A Chinese Language Expert System Using Bayesian Learning

Yuan-Yu Wu
Computer Science Department, Western Washington University
Bellingham, WA 98225 U.S.A.

and

Jianna J. Zhang
Computer Science Department, Western Washington University
Bellingham, WA 98225 U.S.A.

Abstract

We present the prototype of a Chinese language expert system that applies Bayesian theorem to generate conversation exercises for a given topic. The knowledge stored in the system was learned from the contents of a Chinese language textbook. Students can initiate conversations with this expert system via a Web-browser-oriented user interface. Based on the initiation, the system automatically generates appropriate sentence(s) to continue the conversation. The experimental result shows that this Chinese language expert system will provide a very useful learning tool for students who wish to study Mandarin Chinese.

Keyword: Bayesian theorem, expert system, Chinese language, Chinese segmentation, probability learning.

Introduction

It has long been evident that the ideal way to learn a language is to be in the target language environment. For more than a decade, we have observed a huge gap between what students have learned and their ability to deploy their knowledge in real life situations [13, 14]. Without an environment to apply their knowledge, students who read and write the language well will still have problems carrying on conversations.

In this paper we present a Chinese language expert system using Bayesian learning. It is intended to provide an interactive learning environment for students who wish to reinforce their classroom knowledge through conversation. The rest of this paper is organized as follows. In the “Background” section, we give a brief introduction to some of the related previous research on computer assisted Chinese language learning and Chinese language segmentation, and give a quick review on Bayesian learning to lay out the foundation of this paper. In the “System Overview” section, we show the general structure of our Chinese language expert system, and then give a detailed description of the Chinese language segmenter, the knowledgebase, the expert systems, and the conversation interface. The step by step instructions on how to build the expert systems is shown in the “Algorithm” section. The preliminary tests and test samples are presented and discussed in the “Experimental Results” section. The conclusion and future plan is presented in the last section of this paper.

Background

With the rapid development of computer technology, many efforts have been made to develop computer programs that assist natural language teaching and learning, particularly, for Chinese language teaching and learning. Many educators in universities or educational institutes have used some computer assisted language learning (CALL) [18] tools to help their students. In this subsection we exam some of the related previous research.

Computer-Assisted Language Learning

Recently, CALL research has been concentrated on online tutorials that provide an interactive learning environment for students since more and more language teachers have realized that CALL is efficient. Staffan et. al. [17] has reported that back in CALL history, attempts were being made to use interactive feedbacks since the 1960’s. The earliest CALL models were mainly of the “drill-and-practice” variety or “equivalent of electronic flashcards” [17]. In the 1970’s, Hartley introduced the basic outline of the intelligent tutoring system and attempts to adjust the difficulty of the problems base on performance [17, 20]. By the 1980’s, many CALL projects were concentrating on the best way to represent a student’s
knowledge in order to optimally adapt to it [17]. Staffan et. al. further proposed a prototype of the “Object-Oriented Intelligent Tutoring” system which is a natural language expert system model designed to assist students with their sentence construction and translation to foreign languages.

In 1995, Wallace started ALICE project [19]. ALICE is a multi-natural language chat-bot which is based on an expert system using an artificial intelligence technique. Input sentences are segmented into single words, and then a collection of rules are used to match each word against some pre-formulated responses. The matched words are displayed to the users. Although there is still a long way to go before CALL systems reach to satisfactory standards, they have shown a lot of promising results [17, 18, 19, 20].

### Chinese Language Segmentation

Unlike alphabetical languages, pictographic languages such as Chinese do not have explicit white space between characters ( civilized) or words ( simplified). To process Chinese languages, the first problem raised is word recognition or segmentation [2, 8, 12, 16]. There are three approaches available [22]: statistical method [27], dictionary-based method [4, 5], and a hybrid approach using a combination of these two [15]. Chih-Hao Tsai has developed MMSEG [4], a word identification system for Chinese using the dictionary-based approach. Tsai’s segmenter applied two variations of the dictionary-based approach, and successfully achieved 99.69% accuracy.

Recently, Fu [9] presented a two-stage statistical word segmentation system for Chinese language, which employed word bigram models to segment known words in the database in the first stage, and a hybrid algorithm concerning word contextual information and word-formation patterns to identify unknown word in the second stage. A dictionary-based Chinese segmenter by Peterson [6] is implemented in Perl language. Because this segmenter is a free ware and compatible with our system, it is used in our system with some modification.

