Automatic Chinese Factual Question Generation

Ming Liu, Vasile Rus, and Li Liu

Abstract—Question generation is an emerging research area of artificial intelligence in education. Question authoring tools are important in educational technologies, e.g., intelligent tutoring systems, as well as in dialogue systems. Approaches to generate factual questions, i.e., questions that have concrete answers, mainly make use of the syntactical and semantic information in a declarative sentence, which is then transformed into questions. Recently, some research has been conducted to investigate Chinese factual question generation with some limited success. Reported performance is poor due to unavoidable errors (e.g., sentence parsing, name entity recognition, and rule-based question transformation errors) and the complexity of long Chinese sentences. This article introduces a novel Chinese question generation system based on three stages, sentence simplification, question generation and ranking, to address the challenge of automatically generating factual questions in Chinese. The proposed approach and system have been evaluated on sentences from the *New Practical Chinese Reader corpus*. Experimental results show that ranking improves more than 20 percentage of questions rated as acceptable by annotators, from 65 percent of all questions to 87 percent of the top ranked 25 percent questions.

Index Terms—Educational technology, natural language processing, authoring system

1 Introduction

Questions are useful to support reflection and improve learning. Factual questions are questions whose answers are specific facts; typical examples of factual questions are of type who?, what?, where?, and when? [1]. Many studies [2], [3] showed that factual questions are more useful for elementary school students to learn. For example, an elementary school teacher might ask his or her students basic factual questions while they are still learning to read as a way to stimulate their processing of a story they might have just read and also to assess their reading skills based on which appropriate pedagogical strategies could be triggered. It should be noted that asking questions by itself is a good pedagogical strategy to be used by teachers in order to model for their students how to ask questions. Such a pedagogically strategy is highly recommended for teachers to use as there is plenty of evidence from many studies that have shown that students have problems recognizing their own knowledge deficits [4] and ask very few questions during instruction [5]. Furthermore, there is an acute need for questions for developing advanced educational technologies such as intelligent tutoring systems [1]. Last but not least, questions are key elements in assessment instruments such as reading comprehension tests. The authoring of good questions by human experts to be included in manuals for teachers, computer tutors, or reading comprehension assessment instruments, is an expensive, time consuming and effortful process. We propose here a method to automate the process.

- *M. Liu is with the School of Computer and Information Science, Southwest University, Chongqing 400716, P.R. China. E-mail: mingliu@swu.edu.cn.*
- *V. Rus is with the Department of Computer Science, University of Memphis, Memphis, TN 38152. E-mail: vrus@memphis.edu.*
- *L. Liu is with the School of Software Engineering, Chongqing University, Chongqing 400044, P.R. China. E-mail: dcsliuli@cqu.edu.cn.*

Manuscript received 27 Nov. 2015; revised 25 Apr. 2016; accepted 3 May 2016. Date of publication 10 May 2016; date of current version 16 June 2017. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TLT.2016.2565477

Our proposed approach to automatically generate questions will have an impact on the broad area of intelligent authoring tools which are needed in order to help teachers save time and effort for creating pedagogical content and to assist educational technology developers reduce development costs. To emphasize this point, we illustrate next authoring tools in the areas of intelligent tutoring systems and computer-assisted language learning. Ritter [6] describes an authoring tool for automatically parsing the text of an algebra word problem into a formal semantic representation that could be loaded into a cognitive tutor. Aleahmad et al. [7] presents a crowd-sourcing approach to the problem of generating content.

There has also been work on automatically creating content in the area of computer-assisted language learning. For example, Meurers et al. [8] describe a system that takes arbitrary texts as input and, with natural language processing (NLP) technologies, highlights specific grammatical constructions and automatically creates grammar practice exercises. Also, Heilman et al. [9] describe a system that uses NLP and text retrieval technologies to help English as a Second Language teachers find pedagogically appropriate reading practice materials (e.g., texts at an appropriate reading level) for intermediate and advanced language learners. The automated question generation system can be helpful for creating hint and prompts in an intelligent tutoring system [1] and constructing questions for English language learning [10].

The generation of questions by humans has long motivated theoretical work in linguistics since 1967 [11], particularly work that considers questions as transformations of canonical declarative sentences [12]. In recent years, with the advance of NLP techniques (e.g., syntactic parser, named entity recognizer, and shallow semantic labeling), automatic question generation tools has been proposed to generate specific questions about facts (where, when, who), and intentions and mental states of characters to support reading comprehension and vocabulary assessment [10], [13], [14], [15], [16]. In addition, some researchers focus on generating multiple choice questions about

lexical relationships [17] based on WordNet [18], and related words based on distributed similarity techniques [19]. Furthermore, other researchers emphasize generating questions to create dialogues from monologues as instructional content [20]. Our previous work on automatic question generation focused on generating trigger questions for academic writing support in English [21], [22]. The questions were generated from academic essays and literature reviews written by university students. The current question generation system focused on generating factual questions in Chinese, which could be useful for reading comprehension. The questions were generated from learning materials. Like our previous work, the current study adapted the three stages of question generation approach including, sentence simplification, question generation and ranking, which are described in detail in Section 3.

Chinese Question Generation is a relatively new research area, which has been investigated by a number researchers [23], [24]. Professor He et al. [23] from TianJing University first trained a maximum entropy classifier to identify the semantic chunk of a noun phrase, including entity, attribute, attribute value, event and role, within a declarative sentence. Then, they defined question generation rules based on semantic and syntactical information to generate four types of questions: 1 - Yes/No question, 2 - Affirmative-negative question 3, - Choice question 4, Wh-questions. In the experiment, they used 30 sport articles from the Chinese Wikipedia including 100 sentences for evaluation. The reported average accuracy was 53.34 percent for wh-questions. One major drawback of this approach is that the maximum entropy classifier yielded poor performance due to the small training dataset (30 articles) and lack of a sufficient feature set.

The most relevant work to ours is the question generation system developed by You [24] from Harbin Institute of Technology. This system focuses on generating factual wh-questions based on the semantic and syntactical information extracted. In their evaluation, they selected 100 sentences from Chinese Wikipedia, called Baidubaike (http://baike.baidu.com/), to generate questions. The study results showed that the precision was 0.514 while the recall was 0.500. The major reason for this poor performance is the complex, long sentences in Chinese which often cause problems for NLP steps such as sentence parsing and named entity recognizing. Furthermore, it is unclear how they defined the question generation rules.

In this article, we present a novel automatic Chinese factual question generation system that includes three major stages: sentence simplification, question generation and ranking. Compared to previous studies [23], [24], the present study describes an improved question generation system based on an overgeneration-and-ranking strategy and its evaluation in an educational context. Specifically, the major contributions of this paper are the following:

1. Applied a learning-to-rank approach to the task of generating and ranking Chinese natural language questions. The results of the evaluation of the approach show that the ranking of questions is improved by including features beyond surface characteristics.

