ARTIFICIAL INTELLIGENCE AND NATURAL LANGUAGE PROCESSING IN A CHINESE LANGUAGE LEARNING GAME

by

MATTHEW SAMUELS

Advisor
DR. DANIEL PLANTE

A senior research [proposal/paper] submitted in partial fulfillment of the requirements
for the degree of Bachelor of Science
in the Department of Mathematics and Computer Science
in the College of Arts and Science
at Stetson University
DeLand, Florida

Fall Term
2010

TABLE OF CONTENTS

ABSTRACT

Mandarin has been classified among the most difficult languages for foreigners to learn as a second language. Not only is the vocabulary considerably different from that of English, but also the grammar and Chinese writing system are considerably complex. For this project, known rules of grammar will be used in order to generate sentences in Mandarin that can be used by a foreigner to lessen the difficulty of learning this language. These sentences will then be used in a game that will aid the learning of Mandarin Chinese. Different types of questions will be created and artificial intelligence will be incorporated into dividing questions into levels of difficulty based on grammar, character, or possibly overall sentence complexity. Expert systems including real world data will be used in order to train the algorithm to generate correct sentences.

I. Introduction

Mandarin is one of the most complex languages of the world. It does not use standard letters to express ideas like the letters of the English alphabet. Instead, words are made up of one, two or more syllables and each syllable is represented by a character, which can be used by itself or in combination with other characters to represent words. Chinese characters are called hanzi (汉字 hànzi). Each hanzi is drawn using a set of strokes and there are twelve different kinds of strokes that may be used in a given hanzi.

Before the more official writing system that exists today, early writing involved drawing pictures that were representative of their meaning. However, these pictures have become more stylized and much less resembling of what they actually represent. Furthermore, hanzi are written together without spaces but do include punctuation, so it is important to understand the meaning of hanzi and to know when they should be grouped together to determine the correct meaning of a given sentence.[1]

Because the Chinese writing system consists of hanzi, a standard alphabet cannot be used to represent these pictographs. However, in order to teach Mandarin there is a Romanized system of representing the characters, called pinyin (拼音 pīnyīn), which shows how a character is spoken by writing the character as a word using a subset of the letters that exist in the English alphabet. Since Mandarin is a tonal language, with four different tones as well as a fifth neutral tone, the pinyin includes tone markers in each word to illustrate exactly how each word is to be pronounced. When typing pinyin it can be difficult to type a letter with its tone marker. Therefore, pinyin may also be written with numbers in place of tone markers. For example "pīnyīn" would be written "pin1yin1" because the tone marker over each 'i' in these two characters represents the first tone. Similarly, a 2, 3, 4 or 5 can be used to represent the other tones [4]. In this respect, an alphabet is associated with the pinyin system of Mandarin in order to teach how to pronounce every letter from the English alphabet that is used in this system. Once this alphabet is learned, a student of Mandarin can learn to speak and write Chinese characters based on the tone, pronunciation, and meaning of the character. Furthermore, once a student is able to write characters, he is then able to better focus on learning how to form sentences and use the language correctly, both in the written and spoken language, though, to be able to read most Chinese magazines and newspapers, knowledge of about three thousand characters is necessary [1].

The complexity of a Chinese language such as Mandarin lends itself to many methods of learning, such as focusing on specific areas such as grammar, pinyin, writing characters, or perhaps just learning vocabulary, or combining some of these areas of studying into the learning curriculum. Due to the various areas of focus one can choose when studying Mandarin, there are many ways to go about learning, including attending Mandarin classes, listening to Chinese audio, writing characters, and doing drills or playing games specifically made for learning the language. There are many resources online for learning grammar and vocabulary. There are games specifically for testing knowledge of the four tones of the language when given a word in character or pinyin form, flash card games involving characters or pinyin and their meanings, and quizzes for selecting the correct choice to complete a sentence. One website, www.xuezhongwen.net, provides services such as Chinese to English and English to Chinese translation, character search by radical (a smaller part of the character that primarily determines the character and its meaning), character search by drawing the character via user input, as well as Mandarin practice in preparation for the Hànyǔ Shuǐ píng Kǎoshì (HSK), which is a Mandarin proficiency exam.

