2012 Conference on Technologies and Applications of Artificial Intelligence

A Learning-based Framework to Utilize E-HowNet Ontology and Wikipedia Sources to Generate Multiple-Choice Factual Questions

Min-Huang Chu[1], Wen-Yu Chen[2], Shou-De Lin[3]
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
r96943077@ntu.edu.tw[1], fifi9714@gmail.com[2], sdlin@csie.ntu.edu.tw[3]

Abstract—**This paper proposes a framework that automatically generates multiple-choice questions. Unlike most other similar works that focus on generating questions for English proficiency tests, this paper provides a framework to generate factual questions in Chinese. We have decomposed this problem into several sub-tasks: a) the identification of sentences that contain factual knowledge, b) the identification of the query term from each factual sentence, and c) the generation of distractors. Learning-based approaches are applied to address the first two problems. We then propose a way to generate distractors by using E-HowNet ontology database and Wikipedia sources. The system was evaluated through user study and test theory, and achieved a satisfaction rate of up to 70.6%.**

Keywords – multiple-choice questions, distractor, ontology, E-HowNet, Wikipedia

I. Introduction

Life-long learning is becoming a new trend. With the advancement of information technology, people can reach an abundance of Internet resources which allow them to learn anything in anywhere at any time. The popularity of E-learning brings about an important task which is to test whether the students have indeed grasped the concept he or she was supposed to learn. To enable such process, one would require some form of test questions to evaluate a learner.

Multiple-choice questions are one of the most popular ways of conducting tests. Students are usually asked to pick one correct answer out of typically 4 or 5 options. It is also a preferred type of test for E-learning because the grading is trivial and deterministic. The process of generating multiple-choice questions, however, is time-consuming and labor-intensive. Mitkov had demonstrated that computers can help teachers to generate multiple-choice questions [6]. Although most of their output results need to be modified by humans, it shows the total used time of generating questions by computers with slight human modification is significantly lower than generating questions by only humans.

In this paper, we aim to build a multiple-choice question generation system (assuming only one correct answer) that extracts factual knowledge[1] from sentence level corpuses to generate questions automatically. To do that, we need to accomplish several sub-tasks. First, we need to extract factual sentences from a corpus or from the Web. Second, we need to know which section of the sentence can be identified as the answer to this question. Third, we need to generate some wrong answers, which are called distractors, to be chosen for multiple-choice question generation.

To tackle the first sub-task, we extract candidate sentences from the Web or a corpus, and then train a sentence classifier to identify whether a sentence contains factual knowledge. As every multiple-choice question requires a target answer, we then train a classifier to decide which noun phrase in the sentence can be the answer to be provided. Next, using E-HowNet and Wikipedia sources, we can generate distractors as the candidates to be chosen in a multiple-choice question. We subsequently make use of a search engine to help filter out improper distractor candidates. Finally, we use simple rules to transform the selected sentences into questions. The following is an example question generated by our system.

________ is called a living fossil.
(1) Mandarin orange
(2) Chinese pear
(3) Purple grapes
(4) Ginkgo

We evaluate our system through user study and test theory. We invite people to answer questions and evaluate if they are suitable for testing whether people have related knowledge or not. The results show an overall satisfaction rate of 70.6%.

The main contribution of this paper is that this is the first-ever system to our knowledge that is capable of performing such tasks in a fully automatic way.

II. Related Work

A. Question generation system

Nielsen's work provided us with an overview of a variety of kinds of questions for testing [2]. Many different types of questions have been studied such as sentence reconstruction [4], vocabulary assessment [7], cloze test [3] and reading comprehension [5]. In this paper, we aim to generate cloze style multiple-choice questions.

[1] A sentence is considered containing factual knowledge if it has factual information which can be used to test whether people have relevant knowledge.

978-0-7695-4919-4/12 $26.00 © 2012 IEEE
DOI 10.1109/TAAI.2012.38

Mitkov's work is the first study to focus on the automatic generation of multiple-choice questions [6]. Unlike our proposal, they applied rules that derived from natural language processing techniques without employing machine learning models to generate questions. In addition, most of Mitkov's output questions need post-modification, while we aim to building an automatic framework to achieve such goal.