#### Some Bayesian Theorem Applications

Bayesian theorem has provided a powerful probabilistic approach for many applications. It is based on the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data [24]. Researchers have shown that Bayesian theorem is not only powerful but also compatible with other machine learning algorithms [10, 11, 24, 28]. Bayesian learning algorithms, such as naive Bayes text classifiers reported in [1, 23, 24] have achieved accuracies up to 89% for predicting unseen instances [23, 24]. Paek [25, 26] has constructed a Bayesian network model by observing interactions between a user and a spoken dialog system. The system employs a set of Bayesian learning models to interpret the goals of speakers given evidence gleaned from a natural language parse of their utterances [25, 26]. Bayesian learning is the key method in our system which is used to calculate the probability between the initiation and response of conversation exercises.

### System Overview

We have collected 35 basic Chinese grammar patterns for the initiations and their corresponding grammar patterns for responses. These patterns are used as the base of the training data. Currently, the conversation environment of this Chinese language expert system is limited to a simple birthday party environment. Thus, the vocabulary is also limited to those that could be used in this birthday party situation. This list of words and their parts of speech were collected from the following resources: (1) a Chinese text book, Integrated Chinese [21]; (2) Peterson’s Chinese segmenter [6]; (3) a book called Practical Chinese Reader [3]; and (4) a Chinese textbook, Beginner’s Chinese [29]. The training data is a list of possible conversational sentences that can be used in a birthday party, which were collected from Integrated Chinese [21]. Practical Chinese Reader 1 [3], Beginner’s Chinese [29], and Outline of Chinese Grammar [7].

The current system is composed of five parts: (1) Chinese Segmentator; (2) Expert System I (ESI); (3) Expert System II (ESII); (4) Conversation interface; and (5) Knowledgebase (see Figure 1). First, the system must separate training data (Chinese text) into words though the Chinese Segmenter. Then ESI uses the segmented training data to compute prior probabilities for each input and saves the results in the Knowledgebase. When a user initiates a conversation through the user interface, Chinese Segmenter sends the segmented input to ESII. Based on Bayesian theorem, ESII calculates the probabilities and chooses an appropriate grammar pattern corresponding to the input grammar pattern according to the probabilities. Finally, a sentence most likely to carry on the conversation correctly is displayed on the Conversation Interface.

![Figure 1. System Overview](image)

**Chinese Segmenter**

The Chinese segmenter is based on Peterson’s Chinese segmenter [6] with modifications. Peterson’s segmenter is a
A dictionary-based system that searches the existence of the words in the database based on two-word segments. If the current two-word segment doesn’t exist in the dictionary, it combines the current segment with the next segment. If the combined segments still do not exist, the current segment is treated as a valid word.

Peterson’s algorithm is simple and powerful, but the word database consists of some uncommon and misused segments. Because the goal of our expert system is to help students reinforce classroom learning, the uncommonly used vocabulary were eliminated from the system database. We also adopted another variant of the dictionary-based approach called maximum matching technique \[4\] which matches the longest possible segments.

Our segmenter looks at one sentence at a time. The maximum segment length is set to four characters. For each character, a segment from the current character to its fourth character is taken and used to search for its existence in the database. If the segment exists, it is saved for later use. If it doesn’t exist, the segment from the current character to its third character is taken and checked for its existence and so on until a single character is left.

The word database is a list of pairs of segments and its parts of speech. The parts of speech for the system are divided into seven categories: noun, verb, adjective, adverb, conjunction, particle, article and mark. Each segment from the segmented result is also paired with its part of speech and saved in the knowledgebase.

Figure 2 shows an example of the segmentation process. A sample sentence “这是你的生日礼物 [This is your birthday present]” is sent to the segmenter. After segmentation, it is formed into a list of Chinese segments. Each segment is paired with its part of speech (POS) by ESI.

The expert systems (ESI and ESII) make decisions based on probability calculated using Bayesian theorem. Let \( H \) be the hypothesis space, \( h \) be one of the hypothesis, and \( D \) be the observed training data. We can write \( P(h) \) as the prior probability, which is the initial probability that hypothesis \( h \) holds before we observe the training data \[24\]. \( P(D|h) \) denotes the probability of observing data \( D \) given some world in which hypothesis \( h \) holds. \( P(D) \) denotes the probability that training data \( D \) will be observed. We are interested in the probability that \( h \) holds given the observed training data \( D \), which is denoted as \( P(h|D) \).

Let \( P(h|D) \) be the probability of observing data \( D \) given \( h \) holds. We can write the posterior probability \( P(h|D) \) from its prior probability \( P(h) \) together with \( P(D) \) and \( P(D|h) \):

\[
P(h \mid D) = \frac{P(D \mid h) P(h)}{P(D)}
\]

We define initiations to be sentences initiated by users and responses to be sentences generated by ESII to match given initiations. There are two types of initiations: declarations and questions. Questions are sentences which ask for information. For instance, “要带什么吗 [want bring what]？” All initiations other than questions are categorized as declarations. They can be a greeting, “你好 [Hi]”, or a plain sentence, “我要带果汁 [I want bring juice]”.

