DEVELOPING AND EVALUATING A COMPUTER-ASSISTED NEAR-SYNONYM LEARNING SYSTEM USING MULTIPLE CONTEXTUAL KNOWLEDGE SOURCES

Assistant Professor Dr. Liang-Chih YU (Corresponding author)
Department of Information Management, Yuan Ze University, Chung-Li, Taiwan, R.O.C.
E-mail: lcyu@saturn.yzu.edu.tw

Kai-Hsiang HSU
Department of Information Management, Yuan Ze University, Chung-Li, Taiwan, R.O.C.
E-mail: s986220@mail.yzu.edu.tw

ABSTRACT
Despite their similar meanings, near-synonyms may have different usages in different contexts. For second language learners, such differences are not easily grasped in practical use. In this paper, we develop a computer-assisted near-synonym learning system for Chinese English-as-a-Second-Language (ESL) learners to better understand different usages of various English near-synonyms in a range of contexts. To achieve this goal, we implement the system using two automatic near-synonym choice techniques: pointwise mutual information (PMI) and *n*-grams to provide useful contextual information for learning frequent and discriminative words and *n*-grams occurring in the context of different near-synonyms. The system is evaluated using a vocabulary test with near-synonyms as candidate choices. Participants are required to select the best near-synonym for each question both with and without use of the system. Experimental results show that both techniques can improve participants' ability to distinguish different usages of various near-synonyms in a range of contexts, and use them appropriately. In addition, participants are found to prefer to use the PMI in the test, despite *n*-grams providing more precise information.
Keywords: Computer-assisted language learning, Near-synonym choice, Natural language processing, N-gram, Mutual information

INTRODUCTION
Vocabulary learning provides essential knowledge for recognizing the meanings of individual words, thus contributing to the development of various language skills (Jia, Chen, Ding, & Ruan, 2012; Fehr et al., 2012). Previous studies have demonstrated the use of various techniques for vocabulary learning such as multimedia instruction (Kim & Gilman, 2008), game-based learning (Neville, Shelton, & McInnis, 2009), animations (Kayaoglu, Dagakbas, & Ozturk, 2011), collaborative learning (Lin, Chan, & Hsiao, 2011) and ubiquitous learning (Huang, Huang, Huang, & Lin, 2012). In addition to individual word learning, identifying a group of words with similar meanings (i.e., near-synonyms) can also be an effective way of learning a language. Sun et al. (2011) developed a method to discover near-synonyms and similar-looking words from WordNet (Fellbaum, 1998), and rich sources of near-synonyms for different languages have been found in other lexical ontologies such as EuroWordNet (Rodríguez et al., 1998), HowNet (Dong and Dong, 2006), and Chinese WordNet (Huang et al., 2008). These collections of near-synonyms are useful knowledge resources for both computer-assisted language learning (CALL) (Cheng, 2004; Inkpen & Hirst, 2006; Inkpen, 2007; Ouyang, Gao, & Koh, 2009; Wu, Liu, Matthew, & Yu, 2010) and natural language processing (NLP) applications such as information retrieval (IR) (Moldovan & Mihalcea, 2000; Navigli & Velardi, 2003; Shlrl & Revle, 2006; Bhogal, Macfarlane, & Smith, 2007; Yu, Wu, & Jang, 2009) and (near-)duplicate detection for text summarization (Vanderwende, Suzuki, Brockett, & Nenkova, 2007). For example, in composing a text, near-synonyms can be used to automatically suggest alternatives to avoid repeating the same word in a text when suitable alternatives are available in the near-synonym set (Inkpen and Hirst, 2006; Inkpen, 2007). In information retrieval, systems can perform query term expansion to improve the recall rate, for example through recognizing that the weapon sense of "arm" corresponds to the weapon senses of "weapon" and "arsenal".

Although the words in a near-synonym set have similar meanings, they are not necessarily interchangeable in practical use due to their specific usage and collocational constraints (Wible, Kuo, Tsao, Liu, & Lin, 2003). Consider the following examples.

(E1) {strong, powerful} coffee
(E2) ghastly {error, mistake}

Examples (E1) and (E2) both present an example of collocational constraints for the given contexts. In (E1), the word "strong" in the near-synonym set {strong, powerful} is more suitable than "powerful" in the context of "coffee", since "powerful coffee" is an anti-collocation (Pearce, 2001). Similarly, in (E2), "mistake" is more suitable than “error” because “ghastly mistake” is a collocation and “ghastly error” is an anti-collocation (Inkpen, 2007). These examples indicate that near-synonyms may have different usages in different contexts, and such differences are not easily captured by second language learners. Therefore, this study develops a computer-assisted near-synonym learning system to assist Chinese English-as-a-Second-Language (ESL) learners to better understand different usages of various English near-synonyms and use them appropriately in different contexts.

To this end, this work uses NLP techniques such as automatic near-synonym choice techniques (Edmonds, 1997; Inkpen, 2007; Gardiner & Dras, 2007; Islam & Inkpen, 2010; Wang & Hirst, 2010; Yu, Wu, Chang, Liu, & Hovy, 2010; Yu, Chien, & Chen, 2011) to verify whether near-synonyms match the given contexts. The problem of automatic near-synonym choice has usually been formulated as a “fill-in-the-blank” (FITB) task, as shown in Figure 1.

English Sentence:	This will make the _____ message easier to interpret.
Original word:	error
Near-synonym set:	{error, mistake, oversight}

Figure 1. Example of FITB evaluation for automatic near-synonym choice.

