Design of Word Association Games using Dialog Systems for Acquisition of Word Association Knowledge

Yuichiro Machida[†,1] **Daisuke Kawahara**[†] **Sadao Kurohashi**[†] **Manabu Sassano**[‡]

[†]Graduate School of Informatics, Kyoto University
[‡]Yahoo Japan Corporation

machida@nlp.ist.i.kyoto-u.ac.jp, {dk, kuro}@i.kyoto-u.ac.jp, msassano@yahoo-corp.jp

Abstract

We present a design for acquiring word association knowledge of high quality on the basis of a game with a purpose (GWAP). We evaluate automatically acquired word associations using a word association game as a GWAP. In the word association game, a player is given a set of associated words as a hint and is asked to answer a word that can be associated with the hint. If many players can answer the correct keyword, we judge the set of associated words to be of high quality. This word association game was implemented in a smartphone-based dialog system, which has been installed into more than one million smartphones. Our analysis of numerous game logs demonstrated that our framework can effectively select word associations of high quality.

1 Introduction

In recent years, semantic analysis has received attention as an active research area in natural language processing (NLP). It is indispensable to build language resources for accurate semantic analysis. One such resource is word associations, e.g., "glass" has an association with "fragile," "cup," and "reflect." Compiling large-scale word association knowledge of high quality is important to capture the semantic relations among words in texts, enabling deep and accurate discourse analyses. Hereafter, we designate the target word as **keyword**[2]. Words with which the keyword has an association are designated as **associated words**.

There are manually crafted resources of word associations, such as thesauri and ontologies, which have been compiled by lexicographers and which have contributed to many studies in NLP. However, it takes long times and high costs to create a large thesaurus. Furthermore, it is difficult to update it continuously to adapt to neologisms and the changing use of a word.

To reduce the creation cost, methods for automatically acquiring word associations from a large corpus have been studied (e.g., (Lin, 1998; Mikolov et al., 2013)). Although such methods can acquire large-scale resources of word associations, they tend to have lower precision than manually created ones, which would harm the performance of subsequent NLP applications.

To cope with potential difficulties existing both in manual methods and automatic ones, it is necessary to combine the two methods, taking their respective benefits. This paper presents a method for compiling large-scale word association knowledge of high quality by evaluating automatically acquired word associations manually, not by linguistic experts but by the wisdom of crowds.

To make use of the wisdom of crowds, crowdsourcing has been employed widely (e.g., (Snow et al., 2008; Kawahara et al., 2014)). Crowdsourcing is a low-cost service that enlists numerous human workers to make judgments that are difficult for computers.[3] However, costs are still high when we

[1]The first author is now affiliated with Recruit Lifestyle Co., Ltd.

[2]A keyword can be a word or a phrase, but we call both "keywords."

[3]In this paper, we refer to microtask crowdsourcing as "crowdsourcing."

conduct a very large-scale task by aggregating many small-cost microtasks. We employ a game with a purpose (GWAP) (von Ahn, 2006) to use the wisdom of crowds. GWAP dispatches tasks as a game, and thus it is the process by which a game play implicitly corresponds to the execution of another task. Since game players do not want money but instead want fun, the use of GWAP engenders greater cost reductions than crowdsourcing.

We perform a **word association game** using a dialog system on smartphones as GWAP. We evaluate the quality of automatically acquired Japanese word associations based on logs obtained from game players. For example, if players correctly answer the keyword "glass" for the given associated words "fragile," "cup," and "reflect," these associated words can be regarded as high quality for the keyword. We use such a game to evaluate automatically acquired word associations at no cost.

2 Related Work

In recent years, crowdsourcing has been used for data construction and evaluation in NLP (e.g., (Snow et al., 2008; Hill et al., 2015; Schnabel et al., 2015)). Snow et al. (2008) demonstrated that annotations by crowdworkers have almost identical quality with those by experts in various NLP tasks. The motivation of crowdworkers in crowdsourcing is monetary.

Another type of wisdom of crowds is GWAP, for which a player's motivation differs from that of crowdsourcing. Their motivation is "enjoying the game." Therefore, we need not pay for the players, and can reduce the number of low-quality or dishonest workers. Many approaches using GWAP have been proposed in the field of NLP, such as anaphora resolution (Hladká et al., 2009), paraphrasing (Poesio et al., 2013), constructing semantic network (Lafourcade, 2007), and word sense disambiguation (Venhuizen et al., 2013). However, they are designed in a text-based style. Text-based games are probably less enjoyable for players.

