Natural Language Processing for Building Computer-Based Learning Tools

Haodong Wu
Department of Language and Culture
Dokkyo University

<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 using on-line resources.

1. Computer-Based Learning Tool

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 to learn foreign languages such as English. 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.

1.1 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.

1.2 Automatic question generation automatically/semi-automatically produce 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.

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

1.4 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.

1.5 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, they 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.

1.6 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 let each learner accomplish his/her learning goals.

2. Natural Language Processing for Computer-Based Learning

Our goal is creating CBL system for learner mainly by drill and exercise learning. Here, we aim 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.

- **Word Associations and Mutual Information**

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

- **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. 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*).

- **Selectional Restriction**

  Selectional restriction (SR) 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. SR is a semantic restriction imposed on lexical items when forming a sentence. 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-occurrences are 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.

3. 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.

3.1 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.

3.2 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.

3.3 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 to the context. Typically, the answer is justified using the context and the listener's knowledge and/or experience. 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:

His car was involved in a heavy accident. Input your answer: **big**

The system: The correct answer is **serious**, not **big**.

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

3.4 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 be generated using corpora as the resources. What words can be missed, however, is quite difficulty. In many cases, the produced question should be checked 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.**

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.

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

(1) All of the newly elected council members introduced theirselves 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)

3.6 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 Corpus[1], 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 for helping learners 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.

Preferences:

1) Alessi, S.M. & Trollip, S.R.: *Computer-Based Instruction: Methods and Development.* Englewood Cliffs: Prentice Hall, 1991

2) Church, K. W. & Hanks, P.: "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics*, Vol.16, No.1, pp:22-29.

3) Horton, W.: Designing Web-Based Training. Robert Ipsen, 2000.

4) Morgan, S.: "Computers and Academic Development in English at South African Universities". Paper for SAAAD conference 1997.

5) Turton, N.D., & Heaton, J.B.: *Longman Dictionary of Common Errors*. Bilingual English/Japanese edition, Pearson Education Japan, 2001.

6) Wu, H.: "A Hybrid Computational Model for Resolving Structural Ambiguities in Natural Language Processing". Mathesis Universalis. Vol.1, No.2, pp.392-449, 2000.

[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.