Guess What:
A Question Answering Game via On-demand Knowledge Validation

Yu-Sheng Li
National Taiwan University
Taipei, Taiwan
b03902086@ntu.edu.tw

Chien-Hui Tseng
National Taiwan University
Taipei, Taiwan
r05725004@ntu.edu.tw

Chian-Yun Huang
National Chiao Tung University
Hsinchu, Taiwan
yun626.cs02@nctu.edu.tw

Wei-Yun Ma
Institute of Information Science, Academia Sinica
Taipei, Taiwan
ma@iis.sinica.edu.tw

Abstract

In this demo, we propose an idea of on-demand knowledge validation and fulfill the idea through an interactive Question-Answering (QA) game system, which is named Guess What. An object (e.g. dog) is first randomly chosen by the system, and then a user can repeatedly ask the system questions in natural language to guess what the object is. The system would respond with *yes/no* along with a confidence score. Some useful hints can also be given if needed. The proposed framework provides a pioneering example of on-demand knowledge validation in dialog environment to address such needs in AI agents/chatbots. Moreover, the released log data that the system gathered can be used to identify the most critical concepts/attributes of an existing knowledge base, which reflects human's cognition about the world.

1 Introduction and Script Outline

Knowledge validation (Merlevede and Vanthienen, 1991; Nazareth, 1989) aims to validate newly acquired knowledge. Most research work addresses the issue on text domain other than dialog environment. As the techniques and applications of AI agent and chatbot become mature and practical these days, the need of on-demand knowledge validation in the dialog environment is critical as the system needs to validate new knowledge acquired from users' words. Therefore we propose an interactive QA game between system and users, named Guess What[1] to fulfill the need in dialog environment. The demo presentation will be utilizing this web site to showcase our system in either Chinese or English version. Guess What is a variant of Twenty Questions game, which involves players taking the roles of the answerer and the questioners. The answerer chooses an object and conceal it to the other players. The questioners then ask *yes/no* questions to narrow down the wide range of the categories to which the object belongs. The question can be: "Is it animal?" or "Can it fly?", etc. The game terminates when the correct object is guessed by the questioners. Guess What is a kind of Chinese-based Twenty Questions game, where the system serves as the answerer and users as questioners. The answer set of the system currently contains 200 terms, which are general concepts such as dog, cat, boat, computer, etc. Figure 2 shows a running example of Guess What system.

The framework involves different research topics, such as question answering (Berant et al., 2013; Kwok et al., 2001) and relation prediction (Xu et al., 2016). The techniques include understanding the questions and identifying whether the object fits the description of the users' questions. Since most descriptions are based on the existence of a relationship between two entities, such as "Is it an animal?" or "Can it fly?", the latter mission turns out to be identifying whether a certain relationship between entities holds or not, which is a kind of on-demand knowledge validation.

Guess What goes through the following procedures: Parsing the user's question, followed by extracting knowledge and reasoning from metadata of Wikipedia[2] and a lexical semantic representation model named E-HowNet[3] (Ma and Chen, 2009; Chen et al., 2005). If the related knowledge

[1] http://guess-what.com.tw

[2] http://www.wikipedia.org/

[3] http://ehownet.iis.sinica.edu.tw/index.php

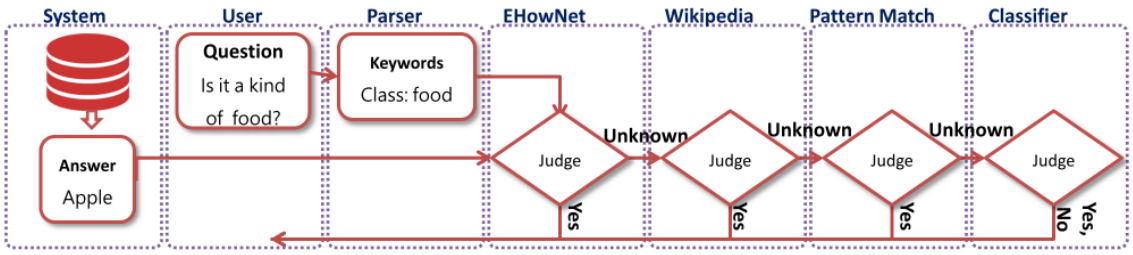

Figure 1: The process flow of Guess What system

Figure 2: A running example (screenshot) with lion as the answer in Guess What.

cannot be found, a pattern matching procedure and a classifier trained with online textual resources, such as Google Search results, are further applied. Figure 1 shows the process flow.

