An intelligent tutoring system for learning Chinese with a cognitive model of the learner

Michał Kosek[1] and Pierre Lison[2]

Abstract. We present an intelligent tutoring system that lets students of Chinese learn words and grammatical constructions. It relies on a Bayesian, linguistically motivated cognitive model that represents the learner's knowledge. This model is dynamically updated given observations about the learner's behaviour in the exercises, and employed at runtime to select the exercises that are expected to maximise the learning outcome. Compared with a baseline that randomly chooses exercises at user's declared level, the system shows positive effects on users' assessment of how much they have learnt, which suggests that it leads to enhanced learning.

Keywords: intelligent tutoring systems, cognitive model, Bayesian networks, zone of proximal development.

1. Introduction

We present an intelligent tutoring system with a probabilistic model of user's knowledge of words and constructions, which chooses exercises that are most likely to maximise the learning outcome. It consists of English-to-Chinese translation tasks, which accept a large number of alternative translations and give interactive feedback when the provided answer is incorrect.

1. Department of Informatics, University of Oslo; michalkk@student.iln.uio.no.

2. Department of Informatics, University of Oslo; plison@ifi.uio.no.

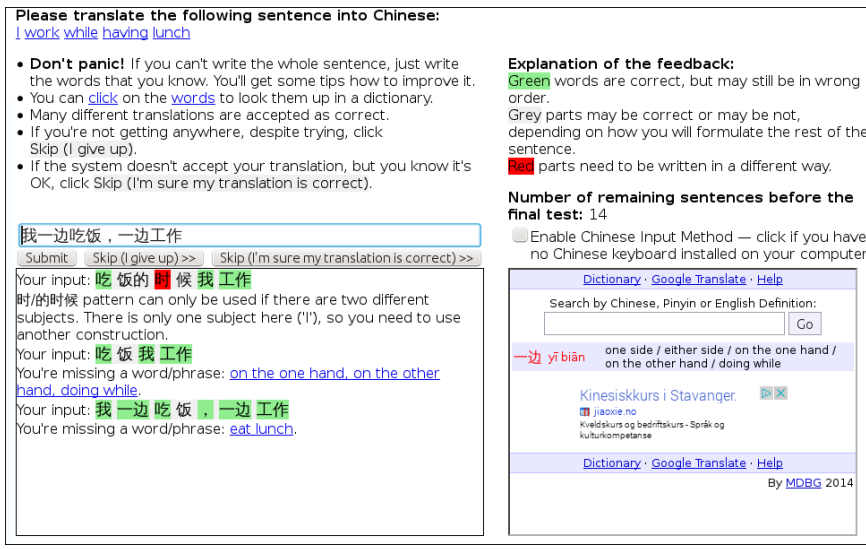
How to cite this article: Kosek, M., & Lison, P. (2014). An intelligent tutoring system for learning Chinese with a cognitive model of the learner. In S. Jager, L. Bradley, E. J. Meima, & S. Thouësny (Eds), *CALL Design: Principles and Practice*; *Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands* (pp. 179-184). Dublin: Research-publishing.net. doi:10.14705/rpnet.2014.000214

1.1. Structure of an exercise

Figure 1 shows an ongoing session with the program. Underlined words can be clicked to look them up in the MDBG English-Chinese dictionary[3]. In the first attempt the student used an incorrect construction, so the relevant fragment was highlighted in red.

After the second attempt, the system indicated a construction that was missing. The hint contained a hyperlink to the dictionary, which was clicked by the student, and a dictionary entry showed up on the right-hand side. Then the student used the construction from the dictionary. The feedback shows another missing construction; the exercise is unfinished.

Figure 1. Exercise in progress



The system leads the user towards an answer that is closest to the input according to the BLEU score (Papineni, Roukos, Ward, & Zhu, 2002). A large number of correct answers are accepted: different possible orders of constituents are allowed, and synonyms are recognised. The user can skip the exercise when she doesn't know the correct answer despite the hints, and when she knows that her answer is correct, despite the system saying otherwise.

3. http://www.mdbg.net

1.2. Selecting next exercise

Our research focuses on selecting exercises that are most beneficial to the user. This requires modelling the user's knowledge of words and grammatical constructions (both called *constructions* here), understood as pairs of one or several forms and a specific meaning. Users' knowledge is only partially observable through their interaction with the program; therefore a probabilistic student model is used. Its core is a set of random variables, one per construction. The probability of knowing a construction is updated as the program gathers indirect evidence.

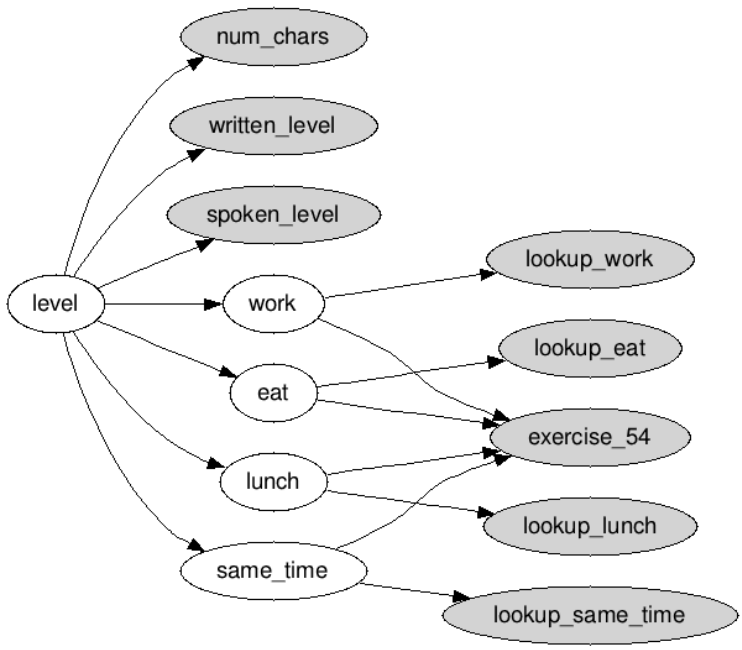
Before starting the exercises, the users assess their Chinese proficiency, and take a character recognition test, which determines the approximate number of Chinese characters they know. These data provide indirect evidence about which constructions are likely to be known –someone who scores high in the test will more likely know difficult words than someone who scores low. Afterwards, every exercise provides additional evidence. The user clicking on a word to check its translation indicates that she probably does not know it. Giving up and skipping to the next exercise indicates that probably some of the constructions in the sentence are unknown.