Given a near-synonym set and a sentence containing one of the near-synonyms, the near-synonym is first removed from the sentence to form a lexical gap. The goal is to predict an answer (i.e., best near-synonym) to fill the gap from the near-synonym set according to the given context. Among approaches to automatic near-synonym choice, *pointwise mutual information (PMI)* (Inkpen, 2007; Gardiner & Dras, 2007), and *n-gram* based methods (Islam & Inkpen, 2010; Yu et al., 2010) are the two most commonly-used approaches to automatic near-synonym choice. The two approaches provide different aspects of contextual information including individual and contiguous relationships between near-synonyms and their context words. PMI is used to measure the strength of co-occurrence between a near-synonym and individual words appearing in its context, while *n*-grams can capture contiguous word associations in the given context. Both aspects are useful sources for learning different usages of various near-synonyms in a range of contexts. Therefore, this study uses both techniques to implement a system with which learners can practice learning useful contextual information from various aspects.

In the remainder of this paper we first introduce the computer-assisted near-synonym learning system, including its main components PMI and N-gram; we then summarize the experimental results of user evaluations, with conclusions and future directions presented at the end.

SYSTEM DESCRIPTION

Main Components

PMI-based Method

The pointwise mutual information (Church & Hanks, 1990) used here measures the co-occurrence strength between a near-synonym and the words in its context. A higher PMI score indicates that the near-synonym fits well in the given context. Let w_i be a word in the context of a near-synonym NS_j. The PMI score between w_i and NS_j is calculated as

$$PMI(w_i, NS_j) = \log_2 \frac{P(w_i, NS_j)}{P(w_i)P(NS_j)},$$

(1)

where $P(w_i, NS_j) = C(w_i, NS_j)/N$ denotes the probability that w_i and NS_j co-occur; $C(w_i, NS_j)$ is the number of times w_i and NS_j co-occur in the corpus, and N is the total number of words in the corpus. Similarly, $P(w_i) = C(w_i)/N$, where $C(w_i)$ is the number of times w_i occurs in the corpus, and $P(NS_j) = C(NS_j)/N$, where $C(NS_j)$ is the number of times NS_j occurs in the corpus. All frequency counts are retrieved from the Web 1T 5-gram corpus released by the Linguistic Data Consortium. Therefore, Eq. (1) can be re-written as

$$PMI(w_i, NS_j) = \log_2 \frac{C(w_i, NS_j) \cdot N}{C(w_i)C(NS_j)}.$$

(2)

The PMI score is then normalized as a proportion of w_i occurring in the context of all near-synonyms in the same set, as shown in Eq. (3).

$$\widetilde{PMI}(w_i, NS_j) = \frac{PMI(w_i, NS_j)}{\sum_{j=1}^{K} PMI(w_i, NS_j)}, \quad (3)$$

where $\widetilde{PMI}(w_i, NS_j)$ denotes the normalized PMI score, and K is the number of near-synonyms in a near-synonym set. For example, in Figure 1, the normalized PMI scores between the context word *message* and the near-synonyms *error*, *mistake*, and *oversight* are 0.99, 0.01, and 0, respectively, indicating that for all co-occurrence frequencies between *message* and the three near-synonyms, the proportion of the co-occurrence of (message, error) is 99%, while the remaining 1% is (message, mistake) and (message, oversight) is 0. Through the proportion-based PMI scores, learners not only learn frequently co-occurring context words for individual near-synonyms, but also learn context words (e.g., *message*) that are useful for discriminating among near-synonyms through comparing their co-occurrence proportions among the near-synonyms.

N-gram

A *n*-gram represents *n* contiguous words such as “error message” (bi-gram), “error message easier” (tri-gram), “error message easier to” (4-gram), and “error message easier to interpret” (5-gram). This component retrieves *n*-gram frequencies (*n*: 2~5) from the Web 1T 5-gram corpus.

System Implementation

Based on the contextual information provided by the PMI and N-gram, the system implements two functions: contextual statistics and near-synonym choice, both of which interact with learners. The system can be accessed at http://nlptm.mis.yzu.edu.tw/NSLearning.

Contextual statistics

This function provides the contextual information retrieved by PMI and N-gram. This prototype system features a total of 21 near-synonyms grouped into seven near-synonym sets, as shown in Table 1. This dataset has been widely used in previous work on automatic near-synonym choice.

Table 1. Near-synonym sets.

No.	Near-Synonym sets
1	difficult, hard, tough
2	error, mistake, oversight
3	job, task, duty
4	responsibility, burden, obligation, commitment
5	material, stuff, substance
6	give, provide, offer
7	settle, resolve

Figure 2 shows a screenshot of the interface for contextual information lookup. Once a near-synonym set is selected, the 100 top-ranked context words and *n*-grams are retrieved for each near-synonym in the set. For PMI, both proportion-based PMI scores (Eq. (3)) and co-occurrence frequencies between near-synonyms and their context words are presented. For N-gram, the 100 top-ranked *n*-grams with their frequencies are presented. Through this function, learners learn to determine the most frequently co-occurring and discriminative words and *n*-grams for different near-synonyms.

Near-synonym choice

This function assists learners in determining suitable near-synonyms when they are not familiar with the various usages of the near-synonyms in a given context. Learners can specify a near-synonym set and then input a sentence with “*” to represent any near-synonym in the set. The system will replace “*” with each near-synonym, and then retrieve the contextual information around “*” using PMI and N-gram. Figure 3 shows a sample sentence (the original word *substance* has been replaced with *) along with its contextual information retrieved by the system. For PMI, at most five context words (window size) before and after “*” are included to compute proportion-based PMI scores for each near-synonym. In addition, the sum of all PMI scores for each near-synonym is also presented to facilitate learner decisions. For N-gram, the frequencies of the *n*-grams (2~5) containing each near-synonym are retrieved. In the example shown in Figure 3, learners can learn useful word pairs such as (substance, matter) and *n*-grams such as “substance of the matter”, thus learning to discriminate between *substance*, *material* and *stuff*.

