A Framework to Build Computer-Based Learning Tools for Studying Foreign Languages

Haodong Wu

〈Abstract〉

This paper outlines a framework to use computer and natural language techniques for various levels of learners to learn foreign languages in Computer-Based Learning environment. We propose some ideas for using the computer as a practical tool for learning foreign language where the most of courseware is generated automatically. We then describe how to build Computer Based Learning tools, discuss its effectiveness, and conclude with some possibilities for the future.

〈key words〉

computer-based learning, foreign language, natural language processing, mutual information corpus, machine-readable dictionary, lexical database, student model

1. Introduction

Interactive software, which ensures that learners can learn at their own pace, provides options for individualized exploration and control of the learning process. This, according to Alessi and Trollip (1991), has significant advantages, when compared with predominant traditional learning modes, in the areas of: perception and attention, memory, comprehension, active learning, motivation, locus of control, transfer of learning, and individual difference. Computers can also empower the learner to get around the many overt and covert agendas of educational monopolies through programs and applications not controlled by institutional bureaucracy. In many educational areas such as foreign language learning, computers, as a learning tool can provide self-learning environment for each learner according to his/her abilities, preferences and paces, etc. which is impossible for human instructors.

One charge consistently leveled against the rising academic interest in technology has been that it isolates learners from each other and from the instructor. However, it has also been argued that technology is a tool and has no inherent or required mode of application. The role of technology in education is always under the control of the instructor and is isolating only if learners are required to work on projects or assignments alone. In some cases independent work is appropriate, and in other cases cooperative work is desirable. Through active and adaptive learning; problem-solving and thinking skills; project-based learning; reflective thinking and productivity; task-relevant skills; authentic activities and cognitive apprenticeships, CBL can facilitate a constructivist and cooperative model of learning.

Now, many computer-based learning (CBL) systems are designed as Web-based Training (WBT) systems. As pointed by William Horton (2000), the most important features that are required for ideal learners are who:

- Learn independently and view learning positively
- Are self-disciplined, manage time well, and enjoying working alone
- Have a definite goal, such as certification, a degree, or the ability to perform a specific task
- Are moderately experienced in a field and already understand the basic concepts of that field.

We in this paper will discuss how the language resources (e.g., machine-readable dictionaries (MRDs), large scale corpora, lexical databases, grammar books, etc.) can be used to build personal-based question-answering tools for learners/learners to learn foreign languages such as English.

2. **Computer Tools for Learners**

In this research, we study how to employ various on-line linguistic resources to build CBL tools for training learners/learners to learn foreign languages, we in this paper limit our discussion to English. These tools will be embedded in a CBL system to improve learners' abilities in syntax and vocabulary. Unlike traditional classroom training and over disk-based computer-based training which are unable to corresponding to individual differences and needs efficiently, ours aims to provide a dynamic learning environment in which different learner will answer different questions and learn language knowledge by one pace.

2.1 **Linguistic knowledge** is automatic extracted from on-line machine-readable dictionaries and thesaurus, lexical databases, plain corpora and annotated corpora, grammar books. From these resources, we can automatic extract collocations, co-occurrences, conceptual relations, semantic similarities, contextual similarities, selectional restrictions, the definition and real usage of vocabulary.

2.2 **Automatic question generation** automatically/semi-automatically produces various types of questions and answers from online corpora or dictionaries. These questions will be stored in a question base. This topic will be discussed in Section 4 in detail.

2.3 **Error detections** find and correct misspelling, word or phrase misusing, grammar misunderstanding, etc. and provide this information to learners.

2.4 **Tutorials** for language, grammar, spelling and vocabulary deliver onscreen, sequential instruction that simulates the presentations instructors typically make to a whole class. In general they are self-paced, they are interactive, they provide some drill and practice, and they test learners learning periodically. Some have an automatic branching capacity that moves learners ahead if they perform well or takes them to a screen containing additional explanations, examples, or exercises if they're having difficulty. Good quality tutorial software can facilitate and multiply guided practice and remediation with a tutor.