Goto et al used preference learning and conditional random field (CRF) to train their model to create a cloze test to evaluate learners' English proficiency [3]. They used TOEIC workbooks as inputs and use statistical methods to generate some distractors. Many other related works also focus on generating proper English proficiency test questions [3, 7, 12] rather than factual knowledge questions. Since we are not generating questions to test the subjects' language proficiency, we do not rely on grammatical rules to generate distractors. Instead, we use common-sense ontology database and make use of the semantic relationship among concepts to generate distractors.

, Heilman extracted simplified statements from complex sentences with factual information to generate factual questions [1]. Their work, however, assumes every input sentence contains factual information, which can be problematic at times. Distinctive from their models, we train a classifier to distinguish which sentences contain factual knowledge and are suitable to become cloze style multiple-choice questions. Agarwal aimed to select informative sentences from text without using external resources [11]. The concern is that they rely on only syntactic and lexical features and did not consider semantic relationship or semantic similarity like we do through a semantic network.

B. *Ontology database*

Ontology databases are defined to record the relationship between different entities and meanings of different words or concepts with the goal to enable computers to understand human language or behavior. In this paper, we used ontology to find the similar but not identical concepts as distractors.

WordNet is a highly-exploited English ontology database, and has been widely used in the question generation system [3, 6]. Papasalouros studied how to use web-based ontology standards to generate distractors by rules between classes, properties and terminologies [13]. In this paper, we used E-HowNet ontology [9] as our database, which is similar to WordNet but defined in Chinese. Since languages evolve over time, it is impossible for us to find all concepts through an ontological database. Here we tried to overcome such limitation by the use of Wikipedia.

III. Methodology

A. *Overview*

Our goal is to generate factual questions from Chinese sentences automatically. We transformed selected sentences into cloze style multiple-choice questions, each of which contains only one correct answer out of four possible candidates. There is a blank in each question for test takers to choose the most suitable answer to fill in.

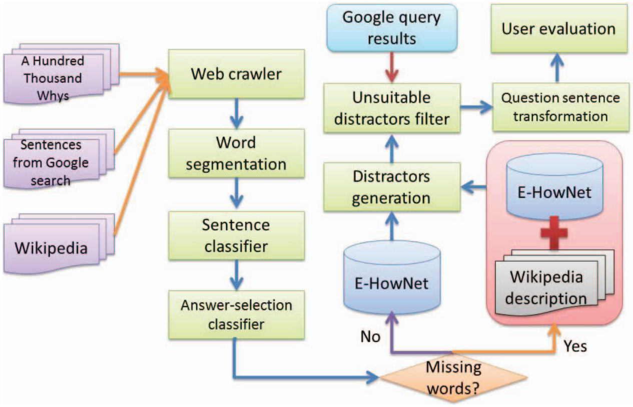


Figure 1. System diagram of automatic multiple-choice questions generation system.

Not all sentences are suitable for transformation into factual questions. We focus on selecting sentences containing factual knowledge. It is generally non-trivial to formulate rules to decide which kinds of sentences contain factual knowledge. Our solution is to train a classifier to identify which sentences contain factual knowledge and are eligible for transformation into questions. To transform a factual sentence into a question, we need to determine which part of the sentence should become the query to be asked. As there could be multiple noun-phrases to serve as the query, here we train another classifier to identify which noun phrases are suitable to be an answer. Finally, we need some distracting candidates to be the distractors for multiple-choice questions. To generate distractors with semantic relationship to the answers, we applied E-HowNet ontology and combine Wikipedia sources to deal with unseen terms in E-HowNet ontology. We further make use of a search engine to filter distractors with high possibility to be correct answers. The distractors which generate high search counts with the factual sentence might not be ideal because they could indeed be the correct answer. Finally, we transform the sentences into questions using some simple rules.

Fig. 1 shows the system diagram of the automatic multiple-choice questions generation system. We describe details of each component in the following sections.

B. *Corpus*

The corpus to generate candidate sentences is collected from three different sources.