Near-synonym set job task duty

job			task			duty		
Context	PMI_score	Frequency	Context	PMI_score	Frequency	Context	PMI_score	Frequency
teen	1	1,816,316	trivial	1	78,286	cycle	1	342,491
seekers	1	1,479,452	committees	1	75,321	breach	1	336,947
listings	1	1,473,629	pane	1	52,660	fiduciary	1	325,983
opportunities	1	1,416,347	privileged	1	49,161	tour	1	240,835
openings	1	1,071,984	force	0.99	2,874,435	stamp	1	172,109
interview	1	858,245	impossible	0.99	186,108	trucks	1	170,757
vacancies	1	674,304	ending	0.99	162,489	steel	1	165,604
cum	1	573,873	faced	0.99	70,166	statutory	1	155,827
postings	1	570,120	complex	0.98	142,552	customs	1	154,790

Near-synonym set job task duty N-gram 4 Submit

job		task		duty	
to do the job	438,769	the task at hand	172,859	have a duty to	202,441
did a great job	425,841	with the task of	167,026	is the duty of	191,024
a good job of	412,589	not an easy task	145,907	be the duty of	178,800
do a better job	357,618	up to the task	143,106	shall be the duty	161,951
link to save job	346,000	of the task force	122,120	the line of duty	160,011
link to job advert	342,589	the task of the	102,082	the call of duty	143,000
Direct link to job	342,236	be a daunting task	86,773	has a duty to	140,161
to get the job	340,263	a task force to	77,369	it is the duty	88,590
do a good job	320,218	is no easy task	71,692	It is the duty	87,750

Figure 2. Screenshot of contextual statistics.

Near-Synonym set material stuff substance

It was found that the * of the matter and not only mere theory was to be regarded submit

Window size 3

PMI	material	stuff	substance
found	0.35	0.34	0.31
that	0.24	0.49	0.28
the	0.31	0.27	0.42
of	0.28	0.28	0.44
the	0.31	0.27	0.42
matter	0.15	0.08	0.77
	1.64	1.73	2.64

Bi-gram	the material	7,488,173	the stuff	2,581,457	the substance	1,319,583
	material of	817,776	stuff of	392,805	substance of	848,188
Tri-gram	that the material	237,330	that the stuff	25,305	that the substance	52,803
	the material of	129,580	the stuff of	254,931	the substance of	545,206
	material of the	179,643	stuff of the	31,130	substance of the	341,962
4-gram	found that the material	1,706	found that the stuff	211	found that the substance	623
	that the material of	3,082	that the stuff of	910	that the substance of	15,240
	the material of the	48,148	the stuff of the	9,307	the substance of the	242,832
	material of the matter	0	stuff of the matter	0	substance of the matter	6,205
5-gram	found that the material of	0	found that the stuff of	0	found that the substance of	110
	that the material of the	1,229	that the stuff of the	150	that the substance of the	7,898
	the material of the matter	0	the stuff of the matter	0	the substance of the matter	5,427

Figure 3. Screenshot of near-synonym choice.

EXPERIMENTS AND EVALUATIONS

To evaluate the system, we designed a vocabulary test with near-synonyms as candidate choices to examine whether the system can provide useful contextual information to assist participants in discriminating among near-synonyms for each question. In the following subsections, we first introduce the design of the vocabulary test questions, and the test procedure including a pre-test and post-test. The test results are then presented, followed by a comparison of PMI and N-gram.

Experimental Setup

- *Question design*: The vocabulary test consisted of 50 questions with a single correct answer for the 21 near-synonyms, where each near-synonym had at least two questions. The remaining eight randomly selected near-synonyms had three questions each. Each question was formed from a sentence selected from the British National Corpus (BNC). Figure 4 shows a sample question. For each question, the original word was removed from the question sentence and held as the correct response. The original word and its near-synonyms in the same set were then supplied as candidate choices. We then followed the FITB evaluation procedure presented in Figure 1 to determine which method, PMI or N-gram, is correct for each question (i.e., which method can propose the original word as the answer). Table 2 shows the evaluation results for the 50 questions. Neither method produced perfect results. The respective accuracies of PMI and N-gram were 64% (32/50) and 68% (34/50).

Question:	He wanted to do a better ______ than his father had done with him.
	A. job B. task C. duty
Questionnaire 1:	How much did you depend on the system to answer the question?
	□ 1 (Not at all dependent) □ 2 □ 3 □ 4 □ 5 (Completely dependent)
Questionnaire 2:	Which method did you use in the test? □ PMI □ N-gram

Figure 4. Sample question in the vocabulary test. The original word in the sample question sentence is *job*.

Table 2. Number of questions correctly and incorrectly identified by PMI and N-gram.

