

# PAL: A Chatterbot System for Answering Domain-specific Questions

Yuanchao Liu<sup>1</sup> Ming Liu<sup>1</sup> Xiaolong Wang<sup>1</sup> Limin Wang<sup>2</sup> Jingjing Li<sup>1</sup>

<sup>1</sup> School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

<sup>2</sup> School of public health, Harbin Medical University, Harbin, China

{lyc,mliu,wangxl,jjl}@insun.hit.edu.cn, wanglimin2008@163.com

## Abstract

In this paper, we propose PAL, a prototype chatterbot for answering non-obstructive psychological domain-specific questions. This system focuses on providing primary suggestions or helping people relieve pressure by extracting knowledge from online forums, based on which the chatterbot system is constructed. The strategies used by PAL, including semantic-extension-based question matching, solution management with personal information consideration, and XML-based knowledge pattern construction, are described and discussed. We also conduct a primary test for the feasibility of our system.

## 1 Introduction

A wide variety of chatterbots and question-and-answer (Q&A) systems have been proposed over the past decades, each with strengths that make them appropriate for particular applications. With numerous advances in information construction, people increasingly aim to communicate with computers using natural language. For example, chatterbots in some e-commerce Web sites can interact with customers and provide help similar to a real-life secretary (DeeAnna Merz Nagel, 2011; Yvette Colón, 2011).

In this paper, we propose PAL (Pychologist of Artificial Language), a chatterbot system for answering non-obstructive psychological questions. Non-obstructive questions refer to problems on family, human relationships, marriage, life pressure, learning, work and so on. In these cases, we expect the chatterbot to play an active role by providing tutoring, solution, support, advice, or even sympathy depending on the help needed by its users.

The difference of PAL from existing chatterbots lies not only in the specific research focus of this paper but also in the strategies we designed, such as P-XML templates for storing a knowledge base, comprehensive question matching method by considering both index and semantic similarities, and solution management by considering personal information. In the following sections, we will briefly discuss related work and then introduce our system and its main features.

## 2 Related Work

A number of research work on chatterbots (Rafael E. Banchs, Haizhou Li, 2012; Ai Ti Aw and Lian Hau Lee, 2012), Q&A systems (Shilin Ding, Gao Cong, Chin-Yew Lin, 2008; Leila Kosseim, 2008; Tiphaine Dalmas, 2007), and related natural language understanding technologies have recently been conducted (Walid S. Saba, 2007; Jing dong, 2009). Several studies on the application of natural language processing technologies for non-obstructive psychological Q&A systems have also been published (Hai-hu Shi, 2005).

Several online psychology counselling Web sites with service provided by human experts have also been established recently (DeeAnna Merz Nagel, 2011; Yvette Colón, 2011). For these Web sites, when the visitors ask similar questions, the expert may provide the same or very similar answers repeatedly. Based on this observation and consideration, we collected a large number of counselling Q&A pairs to extract common knowledge for the construction of a chatterbot system. Advances in automatic language analysis and processing are used as the bases for the emergence of a complex, task-oriented chatterbot system.

### 3 Basic Framework of PAL

A running screenshot of PAL is shown in Figure 1, and its basic system structure is demonstrated in Figure 2. As shown in Figure 2, the basic principles of PAL are as follows:

1) All interactions between system and users are

scheduled by control logic;

2) When the user inputs a question, the system will search through its knowledge base for the matching entry, and then

3) The system will respond with an appropriate answer by analysing both the matched entry and the dialogue history.