1) *十萬個為什麼 (A Hundred Thousand Whys)*

We observed that sentences begin with *why* are more likely to contain factual knowledge. Therefore, we extracted questions from *A Hundred Thousand Whys*. Notice that not all questions here begin with *why*.

2) Sentences from search engine outputs

We used query terms "why+noun" such as *why fish* or *why earth* to retrieve sentences such as *why fish lay so many eggs* from search results.

3) Sentences from Wikipedia description

To extend beyond the limit of question sentences, we collected declaration sentences from Wikipedia. We selected noun words from ontology database and retrieve descriptions from Wikipedia. We only extract the first sentence of the description, because it is usually the definition of the given query term. We also perform segmentation and POS tagging on the sentences.

C. Sentence classifier

To train a classifier, we manually annotate 2100 positive sentences (i.e. sentence that contains factual knowledge) and 2551 negative ones. Our goal is to choose sentences contain factual knowledge and to transform them into questions. In other words, sentences contain subject opinions is not preferable. Below we first show some examples of positive and negative sentences:

1) Sentences labeling guidelines

a) Why is Amazon River the biggest river in the world?

b) Why are Taiwan black bears going to extinct?

c) Why do some buildings have strange appearance?

d) How to train your muscle?

Sentence (a) contains factual knowledge that *Amazon River is the biggest river in the world*. Sentence (b) contains factual knowledge that *Taiwan black bears are going to extinct*. Both (a) and (b) are considered as positive instances. In sentence (c), we don't know what kinds of buildings are specified here. Moreover, *strange* is a subjective opinion. In sentence (d), it is a simple question without any factual knowledge. Both (c) and (d) are considered as negative instances.

2) Feature extraction

We employ a supervised learning model to classify sentences. To learn such model, we need to first extract meaningful features. Each sentence is regarded as an instance, and we extract the following features to learn a classifier. We used LIBSVM as the tool for learning [15].

a) Does the sentence has "N + Vt + N" structure?

b) Does the sentence contain "是" (be) verbs?

Features (a) and (b) are informative signs to indicate whether a sentence mentions certain specific things. If so, it tends to contain factual information.

c) Does the sentence contain question words?

They are *how*, *which*, *what*, *where* and some interrogative auxiliary words in the end of a sentence such as 嗎 and 呢 in Chinese. They tend to imply no specific declarative information in a sentence, and are indicators for negative instances.

d) How many collective words does a sentence contain?

Here we consider words such as *children*, *student*, *animals*, *plant*, *country* and *people* as collective words. They usually indicate that the subject of a sentence is not particularly specified, so the sentence tends to lack precise factual information.

e) How many pronoun words does a sentence have?

Words such as *we*, *they* and *you* usually appear in a sentence that does not have factual information.

f) How many subjective terms does a sentence contain?

We can use the semantic relationship defined in E-HowNet ontology to retrieve subjective terms such as *wonderful*, *disgust*, *like* and *spectacular*. Because those terms usually have strong subjective opinion, and therefore tend to indicate a sentence without factual knowledge.

g) Sentence length.

h) Number of segmentations of each sentence.

i) Number of verbs of each sentence.

j) Number of nouns of each sentence.

k) Tree depth.

l) Number of nodes.

m) Number of branches of IP node (not including PU).

We also extract structure features to catch the characteristics of a sentence. Features (h), (i) and (j) are generated by the segmentation and POS tagging tool[2]. Feature (k), (l) and (m) are generated by the Stanford parser[3].

D. Answer-selection classifier

In a cloze test, the main concept of a sentence is chosen to be filled, which is considered as the answer. For each sentence, we want to capture the main concept or proper noun phrase for the answer. Notice that each sentence might contain more than one candidate as the answer for cloze test. For example, *why is the polar bear the king of animals in North Pole?* In the sentence, *polar bear* and *North Pole* are both suitable candidates. That is, each sentence can potentially be transformed into more than one question (e.g. which is the king of animals in North Pole? or polar bear is the king of animals in where?). Therefore, we designed a classifier to identify whether a noun phrase is suitable to become the missing piece of the puzzle. Here we train the classifier using LIBSVM. We treat each noun phrase as an instance and make use of the following features.