		N-gram		Sum
		Correct	Incorrect	
PMI	Correct	19	13	32
	Incorrect	15	3	18
Sum		34	16	50

- *Test procedure*: The test procedure included a pre-test and post-test. In the pre-test, participants were presented with each question (excluding the two questionnaire items) and asked to propose an answer from the candidate choices without using the system. Before the post-test, participants could practice using the system to learn different aspects of contextual information for near-synonyms by either using the lookup interface or keying in sentences. In the post-test, the participants were presented with each question sentence along with the two questionnaire items and asked to answer the questions using the system. After completing each question, the participants were asked to provide two feedback items. The first questionnaire item, as shown in Figure 4, is a 5-point scale measuring the degree to which the participant felt reliant on the system during the test, and reflects participants' confidence in answering questions. A response of 1 indicates the participant felt highly confident in his/her proposed answers, and thus chose to answer questions without use of the system, while higher responses indicated increasing reliance on the system in answering questions. In the second item, participants were asked to indicate which method, (i.e., PMI, N-gram, both or neither) provided the most useful contextual information. Analysis of participant responses indicated their preferred methods and which method best contributed to near-synonym learning.

Evaluation Results

A total of 30 (16 males and 14 females) non-native English speaking graduate students volunteered to participate in the vocabulary test. Each completed the 50 questions in both the pre-test and post-test, and the two questionnaire items for each question in the post-test. Table 3 shows the percentage of correctly answered questions in the pre-test and post-test, averaged over the 30 participants. The results show that, on average, around 44% of questions were correctly answered in the pre-test. After using the system, this increased substantially to 70%. The performance difference between the pre-test and post-test was statistically significant (t-test, $p<0.001$). This finding indicates that the use of the system improved participants' ability to distinguish different usages of various near-synonyms in a range of contexts, and use them appropriately.

Table 3. Pre-test and post-test results.

	Correctly answered questions			
	Average	Min.	Max.	Std.
Pre-test	44%	30%	66%	9.62
Post-test	70%*	60%	84%	5.60

*statistically significant ($p<0.001$)

We performed a cross analysis of the two questionnaire items against the 1500 answered questions (i.e., 30 participants each answering 50 questions) in both the pre-test and post-test, with results shown in Table 4. The columns labeled C_{pre}/C_{post}, $C_{pre}/\overline{C}_{post}$, $\overline{C}_{pre}/C_{post}$ and $\overline{C}_{pre}/\overline{C}_{post}$ respectively represent four groups of questions: correct in both the pre-test and post-test, correct in the pre-test and incorrect in the post-test, incorrect in the pre-test and correct in the post-test, and incorrect in both the pre-test and post-test. The rows labeled "Without_system" and "With_system" indicate whether or not questions in the post-test were answered using the system, where "Without_system" was identified based on participants' responding 1 or 2 on the first questionnaire item, and "With_system" based on responses of 3~5. For "Without_system", Table 4 shows that around 36% (536/1500) questions in the post-test were answered without use of the system due to high confidence on the part of participants. As shown in Figure 5, around 59% (315/536) of these questions were answered correctly in both the pre-test and post-test, while only 28% (151/536) were answered incorrectly in both the pre-test and post-test, indicating that participants' confidence in their ability to answer certain questions correctly was not misplaced. The remaining 13% of questions in "Without_system" provided inconsistent answers between the pre-test and post-test.

Table 4. Cross analysis of questionnaire items against answered questions.

	C_{pre}/C_{post}	$C_{pre}/\overline{C}_{post}$	$\overline{C}_{pre}/C_{post}$	$\overline{C}_{pre}/\overline{C}_{post}$	Total	
Without_system	315	21	49	151	**536**	**1500**
With_system	244	78	448	194	**964**	
PMI	91	51	239	100	**481**	**824**
N-gram	93	19	177	54	**343**	

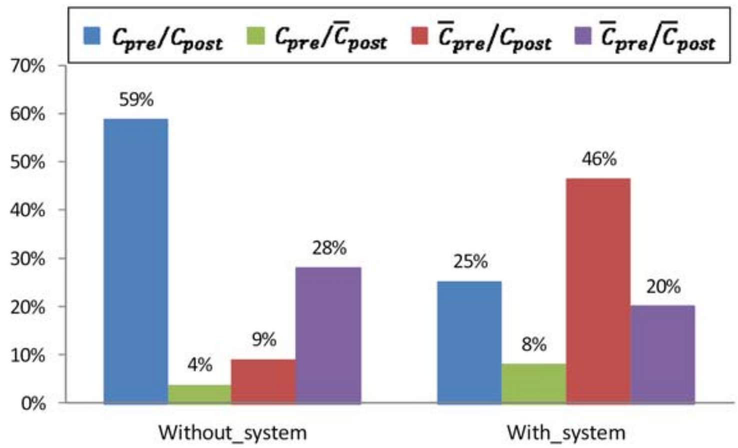

Figure 5. Histograms of with and without system.

For "With_system", Table 4 indicates around 64% (964/1500) questions answered using the system in the post-test. Of these questions, around 46% (448/964) were answered incorrectly in the pre-test but were corrected in the post-test, indicating that participants had learned useful contextual information from the system. Around 25% (244/964) of questions answered correctly in the pre-test were also answered correctly in the post-test because participants became more confident after double-checking their proposed answers with the system. Only 8% (78/964) of questions answered correctly in the pre-test were answered incorrectly in the post-test, and the remaining 20% (194/964) of questions answered incorrectly in the pre-test were still incorrect in the post-test. A possible explanation is that the system does not always provide perfect results, as indicated in Table 2. In some circumstances the system may provide ambiguous contextual information, such as when the given context is too general. In such cases, participants may propose incorrect answers despite having used the system.