2.5 **Human interfaces** are designed for learners, tutors and designers. For learners, they can evaluate these tools by their degrees of satisfaction, and give some suggestions that may be useful to improve the CBL system. For tutors or teachers, them can check the question base and edit it, assign some important syntactic patterns, and select particular source(s) that is used to generate questions, answers, and explanations if possible. For designers, they can make the system friendlier to the users of the system and extend its applications to other areas or other language.

2.6 **Student models** offer an efficient, time-saving method of keeping track of each learner's progress and record each learner with the items of tests' scores and his degree in learning. The student model of a learner can help the system to generate individual oriented courseware and organize individual learning sequences that lets each learner accomplish his/her learning goals.

2.7 **Games** if tests are viewed as game-like and unthreatening learners usually look forward to them. A number of products are designed to test learners comprehension of specific literary works with quiz programs, multiple choice, true/false and matching.
and various game formats. Electronic Bookshop (with over two hundred book titles) is an example of a computerized quiz program that tests learners' reading comprehension.

3. Natural Language Processing for Computer-Based Learning

Our goal is creating CBL system for learner mainly by drill and exercise learning. Here, we aims to build CBL environment to enrich learners' vocabulary, correct their errors both in syntax and vocabulary. Here, we acquire various kinds of information that is vital for designing our CBL tools. Here, we discuss below several most important cues helpful to realize our aims.

3.1 Word Associations and Mutual Information

We use mutual information (church et al. 1990) as a measurement to measure word association norms.

If two events \( x \) and \( y \), have probabilities \( P(x) \) and \( P(y) \), and co-occurrence probability \( P(x,y) \), then their mutual information, \( I(x,y) \), is define to be

\[
I(x, y) = \log \frac{P(x, y)}{P(x)P(y)}
\]

The value of mutual information has some features corresponding to the relationship between event \( x \) and event \( y \).

1. If there is a genuine association between \( x \) and \( y \), i.e., \( P(x,y) \gg P(x)P(y) \), and consequently \( I(x,y) \gg 0 \).
2. If there is no interesting relationship between \( x \) and \( y \), \( P(x,y) \approx P(x)P(y) \), and thus, \( I(x,y) \approx 0 \).
3. If \( x \) could hardly co-occur \( y \), it means \( x \) and \( y \) is in complement distribution, then \( P(x,y) \ll P(x)P(y) \), forcing \( I(x,y) \ll 0 \).

In our research, word probabilities \( P(x) \) and \( P(y) \) are estimated by counting the number of observations of \( x \) and \( y \) in a corpus, \( f(x) \) and \( f(y) \), and normalizing by \( N \), the size of the corpus. (In our research, we use a number of different corpus, 100 million words of Brown Corpus, on-line text Corpus, and so on). Joint probabilities, \( P(x,y) \), are estimated by counting the number of times that \( x \) is followed by \( y \) in a window of \( w \) words, \( fw(x,y) \), and normalizing by \( N \).

3.2 Semantic Similarity

Semantic Similarity is vital to generate both questions and/or their answers for many types of questions. Here we use the WordNet as a tool that measures the semantic similarities between two words or phrases.

The WordNet is a lexical database organized on psycholinguistic principles (Miller 1990; Beckwith et al.1991). It has a thesaurus-like structure in which lexical information about nouns, verbs, adjectives and adverbs is put in terms of word meanings, rather than word forms.

The most important part is a word sense network. A word sense node in this network is a group of strict synonyms called "synset" which is regarded as a basic object in the WordNet. Each sense of a word is mapped onto a synset and all word sense nodes in the WordNet are linked by a variety of semantic relationships like IS-A (hyponymy/hypernymy) and PART-OF (meronymy/holonymy). The IS-A relationship being dominant, "synset's" are grouped into hierarchies as is shown in Figure 1.

Two words are considered to be similar if their distance in the WordNet is 0 (synonyms) or 1 in semantic relationships of antonyms (e.g., man and woman, boy and girl, black and white), IS-A (e.g., red and color, man and human), PART-OF (e.g., head and body, Japan and Asia).

3.3 Selectional Restriction

Selectional restriction governs a semantic relation that occurs between or among the constituents in a phrase or sentence. It can be expressed as semantic constraints in the semantic interpretation process.

