A Selection Strategy to Improve Cloze Question Quality

Juan Pino, Michael Heilman, and Maxine Eskenazi

Language Technologies Institute
Carnegie Mellon University, Pittsburgh PA 15213, USA
{jmpino,mheilman,max}@cs.cmu.edu

Abstract. We present a strategy to improve the quality of automatically generated *cloze* and *open cloze* questions which are used by the REAP tutoring system for assessment in the ill-defined domain of English as a Second Language vocabulary learning. Cloze and open cloze questions are fill-in-the-blank questions with and without multiple choice, respectively. The REAP intelligent tutoring system [1] uses cloze questions as part of its assessment and practice module. First we describe a baseline technique to generate cloze questions which uses sample sentences from WordNet [2]. We then show how we can refine this technique with linguistically motivated features to generate better cloze and open cloze questions. A group of English as a Second Language teachers evaluated the quality of the cloze questions generated by both techniques. They also evaluated how well-defined the context of the open cloze questions was. The baseline technique produced high quality cloze questions 40% of the time, while the new strategy produced high quality cloze questions 66% of the time. We also compared our approach to manually authored open cloze questions.

1 Introduction

This paper describes a strategy to generate cloze (fill-in-the-blank) and open cloze (without multiple choice) questions, which are part of the assessment and practice module in the REAP system [1]. REAP is an intelligent tutoring system for English as a Second Language (ESL) vocabulary learning that provides a student with authentic documents retrieved from the web that contain target vocabulary words. These documents go through text quality filtering and are annotated for readability level [3] and topic. The system uses these annotations to match documents to the student's level and interests. After each reading the system provides a series of practice exercises for focus words that were in that reading. Following the practice session, the system updates the learner model for the student based on his or her performance. In the following reading session, the system searches for texts according to the updated model, in order to give the student individualized practice. The REAP system uses several types of questions, for example *synonym* or *related word* questions [4], and in particular cloze questions. A cloze question consists of a *stem*, which is a sentence

with a target word removed, and of *distractors*, which are words semantically or grammatically related to the target word. Examples of good and bad quality cloze questions are shown in Fig. 1 and 2. Brown et al. [5] successfully generated several different types of questions, for example synonym and cloze questions, by using WordNet [2]. However, teachers whose students were using the tutor judged the overall quality of the cloze questions to be insufficient to be used in their class. As a result, teachers generated the cloze items manually and without authoring support, which is time-consuming. This paper presents a strategy to improve the quality of automatically-generated cloze and open cloze questions.

Vocabulary learning can be considered an ill-defined domain due to the fact that there is no absolute measure of vocabulary knowledge. The exact nature of individual word knowledge is an unresolved research question, with various possible theories [6–8]. Although there exist ways of encoding lexical knowledge such as WordNet, Latent Semantic Analysis [9], and thesauri which can in some sense be viewed as models, there is no clear consensus on how to model lexical information. Even when assuming that an individual word is known, it is still challenging to make inferences about other, possibly related, words, because words appear in very different contexts [12]. On the contrary, well-structured domains have fully-developed cognitive models which have been applied successfully in intelligent tutoring. Examples include the work of VanLehn et al. [13] in physics and Anderson and Reiser in computer programming [10]. Also, many practice tasks for vocabulary learning, such as writing a sentence containing a vocabulary word (sentence production), lack single, definite correct answers. In fact, the sentence production problem can have an infinite number of correct answers: words can be combined in novel ways since human language is productive.

Although sentence production is a very good exercise for vocabulary learning and provides a rich measure of the knowledge of a word, it is very hard to assess automatically, precisely because it has many solutions. One could imagine having a model solution and compare to it a student's solution as in the entity relationship modelling tutor KERMIT [11]. Again this is hard to do automatically because the student's solution could be very far, for example semantically, from the model and yet correct. On the other hand, the reverse problem, posed by cloze questions, of completing a sentence with a missing word is less open-ended but still valuable for practice and assessment. One advantage is that cloze questions can be scored automatically. Even though the process of grading is simpler, cloze questions can still assess various aspects of word knowledge. Nation divides the concept of word knowledge into *receptive knowledge* and *productive knowledge* [6]. He lists ten criteria for receptive knowledge and nine criteria for productive knowledge. Multiple-choice cloze questions appear to involve five of the categories for receptive knowledge: students have to identify the written form of the target word and of the distractors. They also have to know at least one meaning of the word, namely the meaning of the word in a given context. They can make use of related words if such words are found in the stem. Finally, they need to recognize the correct grammatical usage of the word. However, there is no test of morphological knowledge. Students can also answer the question if