Comparison of PMI and N-gram

Table 4 shows that there were a total of 824 (<=964) questions with feedback on the second questionnaire item, where 58% (481/824) questions were answered based on PMI, and the remaining 42% (343/824) based on N-gram, indicating that participants had a preference for PMI in the test. But, in fact, previous studies have shown that the 5-gram language model has an accuracy of 69.9%, as opposed to 66.0% for PMI (Islam & Inkpen, 2010), thus N-gram provides more precise information. Our evaluation results of the 50 questions, as shown in Table 2, were consistent with this discrepancy, and the respective accuracies of N-gram and PMI were found to be 68% and 64%.

Figure 6 shows the comparative results of PMI and N-gram. The percentages of both C_{pre}/C_{post} and $\overline{C}_{pre}/C_{post}$ for N-gram were higher than those for PMI, and the percentages of both $C_{pre}/\overline{C}_{post}$ and $\overline{C}_{pre}/\overline{C}_{post}$ for N-gram were lower than those for PMI. Overall, N-gram use resulted in a correct/incorrect ratio of 79:21 in the post-test, as opposed to 69:31 for PMI, indicating that N-gram can assist participants in correctly answering more questions and producing fewer errors caused by ambiguous contextual information.

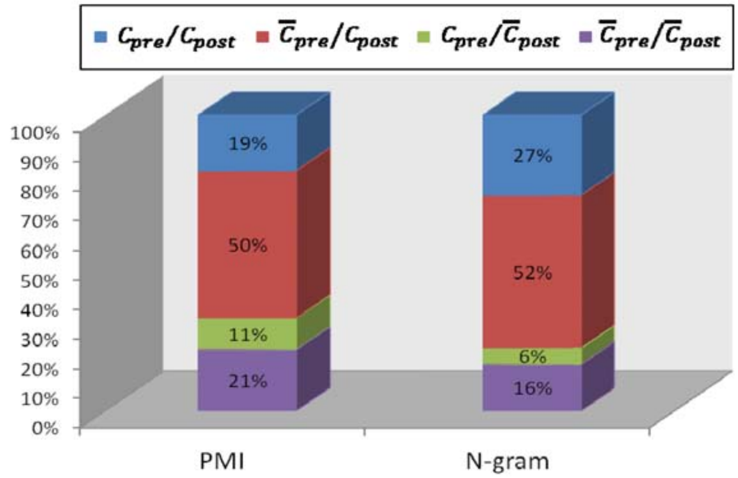

Figure 6. Comparative results of PMI and N-gram.

DISCUSSIONS

The evaluation results for each question are summarized in Table 5. Columns 1~5 indicate the proportions of each scale rated by the 30 participants in answering each question, columns C_{pre}/C_{post}, $C_{pre}/\overline{C}_{post}$, $\overline{C}_{pre}/C_{post}$ and $\overline{C}_{pre}/\overline{C}_{post}$ indicate the respective proportions of the four groups of questions, the columns PMI and N-gram indicate the proportions of participants' choices for each method, and the column Correctness indicates which method (i.e., PMI, N-gram, Both, or None) is correct for each question. The proportions of ratings 1~5 shows that every rating was used to a certain degree for each question because participants could choose to use the system or not in answering questions in the post-test depending on their level of confidence. Overall, the cross-analysis results shown in Table 4 shows that around 36% (536/1500) of questions in the post-test were answered without use of the system, while Table 5 (bottom row) indicates that this 36% is made up of 23% for rating 1 and 13% for rating 2. Although every rating was used to a certain degree for each question, several questions such as Q5, Q17, Q18, and Q47 received a high proportion of rating 1 + rating 2, indicating that a large proportion of participants felt confident enough to answer these questions without the system. In fact, the majority participants answered correctly on these questions without the system in both pre-test and post-test, thus yielding a higher ratio of C_{pre}/C_{post} for these questions. Overall, the cross-analysis results in Table 4 show that, for "Without_System", the proportion of C_{pre}/C_{post} is 59% (315/536), which confirms that participants are usually able to correctly answer questions for which they feel confident.

When using the system, both PMI and N-gram can provide useful contextual information for learning. As indicated in Method Correctness in Table 5, except for three questions marked "None" (i.e., Q3, Q9, and Q27), at least one method (PMI or N-gram) can provide correct contextual information for nearly every question. For example, for Q21, the correct method is N-gram which provided useful contextual information, as shown in Figure 7.

Q21: In the past year, the number of those contracted to be on _____ for more than 83 hours has halved, but we must go much further.
A. job B. task C. duty (Original word: duty)

PMI	job	task	duty
to	0.19	0.34	0.48
be	0.11	0.55	0.35
on	0.25	0.25	0.51
for	0.43	0.32	0.26
more	0.52	0.38	0.1
than	0.38	0.31	0.31
	1.88	2.15	2.01

Bi-gram	on job	278,109	on task	155,719	on duty	854,626
	job for	1,321,997	task for	605,355	duty for	260,154
Tri-gram	be on job	566	be on task	1,334	be on duty	40,627
	on job for	478	on task for	2,803	on duty for	22,688
	job for more	34,160	task for more	1,176	duty for more	6,756
4-gram	to be on job	146	to be on task	705	to be on duty	13,070
	be on job for	0	be on task for	0	be on duty for	3,112
	on job for more	0	on task for more	0	on duty for more	1,273
	job for more than	6,933	task for more than	578	duty for more than	6,311
5-gram	to be on job for	0	to be on task for	0	to be on duty for	1,155
	be on job for more	0	be on task for more	0	be on duty for more	223
	on job for more than	0	on task for more than	0	on duty for more than	1,171

Figure 7. Contextual information provided by PMI and N-gram for Q21.