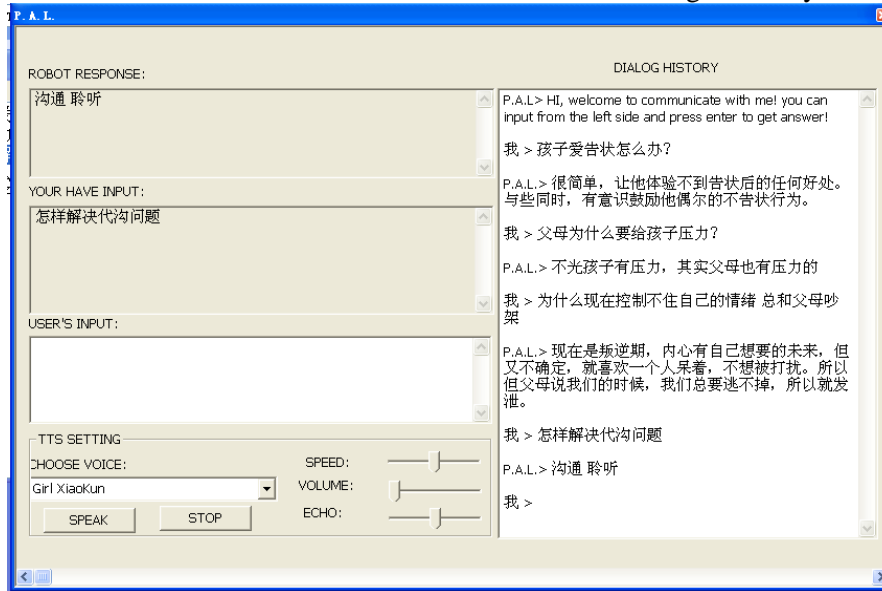


Figure 1. Running Screenshot of PAL

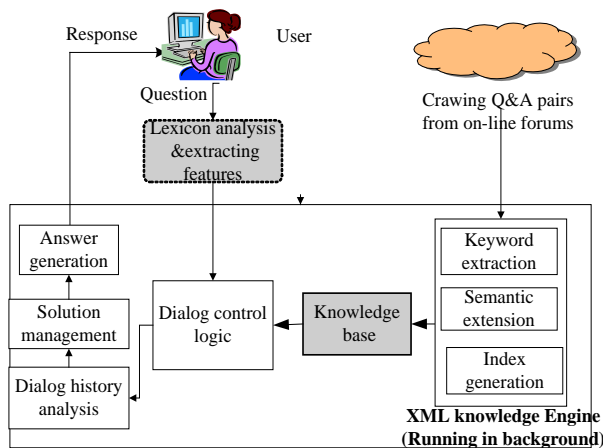


Figure 2. Basic Framework of PAL

### 4 Conversation Control Strategy of PAL

The Q&A process of the PAL system is coordinated by control logic to communicate with users effectively. The basic control logic strategy is shown in Figure 3.

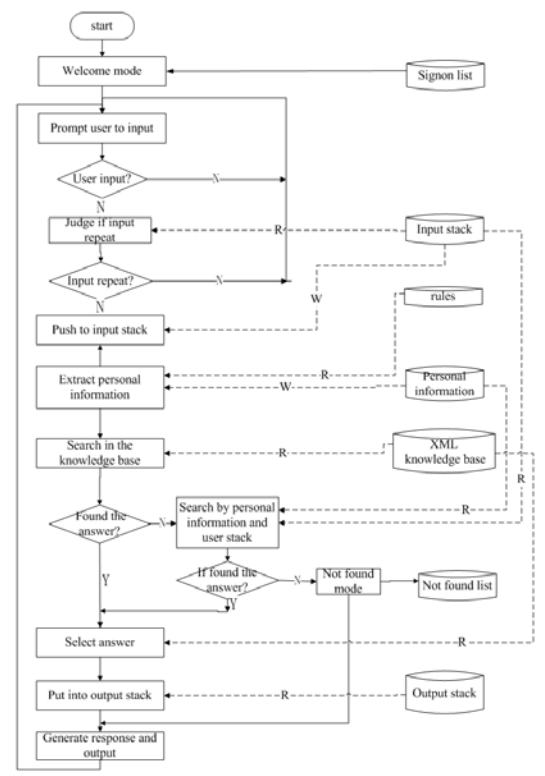


Figure 3. Basic Control Logic of PAL

As shown in Figure 3, the initial state is set to welcome mode, and the system can select a sentence from the “sign on” list, which will then provide a response. When users enter a question, the system will conduct the necessary analysis. The system’s knowledge base is indexed by Clucene<sup>1</sup> beforehand. Thus, the knowledge index will be used to search the matched records quickly. If the system can find the matched patterns directly and the answer is suitable for the current user, one answer will be randomly selected to generate the response. Historical information and personal information will be analysed when necessary. We mainly adopted the method of ELIZA<sup>2</sup>, which is an open-source program, to consider the historical information. A “not found” response list is also set to deal with situations when no suitable answers can be identified. Both system utterance and user input will be pushed into the stack as historical information. Given that user questions are at times very simple, the combination with historical input may also be required to determine its meaning. This step can also avoid the duplication of utterances.

## 5 Knowledge Construction and Question Matching Method

We design P-XML to store the knowledge base for PAL, as shown in Figure 4. The knowledge base for PAL is mainly derived from the Q&A pairs in the BAIDU ZHIDAO community<sup>3</sup>. One question usually has many corresponding answers.

```
<?xml version="1.0" encoding="GB2312"?>
<domain name="*"> <qapair speaker="*">
<zhidao_question_title>*</zhidao_question_t
itle>
<zhidao_question_content>*</zhidao_question
_content><zhidao_other_answer
intersection_number="4">
<entity_and_problemword>*</entity_and_prob
emword> <peopleword>*</peopleword>
</zhidao_other_answer>
<title_extension>*</title_extension>
</qapair>
...
</domain>
```

Figure 4. The Structure of P-XML

<sup>1</sup> <http://sourceforge.net/projects/clucene/>

<sup>2</sup> <http://www.codeforge.cn/article/191554>

<sup>3</sup> <http://Zhidao.baidu.com>

An effective method of capturing the user’s meaning accurately is to create an extension for questions in the knowledge base. In this paper, the extension is primarily a synonym expansion of the keywords of questions, with CILIN (Wanxiang Che, 2010) as extension knowledge source.

The questions are indexed by Clucene to improve the retrieval efficiency of the search for a matched entry in the knowledge base. During the knowledge base searching step, both the index of the original form and the extension form of the problem are used to find the most possible matched record for the user’s question, as shown in algorithm 1. Algorithm 1 is used to examine the similarity between user input and the record returned by Clucene, including traditional and extension similarities.

### Algorithm 1. Problem-matching method

*Begin*

- 1) User inputs question Q;
- 2) Search from the index of original questions and obtain the returned record set RS1;
- 3) For the highest ranked record R1 in RS1,
  - a) compute the similarity sim1 between question R1 and Q;
  - b) compute the extension similarity sim2 between the question extensions of R1 and Q;
- 4) If sim1 is greater than the threshold value T1 or sim2 is greater than the threshold value T2, go to the solution management stage and obtain the answers of R1, and then find the candidate answer using algorithm 2;
- 5) Otherwise, a “not found” prompt is given.

*End*

## 6 Response Management Method

One question usually has many corresponding answers in the knowledge base, and these answers differ in explanation quality. Thus, the basic strategy employed by solution management is to select a reliable answer from the matched record as response, as shown in algorithm 2.

Personalised information includes name entity, gender, marital status and age information. PAL maintains some heuristics rules to help recognize such information. Based on these rules, if one answer contains personal information, it will be selected as the candidate answer only when the personal information is consistent with that of the current user. Very concise answers that do not

contain personal information can generally be selected as a candidate answer.

<p><b>Algorithm 2. Answer-selection method</b></p> <p><b>Begin</b></p> <ol style="list-style-type: none"> <li>1) User inputs one question Q;</li> <li>2) The system extracts the speaker role S and personal information from Q;</li> <li>3) Use Q as query to conduct information retrieval from the index and knowledge base and obtain the top matched record set R;</li> <li>4) For each matched question Q' in R, test the following conditions: <ol style="list-style-type: none"> <li>a) (condition 1) extract the speaker role S' in Q', and examine if S' is equal to S;</li> <li>b) (condition 2) extract personal information in Q', and examine if they are equal to that of in Q;</li> <li>c) For each answer A' of Q' <ol style="list-style-type: none"> <li>i. If no personal information is found in A', A' will be pushed into response list;</li> <li>ii. If personal information is contained in A' and if both conditions 1 and 2 are true, A' will be pushed into response list;</li> </ol> </li> <li>d) End for</li> </ol> </li> <li>5) End for</li> </ol> <p><b>End</b></p>
--

## 7 Experiments

For the current implementation of PAL, the size of the knowledge base is approximately 1.2G and contains six different topics: “Husband and wife”, “Family relations”, “Love affairs”, “Adolescence”, “Feeling and Mood”, and “Mental tutors”. Dialogue data collection used in PAL is mainly crawled from <http://zhidao.baidu.com>, which is one of the largest Chinese online communities. The criterion for choosing these six categories is also because they are the main topics in BAIDU communities about psychological problems. Some information on the knowledge base is given in Table 1, in which “Percent of questions matched” denotes the number of similar questions found when 100 open questions are input (we suppose that if the similarity threshold is bigger than 0.5, then a similar question will be deemed as “hit” in the knowledge base).

In 7.1, we examine the feasibility of using downloaded dialogue collection for constructing the knowledge base. Some dialogue examples are given in 7.2.

Domain	Avg. ques. length	Num. of unique Terms in ques.	Avg. ans. length	Num. of unique terms in ans.	Percent of questions matched (similarity threshold: 0.5)	Size(MB)
QS1	58.69	11571	64.13	27312	25	125
QS2	54.96	10918	64.92	25185	24	292
QS3	59.66	13530	49.52	13664	15	53
QS4	42.41	8607	47.11	23492	22	224
QS5	63.57	11915	48.86	26860	26	276
QS6	31.82	10009	98.55	20896	25	216

Table 1. Information of the knowledge base

### 7.1 System Performance Evaluation

Additional questions and their corresponding answers beyond the knowledge base are also used as a test set to evaluate system performance. Concretely, suppose question Q has  $|A|$  answers in the test set. Q is then input into the system. Suppose the system output is O, we examine if one best answer exists among  $|A|$  answers that are very similar to O (the similarity is greater than threshold T3). If yes, we then assume that one suitable answer has been found. In this way,

precision can be calculated as the number of questions that have very similar answers in the system divided by the number of all input questions.

The performance evaluation results are shown in Figure 5. The horizon axis denotes the similarity threshold (T1 for sim1 and T2 for sim2) between a user’s input and the questions in the knowledge base. Sim1 is the original similarity, whereas sim2 is the semantic extension similarity. Different thresholds were used (0.5 to 0.9). The similarity threshold T3 denotes the similarity

between the answer in the test set and system output O. From Figures 5 (A) and (B), different T3 values were used (0.5 to 0.8).

Some observations can be made from Figure 5. The average system precision is approximately 0.5, and the range is from 0.2 to 0.9. Basically, when T3 is bigger, the system's performance tends to decrease because a high T3 value denotes a strict evaluation standard. Performance also

differs between different areas, such that D4, D5 and D6 outperform than D1, D2 and D3.

When only index is used and both sim1 and sim2 are below the corresponding threshold T1 or T2, the system can still return record set RS2, but the returned answer may be inconsistent with user's question. Thus, incorporating semantic search shown in algorithm 1 is necessary.

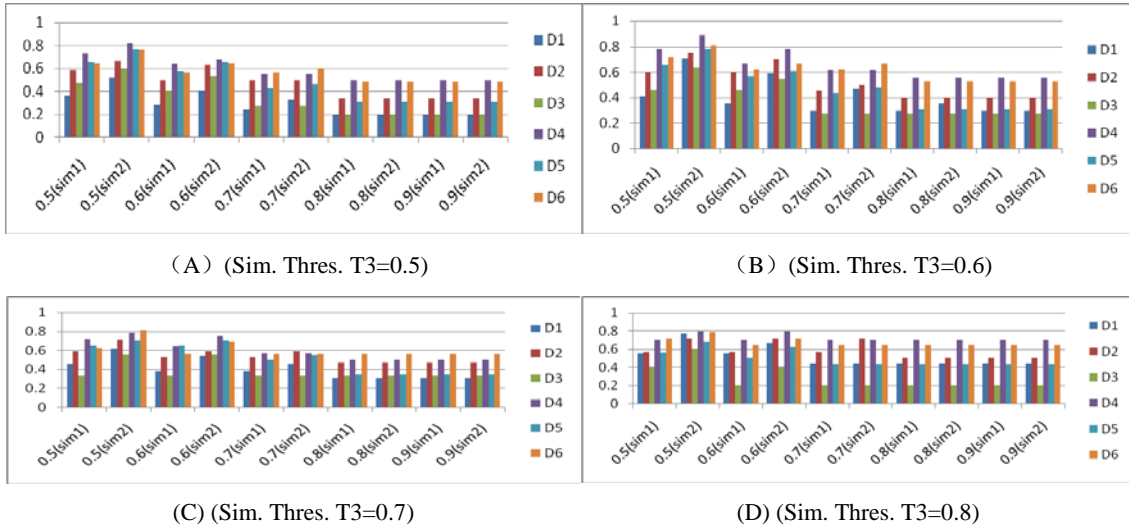


Figure 5. Precision of PAL with different similarity thresholds T3 (The X axis denotes different thresholds for sim1 (T1) and sim2 (T2). For example, 0.5(sim1) indicates that the threshold for sim1 is 0.5, whereas 0.5(sim2) suggests that the threshold for sim2 is 0.5. The Y axis stands for the precision value of different T1, T2 and T3 values for different areas).

## 7.2. Some Interaction Examples

Table 2 provides an interaction example of PAL in English. As users input their questions, the system can output the corresponding responses. The entire dialogue history can also be displayed. Basically, when users describe their troubles, PAL can provide some explanation or even try to console the user.

Compared to other chatterbots like Eliza, PAL is domain-specific and may provide help to some extent. Many different answers can usually be identified. Thus, the user could input the same question many times, with the system providing suggestions from different angles. PAL can also output voice by using Xiao-kun and Microsoft Anna as the text-to-speech (TTS) engine. The speed, volume and echo properties of the voice can be adjusted.

Dialog for Figure 1 (in English)	
PAL>	Hello, you are welcome to communicate with me! You can input from the left side and press enter to get answer!
I>	How do I deal with the children who like to complain very frequently?
PAL>	Very simple. Don't let him experience the benefits of complaining very frequently.
I>	Why do parents give children so much pressure
PAL>	Children have pressure indeed; but parents also have a lot of pressure.
I>	Why am I unable to control my emotions, and often quarrel with my parents?
PAL>	Now you may be in the stage of treason. And have your own uncertain plan for the future, want to be yourself, and do not want to be disturbed. So when parents criticize, you will be unhappy.
I>	How to solve the gap problem?
PAL>	Communication, and listening .

Table 2. Example of an interaction with PAL (in English)

## 8 Conclusions

In this paper, we introduce a simple chatterbot for answering non-obstructive psychological questions. The knowledge behind the chatterbot is learned from the Q&A pairs derived from an online forum using several extraction strategies. The historical and personal information from the dialogues are also incorporated to output an appropriate answer.

For future work, we expect to add more features to PAL, e.g., enabling the system to ask questions actively and further improving P-XML to form richer patterns for storing Q&A knowledge. Another interesting aspect would be to add speech input as well as TTS and to transform PAL into a mobile platform for widespread use.

## Acknowledgments

This research was supported by the project of The National High Technology Research and Development Program (863 program) of PR China under a research Grant No.2007AA01Z172 , Youth Funds of China social & humanity science (10YJCZH099), and Key Laboratory Opening Funding of China MOE—MS Key Laboratory of Natural Language Processing and Speech (HIT.KLOF.2009022).

## References

- Ai Ti Aw and Lian Hau Lee. Personalized Normalization for a Multilingual Chat System. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, 8-14 July 2012, pages 31–36,
- DeeAnna Merz Nagel, Kate Anthony. Text-based Online Counseling Chat. *Online Counseling (Second Edition)*, 2011, Pages 169-182
- Hai-hu Shi, Yan Feng, LI Dong-mei, HU Ying-fei. Research on on-line psychology consultation expert system based on man-machine interaction technique. *Computer Engineering and Design*. 2005, 26(12):3307-3309
- Jing dong. Research of sentiment model based on HMM and its application in psychological consulting expert system. *Master's thesis. Capital normal university (china)*, 2009.
- Leila Kosseim, Jamileh Yousefi. Improving the performance of question answering with

semantically equivalent answer patterns. *Data & Knowledge Engineering*, 2008, 66(1):53-67

- Rafael E. Banchs, Haizhou Li. IRIS: a Chat-oriented Dialogue System based on the Vector Space Model. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, 8-14 July 2012. pages 37–42
- Shilin Ding, Gao Cong, Chin-Yew Lin, Xiaoyan Zhu. Using Conditional Random Fields to Extract Contexts and Answers of Questions from Online Forums. *Proceedings of 2008 Association for Computational Linguistics*, Columbus, Ohio, USA, June 2008. pages 710–718
- Tiphaine Dalmas, Bonnie Webber. Answer comparison in automated question answering. *Journal of Applied Logic*, Volume 5, Issue 1, March 2007, Pages 104-120
- Walid S. Saba. Language, logic and ontology: Uncovering the structure of commonsense knowledge. *International Journal of Human-Computer Studies*, Volume 65, Issue 7, July 2007, Pages 610-623
- Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. *In Proceedings of the Coling 2010: Demonstrations*. August 2010, pp13-16, Beijing, China.
- Yvette Colón, Stephanie Stern. Counseling Groups Online: Theory and Framework. *Online Counseling (Second Edition)*, 2011, Pages 183-202.