Selectional restriction (SR) is a semantic restriction imposed on lexical items when forming a sentence.
Consider the sentence (1).

(1) Tom hits a ball.

With the verb *hit*, one must use the subject that belongs to human beings or higher animals; the object of the verb *hit* should be a concrete entity like *ball*. There are many exceptions to, or irregularities in, the rules, however. One may say:

(2) Tom hits on an interesting idea.

Nevertheless, SRs are quite effective in producing error-detection questions that is an important instrument to train students. See an example:

(3) Peter likes **green** music and often discusses it with his friends.

We know in (3) that *green* as a color can be used to modify concrete entities like *vegetables*, but not abstract ones such as *music*.

We view SR as negative information, a constraint between two words. We acquire information on SRs

from corpora: for a particular adjective and a particular noun, we try to find 'similar' words to them and then check if they co-occur in the corpora. If no co-occurrence is observed in the corpora, then we consider that there is a SR between two words.

Selectional restrictions are used as a semantic constraint when we generate questions.

4. Question Generation

In this Section, we shall define five basic question forms for learning English, in which each question is automatically produced using on-line language resources. From these produced questions, instructors can choose and store them in question bases for different level learners, add explanation to the questions if necessary and sort the questions into a unit of a test or a lesson. The number of the question done by a learner and its result are also stored in the learner's student model.

4.1 Types of Question

4.1.1 Multiple-choice questions

Multiple-choice questions display a list of answers for learners to choose from. We generate multiple-choice questions from online linguistic resource such as a dictionary definition randomly selected from on-line machine-readable dictionaries (MRDs) or thesauri like Collins Cobuild English Dictionary, WordNet, etc. The word being defined has been blanked out from the definition. Each missing word is indicated by underline.

Example 1: Fill the missing word by choosing the right word listed.

When you make a _______ you choose what should be done or which is the best of various possible actions.

(a) plan (b) selection (c) decision (d) adjustment

The candidate words excluding the correct answer (i.e., the missing word) are produced using the words that co-occur the key word(s) and their mutual information is larger than 0. In this example, the key word is *make*.

Multiple-choose questions are easy to construct and easy to understand. Unfortunately, some learners attempt to make a guess rather than think, however.
4.1.2 Answer-input questions

Answer-input questions demand the learner to type in the answer to a question. Typically, the answer is dependent to the context. Answer-input questions are produced by the similar way of multiple-choose questions that use definition of the missing word or phrase on-line dictionaries.

Example 2: Can you guess what word or phrase is being defined?

Some people use _____ to call the partner of a married couple that they love each other very much.

Use answer-input questions to verify whether the learners have truly learnt the meanings of words or phrases. The answer may not be unique. Say, in some case, a word should be replaced by its synonyms.

4.1.3 Correct-misusage questions

Correct-misusage questions demand the learner to use a synonym to replace an underlined word or a phrase that is not suitable the context. Typically, the answer is justified using the context.

This kind of question is generated using the sentence from some corpus or online grammar book where the modifier is replaced by its similar word defined in WordNet and the mutual information between the replaced word and the word that is modified by it is great less than 0, which means that there is no modifier-modificant relation between the two words and there is a selectional restriction between them.

Example 3: Correct the miss in the following sentence:

(1) His car was involved in a heavy accident.
   The system: The correct answer is serious, not big.
   Input your answer: big

(2) The medium score was about 73 out of 100.
   Your answer: average
   The system: You are correct.

Correct-miss questions are helpful to building and checking learners' vocabulary efficient.

4.1.4 Fill-in-the-blanks questions

Fill-in-the-blanks questions require the learner to supply missing words in several places in a paragraph of text or a segment in a dialog. Fill-in-the-blanks questions are also called cloze questions. Such questions have a long history and are staple of education.

Example 4: Fill in the missing word in underline.

(1) _____ are you?
   ___ am ____, Thank ____, and ____?

(2) Suddenly the spring ____, the bad ____ weather ____ gone. It ____ May 1945.