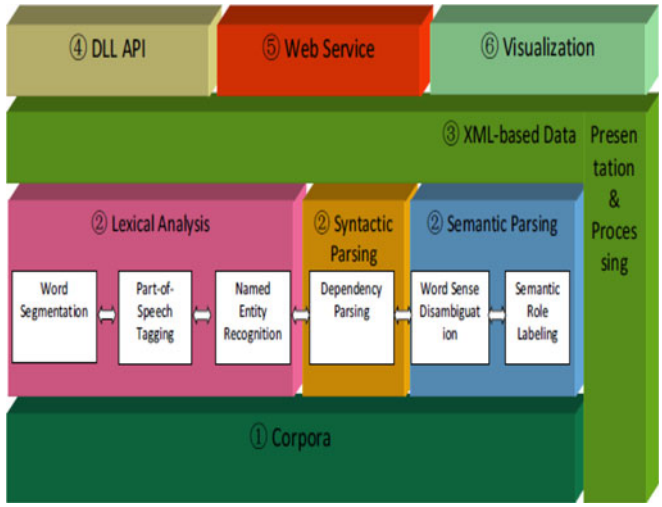

Fig. 1. The LTP architecture. This image is taken from [24].

2. Used actual instructional materials (*New Practical Chinese Reader corpus*), and a relatively large dataset (1,816 generated questions), to evaluate the system.
3. Incorporated linguistic knowledge in the sentence simplification step and in the question generation rules such as abstract temporal adverb constraints, discourse relations in long Chinese sentences [25], and interrogative pronoun classification [26].

2 Related Work

2.1 Chinese Natural Language Processing Platform

Chinese NLP research has been going on for more than 50 years and delivered some success in morphological, syntactic and semantic analysis [27]. Lack of clear delimiters between words in a Chinese sentence illustrates Chinese NLP's uniqueness relative to Western languages, e.g., English. For this reason, automatic word segmentation is a major step in Chinese morphological analysis. During the last decades, Chinese NLP systems were developed such as those developed by the Institute of Computing Technology, the Chinese Lexical Analysis System (ICTCLAS: http://ictclas.nlpir.org) and the Language Technology Platform (LTP: http://www.ltp-cloud.com/). A typical Chinese NLP system includes lexical analysis (word segmentation, part-of-speech tagging, named entity recognition) and syntactic parsing modules. These systems can now perform quite well; for instance, ICTCLAS can reach a precision of 98.45 percent in word segmentation. More recently, LTP [28], developed by Harbin Institute of Technology, added a semantic parsing module (word sense disambiguation, semantic role labeling) and achieved excellent results in word segmentation ($F_1 = 97.4$), named entity recognition ($F_1 = 92.25$), syntactic parsing (LAS = 78.23%) and semantic parsing ($F_1 = 77.18\%$) at CoNLL and SemEval. Fig. 1 shows the system architecture of LTP. It uses XML to exchange information and provides visualization facilities to display sentence processing results. Fig. 2 presents an example of the processing result of the sentence below:

雷锋是最可爱的人 (LeiFeng is the most lovely man).

In this example, the syntactic dependecy parsing result is shown on top together with syntatic relation labels. Rows 1

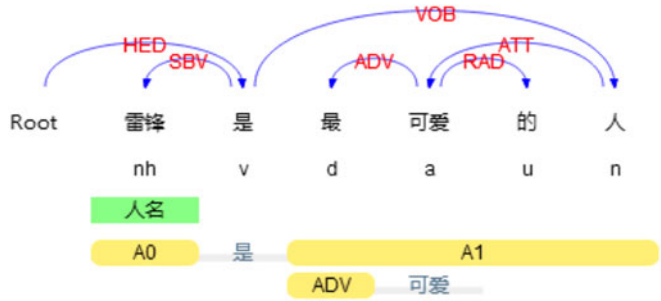

Fig. 2. An example of sentence processing result in LTP.

to 4 show the word segmentation, part of speech tagging, named entity recognition, and semantic relation labelling results, respectively. LTP provides open access to this service through web service and sends the processing result back in a number of formats, such as XML and JSON. Professor You [24] used an early version of LTP to develop a Chinese question generation system. In this study, we extend their work by defining more explicit question generation rules, implementing a sentence simplification module and a question ranking module.

2.2 Learning-to-Rank

Learning to rank is a relatively new research area, which received increasing attention in both the Information Retrieval and Machine Learning research communities, during the past decade. Most of approaches to learning to rank are designed as supervised machine learning approaches, i.e., learning a target concept from expert-labeled instances. Typically, instances are assigned a (binary or ordinal) score or label indicating their relevance to the target concept as decided by an independent, expert judgment. In the training phase, a ranking function is learned based on a set of features the expert labels. In the testing phase, the ranking function is used to rank a new set of instances and generate a ranked order.

According to how they treat the sets of ratings and loss functions used during training, Cao et al. [29] classify learning-to-rank approaches into three categories: 1) Pointwise Approach: learning to classify instances, i.e., questions in our case, according to their label individually (e.g., positive or negative category), 2) Pairwise Approach: classifying pairs of rated questions into two categories (correctly ranked or incorrectly ranked), and 3) Listwise Approach: optimizing the loss function for ordering a set of candidate questions. In the information retrieval literature, the Pointwise approach is viewed as the weakest of the three learning-to-rank approaches because it ignores the cluster of answer instances per query. An answer refers to a searching result for a given query. Machine learning techniques that can be used in conjunction with the Pointwise approach are classifiers (e.g., Naïve Bayes and Support Vector Machine) and regressors (e.g., Logistic Regression and Support Vector Machine (SMV) Regression; [30], [31]). In the case of classifiers, the classifier is trained based on each instance label and predicts a score as a ranking value for each instance, expressing the probability that it should be classified as relevant. The Pairwise approaches are considered more effective than Pointwise approaches because pairs of answer instances are considered. The algorithms used in Pairwise approaches are RankSVM [32], RankBoost [33] and RankNet [34]. Listwise approaches are more recent developments. Liu [35] shows that the Listwise techniques, such as Adarank [36], reach scores similar to or better than Pairwise techniques.

The general idea of ranking the output of a system using learning-to-rank approach has been explored in sentence parsing, natural language generation and dialogue systems. Collins and Koo [37] presented methods for re-ranking syntactic parse trees from a generative parsing model using a discriminative ranker that can consider complex syntactic features. Langkilde and Knight [38] describe methods for efficiently using n-grams statistics gathered from a large general-purpose corpus to rank the outputs from a rule-based natural language generation system in order to improve fluency. Working on a spoken dialogue system for the travel domain, Walker et al. [39] use a statistical ranker to select sentence plans, which are formal graph-based representations of potential outputs that their system can generate. They created a tailored dataset to train this ranker by gathering ratings of such sentence plans from trained experts. Probably the most similar overgeneration-and-rank approach to our own solution is that of Heilman and Smith [14]. They applied a linear regression-ranking model to rank the quality of English questions generated from articles in Wikipedia. In our study, we applied and evaluated learning-to-rank (RankSVM and Linear Regression) in our question-ranking model.

3 System Framework

This section presents an overview of the system's pipeline architecture (see Fig. 3), describing each step and emphasizing the question ranker. The three major stages of the pipeline are: sentence simplification, question transformation and ranking. Preprocessing is also needed but we do not consider it a major stage but rather a necessary preparation step.

3.1 Pre-Processing

Each sentence extracted from a given article is parsed using the LTP software. Specifically, in this step our system performs word segmentation, part of speech tagging, named entity recognition and dependency parsing and semantic role label parsing (an example is shown in Fig. 2). This information is essential for sentence simplification and question generation, described next.