Through the provided contextual information, participants can learn useful *n*-grams such as "on duty" and "to be on duty" to discriminate between *duty*, *job* and *task* in the context of this question sentence. The results shown in Table 5 indicate that around 73% of participants answered Q21 using the system, and most of them (77%) used N-gram, which is the correct method for this question. Therefore, this question had a high proportion of correct answers in the post-test (i.e., $C_{pre}/C_{post} + \overline{C}_{pre}/C_{post}$) because most participants had learned useful contextual information from the correct method N-gram. This allowed learners who answered incorrectly in the pre-test to offer the correct answer in the post-test, while those who answered correctly in the pre-test can also offer the correct answer by double-checking with the system. In addition to Q21, using the correct methods provides participants the opportunity to learn with most other questions, thus yielding a high proportion of $C_{pre}/C_{post} + \overline{C}_{pre}/C_{post}$. The incorrect method may provide ambiguous contextual information for a given question, and participants learned from the incorrect methods for only for a few questions (e.g., Q6, Q19, Q38 and Q41), thus yielding a low proportion of $C_{pre}/C_{post} + \overline{C}_{pre}/C_{post}$ (or a high proportion of $C_{pre}/\overline{C}_{post} + \overline{C}_{pre}/\overline{C}_{post}$). Overall, the cross-analysis results in Table 4 shows that, of the 64% (964/1500) of questions answered using the system, the respective proportions of C_{pre}/C_{post} and $\overline{C}_{pre}/C_{post}$ were 46% and 25%, while those of $C_{pre}/\overline{C}_{post}$ and $\overline{C}_{pre}/\overline{C}_{post}$ were 8% and 20%, yielding a correct/incorrect ratio of 72:28 in the post-test. Conversely, as indicated in Table 3, the average percentage of correctly answered questions in the pre-test was just 44%, and even the strongest participant only scored 66% on the pre-test.

CONCLUSIONS

This study developed a computer-assisted near-synonym learning system using two automatic near-synonym choice techniques: PMI and N-gram, which can capture the respective individual and contiguous relationship between near-synonyms and their context words. User evaluations of the system show that learning such useful and different aspects of contextual information can improve learners' ability to distinguish different usages of various near-synonyms in a range of contexts, and to use them appropriately. While participants had a preference for PMI, N-gram can provide more precise information. Future work will be devoted to enhancing the system by including more near-synonym sets and incorporating other useful contextual information provided by supervised learning methods such as latent semantic analysis (LSA) (Liao, Kuo, & Pai, 2012).

ACKNOWLEDGMENT
This work was partially supported by Aim for the Top University Plan, Ministry of Education, Taiwan, R.O.C. and National Science Council, Taiwan, R.O.C (NSC99-2221-E-155-036-MY3 and NSC100-2632-S-155-001).

REFERENCES
Bhogal, J., Macfarlane, A., & Smith, P. (2007). A review of ontology based query expansion. *Information Processing & Management*, 43(4), 866–886.
Cheng, C.-C. (2004). Word-focused extensive reading with guidance. *Proceedings of the 13th International Symposium on English Teaching*, 24–32.
Church, K., & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1), 22–29.
Dong, Z., & Dong, Q. (2006). *HowNet and the Computation of Meaning*. New Jersey: World Scientific Publishing.
Edmonds, P. (1997). Choosing the word most typical in context using a lexical co-occurrence network. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, 507–509.
Fehr, C. N., Davison, M. L., Graves, M. F., Sales, G. C., Seipel, B., Sekhran-Sharma, S. (2012). The effects of individualized, online vocabulary instruction on picture vocabulary scores: an efficacy study. *Computer Assisted Language Learning*, 25(1), 87–102.
Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
Gardiner, M., & Dras, M. (2007). Exploring approaches to discriminating among near-synonyms. *Proceedings of the Australasian Language Technology Workshop (ALTA-07)*, 31–39.
Huang, C.-R., Hsieh, S.-K., Hong, J.-F., Chen, Y.-Z., Su, I.-L., Chen, Y.-X., & Huang, S.-W. (2008). Chinese Wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Proceedings of the 9th Chinese Lexical Semantics Workshop*.
Huang, Y.-M., Huang, Y.-M., Huang, S.-H., & Lin, Y.-T. (2012). A ubiquitous English vocabulary learning system: Evidence of active/passive attitudes vs. usefulness/ease-of-use. *Computers & Education*, 58(1), 273–282.
Inkpen, D. (2007). A statistical model of near-synonym choice. *ACM Trans. Speech and Language Processing*, 4(1), 1–17.
Inkpen, D., & Hirst, G. (2006). Building and using a lexical knowledge-base of near-synonym differences. *Computational Linguistics*, 32(2), 1–39.
Islam, A., & Inkpen, D. (2010). Near-synonym choice using a 5-gram language model. *Research in Computing Science: Special issue on Natural Language Processing and its Applications*, 46, 41–52.
Jia, J., Chen, Y., Ding, Z., Ruan, M. (2012). Effects of a vocabulary acquisition and assessment system on students' performance in a blended learning class for English subject. *Computers & Education*, 58(1), 67–76.
Kayaoglu, M. N., Dagakbas, R., Ozturk, Z. (2011). A small scale experimental study: using animations to learn vocabulary. *The Turkish Online Journal of Educational Technology*, 10(2), 24–30.
Kim, D., & Gilman, D. A. (2008). Effects of Text, Audio, and Graphic Aids in Multimedia Instruction for Vocabulary Learning. *Educational Technology & Society*, 11(3), 114–126.
Liao, C.-H., Kuo, B.-C., Pai, K.-C. (2012). Effectiveness of automated Chinese sentence scoring with latent semantic analysis. *The Turkish Online Journal of Educational Technology*, 11(2), 80–87.
Lin, C.-C., Chan, H.-J., Hsiao, H.-S. (2011). EFL students' perceptions of learning vocabulary in a computer-supported collaborative environment. *The Turkish Online Journal of Educational Technology*, 10(2), 91–99.
Moldovan, D., & Mihalcea, R. (2000). Using Wordnet and lexical operators to improve internet searches. *IEEE Internet Computing*, 4(1), 34–43.
Navigli, R., & Velardi, P. (2003). An analysis of ontology-based query expansion strategies. *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM-03)*.
Neville, D. O., Shelton, B. E., & McInnis, B. (2009). Cybertext redux: using digital game-based learning to teach L2 vocabulary, reading, and culture. *Computer Assisted Language Learning*, 22(5), 409–424.
Ouyang, S., Gao, H.-H., & Koh, S.-N. (2009). Developing a computer-facilitated tool for acquiring near-synonyms in Chinese and English. *Proceedings of the 8th International Conference on Computational Semantics (IWCS-8)*, 316–319.
Pearce, D. (2001). Synonymy in collocation extraction. *Proceedings of the Workshop on WordNet and Other Lexical Resources*.
Rodríguez, H. Climent, S. Vossen, P. Bloksma, L. Peters, W. Alonge, A. Bertagna, F. & Roventint, A. (1998). The top-down strategy for building EuroWordNet: vocabulary coverage, base concepts and top ontology, *Computers and the Humanities*, 32, 117–159.

