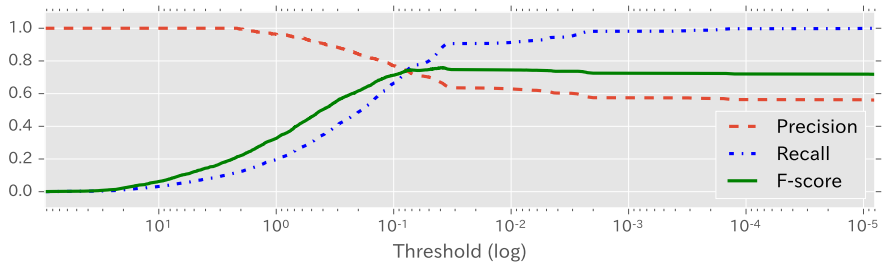

Figure 3: **Relationship between precision/recall/F-score and score threshold.**

maximized the ROC-AUC on the development set. We used word vectors of 500 dimensions that were trained on 9.8 billion Japanese sentences crawled from the Web.

Table 3 shows the ROC-AUC for the evaluation data. It indicates that when facts are ranked by hint distances and semantic similarity to keywords, the performance is better than only considering the number of players that answered the fact, as previous studies did.

4.3 Case Study of Knowledge Acquisition

Although the analysis using crowdsourcing in the previous section is efficient for validating the performance of the scoring methods, it is not sufficient for evaluating the quality of acquired knowledge. Thus, we need a detailed analysis of the acquired knowledge. We first divided the facts into four groups based on their scores and evaluated 100 facts for each group manually.

Figure 3 shows the precision and recall corresponding to thresholds on the scores. The F-score value reaches the highest at around 0.1. Hence, we consider the following four groups: (A) 100 highest-scoring facts, (B) 100 highest-scoring facts whose scores are below 1, (C) 100 lowest-scoring facts whose scores are above 0.1, and (D) 100 highest-scoring facts whose scores are below 0.1.

As the validity of a fact depends on the context in which it is used, we classified a fact into ASR errors and five classes: (5) always true, (4) true in many contexts, (3) true in several contexts, (2) true in only a few contexts, and (1) false.

Table 4 shows the evaluation results. We observe that the top-ranked facts contain knowledge that is true in general contexts, whereas the facts in the low-scoring groups include many context-dependent or incorrect facts. Table 5 provides examples for each group. In group (A), most of the 100 facts contain keywords or their synonyms and already exist in ConceptNet. This is because players tried to answer the keywords in the game. In contrast, 61, 90, and 92 out of 100 facts do not exist in ConceptNet in groups (B), (C), and (D), respectively.

As we expected, our quiz game obtained knowledge that is not likely to appear in texts. For example, (ladle, *AtLocation*, kitchen), (restaurant, *RelatedTo*, work), and (marriage proposal, *RelatedTo*, meal) cannot be found in Wikipedia, a corpus often used for knowledge acquisition.

However, low-scoring groups (C) and (D) contain some incorrect facts. The false facts include (Wu long tea, *MadeOf*, beans) in (C) and (tiger, *HasProperty*, herbivore) in (D). ASR errors appear most in group (C) because their hint distances are close to zero even though their frequencies and semantic similarities to keywords are low. Tackling these problems is left for future work.

4.4 Discussion

Our game gives a hint (*relation*, tail) to players and only obtains the head corresponding to the hint (Section 3.2). If the number of collected facts increases, we can analyze this knowledge for further details. Suppose we obtained the head “strawberry” from two hints “this is at supermarkets” and “this is

Group	5	4	3	2	1	ASR error
(A)	93	5	2	0	0	0
(B)	85	9	4	1	0	1
(C)	48	11	13	8	7	13
(D)	50	12	14	11	9	4

Table 4: **Evaluation results.**

	Keyword	Collected fact	Score	Judgment
(A)	hair dryer	(ドライヤー (hair dryer), *UsedFor*, 髪を乾かす (dry hair))	93.22	5
	hair dryer	(ハサミ (scissors), *AtLocation* 床屋 (barbershop))	15.27	5
	TV	(テレビ (TV), *AtLocation*, リビングルーム (living room))	12.96	4
(B)	money	(バット (bat) *MadeOf*, 金属 (metal))	0.987	4
	kitchen knife	(おたま (ladle), *AtLocation*, キッチン (kitchen))	0.911	5
	cafe	(レストラン (restaurant), *RelatedTo*, 仕事 (work))	0.803	3
(C)	cafe	(プロポーズ (marriage proposal), *RelatedTo*, 食事 (meal))	0.107	4
	cellphone	(テレビ (TV), *IsA*, 電話 (telephone))	0.105	2
	coffee	(ウーロン茶 (Wu long tea), *MadeOf*, 豆 (beans))	0.106	1
(D)	cake	(コーヒー牛乳 (coffee-flavored milk), *Causes*, 虫歯 (tooth decay))	0.099	4
	farmer	(おまわりさん (police officer), *IsA*, 仕事 (job))	0.098	5
	giraffe	(虎 (tiger) *HasProperty*, 草食 (herbivore))	0.098	1

Table 5: **Examples of collected facts.**

at farms." We can undertake further analyses such as comparing the frequencies of (strawberry, *AtLocation*, supermarket) and (strawberry, *AtLocation*, farm) to learn where people mostly think strawberries are. This would be beneficial for computers to understand humans' social communications.