3.2 Sentence Simplification

In order to reduce the complexity of question generation, we perform a sentence simplification step in the first stage. This includes sentence splitting and sentence compression because Chinese sentences are usually very long and often connect two or more self-complete sentences together. Researchers focus on preserving the critical information of the sentence for summarization [40], [41]. In our approach, a set of transformation operations derive a simpler form of the source sentence by removing parentheses (The elements in a sentence which function as the explanatory or qualifying remarks and have no clear dependent relations with the other constituents of a sentence.), adverbial modifiers

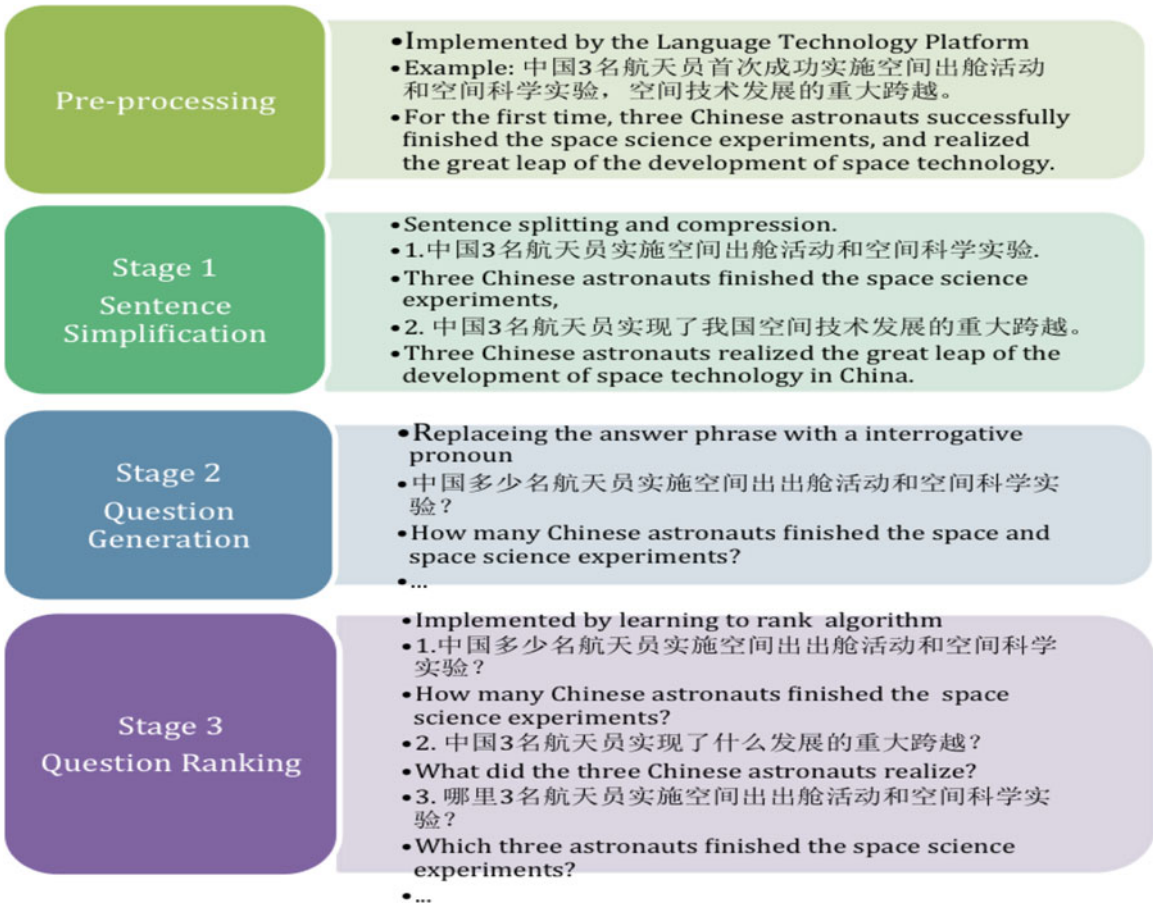

Fig. 3. Three stages question generation system.

between commas and phrase types such as sentence-level modifying phrases (e.g., manner adverb). But, in some cases, we keep some adverbial modifiers if they contain information about a person name, place, number, and time because this information can generate potential questions.

To enable questions about syntactically embedded content, our system splits a complex or compound sentence into a set of simple sentences. Wu et al. [25] classified Chinese discourse relations within a long sentence into four main categories: conjunction (e.g., coordinate and temporal), comparison (e.g., contrast and concession), contingency (e.g., cause, result, condition) and expansion. In our work, we defined a set of rules based on the semantic labels and dependency relations between main verbs. In the current version of LTP, all these relations are denoted as Coo (Main Verb1, Main Verb2), depicting a relation between embedded sentences including conjunction, comparison and cause relations between two sentences. Table 1 shows two sentence simplification rules for splitting a complex sentence into two simple sentences. The first complex sentence has been split into two simple sentences based on the conjunction relation while the second complex sentence was split based on the cause relation.

3.3 Question Transformation

In this stage, the simplified declarative sentences derived in stage 1 are transformed into a set of questions based on predefined question generation rules showed in Table 2. A key subtask of question generation is target content selection, i.e., what is the target content the question is asking about. In our case, we identify answer phrases in the input declarative sentence as potential targets for generating questions about. In Chinese, a question is generated by using an interrogative pronoun to replace the target answer phrase in the declarative sentence. Unlike question generation in English, it does not require subject-auxiliary inversion and verb decomposition. In this respect, the question generation process in Chinese is simpler.

Zhang [26] identified 11 interrogative pronouns in Chinese including 谁 (who), 什么 (what), 哪 (which), 哪儿 (where), 哪里 (where), 怎么 (how), 怎么样 (how to), 多少 (how many), 多 (how), 几 (how many | how much) and 为什么 (why). They classified them into three groups, nominal interrogative pronoun, predicate interrogative pronoun and "which" interrogative pronoun. Nominal interrogative pronouns include who, what and where. The predicate interrogative pronouns include how many, how, how to and why. "Which" interrogative pronoun only includes which. Moreover, they identified the common use of each interrogative pronoun and found that an interrogative pronoun often acted as subject, object and adverbial modifier of time or location in a question. In our implementation, rules (shown in Table 2) are defined to extract answer phrases used as subject, object and attribute for 谁 (who), 哪里 (where), 什么 (what) and多少 (how many) questions based on Zhang et al' s work. For example, the 谁 (who) pronoun can often be used as subject, object, or attribute.

TABLE 1
Examples of Sentence Simplification Rules

Description	Pattern	Derived Sentence with Main Elements
Rule 1: A complex sentence embedded with conjunction relation	$\{A0_1\}+\{\text{Main Verb}_1\}+\{A1_1\}+$ Coo (Main Verb_1, Main $\text{Verb}_2)+\{A1_2\}$	$\{A0_1\}+\{\text{Main Verb}_1\}+\{A1_1\}$ $\{A0_1\}+\{\text{Main Verb}_2\}+\{A1_2\}$
	Example: 西安既是一座历史悠久的古都，又是一座现代化的城市。 Xi'an is not only an ancient capital with a long history, but also a modern city. Simple Sentences: 1 西安是一座历史悠久的古都。 Xi'an is an ancient capital with a long history. 2 西安是一座现代化的城市。 Xi'an is a modern city.	
Rule 2: A complex sentence embedded with cause relation	$\{A0_1\}+\{\text{Main Verb}_1\}+\{A1_1\}+$ $\{A0_2\}+$Coo(Main Verb_1, Main $\text{Verb}_2)+\{A1_2\}$	$\{A0_1\}+\{\text{Main Verb}_1\}+\{A1_1\}$ $\{A0_2\}+\{\text{Main Verb}_2\}+\{A1_2\}$
	Example: 因为雷锋叔叔为人民做许多好事，所以人民永远怀念他。 Because Uncle Lei Feng does a lot of good things for the people, people will remember him forever. Simple Sentences: 1 雷锋叔叔为人民做许多好事。 Uncle Lei Feng does a lot of good things for the people 2人民永远怀念他。 People will remember him forever.	