Some studies have specifically examined non-text-based games, i.e., video games (Vannella et al., 2014; Jurgens and Navigli, 2014). Video games are familiar for ordinary people and are much more enjoyable than text-based games. For example, Vannella et al. (2014) developed a video game to validate the associations between images and senses. They reported that the annotation quality using the video game is better than crowdsourcing. The enjoyable game design improves the quality of annotations by crowds. However, because playing a video game requires a certain amount of time, it is a bit difficult to play it in one's spare-time. Furthermore, developing an attractive video game would be a time-consuming task.

Our word association game works on a dialog system to encourage player motivation. The game progresses interactively, so that many players can play easily. Moreover, we need not spend much time to develop it because the game system is simple.

3 Automatic Acquisition of Word Associations

We first explain our model for acquiring word associations. Then, we describe a method for clustering the acquired word associations to efficiently make questions for the word association game.

3.1 Definition and Collection of Word Associations

In general, many relations exist among words, such as hypernym-hyponym relations and part-whole relations. However, we do not care about the kind of relation but the strength between words. Matsuo et al. (2006) used pointwise mutual information (PMI) and chi-square as the strength measures of relations, and acquired associated words using graph clustering. Our method is based on this method, but uses word frequencies and PMI as the strength measures as proposed by Shin and Kurohashi (2014). We used a Japanese Web corpus of 4.2 million sentences and Japanese Wikipedia to acquire associated words for nouns, verbs, and adjectives.

3.2 Clustering Word Associations

Next, we cluster the acquired associated words according to their basic meanings. By clustering associated words, we can not only structure associated word knowledge, but also efficiently produce questions for the word association game to evaluate the acquired associated words. Presume that there is a fruit cluster like $\{$fruit, banana, orange, $\cdots\}$ among

associated words of "apple." We can evaluate the quality not from all words in the fruit cluster but only from a few words in the cluster. We can reduce the number of questions for a keyword using clustering. We perform clustering based on the Girvan-Newman algorithm (Girvan and Newman, 2002).

4 Design of Word Association Game

To evaluate the quality of a set of automatically acquired word associations, we designed a word association game.

Our key idea is that humans can associate a given set of associated words with a keyword or its similar words if these associated words are of high quality. For example, if players answer the keyword "glass" or its similar word "window" for the given associated words "fragile," "cup" and "reflect," these associated words can be regarded as high quality for the keyword. Based on this idea, we conduct a word association game in which a player is given a set of automatically acquired associated words as a hint and is asked to answer a word that can be associated with the hint. We compare the answers of players with the keyword, and judge that the acquired associated words are of high quality if many answers are the keyword or similar words.

In this way, we do not evaluate the quality of each associated word for a keyword. We instead acquire large-scale word association knowledge of high quality verified as a whole. By using this strategy, we do not require a large number of annotations, i.e., game plays, for the quality judgment of a keyword. For keywords that are judged to be of low quality using our method, it is possible to conduct another stage of crowdsourcing to evaluate each associated word.

The word association game operates on a dialog system on a smartphone application as shown in Figure 1. We adopted a dialog system as a platform because enjoyment of conversations can promote game plays and vice versa. We developed a word association game as a dialog chatting function in Yahoo! Onsei Assist (Yahoo Japan's Voice Assistant, hereinafter YOA),[4] which is a spoken dialog smartphone application provided by Yahoo Japan.

[4]http://v-assist.yahoo.co.jp/

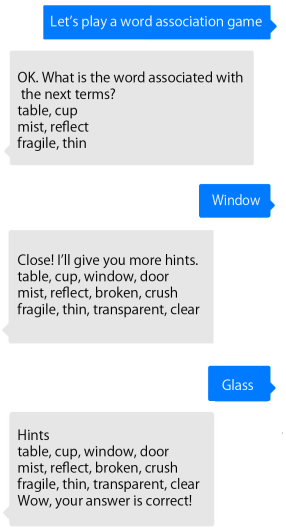

Figure 1: Word association game. Blue balloons show utterances of the player. All utterances of the player and the system are translated into English.

4.1 Method for Making Word Association Game

The word association game is executed as follows.