We obtained more than 60,000 facts within a month using about 200 hints in March 2016. We expect to collect millions of items of knowledge over a year by continuously updating the hints. The collected resources will be freely available.

5 Conclusion

We developed a quiz game as a module in a widely used Japanese spoken dialogue system to obtain large-scale and high quality commonsense knowledge from many humans. We released the game in December 2016 and so far have collected over 150,000 unique facts from more than 70,000 players. In this paper, we reported the speed and quality of the knowledge acquisition process using the dialogue system quiz game. We also addressed the problem of aggregating the collected facts to obtain correct knowledge. We presented a simple scoring method that considers hint distances and semantic similarities between a player's guesses and the answer of the quiz. The experiments showed that when facts are ranked by using the scoring method, the performance is better than when only the number of players that answered the fact is considered, as previous studies did.

As future work, we will develop further acquisition and validation methodologies to obtain accurate commonsense facts. ASR errors are hard to avoid in a spoken dialogue system, and we must develop a more sophisticated approach to tackle this problem. Furthermore, although our current quiz game selects a quiz and hint at random, it would be more effective to select them based on a strategy. For example, Kuo and Hsu (2011) attempted to utilize the English ConceptNet to generate effective quiz games.

We will collect additional knowledge by updating the quizzes continuously, and expect that the number of acquired facts will reach more than one million in the near future, which would be significantly beneficial for various Japanese NLP applications.

Acknowledgments

We would like to thank participants in our game and developers of Yahoo! Voice Assist. We are also thankful for the valuable comments from the anonymous reviewers. This research was supported by the Leading Graduates Schools Program, "Collaborative Graduate Program in Design" by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 1247–1249, New York, New York, USA, May. ACM Press.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of The 3rd Workshop on Automated Knowledge Base Construction (AKBC)*. ACM Press, October.

Amaç Herdağdelen and Marco Baroni. 2012. Bootstrapping a game with a purpose for commonsense collection. *ACM Transactions on Intelligent Systems and Technology*, 3(4):1–24.

Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*, pages 506–516, New York, New York, USA, May. ACM Press.

Hayato Kobayashi, Kaori Tanio, and Manabu Sassano. 2015. Effects of game on user engagement with spoken dialogue system. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 422–426, Prague, Czech Republic, September. Association for Computational Linguistics.

Yen-Ling Kuo and Jane Yung-Jen Hsu. 2011. Resource-bounded crowd-sourcing of commonsense knowledge. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2470–2475, Barcelona, Spain, July. AAAI Press.

Yen-ling Kuo, Jong-Chuan Lee, Kai-yang Chiang, Rex Wang, Edward Shen, Cheng-wei Chan, and Jane Yung-jen Hsu. 2009. Community-based game design: Experiments on social games for commonsense data collection. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)*, pages 15–22, Paris, France, June. ACM Press.

Edith Law and Luis von Ahn. 2011. *Human computation*, volume 5 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers.

Douglas B. Lenat. 1995. Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, August. Association for Computational Linguistics.

Henry Lieberman, Dustin a Smith, and Alea Teeters. 2007. Common consensus: a web-based game for collecting commonsense goals. In *Proceedings of IUI 2007 Workshop on Common Sense for Intelligent Interfaces*, Honolulu, Hawaii, January. ACM Press.

Hugo Liu and Push Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Yuichiro Machida, Daisuke Kawahara, Sadao Kurohashi, and Manabu Sassano. 2016. Design of word association games using dialog systems for acquisition of word association knowledge. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction (AKBC)*, pages 86–91, San Diego, CA, USA, June. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 3111–3119, Stateline, Nevada, USA, December. MIT Press.

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2292–2297, Lisbon, Portugal, September. Association for Computational Linguistics.

Kazuhiro Nakahara and Shigeo Yamada. 2013. Social game the test for Japanesenes for common-sense knowledge acquisition (in Japanese). *Unisys Technology Review*, 32(4):389–401.

Kazuhiro Nakahara. 2011. Development and evaluation of a web-based game for common-sense knowledge acquisition in Japan (in Japanese). *Unisys Technology Review*, 30(4):295–305.

Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano. 2016. Prediction of prospective user engagement with intelligent assistants. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1203–1212, Berlin, Germany, August. Association for Computational Linguistics.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 3679–3686, Istanbul, Turkey, May. European Language Resources Association.

Niket Tandon, Gerard de Melo, Fabian Suchanek, and Gerhard Weikum. 2014. Webchild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the 7th ACM international conference on Web search and data mining (WSDM)*, pages 523–532, New York, New York, USA, February. ACM Press.

Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli. 2014. Validating and extending semantic knowledge bases using video games with a purpose. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1294–1304, Baltimore, Maryland, June. Association for Computational Linguistics.

Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: A game for collecting common-sense facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 75–79, Montréal, Québec, Canada., April. ACM Press.

Liangjun Zang, Cong Cao, Yanan Cao, Yuming Wu, and Cungen Cao. 2013. A survey of commonsense knowledge acquisition. *Journal of Computer Science and Technology*, 28(4):689–719, July.

Dengyong Zhou, Qiang Liu, John Platt, and Christopher Meek. 2014. Aggregating ordinal labels from crowds by minimax conditional entropy. In *Proceedings of The 31st International Conference on Machine Learning (ICML)*, pages 262–270, Beijing, China, June. ACM Press.