Currently, we used the semantic role labels to identify subjects and objects as well as other roles such as 时间副词 (temporal adverbial) and 地点副词 (locative adverbial). In addition, we utilized named entities, such as 人名 (person name), 机构名 (organization name), 地点名 (location name) and part of speech, such as 时间名词 (temporal noun) and 数字 (number), to identify target answer phrases to generate factual questions, including 谁 (who) and 哪里 (where), 什么 (what), 多少 (how many), 哪所机构 (which) and 什么时候 (when).

Not all of the above target answer phrases can be used for question generation. For example, some abstract temporal adverbs cannot be used as answer phrases to generate when questions because these answer phrases do not refer to a specific time. We obtained 130 abstract temporal adverbs identified by Lu and Ma [42] which can be used for question generation. They include 最近 (recent), 此时 (now), 有时 (sometimes), 经常 (often), 偶尔 (occasion), 一会 (moment), 马上 (at once), 立刻 (right away), 以前 (before) and 常常 (often). Moreover, we do not generate questions about the content in double quotes.

We provide next an example to illustrate the whole question generation process including sentence simplification and question generation. The following sentence is first extracted from a source article and then parsed in the pre-processing stage. Fig. 4 shows the parsed tree by the LTP.

中国3名航天员首次成功实施空间出舱活动和空间科学实验，实现了空间技术发展的重大跨越。

For the first time, three Chinese astronauts successfully finished the space science experiments, and realized the great leap of the development of space technology.

This long sentence is split into the following two simple sentences by matching the sentence simplification rule 1 and then simplified by removing the adverbs, 首次 (at the first time) and 成功 (successfully):

1. *中国3名航天员实施空间出舱活动和空间科学实验.*
   Three Chinese astronauts finished the space science experiments.
2. *中国3名航天员实现了空间技术发展的重大跨越。*
   Three Chinese astronauts realized the great leap of the development of space technology.

In the second stage, the question generation rules are applied to each simple sentence. For example, the token 'three' has been correctly detected as a number in the input sentences as it is denoted as m which in turn matches the pattern Contains(m+q, A0) defined in the How Many question type defined in Table 2. This pattern indicates that m+q is a subject (A0) which determines a question to be generated by using the question template Replace(多少, m, A0).

多少名航天员实施空间出出舱活动和空间科学实验?

How many Chinese astronauts finished the space science experiments?

Similarly, other four questions are generated, including the following unacceptable question.

中国3名什么实施空间出舱活动和空间科学实验.

Three Chinese of what finished the space and space science experiments.

This problematical what-question is generated based on the question generation rule that replaces the interrogative pronoun 什么 (what) to the general noun 航天员 (astronaut). The LTP parser is the root cause of this bad question as explained in the discussion and conclusion section.

4 Question Ranking

The previous stages generate questions that vary in their quality with respect to syntax, semantics or importance.

TABLE 2
Examples of Question Generation Rules

Question Type	Description	Pattern	Question template
谁 (Who)	Subject contains person name	Contains(nh|r, A0)	Replace(谁, nh|r, A0)
		E.g. *雷锋是最可爱的人。Lei Feng is the most lovable person.* *谁是最可爱的人？Who is the most lovely person*	
哪家机构 (Which)	Object contains organization name	Contains(ni, A1)	Replace(哪所机构, ni|r, A1)?
		E.g. *张华本科毕业后，就来到了微软公司。After graduation, Zhang Hua came to the Microsoft Corporation.* *张华本科毕业后，就去了哪所机构？After graduation, Which organization did Zhang Hua come to?*	
哪里 (Where)	It serves as an adverbial modifier of place.	Contains(ns|nl, LOC)	Replace(哪里, ns|nl)
		E.g.*第 29 届夏季奥林匹克运动会在北京举行。The twenty-ninth Summer Olympic Games was held in Beijing.* *第 29 届夏季奥林匹克运动会在哪里举行 Where was the twenty-ninth Summer Olympic Games held?*	
多少 (How Many)	Subject contains a number following quantity	Contains(m+q, A0)	Replace(多少, m, A0)
		E.g. *20 支队伍在进行划船比赛。The 20 teams are rowing a boat race.* *多少支队伍在进行划船比赛? How many teams are rowing a boat race?*	
什么时候（When）	It serves as an adverbial modifier of time.	Contains{group(nt), TMP}	Replace(什么时候, TMP)
		E.g. *1942 年新加坡宣布独立。Singapore announced independence in 1942.* *新加坡什么时候宣布独立? When did Singapore announce independence?*	
什么 (What)	Possessive NP	Possessive(possessor, 的)	Replace(什么，possessor, A0|A1)
		E.g. *航空发动机是飞机的心脏。Aero engine is the plane's heart.* *航空发动机是什么的心脏？Whose heart is the Aero engine?*	
	Subject contains	Contains(n|nz, A0)	Replace(什么, NP, A0)?
		风筝做好了。The Kite is ready. 什么做好了？What is ready?	

Note: LOC refers to the adverb of place, TMP the adverb of time, A0 the agent, A1 the object, ns geographical name, nl location name, nt temporal noun, Nz other proper noun, n general noun phrase Q refers to quantity, which can be 小时(hour), 天(day)，周(week)，月(month)，世纪(century). Contains(X,Y) function means that X is in Y. Replace(X,Y,Z) function means that the interrogative pronoun (X) replaces the answer phrase (Y) in a certain part of the sentence (Z). Group(nt) refers to add one or more nts together by looking at att(nt, nt). NP refers to noun phrase which does not include location, person name, organization name and person pronouns. Possessive (owner, 的) means that there is a possessive noun form by detecting the relationship between the possessor (noun phrase) and possession (noun phrase)