1) The order of this none phrase among all noun phrases in the sentence.

2) The order of this noun phrase among all words in the sentence.

Noun phrases appear earlier tend to have more important concept.

[2] Academia Sinica: A Chinese Word Segmentation System with Unknown Word Extraction and Pos Tagging. http://ckipsvr.iis.sinica.edu.tw/

[3] http://nlp.stanford.edu/software/lex-parser.shtml

3) Taxonomy depth in E-HowNet ontology. Noun phrases with deeper depth are usually considered to be defined more precisely, and can be a better candidate.

*4) Term frequency in E-HowNet database.*Noun phrases with high frequency tend to be not as specific, and therefore is less likely to be the candidate answer.

5) Tf-idf value in the corpus. Tf-idf value usually represents the importance of a noun pharse.

E. Distractors generation

The distractors are the wrong answers in a multiple-choice question. We aim to generate distractors that possess some semantic relationships with the correct answer, but are not exactly identical to the answers. In other words, ideal distractors should be similar to some extent to the gold standard, but not too similar to be considered as a valid answer. We design a strategy to extract such distractors using E-HowNet ontology database and Wikipedia. In the following paragraphs, we first introduce the E-HowNet ontology database before we show the process of generating distractors with E-HowNet and Wikipedia sources.

1) E-HowNet ontology database [9]

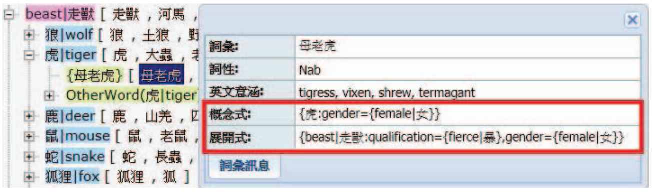
E-HowNet is a lexical knowledge database. It provides lexical definitions and hierarchical information of a common sense ontology. Fig. 2 is an example of querying "母老虎" (female tiger). First, we see "概念式" (definition) in E-HowNet and "展開式" (expansion) in E-HowNet. The definition is "虎" (tiger) and the first term in expansion, "Beast|走獸", is the primitive concept of the query term. The rest in the expansion are attributes of the query term.

2) Generating distractors from E-HowNet ontology database

If the missing answer is defined in E-HowNet ontology database, we then propose to retrieve the primitive concept from its expansion to generate distractors. We first search for words that have the same primitive concept from the database then we filter away words that have the same definition, the exchangeable words. The remaining words are considered as candidate distractors.

3) Clustering missing words via Wikipedia and E-HowNet ontology to generate distractors

If, unfortunately, the missing answer is not defined in E-HowNet ontology, we provide a workaround method to overcome the coverage limitation. First, we retrieve the description of all ontology words from Wikipedia. Second, assuming the description of the missing word can be found from Wikipedia, we can apply PLSA[4] [14] as an unsupervised clustering method to group the missing words with existing ontology words. Finally, we count which primitive concept appears most frequently with this missing concept, and assign this primitive concept as the expansion field of the missing word. Then we can apply the same approach described in previous section to generate other distractors. Fig. 3 shows the clustering process of the missing word – "珠海市" (Zhuhai City) – with Wikipedia sources.



beast|走獸 [走獸 , 河馬 ,
狼|wolf [狼 , 土狼 , 野
虎|tiger [虎 , 大蟲 , 老
{母老虎} [母老虎 ,
OtherWord(虎|tiger
鹿|deer [鹿 , 山羌 , 四
鼠|mouse [鼠 , 老鼠 ,
蛇|snake [蛇 , 長蟲 ,
狐狸|fox [狐狸 , 狐]

詞彙:	母老虎
詞性:	Nab
英文意涵:	tigress, vixen, shrew, termagant
概念式:	{虎:gender={female|女}}
展開式:	{beast|走獸:qualification={fierce|暴},gender={female|女}}

Figure 2. The result of querying "母老虎" (female tiger) in E-HowNet ontology.