These pieces of evidence are represented in a Bayesian network (Pearl, 1988), letting the system reason about the learner's knowledge. Figure 2 presents a fragment of the network, relevant to the exercise shown before. The white nodes are hidden: the user's actual level and her knowledge of words are not directly observable. The grey nodes contain observable evidence: self-assessment, results of the character test, dictionary lookups and the exercise status (completed successfully or skipped). After the self-assessment and the character test, the values in the three top nodes are updated, and these changes are propagated into the hidden nodes. After each exercise, information about word lookups and the exercise status is used to update the hidden nodes.

The information about the user's knowledge is used to select an exercise that that will probably maximise the learning outcome. The system selects sentences that are most likely to lie within the Zone of Proximal Development (ZPD) (Vygotsky, 1978). Here, ZPD consists of sentences that the user would not translate without help, but would translate given the dictionary and system's hints. The sentence cannot be too easy (if everything is known, nothing will be learnt), or too difficult (with many unknown words, the user probably will not remember them). The next sentence is therefore chosen by an influence diagram (Pearl, 1988) that assigns lowest utility to sentences with all known

constructions, and highest utility to sentences with some unknown constructions –but not too many.

Figure 2. Fragment of the Bayesian network



2. Method

2.1. Parameter initialisation

We created 94 exercises, containing 91 constructions selected as learning targets, repeated among different exercises, with 3 constructions per exercise on average. 60 learners of Chinese were invited to assess their level, take the character test and do randomly chosen exercises.

The analysis of self-assessment, character test results, lookup ratio and skip ratio revealed four clusters of users that we called A, B1, B2 and C, to indicate rough correspondence with the CEFR levels (Council of Europe, 2001). There was not enough data to differentiate sub-levels of A and C. We assumed that the character test results were normally distributed within each group, and estimated the distribution of the number of known characters given the user's actual level.

We divided constructions into difficulty classes: for every user level X, class X contains constructions that were looked up by some users whose level is X, but not by those at higher levels. The constructions that were never looked up were removed from the learning target list, being too easy. Common conditional probability tables were created for each class. Classes B2 and C were merged during the evaluation, because the latter contained only three constructions.

2.2. Experiment

The goal of the experiment was to investigate the effects of using the cognitive model to select the exercises. The baseline used users' self-assessment of their written proficiency, and selected random exercises at that level. The level of an exercise was defined as the highest level of a construction that appears in that exercise.

The participants were recruited at Chinese language classes, online forums and, because of relative scarcity of Chinese L2 learners, by snowball sampling. 60 people used the program online, 33 of them went through all selected exercises and submitted a post-test and questionnaire. 24 participants were left after discarding two native Chinese speakers and those who had used the system before the experiment. The participants were randomly assigned to the system or baseline, their level was assessed, and they took the character test, and did 14 exercises. The system used the influence diagram to select the exercise with the highest utility, while the baseline chose a random unseen exercise at the user's declared level.

The post-test contained a stratified random sample of constructions to translate, with 6 random items from each of 3 strata: A, B1, B2+C. The questionnaire asked the users about subjective difficulty of the exercises and how much they had learnt. Post-test results and answers to the former question did not show significant differences, while answers to the latter showed statistically significant differences ($p<.05$) between the users of the system and the baseline, shown in the last row of Table 1.

The subjective effects could be compared for whole populations that submitted the questionnaire. As for the objective measures, the evaluation had to be done separately for every level, to satisfy an assumption that users' prior knowledge is similar. Hence, the lack of difference in the objective measures may have been caused by small sample sizes. An experiment with more participants is needed to investigate this.

The system currently models user's knowledge only during one session. Ways of separating short-term and long-term knowledge in the model must be investigated. No constructions are retained forever without repetition, but repeating same words during every session is suboptimal, therefore a forgetting model is important for a vocabulary tutor.

Table 1. Evaluation results

		System (15 participants)		**Baseline** (9 participants)	
	User level	**Mean**	**SD**	**Mean**	**SD**
Post-test results (number of correct answers)	A	15.33	2.81	14.25	1.48
	B1	16.75	1.09	17.75	0.43
	B2	18.00	0.00	18.00	0.00
	C	18.00	0.00	n/a	n/a
Subjective difficulty assessment (too easy or too difficult=0, right level=1)		0.53	0.50	0.44	0.50
Users' subjective assessment of how many items they've learnt (none=0, few=1, some=2, a lot=3)		**1.53**	**0.88**	**0.89**	**0.57**

3. Conclusions

We have presented a system that stores probabilistic information about users' knowledge of words and constructions on the basis of evidence collected, which is used to select exercises that are most beneficial to the user. Our experiment has shown positive effects of the system on users' assessment of how much they have learnt. Larger and longer-term experiments must be conducted to determine a possible difference in objective measures.

Acknowledgements. We would like to thank Jan Tore Lønning for his help with editing and proofreading the draft version of this article.

References

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311-318).

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco: Morgan Kaufmann.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.