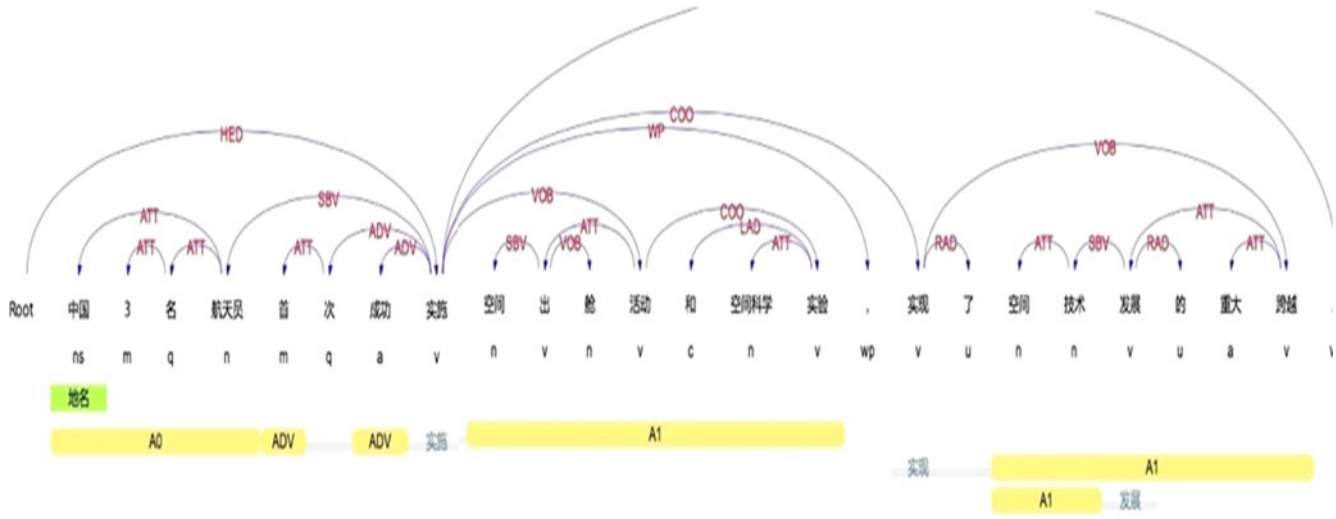

Fig. 4. An example of a parsed long sentence by LTP.

TABLE 3
Dataset Description in New Practical Chinese Reader

Num. of Articles	Num. of Sentences	Num. of Words	Ave. Num. of Sentences Per Article	Ave. Num. of Words Per Article	S.D. of Num. of Sentences Per Article	S.D. of Num. of Words Per Article
99	399	17,572	5.33	177.49	1.57	44.41

This is unavoidable and happens for different reasons, such as errors in sentence parsing, named entity recognition, and sentence simplification. To address this problem, ranking the large pool of questions according to their quality is needed. Stage 3 in our method implements a learning to rank algorithm to meet this challenge.

4.1 Ranking Model

In the ranking model, we used two common learning to rank approaches: Pointwise approach (Logistic Regression [37]) and Pairwise approach (RankSVM [43]). The logistic regression model is learned by fitting training data to a logit function by using the predictor binary variable which indicates whether a question is acceptable or not. A RankSVM model is learned using a Pairwise approach which can naturally specify questions that are of an equivalent rank. Support Vector Machines has been used previously for preference ranking in the context of information retrieval. We adopt the same framework for ranking questions. In this model, given a collection of questions ranked according to preferences between two questions represented by feature vector q_i and q_j, respectively, and a linear learning function f, we can say

$$q_i \succ q_j \Leftrightarrow f(q_i) > f(q_j). \quad (1)$$

Where f indicates that q_i is preferred over q_j. The function f is defined as $f(q) = w \bullet q$, where

$$f(q_i) > f(q_i) \Leftrightarrow w \bullet q_i > w \bullet q_j. \quad (2)$$

In the context of SVM, these weight vectors or support vectors (w) are identified by minimizing the function using slack variables ξ_{ij}:

$$\min_{w,\xi_{ij}} \frac{1}{2}\|w\|^2 + c\sum\nolimits_{ij} \xi_{ij}. \quad (3)$$

Subject to the constraints:

$$\forall(q_i, q_j) : w \bullet q_i \geq w \bullet q_j + 1 - \xi_{ij}$$

$$\forall(i, j) : \xi_{ij} \geq 0.$$

Finding the support vectors and the generalization of the Ranking SVM is done differently [32]. If the data are linearly separable, the slack variables ξ_{ij} are all equal to 0. In this case, we can consider the ranking function as projecting the data points onto the separating hyperplane and the support vectors as the two points q_i and q_j nearest each other on the hyperplane. The generation is accomplished by calculating w to maximize the distance between these closest points. The distance or margin between these two points is formulated as $\frac{w(q_i - q_j)}{\|w\|}$. Like the classification SVM algorithm [44], the margin is maximized when $\|w\|$ is minimized.

4.2 Feature Definition

The features used in the ranking models were developed by an in-depth analysis of questions generated from the training set. Question generation is a pipeline process, where errors could occur at any stage. The goal of our current ranking model is to filter questions with grammatical and semantic errors. If the source sentences are complex and lengthy, it is more likely to generate erroneous questions. In addition, if the sentence simplification is performed, it could cause errors since it is a rule-based approach. Furthermore, the performance of generating different types of questions could be different. In fact, which and where questions are more accurate than what questions. Therefore, these features should indicate the likelihood of generating an acceptable or unacceptable question in terms of the complexity of source sentences (Num. of NamedEntities, Num. of Main Verbs, Num. of Clauses and Length), the transformation performed during the processing (IsCompressed and IsSplit), and question type (InterrogativePronounType, AnswerPhraseType). The source sentence refers to the declarative sentence in the description of the features below. There are 10 features defined as below:

1. Num. of Named Entities: this numeric feature describes the number of named entities in the source sentence.
2. IsCompressed: this Boolean feature detects whether the source sentence is compressed.
3. IsSplit: this Boolean feature detects whether the source sentence is split.
4. InterrogativePronounType: this is a categorical feature that detects which question type is used, such as 什么时候 (when), 哪里 (where) and 谁 (who).
5. IsPossessivePatternMatched: this Boolean feature indicates if the answer phrase is a possessive noun.
6. AnswerPhraseType: this is a categorical feature that describes if the answer phrase in Subject, Object or Attribute.
7. Num. of Clauses: this numeric feature shows the number of clauses in the source sentence. Each clause consists of an argument A0 (subject) followed by a main verb.
8. Num. of Main Verbs: this numeric feature shows the number of main verbs in the source sentences.
9. Negation: This is a Boolean feature for the presence of 不是 (not), 从不 (never), or 不 (no) in the question.
10. Length: this numeric feature describes the number of words in the source sentence.

5 Data and Annotation Process

The corpus was a random sample of articles from the set of featured articles in *New Practical Chinese Reader* textbook with an average number of 5 sentences (See Table 3). The *New*

Practical Chinese Reader textbook was used to teach foreigners Chinese. It provides expository text at a reading level corresponding to elementary education or intermediate second language learning. Expository text exposes you to facts, plain and simple. Unlike narrative text, expository text does not tells a story and involve characters, which often described in a narrative text. For generating factual questions, expository text is easier than narrative text since the expository text contains many factual statements, including time, place and person. If we use other types of text, such as narrative text, only few questions could be generated since they do not contain many factual statements. In fact, different questions generation approaches were proposed for generating "What", "Why" and "How" questions from narrative text [15]. This is reason why we choose expository text to generate factual question. We used the selected texts to generate the training and testing set which consist of 1,216 (generated from 69 articles) and 600 (generated from 30 articles) questions, respectively.

The data was annotated as described next. We asked three Chinese linguistic majors to generate factual questions from the testing set and rate the quality of system generated questions as acceptable or unacceptable from the testing set and training set according to two major criteria, (1) grammaticality and (2) semantic correctness. Grammaticality refers to the presence or absence of grammar errors. Semantic correctness refers that the overall meaning of the generated question is relevant to the context and the question has no vagueness. These two major criteria was also used in the question generation shared task evaluation [45].