Fill-in-the-blanks questions can also generated using corpora as the resources. What words can be missed, however, is quite difficulty. In many cases, the produced question should be checked and modified by human tutors.

According to Horton, W. (2000), we can use fill-in-the-blanks questions to measure the learner's ability to "apply knowledge within a contextual matrix." That means that learners use a partial answer to figure out the complete answer. Use fill-in-the-blanks questions:

- **To test incremental knowledge.** Learners know part of a subject and apply what they know such as world knowledge and language knowledge to
complete the answers.

- **Where context matters.** The correct answer could be inferred from surrounding text.
- **To measure ability to apply verbal knowledge in context.**

**Example 5:** Find the error in the following sentence and circle its mark.

1. All of the newly **elected** council members **introduced** themselves to the audience.
   - (A)  
   - (B)  
   - (C)  
   - (D)

2. John demanded Mary meeting his **parents** at the park as he was too busy to take a rest.
   - (A)  
   - (B)  
   - (C)  
   - (D)

3. We **must not forget** Japan's economic **force**.
   - (A)  
   - (B)  
   - (C)  
   - (D)

**4.1.5 Pick-up-error questions**

Pick-up-error questions are generated by choosing typical syntactic patterns from grammar books, corpus, or from instructors, and then replaced these patterns by change the number of verb and noun, the part of speech of a word, the tense of verb, and so on.

**4.2 Generate Questions for Individual Learner**

To motivate each learner, the question producer should not produce questions too difficult or too easy to the learner. In the recent phase, we consider vocabulary as the standard for define learners' levels. Each learner will decide his/her level by choosing the vocabulary, for example, of three thousand words. In advanced course, the vocabulary will be more than ten thousand words. The CBL tools will sort words with their occurrences in some large scale corpus such as British Nation Corpus¹, and select the words which rank in the $m$th (here, $m$ represents the number of words chosen by a learner).

To activate and motivate learners, the CBL system provides two ways for learners: system-led learning and learner-led learning. The courses can shift from system-led to learner-led during the progress of the course. For example, if a learner has made two many errors, the system will ask if the learner to choose a lower level. To the contrary, if the learner thinks the questions are too easy to him/her, he/her can demand to a higher one.

When the learner's answer is wrong, the system will give the correct answer. If the learner needs more examples, the system will give multiple examples extracted from corpora and other on-line linguistic resources.

**5. Discussion**

Designing Computer-Based Learning systems are normally laborious and also money-consuming work. It costs too much to design courseware and the framework of the system. In our proposal, the framework of the CBL system and courseware are built using natural language processing techniques that employ various machine-readable linguistic resources such as corpora, dictionaries, thesaurus, lexical database, grammar books, and employ various linguistic tools. We think our method if helpful to build CBL system to learn foreign language like English. The CBL system ensures that learners can learn at their own pace, provides options for individualized exploration and control of the learning process.

Since there are more and more free or low-cost linguistic resources come to be usable, the CBL system has good extensibility.

**6. Future Directions**

- **Academic writing tools** using word processors that include spelling, grammatical and style checkers, thesaurus and dictionary facilities in the composition process, accommodates a process-approach to writing. It reduces the drudgery in the revision and refinement of writing.

---

¹ The British National Corpus (BNC) is a 100 millions word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written.
and learners tend to develop a more positive approach to writing. Writing quality is improved by the increase in the number and complexity of revision operations, the writer is freer to experiment and think without committing to paper.

- **Study skills** can benefit from the use of a word processor where for example learners are asked to pick out key or topic sentences from a text or make summaries. Lecture notes can easily be expanded. The relative anonymity and individual nature of CAI is useful where there may be a stigma attached to attendance of 'remedial' sessions.

- **Exploratory Environments** CD-Rom and multimedia technology create visual environments for language learning. Typically the learner using such programs would explore a graphic display and see and hear language associated with objects and concepts in the display. Interactive opportunities exist through typed and spoken input by the learner.

- **Web-Based Integrated Learning Environments.** It supports long distance education.

- **Improvement of student model.** A better student model will make a learner choose the grammar or linguistic phenomenon that the learner has mastered.

### References:


