DATA MINING MODEL IN ANALYZING PORTUGUESE STUDIES AS THE SECOND LANGUAGE ACQUISITION

SAM CHAO¹, FAI WONG¹, CUSTÓDIO CAVACO MARTINS²

¹Faculty of Science and Technology, University of Macau, Taipa, Macau
²Faculty of Social Science and Humanities, University of Macau, Taipa, Macau
E-MAIL: {lidiasc, derekfw, fshcm}@umac.mo

Abstract:
Portuguese is specifically a difficult language with luxuriant tenses, and second language acquisition (SLA) is regarded as highly variable. Many (Chinese) students who learn Portuguese as their second language (L2) are hardly to gain a mastery of it. In this paper, we proposed to employ i+Learning method in building a data mining model to discover the regularities and factors that govern the acquisition process and learning quality of the students. The results reveal that knowledge generated through our model is able to explain the acquisition behavior in practice, which can benefit the subsequent linguistic analysis in SLA and can contribute to studies of Portuguese language acquisition.

Keywords:
Decision trees; Data mining; Second language acquisition (SLA); Natural language processing (NLP)

1. Introduction

Second language is generally used to refer to any language other than the native language. There is no simple answer to the question "What is second language acquisition (SLA)?" The goal of SLA is the description and explanation of the learner's linguistic or communicative competence. To this end, the researcher must examine aspects of the learner's usage or use of the second language (L2) in actual performance, by collecting and analyzing either samples of learner language, reports of learner's introspections, or records of their intuitions regarding what is correct or appropriate L2 behavior [1]. As we enter 21st century, everyday language use is so tied to technology that language through technology has become a fact of life with important implications for all applied linguistics, particularly for those concerned with facets of SLA [2]. Nowadays, computational technology plays a dominant role in linguistics and natural language processing (NLP), in which computational models are sometimes essential tools for improving the linguistic theory, and for gaining better insight into how languages are using for communication.

The development and research of computer applications in second language acquisition (CASLA) has a long history, which can be dated back to 1960s. Since that time, a variety of computer-assisted language learning (CALL) research projects were initiated at many prestigious universities all over the world. On the other hand, as in computer technology aspect, data mining is a burgeoning new technology with a wide range of applications, while its technology is well suited for analysis [3]. Data mining technique emerged since 1990s and stands out above the others, it resolves the bottleneck problem of knowledge acquisition and the black box problem during learning process, hence the extracted knowledge is easier to interpret, analyze and reapply. We found that data mining can be applied in SLA to retrieve implicit language knowledge or identify hidden grammar patterns from great quantity of linguistic data automatically. However, it is hard to find any work on SLA using data mining.

Furthermore, since SLA is regarded as highly variable and Portuguese is specifically a difficult language with luxuriant tenses, such that many Chinese students are hardly to gain a mastery of it. Besides, although we know that both linguistic and non-linguistic factors played a role in the acquisition process, yet it does not give any insight to show how and through what kind of mechanisms these factors interact. Obviously, Portuguese instructors lack proper tools to analyze and evaluate the acquisition processes from their students who are learning Portuguese as L2. All these obstacles motivate us to initiate the related research to model such a process through data mining technique, in order to discover some novel clues in better understanding some of the most important theoretical tenets that have involved in SLA processes. This paper presents a data mining model that aims at analyzing and evaluating the acquisition process of verbal morphology of Portuguese as L2 by the Mandarin and Cantonese speaking students, which is the first phase of our on-going research project, while the entire picture consists of NLP, grammatical inference, and spoken system as well.

The organization of the rest paper is arranged like this:
in the next section, the linguistic theoretical background is reviewed. Then the architecture of data mining model for SLA is illustrated in section 3. In section 4, the construction of the SLA model to replicate the students’ acquisition process is described, and the post-analysis on the generated knowledge is presented as well. The generation of knowledge in specific forms is given in section 5. Finally, conclusions and future direction of our research are summarized.

2. Linguistic theoretical background

This research is innovative in that it focuses both on a specific learning context and also on a specific group of learners who share the same first language (L1). According to the study on [4], [5], [6], we draw several conclusions in learning verbal morphology of Portuguese by Chinese speaking students as below:

- Portuguese learners show great difficulties in the use of the contrast stated by the verb tenses **simple past tense** (pretérito perfeito) and **imperfect past tense** (pretérito imperfeito).
- The marking of past inflectional morphology in obligatory contexts is dependent on the type of task – the presence of a context and time to focus attention on forms seems to have a positive effect on the marking of verbal morphology.
- Learners show greater concern with morphological form and rule application rather than content.
- Semantic information is important when marking **simple past tense** (pretérito perfeito) tense, but not when marking the **imperfect past tense** (pretérito imperfeito).
- Duration of exposure to L2 does not seem to have an effect on the performance of this group of learners.

Embrace all the domain knowledge in mind, our data mining model is intended to reveal that (a) what factors influence the acquisition quality of the students; and (b) whether any latent regularity could be uncovered to explain the acquisition processes.

3. Architecture overview

As the first stage of our research, we start the data mining model by replicating the acquisition process of verbal morphology from students who are learning Portuguese as L2. Then analyze and evaluate the linguistic data by employing data mining methodology, while significant regularities and novel insights could be discovered. Eventually all these new findings can benefit further SLA research and teaching. Figure 1 illustrates the general architecture of the desired model in dealing with all these issues, which contains four essential components: pre-processing, learning & analysis, knowledge generation and evaluation, in addition to the linguistic pre-processing phase.

In general, pre-processing includes both linguistic and non-linguistic pre-processing. The sentences collected from students are all raw data, certain level of linguistic information should be added to the sentences, such as parsing, POS tagging and structure labeling, etc. Besides, the analysis on syntactic, semantic or contextual data should be performed as well. These forms of data should be cleaned or transformed to an appropriate style for further processing afterwards, while data reduction and feature extraction are two of the major processes. Learning & analysis is the kernel phase of the model. In which, the cleaned data are learned and analyzed through various algorithms, where decision tree is our focus in this paper and the other two learning algorithms indicated in Figure 1 are neglected. This results in a set of knowledge in terms of classification regularities, insights and patterns. After that, the acquired knowledge is transformed and organized, according to user's preferences, into various desired forms by the knowledge generation module, such as the decision making rules or importance features ranking. All outputs allow Portuguese instructors in having sufficient evidence to better deploy the future teaching task. Finally, evaluation scheme serves as a self-tuning mechanism, which deals with the correction and improvement of the acquired model according to the repeated evaluation against the new data.

This paper concentrates on the middle two components: learning & analysis and knowledge generation, to discover the regularities and categories in learning Portuguese verbal morphology, which contains luxuriant forms and hardly to gain a mastery of it. These two comprise the crucial part of the acquisition process.
4. Data mining model for SLA learning & analysis

4.1. Data mining and decision tree

Data mining, which also popularly referred as knowledge discovery in database (KDD), is a process to automatically extract interesting patterns from volumes of data [7], [8], [9]. It performs data analysis and may disclose new ways to facilitate optimal decision making [10], which contributes greatly to business strategies, knowledge bases and scientific researches. The area of data mining and knowledge discovery is inherently associated with databases [11], which has gained great concern and thus widely applied in various areas, including business management, finance and marketing management, stock and transaction management, medical diagnosis and medicine management, administrative management, as well as language processing. Daelemans et al. [12] applied Quinlan's C4.5 inductive learning method [13], [14] to test linguistic hypotheses and discover interesting linguistic rules and categories. While Wang et al. [15] proposed an adaptive system analysis by employing decision tree algorithm, in order to optimize the learning sequences in the TESL (Teaching English as a Second Language). In addition, Romero et al. [16] used evolutionary algorithms to discover interesting relationships in students’ usage data, which may be very helpful to both teachers and course designers in maximizing the effectiveness of the course. Furthermore, Hsia et al. [17] used data mining techniques to analyze the course preferences and course completion rates of enrollees in extension education courses at a university in Taiwan. Therefore, data mining technology can be utilized for education and language processing purposes.

Decision tree classifier is one of the most widely used and practical supervised learning methods used for data exploration. Decision trees are considered attractive for many real-life applications, mostly due to their interpretability [18] and they have several other advantages as follows:

- they produce explicit concept descriptions, which have better understanding;
- the generated knowledge can easily be converted into classification rules;
- they are robust to noisy data;
- they have faster execution speed and have proven popular in practice.

Meanwhile, decision tree builders are able to select more appropriate subset of attributes rather than other learning algorithms, such as Bayesian classifiers, ANN and kNN algorithms [19]. The most well-known state-of-the-art decision tree inductive algorithms include C4.5 and its predecessor ID3 [20], [21]. However, basic decision tree algorithms work in batch mode. For a serial learning task that a new instance is introduced dynamically, the conventional non-incremental methods discard the existing rules and regenerate the new rules from scratch, such reproduction wastes the previous knowledge and effort while its cost may be too expensive. Therefore, incremental algorithms are proposed to build and refine a concept in a step-by-step basis each time a new training instance is observed, in which the decision tree acquired from learning on previous training data has been restructured, rather than relearning a new tree from scratch.

In this paper, we employ our previously proposed incremental decision tree approach to construct a SLA infrastructure to replicate the acquisition process of verbal morphology from students, who are learning Portuguese as L2. Such investigation forms a solid base for our future research in SLA with data mining technology.

4.2. Corpus

We collected a corpus of about 12,000 Portuguese sentences from the students’ written and oral story retell. For each of these raw sentences, the following linguistic processes are performed in order to have sufficient information for the subsequent mining process:

1) Label the tense for the sentence as: simple past tense (pretérito perfeito) or imperfect past tense (pretérito imperfeito). These two tenses are regarded as two class labels in our dataset.
2) Label linguistic information for the sentence into a variety of variables, such as verb stem, verb form, grammatical aspect, lexical aspect and syntax, etc.
3) Besides, non-linguistic information are tagged according to the student's characteristics and profiles, for instance, age, gender, mother tongue and number of years studying Portuguese, etc.

The labeled variables in steps 2) and 3) are used as features in the mining process, there are totally 26 features produced in addition to the two class labels in step 1). Figure 2 illustrates two examples of the parsed Portuguese sentences with linguistic annotation. In Figure 2, the original Portuguese sentence is given first; then the corresponding translation into English is followed for better comprehension; while the parsed linguistic data is shown in bold. The first variable (letter) indicates the category of classes, where the missing feature values are displayed with a slash (/) sign.
After the corpus has been created, it is ready to construct the data mining model for SLA learning & analysis. Two objectives are claimed:

I) To discover the regularities governing linguistic behavior for most of the students, through the automatic learning of the sentences that the students generated.

II) To discover the factors mastering the acquisition quality for most of the students, by automatic ranking the importance of each linguistic or non-linguistic variable in the corpus.

To fulfill the above intentions, we employ our previously proposed decision tree algorithm – $i^*$Learning (Intelligent and Incremental Learning) [22], to deploy a SLA model, in which the resulting knowledge is able to explain the acquisition behavior of our students. This method is chosen because of the following benefits:

- $i^*$Learning is an multi-dimensional incremental decision tree algorithm. We can take advantage of this into our further corpus expansion, in order to prevent re-learning all the regularities due to the increment of corpus data, especially when additional linguistic variables are required into analysis in the future.
- $i^*$Learning is superior to other state-of-the-art incremental learning algorithms not only on the classification accuracy, but also enhance the learning capacity without sacrificing the learning performance.
- $i^*$Learning is comparable to those non-incremental benchmark algorithms, such as C4.5. Since batch mode decision tree learning algorithms so far are proved to be able to guarantee the classification accuracy as well as the knowledge comprehensibility.

Incremental learning technique is one possible solution to the scalability problem, like the one described in this paper, where data may be provided time by time and processed in parts without re-training from scratch. $i^*$Learning grows a decision tree model in three phases: (1) Primary Off-line Construction of Decision Tree (POFC-DT); (2) Incremental On-line Revision of Decision Tree (IONR-DT); and (3) Dynamic Incorporating Attributes to Decision Tree (DIA-DT).

However, the construction of the model is out of the scope of this paper, please refer to [22] for details. Here we focus on the learning results from the replication of students' acquisition process.

4.4. Analysis

Analysis of the resulting knowledge for linguists' further usage and reference has equal importance as constructing the linguistic model. Our model results in two kinds of knowledge to reveal the acquisition behavior of students: a set of rules and a list of ranked factors. For better realization, we involve another application GoldVarb [23] into our analysis. It specializes in statistical analysis of linguistic variables, and is generally applied by the Portuguese instructors. Table 1 presents the resulting knowledge regarding factor ranking (only the most decisive ones are shown) by our model against GoldVarb application. It reveals that our model is able to discover the most significant factors, which influence the acquisition quality in learning Portuguese, while the result is similar to that of GoldVarb does, except the second factor.