1. If a player utters a sentence such as "連想ゲームしよう" (Let's play a word association game) in the mode of normal dialog processing, the word association game is booted up.[5]
2. A keyword is selected randomly from a pool of keywords.
3. One noun cluster, verb cluster, and adjective cluster of associated words are randomly selected.
4. A question is created using two words as a hint in order of the degree of association in each selected cluster (six words in total).
5. A player utters an answer to the question.
6. The answer is judged. If it corresponds to the keyword, the game ends and YOA returns to the mode of normal dialog processing; otherwise steps 3 and 4 are called again to add more hints to the question.
7. The player utters an answer to the revised question.
8. The answer is judged. In any case, the game ends and YOA returns to the mode of normal dialog processing.

[5]It was announced to YOA users that they can play a word association game.

For our experiment, we selected 100 keywords ranked between 3,000 and 10,000 in order of word frequency on the Web. Of these, we manually judged that 85 keywords have a high-quality set of associated words; 15 keywords do not.[6] These manual labels are used as a gold standard set in Section 5.2. Since high enjoyment for a game requires a somewhat high ratio of keywords with high-quality associated words, we set this ratio as high in our first-phase experiment for evaluation purposes. To practically operate the game, we would be able to lower the ratio to around half using keywords evaluated to be of high quality at the moment and keywords without evaluation.

4.2 Answer Judgment

At steps 6 and 8 in the procedure described above, the player answer is judged automatically as one of the following three classes: "Exact," "Near" and "Bad." When the answer of the player is the same as the keyword, the judgment is "Exact." When the answer is similar to the keyword, the judgment is "Near." Similarity between words is calculated based on distributional/distributed similarity (Lin, 1998; Mikolov et al., 2013). The answer of a player is judged as "Near" if its similarity to the keyword is greater than a threshold.

5 Evaluating Word Associations with the Word Association Game

We evaluated word associations by analyzing game logs of our word association game. The word association game has been operating since the middle of December 2014. To date, 87 keywords out of 100 have been used for the game. We analyzed game logs of approximately a month and a half.

From the game logs of our word association game, we can obtain the utterance time, player ID, and utterance contents of each player. Table 1 lists statistics of game logs for a month and a half.

5.1 Method for Aggregating Answers

To evaluate a set of associated words for a keyword, we adopted five answers that are highly ranked in order of frequency in the answers for the keyword. The purpose of this is to exclude unintentional utterances and errors of speech recognition caused by spoken dialog. We calculated precision for a keyword using its five answers as follows:

$$\text{precision} = \frac{|Exact| + |Near|}{|Exact| + |Near| + |Bad|},$$

where $|Exact|$ denotes the frequency of answers judged as "Exact." $|Near|$ and $|Bad|$ are defined in the same way.

We evaluated the quality of word associations in terms of the precision defined above. We judged the associated words of a keyword to be of high quality if its precision is higher than 0.3, and vice versa. This threshold was set empirically by considering the remaining noises of speech recognition and similar words that failed to be recognized.

Player IDs	9,997
Plays	19,438
Exact	6,930
Near	4,470
Bad	10,895
Mode response interval	20.0s
Average plays per player	1.9
Max plays per player	59

Table 1: Statistics of game logs of a month and a half.

		human annotation	
		appropriate	inappropriate
Game log	appropriate	70	6
	inappropriate	6	5

Table 2: Comparison with human annotations.

5.2 Acquired Evaluations of Word Associations

We obtained automatic judgments for the target 87 keywords by aggregating answers. We compared these automatic judgments with the gold standard labels described in Section 4.1. Table 2 reports this comparison. We achieved a precision of 0.92, a recall of 0.92, and an F-score of 0.92. From these results, we can see that the quality of automatically acquired word associations can be evaluated precisely using our word association game.

5.3 Analysis of Word Associations with High Precision

High precision basically means high quality of associated words for a keyword. Table 3 shows play-

[6] In the experiment, 87 keywords out of these 100 were actually used for the word association game because of the specifications of YOA.

keyword	question examples	players' answers	prec
event	site, live, information, various inform, festival, boost special, big, official, latest	**event**:41 party:19 concert:16 festival:12 sports festival:9	0.907
	memorial, plan, information, various enjoy, fun, festival, boost one-man, great, safety		
	memorial, plan, information, various inform, festival, boost special, big, official, latest		
ice cream	chocolate, cake, ice, milk taste, smell, serve with, bake delicious, sweet, fresh, plenty	**ice cream**:55 cake:33 ice:32 pudding:19 soft ice cream:17	0.891
	milk, fresh cream, dessert, yogurt mix, melt, eat, make dense, smooth, cold, hot		
	chocolate, cake, milk, fresh cream taste, smell, mix, melt delicious, sweet, dense, smooth		