The purpose of this project is to create a Chinese learning game for studying grammar, syntax, and vocabulary through characters and pinyin. Different types of questions will be created that will be divided into levels of difficulty, such as beginner, novice, and difficult. Difficulty may be determined by complexity of grammar, character or perhaps an entire sentence, based on its grammar or characters or a combination of these two factors. Artificial intelligence will be incorporated into creating good questions from an algorithm trained using correct sentences. Furthermore, expert systems will be used in order to test the quality of the questions produced. Due to the complexity of Mandarin, certain types of questions may not be included due to the inability to create such questions correctly. For example, we have been unable to find a Chinese thesaurus available for download due to the difficulty in creating a list of truly similar Chinese words. While two words may have similar meanings they may be used in entirely different contexts, in which case they are not truly synonyms. Therefore it would not be possible to create good questions dealing with analogies with respect to synonyms.

II. Related Works

1. Term Frequency-Inverse Document Frequency

Various methods exist for information retrieval and data mining. One such method is Term Frequency – Inverse Document Frequency (TF-IDF), which assigns weights to terms based on their importance and relevance to a given document. Terms relevant to a document will appear frequently in that document compared to how often it appears in other documents in a given set of documents. The term frequency (tf) is calculated as a ratio of the number of times a term appears in a specific document to the number of documents in the set. The inverse document frequency (idf) is then calculated as the logarithm of the ratio of the number of documents to the number of documents in which the term being analyzed appears [3]. The tf-idf is then calculated by multiplying the tf by the idf, which can be done for each term to be analysed.

2. Extensions to TF-IDF

Using the TF-IDF model of weighting terms in documents as a starting point, several extensions to this model were proposed in [4] in order to adapt this model to domain specific information retrieval. In this respect an extra value called the “specificity” of a term is added to the tf-idf weight calculated for that term. The specificity is calculated differently depending on the extension to tf-idf that is used and some of the possible extensions proposed are mutual information, information gain, relative frequency ratio and index of peculiarity. Mutual information is used in order to estimate the relevance of a term across the entire collection of documents. Information gain estimates how relevant a term is to a specific document. Relative frequency ratio uses a general corpus and a specific corpus to calculate the relative frequency of a term in the specific corpus and divide this by the relative frequency in the general corpus. This ratio is used to determine the specificity for this extension, based on the range in which this ratio falls, either less than or equal to one, between one and infinity, or equal to infinity [4].

III. Tools for Parsing

1. Stanford Parser

The Stanford NLP (Natural Language Parsing) Group developed a natural language parser [5] and Chinese word segmenter for working with natural language processing and the Chinese language. The parser can be used with English, Chinese and Arabic text and it reads each sentence and determines the grammatical structure of sentences [6]. It thus breaks sentences into subject and object, as well as into phrases, such as noun phrase or verb phrase, and further parses each word in a given phrase into its Part of Speech (POS). The Stanford parser requires sentences to be space segmented in order to parse each word and since English sentences are written so that words are separated by spaces, it will work fine with English. However, since Chinese sentences do not separate words, these sentences cannot be directly input to the parser. Therefore, the Chinese word segmenter is important because it reads in Chinese sentences and segments each sentence by putting a space between characters and thus grouping certain characters that should be used together. The Stanford parser can then parse the resulting spaced Chinese sentence just like an English sentence. Stanford University's NLP webpage has an online version of the parser that visitors can use and downloadable versions of the parser and segmenter, written in Java, are also available freely under the GNU public license.

Sample output from the Stanford parser for English, given the sentence "The Internet, it exists." and using the command:

./lexparser.csh "file.txt"

where the file "file.txt" contains the English sentence to be parsed, displays the following:

```
(ROOT
  (S
    (NP (DT The) (NN internet))
    (, ,)
    (NP (PRP it))
    (VP (VBZ exists))
    (. .)))

det(internet-2, The-1)
nsubj(exists-5, internet-2)
nsubj(exists-5, it-4)

Parsed file: parseEnglish.txt [1 sentences].
Parsed 6 words in 1 sentences (12.58 wds/sec; 2.10
sents/sec).
```

Figure 1: Sample output from the Stanford parser to parse the English sentence "The Internet, it exists." into its POS's.

Given the same sentence in Chinese, “网络它存在”, first passed into the word segmenter, segments the sentence into “网络 它 存在” (‘The Internet’ ‘it’ ‘exists’). Then, using the Stanford parser for Chinese via the command:

```
./lexparser-zh-utf8.csh "file.txt"
```

produces:

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
Parsing [sent. 1 len. 3]: [网络, 它, 存在]
<s id="1">
  <words pos="true">
    <word ind="1" pos="NN">网络</word>
    <word ind="2" pos="PN">它</word>
    <word ind="3" pos="VV">存在</word>
  </words>
  <tree style="penn">
(IP
  (NP (NN 网络))
  (NP (PN 它))
  (VP (VV 存在)))
```

```
  </tree>
<dependencies style="typed">
  <dep type="nsubj">
    <governor idx="3">存在</governor>
    <dependent idx="1">网络</dependent>
  </dep>
  <dep type="nsubj">
    <governor idx="3">存在</governor>
    <dependent idx="2">它</dependent>
  </dep>
</dependencies>
</s>

</corpus>
Parsed file: parseChinese.txt [1 sentences].
Parsed 3 words in 1 sentences (4.44 wds/sec; 1.48
sents/sec).
```

Figure 2: Sample output from the Stanford parser to parse the Chinese sentence "网络它存在" into its POS's.

This output displays the words and parts of speech tags, the parsed sentence tree, and the type dependencies of each word in the sentence, where the output format specified in the command line interface file, "lexparser.csh" or "lexparser-zh-utf8.csh" for English or Chinese, respectively, is "wordsAndTags, penn, typedDependencies". One or some combination of these output types can be chosen depending on the output that is needed. For example, the output format can be changed in the file to specify only "wordsAndTags" and this will produce output of just the words and their parts of speech. From this output we can determine the parts of speech of each word in a given sentence and through some parsing of the output we can store a list of different parts of speech and the words that fall into each part of speech category.

2. Adso Parser

Another example called Popup Chinese provides registered users with various tools for practicing and studying Chinese. This website provides an engine called "adsotrans" that is defined as a "Chinese-English translation, hanzi-to-pinyin conversion and text analysis and segmentation engine." Adso is written in C++ and is an open-source project. It includes features for translation from Chinese to English, conversion from Chinese character to pinyin and word segmentation, similar to the Stanford Chinese word segmenter [7]. A Chinese sentence can be given directly from the command line using the interactive mode flag, '-i', or from a file containing the sentence by using the flag, '-f'. The adso engine also includes a database of Chinese words and information about the words as well as grammar rules that is uses in the services that it provides. These rules are written in an XML format and stored in an external grammar file that can be used by the engine. This allows for extending the existing rules or creating one's own set of rules that could be used to fine tune or just improve the translation ability of the engine.

Sample output from the adso engine to segment the sentence used with the Stanford segmenter ("网络它存在"), using the command

```
./adso –i 网络它存在 --segcn
```

produces the following output.

```
网络 它 存在
```

This is the same segmentation produced by the Stanford parser. Some flags that can be used with adso are "**-cn**" to print the results in Chinese, "**-y**" to display the results in pinyin, "**-t**" to translate from Chinese to English and "**-os [simplified | traditional] -cn**" to set the output script to simplified or traditional Chinese characters and echo the Chinese with the "**-cn**" flag.

Examples:

```
./adso –i 网络它存在 -cn
    网络它存在
./adso –i 网络它存在 -y
    wang3luo4 ta1 cun2zai4
./adso –i 网络它存在 -t
```