<table>
<thead>
<tr>
<th>Rank</th>
<th>SLA model ($i^*$Learning)</th>
<th>GoldVarb</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tempo</td>
<td>Tempo</td>
</tr>
<tr>
<td>2</td>
<td>Forma_Verbal</td>
<td>Caracteristicas_sintacticas_do_enunciado</td>
</tr>
<tr>
<td>3</td>
<td>Aspecto_Gramatical</td>
<td>Aspecto_Gramatical</td>
</tr>
</tbody>
</table>

However, GoldVarb requires several steps to carry out the task and provides only statistical analysis, whereas the process is complicate and the result is hard to understand, especially for non-technical people, such as linguists and language instructors. Our model tackles these problems by a single clicking on the corresponding task and provides comprehensive results in terms of rules other than the ranking factors. Hence, even the non-skilled people are capable of performing the desired tasks through our model. Figure 3 shows three of the resulting rules in explaining the acquisition process according to the sentences produced by the students. Most of the rules are interested by the instructors, and are considered to be useful in their further teaching of L2.
On the other hand, the effectiveness and practicability of our model can be illustrated by comparing the learning performance of various learning algorithms like SVM [24] and neural networks (NN) [25], against \textit{i}+Learning on our corpus. TABLE 2 gives the evaluation results in learning accuracy, which reveals the fact that by employing \textit{i}+Learning method in SLA model, it is able to induce the most appropriate set of rules to explain the acquisition behavior in studying verbal morphology of Portuguese, amongst all the other algorithms. Although C4.5 has the same high learning accuracy as \textit{i}+Learning algorithm in studying the acquisition problem, it is a batch mode method, which lacks the sufficient capability for further expansion in both corpus and the entire SLA model. In addition, all four algorithms achieved relatively high learning accuracy. However, the former two methods (SVM and NN) need longer execution time than the other two. As a conclusion, \textit{i}+Learning is a right method for our SLA problem.

TABLE 2. LEARNING ACCURACY AMONG VARIOUS ALGORITHMS

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>98%</td>
</tr>
<tr>
<td>NN</td>
<td>95%</td>
</tr>
<tr>
<td>C4.5</td>
<td>99%</td>
</tr>
<tr>
<td>\textit{i}+Learning</td>
<td>99%</td>
</tr>
</tbody>
</table>

5. Knowledge generation

One of our practical researches is to adapt the resulting knowledge to specific representations for specific instructor. Normally, the quantity of resulting regularities depends on the number of linguistic factors in the corpus, which may lead to tens of such useful rules. Moreover, almost all the rules are nested in multiple levels. Generally, too many complex rules are hard and time-consuming for instructors to perform further analysis. Thus, our model offers the capability of allowing instructors to dynamically construct a desired rule by picking the required factors and the corresponding values. Figure 4 shows the interface for the factor selection dialogue. In which, instructors can specify the interesting factors and individual values, rather than displaying all rules without any alternative. Such that instructors can concentrate merely on the desired factors and rules that explain the acquisition phenomenon. Hence, the model is able to yield the best of useful regularities and factors in studying and analyzing the SLA problem.

6. Conclusions

This paper presents a data mining model by employing our previously proposed \textit{i}+Learning methodology, which aims at replicating and analyzing the acquisition process of verbal morphology of Portuguese as L2 by the Mandarin and Cantonese speaking students. The results manifest a solid clue that our model is useful for SLA processing and will help our Portuguese instructors in the language teaching for long-run.

Nevertheless, the current investigation is only the first stage of our long-term language processing research, there are several unavoidable weaknesses. The obvious problem is that there is lot of "missing values" in the corpus, since not every sentence can be annotated with all the designated linguistic variables. In our current model, we dealt with such "missing values" as a distinct value \textit{None}, this might not be a suitable way in real situation. Therefore, it would be valuable to explore a practical method for such case in our further research.

Acknowledgements

This research project is supported by the Research Committee of the University of Macau, under the reference No. RG054/08-09S/CS/FST. The authors are also grateful to...
the Faculty of Science and Technology, and the Portuguese Department of the University of Macau for supporting our research work in various aspects.

References