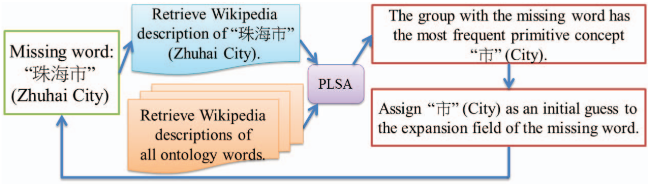

Missing word: "珠海市" (Zhuhai City)
Retrieve Wikipedia description of "珠海市" (Zhuhai City).
Retrieve Wikipedia descriptions of all ontology words.
PLSA
The group with the missing word has the most frequent primitive concept "市" (City).
Assign "市" (City) as an initial guess to the expansion field of the missing word.

Figure 3. Process of clustering missing words with Wikipedia sources.

F. Filtering unsuitable distractors

Since each question has only one correct answer, we should carefully select proper distractors that cannot be mistaken as the correct answers to the questions. Our approach is based on a simple hypothesis: if the candidate tends to be true, then plugging it into to the sentence and use this sentence as a query to a search engine would generate more returns.

Based on this assumption, we used simple rules to determine whether a distractor is proper or not: if the returning search count of a sentence filled in by a distractor is larger than a threshold (we use 50,000 in our experiments), or this count is not much less (exceed 10%) than that of the original sentence, the candidate distractor is discarded.

G. Question sentence transformation

Since most of the input sentences are question sentences, they may contain words such as *which*, *why* or *what*. We used a simple rule-based transformation to transform those sentences into multiple-choice questions. We deleted those *wh*-terms and interrogative auxiliary terms, and added transition words to increase readability. We finished the task by replacing the answer word with "____".

IV. EXPERIMENTS

First, we evaluated sentence classifier and answer-selection classifier using standard cross validation and learning curve. We then randomly generated three test papers (69 questions) and invited people to answer and evaluate the quality of the questions. In addition, we analyzed each of them through classical test theory, which can provide information about question difficulty, discriminating power and the usefulness of distractors.

[4] http://chasen.org/~taku/software/plsi/

A. *Evaluating the components*

We first evaluated the supervised learning models for factual sentence detection and answer generation. We chose to focus on precision because it is a more critical factor for question generation as we don't want to produce bad questions.

1) Sentence classifier

We did k-fold cross validation for the entire corpus. There are a total of 2,100 positive sentences and 2,551 negative ones we annotated for this experiment. Table I shows cross validation results of different k with stable 71%~72% precision. Fig. 4 shows the learning curve.

2) Answer-selection classifier

We randomly selected 395 sentences from corpus to label the position of the answer. Notice that each sentence could have more than one potential answer, so there is a total of 1,679 noun phrase candidates. We labeled 692 out of 1,679 instances as positive. Table II and Fig. 5 show the k-fold cross validation results and the learning curve respectively. Here we have achieved 81~82% precision.

B. *User study*

1) User evaluation

We constructed an evaluation system [5] and invited students from university level or above as testers to evaluate the generated questions. Fig. 6 shows a multiple-choice question and questionnaires for users. In the system, testers were asked to answer each question and respond to two assessment questionnaires for each question. We experimented on a total of 69 questions generated by the system. Table III shows the results of survey. Survey questions are described as follows.

a) Do you think the question have only one correct answer among the four choices? (Only one answer: Yes/No/Unknown)

The feedback shows that 70.7% are believed to be correct while only 4.3% are confirmed to be wrong.

b) Do you think the question is suitable to evaluate whether people possess corresponding knowledge or not? (Quality: Good/Normal/Poor)

The sum of the described quality of good and normal shows an overall acceptable quality rate 70.6%.