Each annotator was asked to read the text of each article, and then rate approximately 180 questions automatically generated from the text. If the article was in the testing set, they had to generate factual questions before rating the system-generated questions. We only chose a subset of the dataset (the testing set: 30 articles) for them to generate questions since it would take much effort to generate questions from 99 articles. For both the training set and testing set, each question was rated independently by three people to provide a more reliable gold standard. To assign final labels to these questions, a question was labeled as acceptable only if a majority of the three raters judged it as acceptable (i.e., grammatical and semantic correctness). An inter-rater agreement of Fleiss's $\kappa = 0.58$ was computed from the datasets acceptability ratings. This value corresponds to "moderate agreement" (Landis and Koch, 1977).

6 Evaluation Results and Discussion

This section describes the results of the experiments we conducted in order to evaluate the quality of the automatically generated questions before and after ranking. The performance metrics we employed are precision and recall defined as follows, which are widely used by the question generation community [46]. For rankings, our metric is the percentage of the top N% labeled as acceptable, for various N,

$$\text{Recall} = \frac{q_h \cap q_s}{q_h} \tag{4}$$

$$\text{Precision} = \frac{q_h \cap q_s}{q_s}. \tag{5}$$

TABLE 4
Question Generation Result in the Testing Set

	Question Type	Recall	Precision
New Practical Chinese Reader Corpus	Who (谁)	0.48	0.69
	Where (哪里)	0.68	0.79
	When (什么时候)	0.74	0.70
	How Many (多少)	0.70	0.71
	Which (哪家机构)	0.38	0.80
	What (什么)	0.87	0.44
	Macro-Average	0.64	0.69

Where q_h is the number of questions generated by our question generation system and q_s is the number of questions generated by human annotators.

6.1 Results for Unranked Questions

We have evaluated the system performance in the testing set, which includes 600 system-generated questions. Table 4 shows Which (Precision: 0.80) and Where (Precision: 0.79) questions get higher precision than others. Wrong entity recognition is the major cause of errors for these questions. For example, the LTP parser always recognizes a country as a place to generate where questions. For instance, China and Japan in the following example refer to a country or nation, rather than a place, which is how the LTP parser wrongly identifies them.

岩田表示，他最大的愿望是通过自己的努力，促进中日友好。

Iwata expressed his best wish to the friendship between China and Japan through his own effort.

Moreover, some entities such as 阿莫尔型小行星 (Amor asteroids) and 基督学说 (Christian doctrine) have been wrongly recognized as an organization. Furthermore, we need to define more fine-grained rules for handling multiple entities in a sentence. In the example below, we should not generate questions about a particular place; instead, we should replace the whole adverb of place with 哪些城市 (which cities).

蝗灾在乌兰察布市、呼仑贝尔盟、通辽市、赤峰市最为严重。

The plague of locusts in Wulanchabu City, Tongliao City, Baer Meng Hu Lun, Chifeng city is the most serious.

What-questions are the most problematic (precision: 0.44). Errors for these questions are caused by the following major reasons:

1. Occupations are recognized as an ordinary noun phrase by the LTP, which can not be used as an answer phrase for what question, such as 记者 (reporter), 工程师 (engineer), 老师 (teacher), 警察 (policy), 航天员 (astronaut)，人员 (staff), 裁判 (judge)，农民 (farmer), 领导 (administrator) and 红军 (Red Army).
2. Abstract nouns can not be used as an answer phrase for what question, such as 成就 (achievement), 事件 (event) and 场景 (scene)
3. Awkward questions were generated from possessive noun patterns. For example, the question 什么的郑双

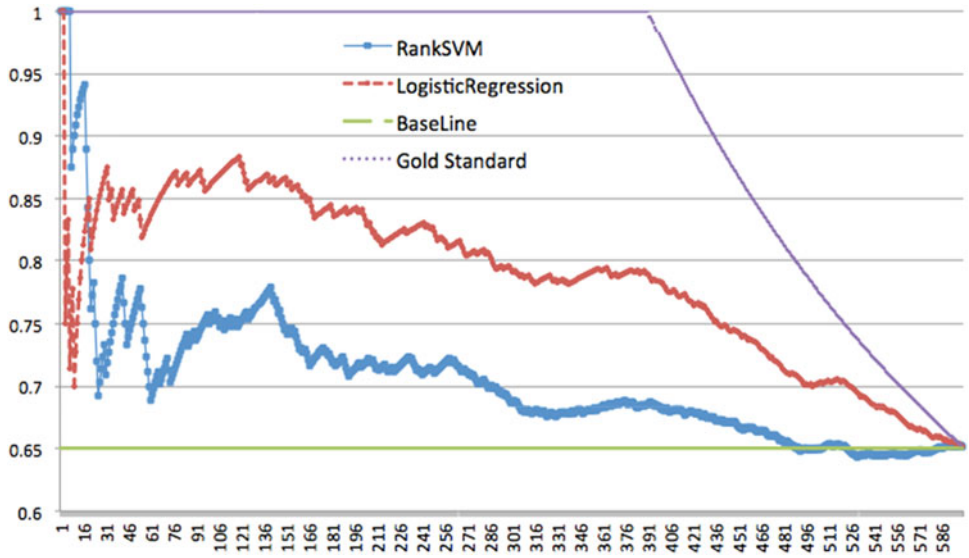

Fig. 5. A graph of the percentage of acceptable questions in the top ranked questions using two ranking models in the testing set of New Practical Chinese Reader corpus.

悦是一名先进共产党员 (Whose Zheng Shuangyue is an outstanding communist party member)?, was generated from the following sentence 草原工作站的郑双悦是一名先进共产党员 (The grassland workstation's Zheng Shuangyue is an outstanding communist party member).

For the how many questions, more question generation rule constraints should be defined. Some numbers refer to special departments, which cannot be used for generating how many questions. For example, the number 120 and 110 mean a medical department and police department, respectively. Besides, the LTP parser generates parsing errors. For instance, the second (refers to time) is always wrongly identified as a number rather than a temporal noun.

Which (Recall:0.38) and Who (Recall:0.48) questions have poor recall because some entities can not be recognized by the LTP parser. These unidentified organization entities include 中国科学研究生院 (Chinese Academy of Sciences) and 领事馆 (consulate). Many person named entities includes the occupation names, such as 记者 (reporter) and 警察 (policeman), cannot be detected.

6.2 Ranking Results

For ranking experiments, we present results for the following ranking models:

RankSVM. This model is implemented by RankSVM algorithm mentioned before. The RankSVM includes all the features described in previous section.

Logistic regression. This model is similar to the approach described by Collins and Koo [37] for ranking syntactic parses. Questions are ranked by the predictions of a logistic regression model of question acceptability.

Baseline. The expectation of the performance if questions were ranked randomly.

Gold standard. The expected performance if all the acceptable questions were ranked higher than unacceptable.

The baseline was 0.65 (65 percent of all test set questions were labeled as acceptable by human annotators). Fig. 5 shows that most of ranking models were unstable for the top 150 questions. The RankSVM and Logistic Regression had very similarly sharp curve, which are higher than baseline overall. Table 5 shows the ranking results of these models for the top 10, 25 and 50 percent of the ranked questions. In the top 60 questions (10 percent) and top 150 questions (25 percent) and top 300 questions (50 percent), Linear Regression got better accuracy (0.83 in top 10 percent, 0.87 in top 25 percent, 0.79 in top 50 percent) than RankSVM (0.70 in top 10 percent, 0.75 in top 25 percent, 0.69 in top 50 percent).

A one-way ANOVA, at a 95 percent confidence level, was conducted to examine whether there are statistical differences among these models. The ANOVA indicated a significant difference, $F(3,2396) = 1787.32$, $p < 0.05$. Fishers' least significant difference (LSD) tests at the 95 percent confidence level were performed to determine whether significant differences occurred between the mean scores for each pair of treatments. Results indicated Logistic Regression got better performance than RankSVM (MD:0.83, $p < 0.05$). Moreover, RankSVM and Logistic Regression significantly outperformed Baseline.

7 Discussion and Conclusion

Automatic generation of natural questions is a challenging task. Previous work [23], [24] in Chinese factual question generation only relied on rule-based approach, and reported a poor performance (average precision:0.50, recall:0.5) based on a small dataset (100 sentences for evaluation) due to the limitations of the parser and the complexity of

TABLE 5
The Percentage of the Top Ranked Questions Labeled Acceptable with Various Models in New Practical Chinese Reader Corpus

Model	Top10%(60)	Top25%(150)	Top50%(300)
RankSVM	0.70	0.75	0.69
Logistic Regression	0.83	0.87	0.79
Baseline	0.65	0.65	0.65
Gold Standard	1	1	1

Chinese long sentences. This article addressed this challenge and presented a novel Chinese question generation system, which includes sentence simplification, question generation and ranking stages, built on the top of a recently developed Chinese natural language processing platform, called Language Technology Platform [28].

In order to evaluate the system performance, we collected 99 articles from New Practical Chinese Reader textbook, a real language learning material, and analyzed the quality of system-generated questions. The experimental result shows that the system performance reaches a recall of 0.64 and precision of 0.69. Most importantly, the results indicated that this approach is effective since the best ranking model (linear regression) improved the acceptability of the top 25 percent questions by more than 20 percent in the dataset. In particular, the statistical question ranking models significantly outperformed baseline, which is consistent with results by Heilman and Smith [14]. Interestingly, the performance of the linear regression ranker that is used in the QG system was significantly different than the performance of RankSVM.

This study has some limitations. The first limitation is that we have not evaluated the effectiveness of the system, which helps a human teacher to generate questions in a real teaching situation. Another limitation is that the question generation rules need improvement. Despite these limitations, we believe that this question generation approach is effective and the evaluation meaningful since we used real learning materials (*New Practical Chinese Reader*) to generate questions and a large number of system generated questions ($N = 1{,}816$) had been evaluated.

Our future work will focus on integrating the question generation tool into a learning management system and evaluating the effectiveness of the system as an authoring tool. Moreover, we will build a shared dataset for evaluating Chinese question generation systems, similar to English question generation [47]. Moreover, we will investigate to generate deep questions based on information extraction or semantic dependency parsing techniques. The performance of named entity recognition in the LTP is one of the major issues for question generation. Since we cannot directly improve the LTP parser, one possible solution to improve the accuracy of the named entity recognition is to use a knowledge base, such as HowNet [48], to re-check the noun phrase outputted by the LTP. HowNet is a knowledge base unveiling inter-conceptual relationships and inter-attribute relationships of concepts [14]. For example, in HowNet, 领事馆 (consulate) can be correctly identified as an institution while 警察 (policy), 记者 (reporter) and 农民 (farmer) can be recognized as an occupation name.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61502397), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (the 50th grants), and Fundamental Research Funds for the Central Universities under Grant Numbers XDJK2014A002 and XDJK2017C024 and SWU114005 and CQU903005203326.

References

[1] V. Rus, Z. Q. Cai, and A. C. Graesser, "Experiments on generating questions about facts," in *Proc. 8th Int. Conf. Comput. Linguistics Intell. Text Process.*, 2007, pp. 444–455.
[2] F. L. Ryan, "Differentiated effects of levels of questioning on student achievement," *The J. Exp. Edu.*, vol. 3, no. 41, pp. 63–67, 1973.
[3] P. H. Winne, "Experiments relating teachers'use of higher cognitive questions to student achievement," *Rev. Edu. Res.*, vol. 49, no. 1, pp. 13–49, 1979.
[4] D. J. Hacker Dunlosky, J. Dunlosky, and A. C. Graesser, *Metacognition in Educational Theory and Practice*. Mahwah, NJ, USA: Erlbaum., 1998.
[5] A. C. Graesser and N. K. Person, "Question asking during tutoring," *Amer Edu. Res. J.*, vol. 31, pp. 104–137, 1994.
[6] S. Ritter, "The authoring assistant," in *Proc. 4th Int. Conf. Intell. Tutoring Syst.*, 1998, pp. 126–135.
[7] T. Aleahmad, V. Aleven, and R. Kraut, "Open community authoring of targeted worked example problems," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2008, pp. 216–227.
[8] D. Meurers, R. Ziai, L. Amaral, et al., "Enhancing authentic web pages for language learners," in *Proc. 5th Workshop Innovative Use NLP Building Educ. Appl.*, 2010, pp. 10–18.
[9] M. Heilman, L. Zhao, J. Pino, et al., "Retrieval of reading materials for vocabulary and reading practice," in *Proc. 3rd Workshop Innovative Use NLP Building Educ. Appl.*, 2008, pp. 80–88.
[10] H. Kunichika, T. Katayama, T. Hirashima, et al., "Automated question generation methods for intelligent english learning systems and its evaluation," in *Proc. Int. Conf. Comput. Edu.*, 2002, pp. 1117–1124.
[11] J. R. Ross, *Constraints on Variables in Syntax*. Cambridge, MA, USA: MIT, 1967.
[12] N. Chomsky, "On Wh-Movement," *Formal Syntax*, P. W. Wasow and A. Akmajian, eds. New York, NY, USA: Academic Press, 1977.
[13] R. Mitkov and L. A. Ha, "Computer-aided generation of multiple-choice tests," in *Proc. HLT-NAACL Workshop Building Educ. Appl. Using Natural Lang. Process.*, 2003, pp. 17–22.
[14] M. Heilman and N. A. Smith, " Good question! Statistical ranking for question generation," in *Proc. Annu. Conf. North Amer. Ch. Assoc. Comput. Linguistics*, 2010, pp. 609–617.
[15] J. Mostow and W. Chen, "Generating instruction automatically for the reading strategy of self-questioning," in *Proc. Int. Conf. Artif. Intell. Edu.*, 2009, pp. 465–472.
[16] P. Mannem, R. Prasad, and A. Joshi, "Question generation from paragraphs at UPenn: QGSTEC system description," In Boyer, Kristy Elizabeth and Piwek, Paul eds. Proceedings of QG2010: The Third Workshop on Question Generation, Pittsburgh, pp. 84–91, 2010.
[17] J. C. Brown, G. A. Frishkoff, and M. Eskenazi, "Automatic question generation for vocabulary assessment," in *Proc. Conf. Human Lang. Technol. Empirical Methods Natural Lang. Process.*, 2005, pp. 819–826.
[18] C. Fellbaum and G. Miller, *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998.
[19] M. Heilman and M. Eskenazi, "Application of automatic thesaurus extraction for computer generation of vocabulary questions," in *Proc. Speech Lang. Technol. Edu.* pp. 65–68.
[20] P. Piwek and S. Stoyanchev, "Question generation in the CODA project," in *Proc. 3rd Workshop Question Generation*, 2010, pp. 1–5.
[21] M. Liu, R. Calvo, and V. Rus, "Automatic generation and ranking of questions for critical review," *Educ. Technol. Soc.*, vol. 17, pp. 333–346, 2014.
[22] M. Liu, R. A. Calvo, and V. Rus, "Automatic question generation for literature review writing support," in *Proc. 10th Int. Conf. Intell. Tut. Syst.*, 2010, pp. 45–54.
[23] Z. Yang, "Research of intelligent inquiry system based on maximum entropy model," *Master Thesis, School Comput. Sci. Technol.*, Tianjing University, 2008.
[24] L. You, "Encyclopedic knowledge based question auto-generation system," *Master Thesis, School Comput. Sci. Technol., Harbin Institue Technol.*, 2012.
[25] Y. Wu, J. Shi, and F. Wan, "Automatic identification of chinese coordination discourse relation," *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 49, no. 1, pp. 1–6, 2013.
[26] Y. Zhang, "Particular referring questions in international chinese teaching," *Master Thesis, College of Humanities, Huazhong University of Science Technol.*, 2013.