```
    network it to exist
./adso –i 网络它存在 -os traditional -cn
    網絡它存在
```

As we see from the second to last example, the translation feature is not perfect. The main problem here is that the verb "to exist" was not conjugated to "exists", but the other words translated correctly. The first word "网络" (wang3luo4) was translated as network, a correct translation, even though our intended translation was "Internet". Though we can use the full term for Internet "国际互联网络"(guo2ji4 hu4lian2 wang3luo4), this shorter version of the word can be used both to mean "network" or "Internet".

IV. Types of Questions

The types of questions that will be used in this learning game include translation from English to Chinese, sentence formation and choosing the correct measure word for an object. Measure words will be defined in the implementation of this project.

Translation type questions will present the player with a sentence in English and some choices of possible translations in Chinese. Ideally, the Chinese sentences will be written using Chinese characters rather than pinyin for the skilled user. Therefore, beginner questions may include only pinyin sentences as possible choices, novice questions may include some questions with choices only in pinyin and some whose choices are only written in characters, and difficult questions may include only questions whose possible translation choices are displayed in characters. This will prevent the translation type questions from being overly difficult and allow a beginner student to practice first with pinyin, then get more accustomed to Chinese characters before being only presented with characters to choose from when harder questions are attempted.

Sentence formation questions will present the user with a group of Chinese words and ask them to form a sentence from the words that are displayed. Choices of possible sentences will be displayed from which the user may choose. This will be used in order to teach Chinese sentence structure. Pinyin or characters will be used depending on the difficulty of the question. In this way, beginner questions will display the broken sentence as well as the choices in pinyin, novice questions will sometimes display only pinyin for the question and choices and other times display only characters for the question and choices, and difficult questions will display both the question and choices in characters.

Questions about measure words will display a number and an object and ask the user to choose a measure word that is associated with the given object. In Chinese, a measure word describes the unit of measurement of an object, such as in the phrase "three cups of water". Measure words do not always represent units of measurement. For example, in the phrase "一位老师"(yí wèi lǎoshī), the word "wèi" is a measure word used with the word "teacher" (lǎoshī) but of course we do not literally measure a teacher in the same sense as we say we measure an amount of water. While in English it is possible to describe an object without specifying the unit, such as in “three persons” or “four pencils”, in Chinese an object must be preceded with a measure word in order to discuss about one or more instances of it. [2] Therefore, it is important to learn the correct measure word to use with an object. Also, many objects have a specific measure word, many may be used with more than one measure word, and many may be used with the more general measure word 个(gè). [2]

V. Levels of Difficulty/Complexity

Levels of difficulty will be determined by the complexity of the question for a given question type. Using the CC-CEDict Chinese-English dictionary, a weight for each character in the dictionary is computed by determining the frequency of the character from all of the character entries in the dictionary and dividing this frequency by the total number of characters in the dictionary. This will result in more frequent characters having higher weights and less frequent characters having lower weights. In order to give less significance to more frequent characters and let less frequent characters have more significance, the inverse of the weights can be used to give frequent characters lower weights and less frequent characters higher weights. Then these weights will be split into ranges to determine the level of difficulty based on complexity, based on the idea that more complex characters are used less often in everyday use of Mandarin, so these characters will have lower frequencies and higher weights. Although it would be more ideal to calculate these weights from a corpus of text that will be used in training, the dictionary being used has almost 100,000 entries and covers many categories of words, rather than being a specialized dictionary that focuses on a single category of words, such as scientific or medical terms. Once a good corpus of text is decided upon for use in this project, another calculation of weights from this corpus can be used to test the validity of those calculated from the dictionary.

The complexity weights of the Chinese characters can be used to determine the complexity of the questions created in the game. For example, the complexity of a sentence may be calculated by summing the complexities of the words used in the sentence. If we know the average Chinese sentence length from the sentences in the corpus, we can determine a range of sentence complexities that should be in each level of difficulty. Also, if a sentence is just a character, such as in an antonym type question, then the complexity of the character itself can be used to determine the difficulty of the question.