2) Classical test theory

Here we used simplified measurement as [6] to analyze the functionality of each question and its distractors. First, we defined two groups. One is called upper group C_U, which includes students who receive the highest scores on the top of one third, and the other one is called lower group C_L, which includes students who receive the lowest scores on the bottom of one third. We evaluated the effective of a multiple-choice question based on the following metrics.

a) Item difficulty (ID)[6]

TABLE I. CROSS VALIDATION RESULTS OF SENTENCE CLASSIFIER

k-fold cross validation	Precision (%)
2	71.45%
3	72.21%
5	72.19%
10	71.83%

TABLE II. CROSS VALIDATION RESULTS OF ANSWER-SELECTION CLASSIFIER

k-fold cross validation	Precision (%)
2	78.4%
3	79.14%
5	79.21%
10	78.2%

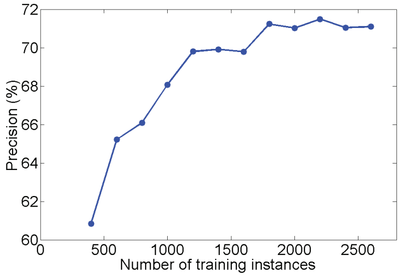


Figure 4. Learning curve of sentence classifier.

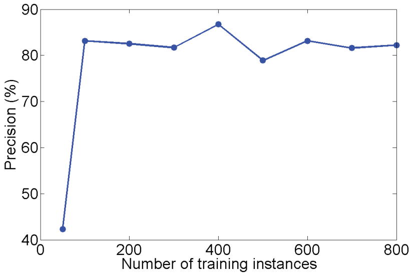


Figure 5. Learning curve of answer-selection classifier.

In test theory, a question is normally referred to as an item. The ID metric is calculated by number of students answered the item correctly divided by total number of students who attempted to answer. Table IV shows the item difficulty results of 69 items. It seems that 68% of the problems are considered as too easy.

b) Discriminating power (DP)[7]

Discriminating power is defined as (C_U – C_L) / (T/2), where T stands for total number of students. A good exam question should have the capability to distinguish testers' ability. Table V shows the DP of three test papers are 0.21, 0.24 and 0.14, which are lower than the maximum value 1.0. It indicates that the questions cannot distinguish testers' ability well.

[5] http://www.csie.ntu.edu.tw/~r96943077/ML_for_Testing/site/contest.php

[6] Here we use ID ≤ 0.15 to be too difficult and ID ≥ 0.85 to be too easy. We use the same measurement as [6].

[7] The maximum value of discriminating power of an item is 1.0. That would have item difficulty become 0.5. Thus, it is suggests test authors generate test papers have item difficulty at 0.5.

TABLE III. USER EVALUATION RESULTS

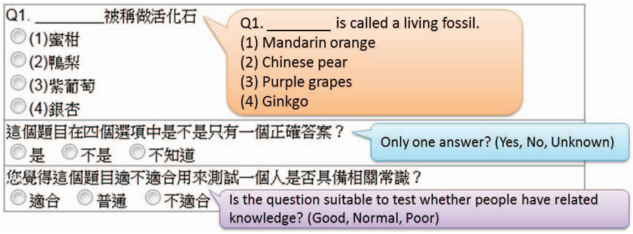
	Option	Average (%)
Does the question have only one answer?	*Yes*	70.7%
	No	4.3%
	Unknown	24.9%
Is the question suitable to test people have related knowledge?	*Good*	**43.6%**
	Normal	**27.0%**
	Poor	29.4%

TABLE IV. ITEM DIFFICULTY RESULTS

	Total items	*Avg. ID*	*Too easy*	*Too difficult*
Item difficulty	69	0.837	47	1

TABLE V. DISCRIMINATING POWER AND USEFULNESS OF DISTRACTORS

Test paper		*Paper 1*	*Paper 2*	*Paper 3*
Number of testers		27	27	30
Number of distractors in total		33	33	27
Avg. discriminating power		0.21	0.24	0.14
Usefulness of distractors	*Poor*	0	0	0
	Not useful	16	18	21



Q1. ________被稱做活化石
○(1)蜜柑
○(2)鴨梨
○(3)紫葡萄
○(4)銀杏

Q1. ________ is called a living fossil.
(1) Mandarin orange
(2) Chinese pear
(3) Purple grapes
(4) Ginkgo

這個題目在四個選項中是不是只有一個正確答案？
○是 ○不是 ○不知道

Only one answer? (Yes, No, Unknown)

您覺得這個題目適不適合用來測試一個人是否具備相關常識？
○適合 ○普通 ○不適合

Is the question suitable to test whether people have related knowledge? (Good, Normal, Poor)

Figure 6. Multiple-choice questions and user evaluation questionnaire.

c) *Usefulness of the distractors*

A useful distractor is supposed to attract students in lower group more than students in upper group. If the situation is reversed, we consider the distractor to be *poor*. If a distractor is not selected by any students, we consider the distractor to be *not useful*. Table V shows no poor distractors but it seems that many distractors don't attract any testers at all. For example, there are 33 distractors in test paper 1 and 16 of them proved to be not useful.