they know the meaning of the word in one specific context only. Finally, they do not have to check if the usage (register, frequency of use) of the word fits in the stem. Furthermore, *open* cloze questions also involve several categories of productive knowledge. To answer an open cloze question, students need to produce a word expressing a certain meaning in the given context. This word could be related to, or form collocations, that is pairs of frequently co-occurring words, with certain words in the stem, it should form a grammatically correct sentence, it should be spelled correctly and its level of formality should be chosen carefully. Therefore, open cloze questions also assess productive knowledge. Having a reliable assessment of individual word knowledge allows the system to know which words it should focus on.

Existing work on cloze question generation such as that of Liu et. al [14] has focused on lexical items regardless of their part-of-speech. Lee and Seneff [15] focused on generating cloze questions for prepositions with a technique based on collocations. Others have generated cloze questions for common expressions. For instance, Higgins [16] generated cloze questions for structures such as "not only the" that assess grammatical skills rather than vocabulary skills. Hoshino and Nagakawa [17] generated questions for both vocabulary and grammatical patterns. Mitkov et al. [18] generated cloze questions about concepts found in a document by converting a declarative sentence into an interrogative sentence.

This work first focused on generating cloze questions for adverbs and was then extended to other parts of speech. Adverbs are considered to stand between open class and closed class words[1], which is why our strategy should be extensible to any part of speech. Additionally, it seems that adverbs with the suffix **-ly** (for example "clearly"), which are the most frequent kind of adverb, can be easily replaced in a sentence by other adverbs of the same kind without producing a grammatically or semantically incorrect sentence, if this sentence does not sufficiently narrow the context. As a consequence, it is important to avoid having several possible answers for cloze questions on adverbs.

This paper concentrates on the quality of the stem and the quality of the distractors. Sumita et al. [19] generate distractors thanks to a thesaurus; if the original sentence where the correct answer is replaced by a distractor gets more than zero hit on a search engine, the distractor is discarded. In [14], the authors use a technique based on word sense disambiguation to retrieve sentences from a corpus containing a target word with a particular sense. Their strategy also uses collocations to select suitable distractors. Our strategy makes use of collocations too by applying it to both stems and distractors in order to ensure their quality.

[1] An open word class is a class of words that tends to evolve rapidly, for example the nouns form an open class because new nouns are often created while other nouns become obsolete; a closed class is stable, for example the set of prepositions is not likely to evolve very quickly.

2 Cloze Question Generation

We first describe the baseline technique based on WordNet sample sentences, which allowed us to investigate linguistic features for good cloze questions.

WordNet is a lexical database in which nouns, verbs, adjectives and adverbs are grouped in synonym sets, or *synsets*. Synsets have a semantic relationship such as synonym, antonym, etc. For each synset, a definition and, most of the time, sample sentences, are provided. The latter are used to produce the stems. We want to generate a cloze question for an adverb w. We assign its most frequent synset as an adverb to w. This synset may have sample sentences, either involving the target word w itself or synonyms in the same synset. If sentences involving the target word are available, we select them. After this first selection, if there are still several possible sentences, the longest one is preferred, because longer sample sentences in WordNet tend to have richer contexts. The distractors are chosen randomly from a list of adverbs. As a result, they have the same part of speech as the target word which is recommended by Haladyna et al. [20] as a way of avoiding obviously wrong answers. Furthermore, Hensler and Beck [21] have shown that higher proficiency readers among students from grade 1 to 6 can use part of speech as a clue to answer a cloze question. Since the REAP system is used by ESL learners who are able to use parts of speech in their native language, we believe they may be able to use them in English as well. Thus all the distractors have the same part of speech.

2.1 Stem Selection

Similar to the baseline, our strategy aims to select the most suitable sentences from a set of sentences containing the target word. However, for both for the stem and the distractors, our selection criteria are more fine-grained and we expect to produce a better quality output.