2 Question Understanding

In order to analyze the question, the system will parse the question through a Chinese parser, named CKIP parser[4], and the parsed question is then used to extract out a representative triple ⟨target, relation, withWhom⟩ via a set of extraction rules. The *target* is the answer term. The set of *relation*s consists of class, attribute, act, subject&act, act&object, location, and time. The *withWhom* is the corresponding term extracted from the question sentence which is in the certain relation to the target. Table 1 shows some examples of questions and their parsed triples where the answer is "bee."

3 Knowledge Validation

The following steps work with the triples parsed by the previous step, trying to figure out whether the relationship represented by each

[4]http://parser.iis.sinica.edu.tw/

Questions	Triples
Is it an animal?	⟨bee,class,animal⟩
Is it red?	⟨bee,attribute,red⟩
Can it fly?	⟨bee,act,fly⟩
Can it gather food?	⟨bee,act&obj, gather&food⟩

Table 1: Questions and their parsed triples

triple holds or not. For example, for the question "Can it gather food?", its triple is "⟨bee,act&object,gather&food⟩", Our goal is to validate the triple.

3.1 E-HowNet

3.1.1 Introduction

Extended-HowNet (E-HowNet) is a frame-based entity-relation model in Chinese and English, annotated by hand. Currently there are more than 100,000 entities in E-HowNet. Take bee for example. The definition of bee on E-HowNet is { InsectWorm : predication = { gather : theme = { food : source = { FlowerGrass } } , agent = { ~ } } }, as illustrated in Figure 3.

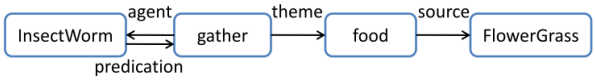

Figure 3: Graphical illustration of definition of bee

The above structure can be phrased in natural language as "bee is an insect whose predication is to gather food from flower." Here we ignore the actual definition of InsectWorm and FlowerGrass for simplicity, and the term *agent* means that bees are the subject of the action gather. The ~ symbol refers backwards to InsectWorm in this case.

3.1.2 Usage

We first use E-HowNet to validate if two entities have a certain relation. For example, for re-

lation "act," , we check out whether there is some predication link to the withWhom term in the E-HowNet definition of the answer. Furthermore, for relation "actobject", we examine if there is some theme or patient link to the withWhom term. Since E-HowNet contains less information about time and location, E-HowNet is not used for the two types of relations.

3.2 Wikipedia

E-HowNet can provide a certain level of common sense, and it sometimes still lack comprehensive common sense and some necessary domain knowledge in order to validate the given questions. This is where Wikipedia can bring the contribution. For almost every Wikipedia title, there are some related categorical hyperlinks at the bottom. If we build edges between these hyperlinked pages and regard the whole Wikipedia categorical hyperlinks as a graph, any given triple can also be validated through the graph. For instance, in the page of bat[5] there are

Bats ; Animal flight ; Pollinators ; Night ; Cave organisms ; Extant Ypresian first appearances ; Animals that use echolocation

Now we know that bats can fly, can pollinate, might be nocturnal, might live in caves, and can use echolocation. Moreover, these are not merely class-type categories, but also information about ability, location, etc.

3.3 From Online Search Texts

The information in knowledge bases is relatively refined but limited while the content on the Internet is relatively rich. Therefore, when searching the knowledge bases is insufficient to claim the relationship between the entities pair doesn't exist, we turn to online resources for more information. For each term in the answer set of the system, we collect textual data from the following sources:

1. top 10 pages of Google search results with the answer term as the query
2. the article of the answer term in Wikipedia
3. the article of the answer term in Baidu Baike[6]
4. the sentences containing the answer term in Academia Sinica Balanced Corpus (ASBC) [7]

With the help of these data, we apply pattern matching and use a classifier to check whether the relationship between the target and withWhom term holds or not.

- SVM Classifier
  We regard the validation of the relationship represented in the triple as a binary classification problem with two classes, *yes* and *no*. For each triple, we extract four features from the textual resource about the target term. There is no difference in different relations in the way extracting features. The four features are listed below.

  Denote $term_q$ = withWhom, $term_a$ = target. Define the the distance of two terms to be the number of words between them

  1. Proportional frequency: Number of sentences containing $term_q$ divided by total number of sentences.
  2. Average distance of $term_a$ and $term_q$
  3. Shortest distance of $term_a$ and $term_q$
  4. Word vectors similarity: By utilizing the word vector model (word2Vec package)[8]trained with ASBC corpus, we can get the vectors in 300 dimensions of terms. We compute the cosine similarity between vectors of $term_a$ and $term_q$ as a feature.

4 Experiments and Discussion

In order to evaluate the performance of each component in the procedure, we designed a testing set with 792 ⟨question, answer, *yes*/*no*⟩ triples, such as ⟨Is it an animal, monkey, *yes*⟩. There are 112 distinct answers and each answer is paired with about 7 questions on average, where questions are manually generated. There are 208 *yes*-labeled and 584 *no*-labeled triples in this testing set. Different kinds of relations are included in these questions as shown in table 2.

Type	class	attr.	act	location	total
Number	181	304	246	61	792

Table 2: Number of each type of questions

[5]http://en.wikipedia.org/wiki/Bat
[6]http://baike.baidu.com/
[7]http://asbc.iis.sinica.edu.tw/
[8]https://code.google.com/archive/p/word2vec/

Table 3 shows performance of each component in our experiments. From the table we can find that E-HowNet, Wikipedia and pattern matching have high precision but low recall, while classifier has relatively low precision and high recall. In summary, in the whole process, E-HowNet, Wikipedia, and pattern matching will be applied first to give reliable predictions. If the corresponding information is not found, the classifier will compensate for the recall. As a result, the whole process achieved the best F1-score.

	Precision	Recall	F1-score
E-HowNet	0.9158	0.4183	0.5743
Wikipedia	1.0000	0.0962	0.1754
Pattern	0.9500	0.0913	0.1667
Classifier	0.7135	0.6587	0.6850
Overall	0.7585	0.7548	0.7566

Table 3: Performance of each component

5 Log Analysis

The system records every question asked by users. Since the latest version of the system was launched, we have recorded 667 games, which contain 5016 questions in total. After removing 277 illegal question sentences (which don't contain 'it' in the sentence) and 274 direct answer term matching, there are 4465 questions in remaining. There are 257 distinct users (identified by their IP addresses) and each user played 2.6 games on average. We summarize the first question which users tend to ask in the game. The top frequently asked types of questions are shown in Table 4, which reflects the most critical concepts/attributes of human's mind.

6 Conclusion

The game system presented in this paper involves a mixture of information extraction techniques. The main contributions include being as a pioneering example of on-demand knowledge validation in dialog environment to address such needs in AI agents/chatbots, and comprehensive analysis of the log data, which can be used to guide the construction of a new knowledge base or be used to identify the most critical concepts/attributes of an existing knowledge based to reflect human's cognition about the world. In the future, we will work on expanding the existing answer set and further develop knowledge inference mechanisms to utilize indirect evidences with the online textual data.

Rank	Type	Examples	Count
1	human beings	Is it human?	96
2	animal	Is it a kind of animal?	59
3	food	Is it a kind of food? Is it edible?	59
4	living beings	Is it a kind of living beings? Is it alive?	50
5	fly	Can it fly?	10
6	occupation	Is it a kind of occupation?	4
7	thing	Is it a kind of thing?	4
8	plant	Is it a kind of plant?	3

Table 4: Types of questions frequently asked as the first one in the game

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, volume 2, page 6.

Keh-Jiann Chen, Shu-Ling Huang, Yueh-Yin Shih, and Yi-Jun Chen. 2005. Extended-hownet-a representational framework for concepts. In *OntoLex 2005-Ontologies and Lexical Resources IJCNLP-05 Workshop*.

Cody Kwok, Oren Etzioni, and Daniel S Weld. 2001. Scaling question answering to the web. *ACM Transactions on Information Systems (TOIS)*, 19(3):242–262.

Wei-Yun Ma and Keh-Jiann Chen. 2009. Lexical semantic representation and semantic composition: An introduction to e-hownet. Technical report, CKIP Group, Academia Sinica.

P Merlevede and Jan Vanthienen. 1991. A structured approach to formalization and validation of knowledge. In *Developing and Managing Expert System Programs, 1991., Proceedings of the IEEE/ACM International Conference on*, pages 149–158. IEEE.

Derek L Nazareth. 1989. Issues in the verification of knowledge in rule-based systems. *International Journal of Man-Machine Studies*, 30(3):255–271.

Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. 2016. Improved relation classification by deep recurrent neural networks with data augmentation. *arXiv preprint arXiv:1601.03651*.