The results show that our questions seemed easy to most testers. The average discriminating power is therefore low and many distractors are considered as not useful. We believe one of the reasons is that most of the knowledge from *A Hundred Thousand Whys* was originally designed for teenagers. The other reason is that the thresholds we have set up to filter distractors are too strict and ad hoc, some of the candidates are not that popular, and therefore testers can eliminate those rare distractors to obtain the correct answer.

V. CONCLUSION

In this paper, we provided an automatic multiple-choice question generation framework that focuses on extracting factual knowledge in Chinese. By employing E-HowNet ontology and Wikipedia sources to generate suitable semantic distractors, we are capable of producing highly accurate and satisfactory questions.

ACKNOWLEDGEMENT

This work was supported by National Science Council, National Taiwan University and Intel Corporation under Grants NSC 100-2911-I-002-001, and 101R7501.

REFERENCES

[1] M. Heilman and N. A. Smith. Extracting Simplified Statements for Factual Question Generation. *In Proc. of the 3rd Workshop on Question Generation*. 2010.

[2] Rodney D. Nielsen, Jason Buckingham, Gary Knoll, Ben Marsh and Leysia Palen. A Taxonomy of Questions for Question Generation. *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge, Arlington, Virginia, September 25-26*, 2008.

[3] Goto, T., Kojiri, T., Watanabe, T., Iwata, T., Yamada, T. Automatic Generation System of Multiple-Choice Cloze Questions and its Evaluation. *Knowledge Management & E-Learning: An International Journal, Vol.2, No.3.* 2010.

[4] Huang, C.-B., Liu, C.-L., Kuo, W.-T., Sun, Y.-T., Lai, M.-H. Computer assisted test-item generation for sentence reconstruction. *The 21nd Conference on Computational Linguistics and Speech Processing (ROCLING)*. 2009.

[5] Huang, C.-S., Kuo, W.-T., Li, C.-L., Tsai, C.-C., Liu, C.-L. Using Linguistic Features to Classify Texts for Reading Comprehension Tests at the High School Levels. *Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING).* 2010.

[6] Mitkov, R. and Le An Ha. Computer-Aided Generation of Multiple-Choice Tests. *HLT-NAACL-EDUC '03 Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing.* 2003.

[7] Jonathan C. Brown, Gwen A Frishkoff and Maxine Eskenazi. Automatic question generation for vocabulary assessment. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 819–82.* 2005.

[8] Ming-Hsiung Ying and Heng-Li Yang. Computer-Aided Generation of Item Banks Based on Ontology and Bloom's Taxonomy. 2008

[9] Huang, Shu-Ling, You-Shan Chung, and Keh-Jiann Chen, "E-HowNet: the Expansion of HowNet," *Proceedings of the First National HowNet workshop, pages 10-22*, 2008.

[10] Roger Levy and Christopher D. Manning. "Is it harder to parse Chinese, or the Chinese Treebank?" *ACL 2003, pp. 439-446.*

[11] Agarwal, M. and Mannem, P. Automatic Gap-fill Question Generation from Text Books. *Proc. of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*. 2011.

[12] Hoshino, A. and Nakagawa, H.. A real-time multiple-choice question generation for language testing: a preliminary study. *Proc. EdAppsNLP 05 Proceedings of the second workshop on Building Educational Applications Using NLP Pages 17-20*, 2005.

[13] Papasalouros, A., Kanaris, K., Kotis, K. Automatic generation of multiple-choice questions from domain ontologies. *International Conference e-Learning.* 2008.

[14] Hofmann, Thomas. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning Volume 42*. 2001.

[15] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.