Complexity may also be determined by the grammar used in a sentence. Starting with a very simple structure such as Subject + Verb + Object and increasing the difficulty of the grammar by using other sentence structures, longer sentences, and structures used with specific characters, the complexity of the sentence can be created by the type of grammar rules that are used. Another metric could be used that involves multiplying the complexity of a sentence that uses a structure specific to some character by the complexity of that character.

Once a metric is determined for calculating the complexity of a question, this metric will require testing in order to check the validity of its results given a set a questions in order to make sure questions are being classified correctly. Artificial intelligence algorithms could be applied here to create a classifier that uses the complexities to determine the level of difficulty of a given question and we are considering some options for their use.

VI. Implementation

1. Chinese corpus

In order to implement this project and train an algorithm using correct Chinese sentences, we will use a Chinese corpus. The corpus we have found comes from a Chinese website called Sougou and contains many documents in Chinese, grouped in folders that are labeled primarily with numbers. Therefore, without reading through many of the documents it is not possible to know the contents of the documents or the difficulty of the contents of the documents. Although this corpus can be used in early stages of the project for testing purposes and generating the algorithm that will create sentences after training from the corpus, it would be better to use a corpus whose contents are known as well as whose contents cover a good variety of areas rather than focusing on a single area. This will allow for a larger variety of types of sentences that can be created.

In using the Sougou corpus it is at least possible to get a sense of the range of difficulty of the sentences contained in this corpus. In this way we take each Chinese document and create a list of all of the sentences, separated by the Chinese period symbol, "。", and their respective sentence lengths. This list can then be sorted by sentence length and sentences for a given sentence length can be further ranked by complexity. These complexities are calculated using the total weight of the sentence, which is calculated by summing the weights of each character in the sentence, where the character weights used are those previously calculated from the CC-CEDictionary. The number of characters in the sentence then divides the sentence weight in order to normalize the weight value for the sentence, which produces the sentence complexity. Using these results we can look at all the sentences of a given length, such as a minimum length of four, and progressively look at sentences of longer length. Analyzing the characters used in the sentences and complexity of the sentences can then aid in determining the contents of the corpus.

2. Chinese grammar

Along with the grammar rules contained in the adso engine, other grammar rules can be gathered from the textbook, *Integrated Chinese*, which is used for teaching students of Mandarin Chinese. This textbook contains a variety of topics, starting with simple greetings and later discussing visiting friends and transportation, with new vocabulary in each lesson and grammar topics progressively increasing in difficulty [8]. A corpus structured similar to this textbook that divides documents into categories and has a range of difficulty would be ideal for training the algorithm and creating sentences similar to those contained in the corpus for this learning game.

3. Implementation of questions

3.1 Translation questions

Translation type questions will be implemented with help from the adso engine. Because it has the feature to translate from Chinese to English, a given Chinese sentence can be input to the engine and translated to determine the correct answer for that specific question. In a given question, the original Chinese sentence will be used as the question and the English translation output from adso along with other incorrect English translations will be displayed as possible choices. The user will then be asked to choose from among these English translations to determine the meaning of the Chinese sentence. Artificial intelligence methods will be used in creating the incorrect English choices in order to create choices that are incorrect interpretations and translations of the original Chinese sentence, either due to incorrect use of the grammar of the sentence or incorrect translation of some words of the sentence. These types of questions will aid the user in improving their understanding of Chinese sentences and sentence structure since a change in structure can change the meaning of the sentence, even though the same words are used.

3.2 Sentence Formation Questions

Sentence formation questions will make use of the word segmentation feature of either the adso engine or the Stanford Chinese word segmenter in order to break a sentence into segments that can then be mixed up such that the user is required to determine the correct order of the segments. Since segments may not necessarily have only one correct order, there may be more than one possible correct answer among the choices presented to the user. In this respect, the choices can be given a weight such that the choice representing the original sentence that was passed through the segmenter has the highest weight and other choices will be weighted according to how closely they resemble this original Chinese sentence, based on the number of characters that are in the correct position. Perhaps the idea of neural networks or Bayes networks as described in [9, 10] can be used in classifying the incorrect choices and giving a probability based on how much the sentence resembles the original.

3.3 Measure Word Questions

Questions about measure words will display a question in the format: "number (measure word) object" and the user will be asked to select the correct measure word that would be associated with the given object. For example, if a question is: "Yi1 (measure word) lao3shi1" then the user could select "wei4" (measure word used for position and people) from the choices since this is a measure word associated with "lao3shi1". This example could be a possible question for beginner difficulty or even novice difficulty, depending on the complexity determined for this sentence, since pinyin is used. More difficult questions would use hanzi. Artificial intelligence could be used here to train an algorithm to classify objects, based on their meanings, into categories, where a category represents a measure word. The meaning of an object could be given specifically for the training algorithm or could be retrieved from a dictionary, such as the

CC-CEDictionary. Then new objects can be tested with the classifier, which will return the associated measure words learned for a given object. Although it is often possible to use the general measure word with objects, it is much more correct if the appropriate measure word is used, and for anyone that wants to learn how to correctly use the Mandarin language, it is important to learn the correct measure words for objects.

VII. Results

After analyzing the CC-CEDictionary and calculating the frequencies of the characters contained in this dictionary, a weight of complexity was calculated by dividing the frequency of each character by the total number of characters. A sample of the frequencies sorted in descending order is shown in table 1.

Character	**Frequency**
不	1718
大	1448
子	1321
人	1311
...	...
分	511
气	500

平	493
发	492
南	477
…	…
㑄	1
尵	1
尶	1
秇	1
巵	1

Higher frequencies suggest that a word is used often and therefore is not a difficult word while lower frequencies suggest that a word is used less often because it is more difficult. Using these frequencies and dividing by a total of 10,395 words gives the following weights for the characters.

Character	**Weight**
不	0.0066946193653
大	0.00564249641499
子	0.00514760895318
人	0.0051086414365

…	…
分	0.00199124010225
气	0.0019483758339
平	0.00192109857223
发	0.00191720182056
南	0.00185875054555
…	…
㑳	3.89675166781e-06
尵	3.89675166781e-06
尶	3.89675166781e-06
[illegible]	3.89675166781e-06
[illegible]	3.89675166781e-06

VIII. Future Work

This project is intended to combine natural language processing and artificial intelligence into a method for studying the Mandarin language. Methods of artificial intelligence such as Bayes networks and decision trees along with other concepts that may fit into this project will be explored in more detail during the implementation phase of this project. Also, the Stanford parser and adso engine will be studied in more detail in order to make the best use of these resources in parsing text and retrieving useful results as appropriate for each type of question of the learning game. Furthermore, once various types of questions can be generated it will be necessary to determine the quality of these questions. Therefore, experts, such as native Chinese speakers can be used to check the questions that are created and rate how good each question is. Ratings will be based on perceived level of difficulty compared to the level determined by the algorithm for determining the level as well as the validity of the question, such as how well formed it is. One possible way of accomplishing this would be to create an online test that presents questions from the developed system and asks users to look at each question and its possible answers, answer the question to see the results, and then rate the question as described. Then this data can be collected and analyzed to determine how well this system works.

References

1. S. Ager, "Chinese script and language," [Online document], 13 Nov. 2010, Available: http://www.omniglot.com/writing/chinese.htm
2. J. Lau, "Chinese Measure Words," [Online document], Jan. 2003, Available: http://www.yellowbridge.com/chinese/measurewords.php
3. P. Klieber, "Document Classification Through Data Mining Social Media Networks," [Online document], 2009, Available: http://www2.stetson.edu/mathcs/people/students/research/pdf/2008/pkleiber/final.pdf
4. C. Fautsch and J. Savoy, "Adapting the tf idf Vector-Space Model to Domain Specific Information Retrieval" in *SAC '10 Proceedings of the 2010 ACM Symposium on Applied Computing.* (New York, NY, 2010, pp. 1708-1712).
5. D. Klein and C. Manning, "Fast Exact Inference with a Factored Model for Natural Language Parsing" in *Advances in Neural Information Processing Systems 15 (NIPS 2002)*. (Cambridge, MA: MIT Press, 2003, pp. 3-10).
6. D. Klein and C. Manning, "The Stanford NLP (Natural Language Processing) Group," [Online document], 2003, Available: http://nlp.stanford.edu/software/lex-parser.shtml
7. Qiu Gui Su, "Pinyin Romanization – Learn Mandarin Chinese with Pinyin

Romanization," [Online document], 2010, Available: http://mandarin.about.com/od/readingmandarin/a/pinyin.htm

8. T. Yao and Y. Liu, *Integrated Chinese: Simplified Character Edition Textbook Extended 2nd Edition*. Boston, MA: Cheng & Tsui Company, Inc, 2005.
9. G. Luger, *Artificial Intelligence: Structures and Strategies for Complex Problem Solving 6th Edition*. Boston, MA: Pearson Education, 2009.
10. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach Second Edition*. Upper Saddle River, NJ: Pearson Education, 2003.