Initiations and responses follow the 35 basic Chinese grammar patterns that we have collected. The grammar parts that do not effect the initiation-response generation, such as adverb, particle, mark and article, are ignored by the expert system. For each initiation, there can be several responses. The prior probability of training data is pre-calculated by ESI and stored in a look up table.

![Figure 2. Segmentation Process](image)

**Figure 2. Segmentation Process**

The Expert Systems

The expert systems (ESI and ESII) make decisions based on probability calculated using Bayesian theorem. Let \( H \) be the hypothesis space, \( h \) be one of the hypothesis, and \( D \) be the observed training data. We can write \( P(h) \) as the prior probability, which is the initial probability that hypothesis \( h \) holds before we observe the training data \[24\]. \( P(D|h) \) denotes the probability of observing data \( D \) given some world in which hypothesis \( h \) holds. \( P(D) \) denotes the probability that training data \( D \) will be observed. We are interested in the probability that \( h \) holds given the observed training data \( D \), which is denoted as \( P(h|D) \). With Bayesian theorem, we can calculate the posterior probability \( P(h|D) \) from its prior probability \( P(h) \) together with \( P(D) \) and \( P(D|h) \):

\[
P(h \mid D) = \frac{P(D \mid h) P(h)}{P(D)}
\]

We define initiations to be sentences initiated by users and responses to be sentences generated by ESII to match given initiations. There are two types of initiations: declarations and questions. Questions are sentences which ask for information. For instance, “要带什么吗 [want bring what]?” All initiations other than questions are categorized as declarations. They can be a greeting, “你好 [Hi]”, or a plain sentence, “我要带果汁 [I want bring juice]”.

Initiations and responses follow the 35 basic Chinese grammar patterns that we have collected. The grammar parts that do not effect the initiation-response generation, such as adverb, particle, mark and article, are ignored by the expert system. For each initiation, there can be several responses. The prior probability of training data is pre-calculated by ESI and stored in a look up table.

![Figure 3. Counting associate probability](image)

**Figure 3. Counting associate probability**

Figure 3 shows a simplified example on the priori probability calculation for “有东西吃吗 [have things eat]”. This initiation is categorized as a question, and follows the “verb-noun-verb” pattern. After the calculation, three responses are generated: (1) 有蛋糕可以吃 [have cake can eat]; (2) 好像有蛋糕吃 [likely has cake eat]; and (3) 应该有点心 [should have desert]. They are categorized into two patterns, verb-noun-verb and verb-noun. ESI counts the frequency of verb and noun segments in their location for each pattern. In the pattern verb-noun-verb, “有 [have]” appears twice as the first verb, “蛋糕 [cake]” appears once as the first noun, “点心 [should have desert]” appears once as the first verb and “点心 [dessert]” appears twice as the second verb. In the pattern verb-noun-verb, “有 [have]” appears once as the first verb and “点心 [dessert]” appears twice as the second verb. Prior probability for verb-noun-verb pattern is recorded and saved in a look up table as in the following:

\[
P(有 [have], 点心) = 100%,
\]
Prior probability for verb-noun pattern is calculated:

\[ P(有[have]verb) = 100\% , \]
\[ P(点心[desert]noun) = 100\% . \]

The conditional probability for each segment related to its initiation is also calculated and recorded. For instance, \( P(有[have] | 有[have]verb) = 100\% , P(点心[thing] | 有[have]verb) = 100\% , P(吃[eat] | 有[have]verb) = 100\% . \) All of the calculated probabilities are saved in the knowledgebase for later use.

**Conversation Generation**

When a user initiates a conversation, the initiation sentence will be changed into a sequence of segments and their part of speech pairs. ESII first determines whether the initiation is a question or a declaration. For each initiation, ESII generates several matched response patterns based on posterior probability calculations, and then randomly selects one to be the desired response.

In figure 4, the initiation “有东西吃吗 [have thing eat]” is the input to the system. The input is segmented and identified as a verb-noun-verb question by ESII. It has two possible response grammar patterns: verb-noun-verb and verb-noun. Assume the verb-noun-verb pattern is chosen at random. For each possible segment in each location, ESII calculates its posterior probability and selects the segment with the highest probability.

Let \( t_i \) be the segment at each position in the initiation, and \( v_j \) be a collection of words appearing as the first verb. The posterior probability for this example is calculated using the equation:

\[ Segment_{verb} = \text{arg max } P(v_j) \prod_{i \in \text{positions}} P(t_i | v_j) \]

For \( \text{verb1} , “有 [have]” \) has the posterior probability of \( P(有[have]verb)*P(有[have] | 有[have]verb)* P(东西[thing] | 有[have]verb)* P(吃[eat] | 有[have]verb)) = 100\% , \) which is currently the highest probability. Therefore, \( “有 [have]” \) is chosen to be in the first verb segment. Using the same equation, a response is formed as: 有点心吃 [have desert to eat].