[27] Z. Chenqing, *Chinese Information Processing: Statistical Natural Language Processing*. Beijing, China: Qinghua University Press, 2013.
[28] W. Che, Z. Li, and T. Liu, "LTP: A chinese language technology platform," in *Proc. 23rd Int. Conf. Comput. Linguistics*, 2010, pp. 13–16.
[29] Z. Cao, T. Qin, T. Y. Liu, et al., "Learning to rank: From pairwise approach to listwise approach," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 129–136.
[30] D. Cossock and T. Zhang, "Subset ranking using regression," in *Proc. 19th Annu. Conf. Learn. Theory*, 2006, pp. 605–619.
[31] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans (eds.), *Advances in Large Margin Classifiers*, Cambridge, MA: MIT Press, pp. 115–132.
[32] T. Joachims, "Training linear SVMs in linear time," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 217–226.
[33] Y. Freund, R. Iyer, R. Schapire, et al., "An efficient boosting algorithm for combining preferences," *The J. Mach. Learn. Res.*, vol. 4, pp. 933–969, 2003.
[34] C. Burges, T. Shaked, E. Renshaw, et al., "Learning to rank using gradient descent," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 89–96.
[35] T. Liu, "Learning to rank for information retrieval," *Found. Trends Inf. Retrieval*, vol. 3, no. 3, p. 7, 2009.
[36] J. Xu and H. Li, "Adarank: A boosting algorithm for information retrieval," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2007, pp. 391–398.
[37] M. Collins and K. Koo, "Discriminative reranking for natural language parsing," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 175–182.
[38] I. Langkilde and K. Knight, "Generation that exploits corpus-based statistical knowledge," in *Proc. 7th Int. Conf. Comput. Linguistics*, 1998, pp. 704–710.
[39] M. A. Walker, O. Rambow, and M. Rogati, "SPoT: A trainable sentence planner," in *Proc. 2nd Meeting North Amer. Chapter Assoc. Comput. Linguistics Lang. Technol.*, 2001, pp. 1–8.
[40] T. Martins and A. Smith, "Summarization with a joint model for sentence extraction and compression," in *Proc. Workshop Integer Linear Program. Natural Lang. Process.*, 2009, pp. 1–9.
[41] C. Zhang, M. Hu, T. Xiao, et al., "Chinese sentence compression: Corpus and evaluation," in *Proc. 12th Int. Conf. Chinese Comput. Linguistics Natural Lang. Process. Based Naturally Annotated Big Data*, 2013, pp. 257–267.
[42] J. Lu and Z. Ma, *Discussion on the Function Words in Modern Chinese*. Beijing, China: Peking Univ. Press, 1985, pp. 89–97.
[43] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 133–142.
[44] N. Christianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[45] V. Rus, B. Wyse, P. Piwek, M. C. Lintean, S. Stoyanchev, and C. Moldovan, "The first question generation shared task evaluation challenge," John D. Kelleher; Brian Mac Namee; Ielka van der Sluis; Anja Belz; Albert Gatt & Alexander Koller, ed., INLG, The Association for Computer Linguistics, 2010.
[46] S. Kalady, A. Elikkottil, and R. Das, "Natural language question generation using syntax and keywords," Boyer, Kristy Elizabeth and Piwek, Paul eds. Proceedings of QG2010: The Third Workshop on Question Generation, Pittsburgh, pp. 1–10, 2010.
[47] V. Rus, B. Wyse, P. Piwek, M. Lintean, S. Stoyanchev, and C. Moldovan, "Overview of the first question generation shared task evaluation challenge," In Boyer, Kristy Elizabeth and Piwek, Paul eds. Proceedings of QG2010: The Third Workshop on Question Generation, Pittsburgh, pp. 45–57, 2010.
[48] Z. D. Z. Dong, and Q. D. Q. Dong, "HowNet - a hybrid language and knowledge resource," in *Proc. Int. Conf. Natural Lang. Knowl. Eng.*, 2003, pp. 820–824.

Ming Liu received the PhD degree in artificial intelligence in education from the School of Electrical and Information Engineering, The University of Sydney, Australia, in 2012. He is an associate professor in the School of Computer and Information Science, Southwest University, China. His main research interests include question generation, learning analytics, and intelligent tutoring system. He participated in national and international projects funded by ARC Linkage (Australia), Young and Well CRC, Office of Teaching and Learning, Google, and Chinese National Fund. He is an author of more than 14 publications papers in prestigious conferences and journals, such as *Intelligent Tutoring Systems*, *IEEE Transactions on Learning Technologies*, and *Journal of Educational Technology and Society*.

Vasile Rus received the bachelor's degree in computer science from the Technical University of Cluj-Napoca in June 1997 with a diploma thesis entitled "Distributed and Collaborative Configuration Management," masterpieced while at LSR Laboratory, INPG, Grenoble, France. He received the master's of science and the doctor's of philosophy degrees in computer science from Southern Methodist University at Dallas, Texas, in May 1999 and May 2002, respectively. He is a professor at the University of Memphis. He has been involved in research and development projects in the areas of computational linguistics and information retrieval for more than 15 years and in open-ended student answer assessment and intelligent tutoring systems for more than 10 years. He has received four Best Paper Awards; produced more than 100 peer-reviewed publications; and currently serves as an associate editor of the *International Journal on Tools with Artificial Intelligence* and a program committee member of the International Conference on Artificial Intelligence in Education (AIED 2015).

Li Liu received the PhD degree in computer science from the Université Paris-sud XI in 2008. He is an associate professor at Chongqing University. He is also serving as a senior research fellow in the School of Computing at the National University of Singapore. He served as an associate professor at Lanzhou University in China. His research interests are in pattern recognition, data analysis, and their applications on human behaviors. He aims to contribute in interdisciplinary research of computer science and human related disciplines. He has published widely in conferences and journals with more than 50 peer-reviewed publications. He has been the principal investigator of several funded projects from government and industry.