Table 3: Examples of players' answers with high precision. All keywords, question examples, and players' answers are translated into English. Bolded words and underlined words are judged as "Exact" and "Near," respectively.

keyword	question examples	players' answers	prec
line	distance, horse, curve, leg reach, loop, get away from, go past vertical, zigzag, long, fast	horse race:101 leg:7 horse:7 race track:6 marathon:5	0.0
	count, length, distance, horse face, run, reach, loop simple, round, vertical, zigzag		
	count, length, curve, leg face, run, get away from, go past simple, round, long, fast		
theme	lecture, this, work, exhibition along, mistake, summarize, drill down forever, important, familiar, main	university,29 space,10 STAP cell,10 ocean,7 chemical,6	0.0
	lecture, this, common, discussion along, mistake, tackle, learn forever, important, various, wide range of		
	research, paper, lecture, this decide, draw, along, mistake ambitious, deep, forever, important		

Table 4: Examples of players' answers with low precision. All keywords, question examples, and players' answers are translated into English.

ers' answers for some keywords that have high precision. Players' answers for the keyword "アイスクリーム" (ice cream) include similar words to the keyword, such as "ケーキ" (cake), "アイス" (ice), and "プリン" (pudding). Since the set of associated words includes appropriate associated words, the high precision actually indicates the quality of associated words. Although "ソフトクリーム" (soft ice cream) included in the players' answers is a similar word to the keyword, it cannot be identified as "Near." It is necessary to increase the coverage of similar word identification in future studies.

Furthermore, it is possible to acquire new associated words from players' answers with high precision. In Table 3, players' answers for the keyword "イベント" (event) include "パーティー" (party), "コンサート" (concert) and "運動会" (sports festival) in addition to the keyword itself. Although "パーティー" (party) and "コンサート" (concert) were automatically acquired as associated words, "運動会" (sports festival) was not acquired but an appropriate associated word. In this way, we can acquire novel associated words from the word association game.

5.4 Analysis of Word Associations with Low Precision

Table 4 shows players' answers for some keywords that have precision lower than 0.3. We can see that the low precision means low quality of associated words for a keyword. For example, the keyword "直線" (line) has low precision. Its players' answers were not similar to the keyword but were words that are related to horse racing. Actually, the associated words for this keyword, such as "距離" (distance) and "コーナー" (curve), can be readily associated with horse racing. Therefore, they are unsuitable for associated words for "直線" (line).

Furthermore, players' answers with low precision help not only find associated words of low quality but also analyze errors in the method for acquiring associated words. The above associated words were associated with horse racing because the domain of the Web corpus was biased to horse racing. In this way, we can speculate about the contexts of the acquired associated words, which correspond to error analysis of the acquisition method.

6 Conclusion

We have presented a method for compiling large-scale word association knowledge of high quality. We first automatically acquire and cluster word associations. Then we evaluate these by taking advantage of GWAP. The framework used for evaluating word associations is implemented as a word association game operating on a smartphone-based dialog system. Our analysis of a large volume of game logs has indicated that our framework can extract word associations of high quality effectively.

References

Michelle Girvan and Mark E. J. Newman. 2002. Community structure in social and biological networks. *Proc. of National Academy of Sciences*, 99(12):7821–7826.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Barbora Hladká, Jiří Mírovský, and Pavel Schlesinger. 2009. Play the language: Play coreference. In *Proc. of ACL-IJCNLP 2009*, pages 209–212.

David Jurgens and Roberto Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464.

Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. 2014. Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proc. of COLING2014*, pages 269–278.

Mathieu Lafourcade. 2007. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th International Symposium on Natural Language Processing*, page 7, Pattaya, Chonburi, Thailand.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of COLING-ACL98*, pages 768–774.

Yutaka Matsuo, Takeshi Sakaki, Kôki Uchiyama, and Mitsuru Ishizuka. 2006. Graph-based word clustering using a web search engine. In *Proc. of EMNLP2006*, pages 542–550.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. of Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3:1–3:44.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proc. of EMNLP2015*, pages 298–307.

Yoshiharu Shin and Sadao Kurohashi. 2014. Graph-based representation of documents based on nominal related words knowledge and its applications. In *Proc. of NLP2014 (in Japanese)*, pages 1007–1010.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP2008*, pages 254–263.

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 *Proc. of ACL2014*, pages 1294–1304.

Noortje J. Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for word sense labeling. In *Proc. of 10th International Conference on Computational Semantics (IWCS 2013)*, pages 397–403.

Luis von Ahn. 2006. Games with a purpose. *IEEE Computer*, 39(6):92–94.