Shlrl, A., & Revle, C. (2006). Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. *Journal of the American Society for Information Science and Technology*, 57(4), 462–478.

Sun, K.-T., Huang, Y.-M., & Liu, M.-C. (2011). A WordNet-based near-Synonyms and similar-looking word learning system. *Educational Technology & Society*, 14(1), 121–134.

Vanderwende, L., Suzuki, H., Brockett, C., & Nenkova, A. (2007). Beyond SumBasic: task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6), 1606–1618.

Wang, T., & Hirst, G. (2010). Near-synonym lexical choice in latent semantic space. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-10)*, 1182–1190.

Wible, D., Kuo, C.-H., Tsao, N.-L., Liu, A., & Lin, H.-L. (2003). Bootstrapping in a language learning environment. *Journal of Computer Assisted Learning*, 19(1), 90–102.

Wu, C.-H. Liu, C.-H., Matthew, H., & Yu, L.-C. (2010). Sentence correction incorporating relative position and parse template language models. *IEEE Trans. Audio, Speech and Language Processing*, 18(6), 1170–1181.

Yu, L.-C., Chien, W.-N., & Chen, S.-T. (2011). A baseline system for Chinese near-synonym choice. *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-11)*, 1366–1370.

Yu, L.-C., Wu, C.-H., Chang, R.-Y., Liu, C.-H., & Hovy, E. H. (2010). Annotation and verification of sense pools in OntoNotes. *Information Processing & Management*, 46(4), 436–447.

Yu, L.-C., Wu, C.-H., Jang, F.-L. (2009). Psychiatric document retrieval using a discourse-aware model. *Artificial Intelligence*, 173(7-8), 817–829.

Table 5. Statistics of the 50 test questions.