First, to apply a selection strategy we need to choose from several sample sentences per word. However, WordNet has zero or one sample sentence per word in any given synset. Therefore, we used the Cambridge Advanced Learner's Dictionary (CALD) which has several sample sentences for each sense of a word. We retained the same selection criterion as for the baseline, namely the length of the sentence, and added new linguistically relevant criteria. Our approach employs the following selection criteria: complexity, well-defined context, grammaticality and length.

Each sample sentence was given a weighted score combining these four criteria. We assessed the complexity of a sentence by parsing it with the Stanford parser [22, 23] and counting the resulting number of clauses. The Stanford parser uses the Penn Treebank syntactic tag set described in [24]. By *clause* we mean the sequences annotated with the following tags: S (simple declarative clause), $SBAR$ (clause introduced by subordinating conjunction), $SBARQ$(direct question introduce by *wh*-word or *wh*-phrase) and $SINV$ (declarative sentence with subject-auxiliary inversion). We chose this selection criterion through analysis of a dataset of high quality manually-generated cloze questions. We noticed that

high quality stems often consist of two clauses, one clause involving the target word, and the other clause specifying the context, as in the following sentence: "We didn't get much information from the first report, but *subsequent* reports were much more helpful." (the target word is italicized). We believe that more clauses tend to make the context more well-defined.

The context of a sentence is considered to be well-defined if it requires the presence of the target word in the sentence and rejects the presence of any another word. A way to assess how well-defined the context is in a sentence with respect to a target word is to sum the collocation scores between the target word and the other words in the sentence. Collocations are pairs, or sometimes larger sets, of words that frequently co-occur, often despite the absence of clear semantic requirements. An example is that tea is much more likely to be called "strong" than "powerful". Conversely, a car is more likely to be "powerful" than "strong". An "argument", however, can be either. The strength of the context around a word is in some sense determined by the presence of very informative collocations. For example, the sentence "I drank a cup of strong (blank) with lemon and sugar" has a very well-defined context for "tea" because of the collocations with "strong", "lemon", "sugar", and "drink". We followed the method described by Manning and Schütze [25] to estimate a set of collocations for each target word. We used a corpus consisting of approximately 100,000 texts appropriate for ESL learners, which are a part of the REAP database of potential readings. The system calculated the frequency of co-occurrence of content words by counting co-occurences within a window of two adjacent words. It then identified salient collocations by calculating a likelihood ratio for each possible pair. For this last step, other metrics such as point-wise mutual information and Pearson's chi-square test were also tried and produced similar estimates of collocations.

We also used the Stanford parser to assess the grammaticality of the sentence. Each parsed sentence was assigned a score corresponding to the probabilistic context-free grammar score. Longer sentences generally have poorer scores and thus we normalized this score with the length of the sentence[2]. Although the parser is applied to any sentence, even ungrammatical ones, the latter receive a lower score than grammatical sentences. Knight and Marcu [26] use a similar technique to estimate the probability of a short sentence in a noisy channel model for sentence compression.

Finally, we used the sentence length as a quality criterion. Indeed, we noticed that the sentences generated by the baseline technique were too short (6 tokens on average) to provide enough context for the target word to be guessed (see Fig. 2), although there were some rare exceptions (see Fig. 1).

The weights for each criterion were determined manually by the authors by examining their effects on performance on training data. We randomly chose 30 adverbs as training data and modified the weights so as to obtain the best sample

[2] The scores are log-probabilities, they are therefore negative, for example a correct sentence of ten words could be assigned a score of -100, a correct sentence of five words could be assigned a score of -50, and after normalizing, both sentence would be assigned a score of -10

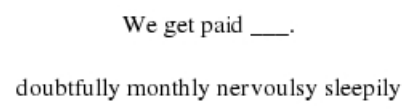

Fig. 1. A rare case of a short sentence with a sufficiently well-defined context for the target word.

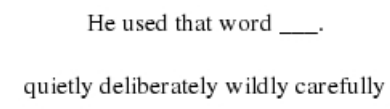

Fig. 2. A short sentence with multiple answers due to an ill-defined context.

sentences for these adverbs. Our criteria to judge the best samples were the same criteria used by ESL teachers for their assessment of the cloze questions (see Sect. 3). Our final weights are 1 for length, 8 for complexity, 0.3 for grammaticality and 0.2 for the collocation criteria[3]. In addition, we filtered out misspelled sentences with the Jazzy spell checker [27]. Although this is not very relevant for the Cambridge Dictionary, it is relevant when using a large corpus of text.

2.2 Distractor Selection

The quality of distractors was also scored. In order to produce a score for each distractor, we replaced the target word with a distractor in the sentence, then rescored the sentence. Only the grammaticality and collocation criteria measured how well the distractors fit in the sentence, both from a syntactic and from a semantic point of view. We selected the distractors with the highest scores, that is the ones that fit best in the sentence. Thus we prevented them from being obviously wrong answers. However we also wanted to avoid having several possible answers, which would happen if the distractors fit too well in the sentence. We therefore selected distractors that were semantically "far enough" from the target word. To compute semantic similarity between two words, we used Patwardhan and Pedersen's method [28]. In this method, two words w_1 and w_2 are associated with their definition in WordNet. Each word d of the definition is associated with a *first order context vector*, computed by counting the co-occurrences of d with other words in a corpus[4]. Then, computing the resultant (i.e. the sum) of these *context vectors* produces a *second order context vector*, which represents the meaning of the word. Finally the dot product of the *second order context vectors* associated with w_1 and w_2 give the semantic similarity between w_1 and w_2. This method, unlike several other methods based on the WordNet hierarchy, handles all parts-of-speech, not just verbs and nouns.

Using distractors that are semantically different from the target word does not guarantee that they will not fit in the sentence. Figure 2 shows that unrelated words can fit in the same sentence. Therefore, this technique will work with a sentence that has few possible answers.

[3] These values do not accurately reflect relative importance of the criteria; they are intended for the reader's information

[4] Patwardhan and Pedersen use the glosses in WordNet as a corpus

2.3 Stem selection using several dictionaries and a raw text corpus

We also applied the technique for stem selection to several dictionaries and a raw text corpus, with all parts of speech included. A corpus provides many sentences per word, therefore producing many good quality cloze questions per word. However, unlike dictionaries where sentences are carefully crafted, a large text corpus will contain many useless sentences that the algorithm ought to discard. We did not generate distractors. This test was mainly designed to evaluate the quality of the sentences independently of the quality of the distractors.

Dictionaries In order to compare sentences from dictionaries and from raw text corpus, we extracted the sample sentences provided by several dictionaries [5] for each word in the Academic Word List [29].

Corpus preprocessing A ten million word subset of the REAP documents, which are gathered from the Web, was filtered for text quality by running a part-of-speech (POS) tagger on each document and computing the cosine similarity between the vectors of POS trigrams in the document and a corpus of English literature, known to be grammatical and of high quality. HTML tags of the documents were stripped out. The raw text was chunked into sentences using the sentence detector of OpenNLP toolkit [30]. Only the sentences containing words from the Academic Word List were retained.

Parameters We used two different sets of weights for dictionaries and for the raw text corpus, shown in Tab. 1. In dictionaries, sentences tend to be too short, therefore the length weight is positive. On the contrary, sentences found in raw text are very long. To avoid long sentences, the length weight is negative. This way, we expect to find a trade-off between well-defined context and length. Furthermore, sentences of more than 30 words were discarded as not suitable for a practice exercise. A nonlinear function of the length could also have discarded sentences that are too short or too long. Dictionary sentences often lack a verb, which is why the complexity weight is high. The grammaticality weight is slightly higher for raw text because dictionary sentences are usually well-formed.

3 Evaluation

A series of three experiments were conducted. In each of them, five ESL teachers, two of whom are native speakers, assessed the quality of a set of questions. This set consisted of manually-generated and automatically-generated questions displayed in random order. Open cloze questions were generated to measure how well-defined the context was independently of the choice of distractors.The experiment settings are detailed in Tab. 2.

[5] Cambridge Advanced Learner's Dictionary, Longman Dictionary of Contemporary English, Dictionary.com, Merriam-Webster OnLine, MSN Encarta, RhymeZone.

Table 1. Different choice of weights for dictionaries and raw text corpus

	Dictionaries	Raw Text Corpus
length	0.4	-0.7
complexity	6	3
grammaticality	0.3	0.4
collocations	0.3	0.3

Table 2. Experiment Settings

Strategy	Baseline	Linguistically Enriched	Linguistically Enriched	
Question Type	Cloze	Cloze	Open Cloze	
Corpus for automatically generated questions	WordNet	CALD	Dictionaries	REAP
Number of automatically generated questions	30	30	34	33
Number of manually generated questions	30	30	30	

For each cloze question, the teachers answered a series of questions, illustrated in Fig. 3. The questions targeted specific reasons regarding the suitability of cloze questions.

For open cloze questions, the teachers assessed the quality of the context on a one-to-four scale. They also indicated if the sentence was too long, too short or of the right length, and if the sentence was too simple, too difficult or at the appropriate level of English for upper-intermediate ESL students.

3.1 Cloze Questions

The same teachers evaluated randomly ordered questions for both the baseline technique and the proposed technique. The target words for these questions were randomly sampled with replacement from a pool a word in both cases. Overall, the teachers judgments had a Fleiss' *kappa*[6] agreement of 0.3. We therefore consider the agreement between the teachers to be "fair" [31]. We expect that agreement would have been higher if better training procedures had been in place, such as allowing teachers to practice on an independent set of items and then discuss their differences in order to better calibrate their judgments. It should be also noted that these teachers teach at a variety of levels, in France and the United States.

Table 3 summarizes the results of the first two experiments about cloze questions. 40.14% of the questions generated by the baseline technique were judged acceptable, while our proposed strategy generated 66.53% of suitable questions.

[6] Fleiss' *kappa*, unlike Cohen's *kappa* measure, is a statistic for measuring agreement between more than two raters.

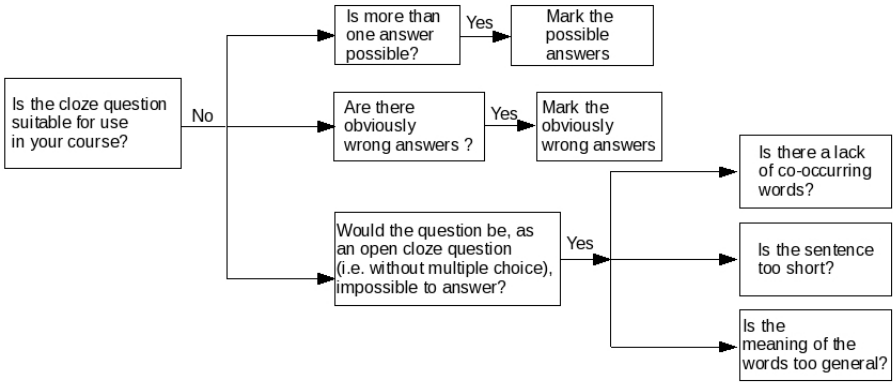

Fig. 3. Flow chart of the cloze question evaluation process

The difference was statistically significant ($t(29) = 2.048$, $p < 0.025$). This improved performance is still too low to plug the generated questions directly in a tutoring system; however, it appears to be high enough to allow us to build an authoring tool to improve authoring efficiency.

The most often cited reason for unacceptable questions was that several answers were possible. It was given for 34.01% of the baseline questions and 12.08% of the questions generated with the new strategy. In the baseline technique, there was no verification to determine if the distractors were semantically far enough from the target word. This flaw was corrected in the new strategy, but there is still room for improvement of the quality of the distractors. The second reason for unacceptable questions was that these questions, as open cloze questions, were impossible to answer, either because of a lack of co-occurring words, overly short sentences, or words with too general meaning. Again, the proposed strategy made a significant improvement over the baseline for all these aspects. However, the *kappa* values for inter-rater reliability were much lower for these points. The presence of obviously wrong answers as a reason for not acceptable questions was not often cited by the assessors either in the baseline strategy, or in the proposed strategy, probably because the distractors had the same part of speech as the correct answer and fit in the sentence at least grammatically.

3.2 Open Cloze Questions

Figures 4 and 5 show the distribution of the questions at levels 3 and 4 of context quality and for each set of question (manual, generated from dictionaries or from raw text). We consider that a sentence at level 3 and 4 of context quality is acceptable for an open cloze question. 61.82% of the automatically-generated questions for dictionaries are at levels 3 and 4 while 71.23% of the manually-generated questions were at levels 3 and 4.

Table 3. Assessment of cloze questions

Strategy	Manual (1st experiment)	Baseline	Manual (2nd experiment)	Linguistically Enriched
Suitable question	90.13%	40.14%	80.67%	66.53%
Several possible answers	3.95%	34.01%	10.67%	12.08%
Obviously wrong answers	0.66%	4.76%	1.33%	3.36%
Multiple choice necessary	1.32%	32.65%	0.67%	6.04%
Lack of co-occurring words	0%	16.33%	0%	2.68%
Too short	0.66%	19.93%	0%	5.37%
Words with too general meaning	0.66%	18.37%	0%	2.01%

When applied to dictionaries, our strategy is comparable to the human method for generating stems. The main difficulty encountered when generating cloze questions was the choice of the distractors. We can therefore focus on this task since the first one, namely generating stems, is relatively well mastered. However, the strategy did not perform as well on raw text. In raw text, the context is usually well-defined over several sentences but it is hard to find a single sentence that contains a rich context in itself. Analysis of length and level of difficulty shows that dictionary sentences tend to be too short but with the right level of difficulty and that raw text sentences tend to be too long and too difficult, as shown in Tab. 4.

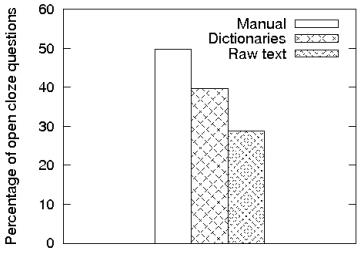

Fig. 4. Distribution of open cloze questions at context level 3 (well-defined context, two or three words can fit in the sentence)

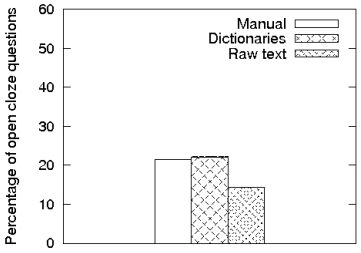

Fig. 5. Distribution of open cloze questions at context level 4 (very well-defined context, only one word can fit in the sentence)

Table 4. Length and difficulty level for open cloze questions

	Manual	Raw text	Dictionaries
Too long	0%	18.38%	2.29%
Too short	5.88%	10.29%	23.66%
Right length	94.12%	71.32%	74.04%
Level too difficult	4.31%	29.46%	9.37%
Level too simple	0.86%	1.55%	5.47%
Right level	94.83%	68.99%	85.16%

4 Conclusion and Future Work

We have developed a strategy for selecting high quality cloze questions. This strategy was prompted by a baseline technique developed within the framework of the REAP system. It was developed by taking into account the weak points of the baseline technique shown by a first evaluation. The evaluation of our strategy showed that we are able to generate stems of good quality, and therefore open cloze questions of good quality. We believe that by generating distractors of better quality, for example using the technique described in [19], we will be able to improve the automatic generation of cloze questions. Our technique can be extended to other languages by using different parsers, dictionaries and corpora. It can also be used as a measure for the amount of context, or information, a sentence provides.

We face two main challenges. First, distractors that fit in a sentence grammatically often also fit semantically. Choosing distractors with a large semantic distance from the correct answer does not always solve this problem. Similarly, open cloze questions rarely have only one possible answer. Second, corpus sentences inherently differ from dictionary sample sentences because the same amount of context is defined in several corpus sentences and in one dictionary sentence only.

At a higher level, we have demonstrated the utility of linguistically motivated, statistical approaches for generating assessment and practice materials in the ill-defined domain of English vocabulary learning. Ill-defined domains often deal with processes, especially linguistic ones, that are probabilistic in nature or at least possess a degree of complexity which make them challenging to model with deterministic algorithms. As such, statistical methods, such as those applied in our approach to generating vocabulary assessments, can be particularly effective for developing educational technology for ill-defined domains.

5 Acknowledgments

We would like to thank Gregory Mizera, Stacy Ranson, Catherine Jutteau, Renée Maufroid, Geneviève Jaslier and Annie Nottebaert for assessing the questions. We also would like to thank Catherine Jutteau and Sarah Carvalho for reviewing

this paper and Le Zhao and Anagha Kulkarni for scientific input. This material is based on work supported by NSF grant SBE-0354420, the research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040063 to Carnegie Mellon University. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsor.

References

1. Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi, M.: Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. Proceedings of the Ninth International Conference on Spoken Language Processing (2006)
2. Fellbaum, C.: WordNet. An electronic lexical database. Cambridge, MA: MIT Press (1998)
3. Collins-Thompson, K., Callan., J.: Predicting reading difficulty with statistical language models. Journal of the American Society for Information Science and Technology, vol. **56**, no. **13** (2005) 1448–1462.
4. Heilman, M., Eskenazi, M.: Application of Automatic Thesaurus Extraction for Computer Generation of Vocabulary Questions. Proceedings of the SLaTE Workshop on Speech and Language Technology in Education. (2007)
5. Brown, J., Frishkoff, G.,Eskenazi, M.: Automatic question generation for vocabulary assessment. Proceedings of HLT/EMNLP. Vancouver, B.C. (2005)
6. Nation, I. S. P.: Learning Vocabulary in Another Language. Cambridge, England: Cambridge University Press (2001) 26–28
7. Dale, E., O'Rourke, J.: Vocabulary building. Columbus, Ohio: Zaner-Bloser (1986)
8. Stahl, S. A.: Three principals of effective vocabulary instruction. Journal of Reading, vol. **29** (1986)
9. Landauer, T. K., Foltz, P. W., Laham, D.: An Introduction to Latent Semantic Analysis. Discourse Processes, vol **25** (1998) 259–284
10. Anderson, J. R., Reiser, B. J.: The LISP tutor: it approaches the effectiveness of a human tutor. Lecture notes in computer science, vol. **174** (1985) 159–175
11. Suraweera, P., Mitrovic, A.: An Intelligent Tutoring System for Entity Relationship Modelling. International Journal of Artificial Intelligence in Education, vol. **14** (2004) 375–417
12. Heilman, M., Eskenazi, M.: Language Learning: Challenges for Intelligent Tutoring Systems Proceedings of the Workshop of Intelligent Tutoring Systems for Ill-Defined Domains (2006)
13. VanLehn, K., Lynch, C., Schulze, K., Shapiro, J.A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., and Wintersgill, M.: The Andes Physics Tutoring System: Lessons Learned. International Journal of Artificial Intelligence and Education, 15 (3). (2005)
14. Liu, C. L., Wang, C. H., Gao, Z. M., Huang, S. M.: Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. Proceedings of the Second Workshop on Building Educational Applications Using NLP. Ann Arbor, Michigan (2005) 1–8
15. Lee, J., Seneff, S.: Automatic Generation of Cloze Items for Prepositions. Proceedings of Interspeech (2007)

16. Higgins, D.: Item Distiller: Text retrieval for computer-assisted test item creation. ETS (2006)
17. Hoshino, A., Nakagawa, H.: Assisting cloze test making with a web application. Proceedings of Society for Information Technology and Teacher Education International Conference. San Antonio, Texas, USA (2007) 2807–2814
18. Mitkov, R., An Ha, L., Karamanis, N.: A computer-aided environment for generating multiple-choice test items. Natural Language Engineering, Cambridge University Press (2006) 1–17
19. Sumita, E., Sugaya, F., Yamamoto, S.: Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. Proceedings of the Second Workshop on Building Educational Applications Using NLP. Ann Arbor, Michigan (2005)
20. Haladyna, T. M., Downing, S. M., Rodriguez, M. C.: A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Applied Measurement in Education. Lawrence Earlbaum (2002)
21. Hensler, B. S., Beck, J.: Better student assessing by finding difficulty factors in a fully automated comprehension measure. Proceedings of the 8th International Conference on Intelligent Tutoring Systems. Jhongli, Taiwan (2006)
22. Klein, D., Manning, C. D.: Fast Exact Inference with a Factored Model for Natural Language Parsing. Advances in Neural Information Processing Systems **15**. Cambridge, MA: MIT Press (2002) 3–10
23. Klein, D., Manning, C. D.: Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics (2003) 423–430
24. Marcus, M. P., Santorini, B., Marcinkiewicz, M. A.: Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics **19** (1993) 313–330
25. Manning, C. D., Schütze, H.: Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press (1999) Chap. 5
26. Knight, K., Marcu, D.: Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. Artificial Intelligence, vol. **139**(1) (2002) 91–107
27. http://jazzy.sourceforge.net/
28. Patwardhan, S., Pedersen, T.: Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts. Proceedings of the EACL Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together. Trento, Italy (2006) 1–8
29. Coxhead, A.: A new academic word list. TESOL Quarterly, vol. **34** (2) (2000) 213–238
30. http://opennlp.sourceforge.net/
31. Landis, J. R., Koch, G. G.: The Measurement of Observer Agreement for Categorical Data. Biometrics, vol. **33**, no. **1** (1977) 159–174