**The Conversation Interface**

The conversation interface is a set of dynamic web pages containing some virtual environments. Users can select a conversation partner from three people (see Figure 5). To enter the selected conversation environment, users must enter their names. During the conversation, users input their sentences in a textbox at the bottom of the page and the conversation is displayed in the center of the page (see Figure 6).

Figure 5 is a screen shot of the conversation page for selecting conversation partners. It shows a boy, a girl, and a Chinese teacher. After entering their names in the textbox, users can click the “enter conversation” link to start the conversation.
Algorithm

The algorithm has three main functions: (1) Segmentation; (2) Probability_Assignment; and (3) Conversation_Generation.

Segmentation ( )
This function segments Chinese sentences into segments.
- Load segment dictionary
- For each character in the sentence
  -- Find the longest segment in the database for the Current character
- For each segment
  Assign the part of speech

Probability_Assignment ( )
This function identifies the sentence pattern, calculates the prior probability for each segment in the response and records in the knowledgebase.
- For each pair of initiation/response sentences in the training database
  1. Categories the initiation/response pattern
  2. If the yielded pattern exists in the database
     For each response
     - Calculate the prior probability for each segment and record it
     - Calculate the conditional probability for each segment and record it
  3. If the pattern does not exist in the database, give a default response.

Conversation_Generation ( )
This function generates a response for each initiation by selecting the segments with the highest probability and displays the output.
- For each initiation
  1. Segment the initiation
  2. Categories the initiation pattern
  3. Select one possible response pattern For each segment in the response pattern
  4. Calculate its posterior probability and select one possible segment.
  5. Form the response with select segments and output to users.

Experimental Results

The preliminary tests are conducted manually though the Conversation Interface. Let $m$ be the number of trails, $n_i$ be the total number of sentences in the $i^{th}$ trial, and $c_i$ be the number of sentences that is correctly generated by the system in the $i^{th}$ trial. The average accuracy is calculated as:

$$\text{Average Accuracy} = \left( \frac{\sum c_i}{\sum n_i} \right) \cdot m$$

Within the birthday party environment, the average accuracy is around 75%. The following are samples of the test results:

Sample 1:
user: 要带什么东西? [What to bring?]
system: 带一些喝的吧! [Bring some drink.]
user: 有什么人会去? [Who will come?]
system: 小林的班上同学. [Xiaolin’s classmates.]
user: 有谁会去? [Who else will come?]
system: 小林的爸爸妈妈还有一些朋友. [Xiaolin’s parents and some friends.]
user: 我带汽水吧! [I will bring some soft drink.]
system: 我讨厌巧克力. [I hate Chocolate.]
user: 我喜欢巧克力. [I like Chocolate.]
system: 我不喜欢,你带鲜奶油. [I don’t, you bring fresh butter.]

Sample 2:
user: 这是我中文班老师. [This is my Chinese teacher.]
system: 好. [Good.]
user: 这是我朋友李安. [This is my friend Lian.]
system: 老师你好! [How do you do teacher!]
Lian: 我不认识你. [I don’t know you.]
system: 我要带汽水,那你带果汁吧! [I want to bring soft drink, you bring juice.]

Sample 3:
user: 小林,生日快乐! [Happy birthday, Xiaolin!]
system: 谢谢你. [Thank you!]
user: 切蛋糕了! [Cut cake!]
system: 好. [Good.]

The conversations went smoothly most of the time. However in Sample 1, when the user says “我带汽水吧! [I will bring some soft drink.]”, the system answered “我讨厌巧克力. [I hate Chocolate.]”. A problem arises when the system faces multiple choices. That is, when there is more than one grammatically correct answer, the system cannot tell which answer is more semantically correct. Thus it simplify chooses the first one encounters.

Conclusion and Future Work

We have constructed the prototype of a Chinese language expert system based on Bayesian theorem to help students learning Chinese. The experimental tests have shown very interesting and promising results although the system is still in its early development stage. The conversations carried on between the user and the system went quite smoothly and reasonably despite some off topic sentences generated by the system. It is a significant first step towards an intelligent Chinese language learning model that can automatically chat with students.

In the future, we want to enlarge the expert system’s vocabulary and add more conversation topics. As the vocabulary gets larger, selecting more meaningful sentences will also be possible although it is difficult to implement. A semantic calculation model is needed to deal with multiple
answers that are all grammatically correct but with some answers more semantically correct than the others.

We also plan to audiolize the conversation so that students can practice spoken language with the system. The future expert system is also expected to be able to detect incorrect grammar so that students will know if they make a mistake. With the help of this interactive Chinese language learning model, students will not only be able to learn the language faster and better but also enjoy learning.

References