No.	without system		with system				C_{pre}/C_{post}	$C_{pre}/\overline{C}_{post}$	$\overline{C}_{pre}/C_{post}$	$\overline{C}_{pre}/\overline{C}_{post}$	Method		
	1	2	3	4	5	Avg.					PMI	N-gram	Correctness
Q1	0.40	0.03	0.07	0.27	0.23	2.90	0.63	0.03	0.33	0.00	0.47	0.53	Both
Q2	0.07	0.07	0.10	0.30	0.47	4.03	0.17	0.00	0.80	0.03	0.41	0.59	Both
Q3	0.20	0.10	0.13	0.17	0.40	3.47	0.10	0.27	0.07	0.57	0.96	0.04	None
Q4	0.27	0.10	0.10	0.17	0.37	3.27	0.60	0.03	0.33	0.03	1.00	0.00	PMI
Q5	0.47	0.20	0.03	0.13	0.17	2.33	0.63	0.03	0.23	0.10	0.50	0.50	Both
Q6	0.20	0.20	0.13	0.20	0.27	3.13	0.17	0.23	0.10	0.50	0.74	0.26	N-gram
Q7	0.37	0.10	0.17	0.13	0.23	2.77	0.27	0.20	0.13	0.40	0.50	0.50	PMI
Q8	0.07	0.10	0.07	0.03	0.73	4.27	0.50	0.00	0.50	0.00	0.44	0.56	Both
Q9	0.23	0.17	0.10	0.13	0.37	3.23	0.13	0.37	0.03	0.47	1.00	0.00	None
Q10	0.33	0.07	0.03	0.30	0.27	3.10	0.40	0.00	0.60	0.00	0.55	0.45	Both
Q11	0.37	0.10	0.17	0.07	0.30	2.83	0.80	0.00	0.20	0.00	0.32	0.68	Both
Q12	0.27	0.17	0.13	0.20	0.23	2.97	0.50	0.03	0.13	0.33	0.64	0.36	PMI
Q13	0.20	0.20	0.20	0.07	0.33	3.13	0.60	0.13	0.17	0.10	0.29	0.71	N-gram
Q14	0.27	0.10	0.07	0.37	0.20	3.13	0.27	0.00	0.30	0.43	0.43	0.57	N-gram
Q15	0.30	0.07	0.07	0.10	0.47	3.37	0.50	0.03	0.47	0.00	0.76	0.24	Both
Q16	0.30	0.03	0.07	0.17	0.43	3.40	0.60	0.03	0.37	0.00	0.95	0.05	PMI
Q17	0.40	0.20	0.13	0.07	0.20	2.47	0.63	0.03	0.13	0.20	0.12	0.88	N-gram
Q18	0.50	0.03	0.07	0.13	0.27	2.63	0.73	0.00	0.23	0.03	0.41	0.59	Both
Q19	0.13	0.20	0.20	0.13	0.33	3.33	0.13	0.13	0.10	0.63	0.71	0.29	N-gram
Q20	0.30	0.07	0.03	0.23	0.37	3.30	0.33	0.00	0.60	0.07	0.61	0.39	Both
Q21	0.13	0.13	0.10	0.30	0.33	3.57	0.17	0.07	0.53	0.23	0.23	0.77	N-gram
Q22	0.33	0.10	0.10	0.10	0.37	3.07	0.70	0.00	0.27	0.03	0.47	0.53	Both
Q23	0.13	0.13	0.07	0.23	0.43	3.70	0.23	0.03	0.37	0.37	0.54	0.46	PMI
Q24	0.20	0.17	0.10	0.20	0.33	3.30	0.13	0.03	0.47	0.37	0.78	0.22	PMI
Q25	0.20	0.03	0.10	0.23	0.43	3.67	0.43	0.00	0.57	0.00	0.32	0.68	Both
No.	without system		with system				C_{pre}/C_{post}	$C_{pre}/\overline{C}_{post}$	$\overline{C}_{pre}/C_{post}$	$\overline{C}_{pre}/\overline{C}_{post}$	Method		
	1	2	3	4	5	Avg.					PMI	N-gram	Correctness
Q26	0.13	0.10	0.10	0.10	0.57	3.87	0.23	0.03	0.47	0.27	0.23	0.77	N-gram
Q27	0.17	0.23	0.07	0.17	0.37	3.33	0.00	0.10	0.00	0.90	0.60	0.40	None
Q28	0.17	0.17	0.10	0.20	0.37	3.43	0.23	0.13	0.20	0.43	0.60	0.40	N-gram
Q29	0.07	0.17	0.10	0.20	0.47	3.83	0.17	0.23	0.20	0.40	0.24	0.76	N-gram
Q30	0.03	0.07	0.03	0.23	0.63	4.37	0.30	0.00	0.70	0.00	0.74	0.26	Both
Q31	0.27	0.03	0.00	0.20	0.50	3.63	0.37	0.00	0.57	0.07	0.83	0.17	Both
Q32	0.20	0.13	0.10	0.20	0.37	3.40	0.37	0.00	0.40	0.23	0.76	0.24	PMI
Q33	0.10	0.20	0.03	0.20	0.47	3.73	0.37	0.00	0.63	0.00	0.62	0.38	Both
Q34	0.17	0.13	0.03	0.20	0.47	3.67	0.30	0.07	0.30	0.33	0.40	0.60	N-gram
Q35	0.17	0.07	0.07	0.23	0.47	3.77	0.30	0.00	0.67	0.03	0.21	0.79	Both
Q36	0.17	0.10	0.07	0.10	0.57	3.80	0.50	0.00	0.50	0.00	0.39	0.61	Both
Q37	0.30	0.20	0.07	0.20	0.23	2.87	0.50	0.00	0.50	0.00	1.00	0.00	Both
Q38	0.27	0.17	0.10	0.20	0.27	3.03	0.27	0.20	0.00	0.53	0.90	0.10	N-gram
Q39	0.20	0.10	0.03	0.17	0.50	3.67	0.30	0.03	0.53	0.13	1.00	0.00	PMI
Q40	0.30	0.10	0.13	0.20	0.27	3.03	0.30	0.13	0.07	0.50	0.55	0.45	N-gram
Q41	0.20	0.13	0.10	0.30	0.27	3.30	0.10	0.00	0.27	0.63	0.38	0.63	PMI
Q42	0.20	0.17	0.03	0.30	0.30	3.33	0.37	0.03	0.47	0.13	0.58	0.42	PMI
Q43	0.13	0.20	0.17	0.30	0.20	3.23	0.23	0.07	0.27	0.43	0.46	0.54	PMI
Q44	0.27	0.03	0.07	0.40	0.23	3.30	0.23	0.27	0.33	0.17	0.26	0.74	N-gram
Q45	0.13	0.17	0.23	0.17	0.30	3.33	0.40	0.03	0.13	0.43	0.46	0.54	PMI
Q46	0.17	0.13	0.17	0.23	0.30	3.37	0.30	0.13	0.13	0.43	0.50	0.50	N-gram

Q47	0.47	0.13	0.03	0.10	0.27	2.57	0.77	0.00	0.20	0.03	0.58	0.42	Both
Q48	0.23	0.10	0.13	0.17	0.37	3.33	0.30	0.00	0.57	0.13	1.00	0.00	PMI
Q49	0.13	0.27	0.17	0.20	0.23	3.13	0.50	0.13	0.00	0.37	0.56	0.44	N-gram
Q50	0.20	0.23	0.13	0.03	0.40	3.20	0.57	0.00	0.43	0.00	1.00	0.00	Both
Avg.	**0.23**	**0.13**	**0.10**	**0.19**	**0.36**	**3.32**	**0.37**	**0.07**	**0.33**	**0.23**	**0.58**	**0.42**	—