Natural Language Engineering **22** (3): 457–489. © Cambridge University Press 2016
doi:10.1017/S1351324915000455

Automatic generation of short answer questions for reading comprehension assessment

YAN HUANG[1] and LIANZHEN HE[2]

[1]*Language Technology Lab, Department of Theoretical and Applied Linguistics, University of Cambridge, 9 West Road, Cambridge CB3 9DA, UK*
e-mail: yh358@cam.ac.uk
[2]*School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, P.R.China*
e-mail: hlz@zju.edu.cn

(*Received 16 June 2014; revised 6 November 2015; accepted 10 November 2015; first published online 13 January 2016*)

Abstract

Writing items for reading comprehension assessment is time-consuming. Automating part of the process can help test-designers to develop assessments more efficiently and consistently. This paper presents an approach to automatically generating short answer questions for reading comprehension assessment. Our major contribution is to introduce Lexical Functional Grammar (LFG) as the linguistic framework for question generation, which enables systematic utilization of semantic and syntactic information. The approach can efficiently generate questions of better quality than previous high-performing question generation systems, and uses paraphrasing and sentence selection to improve the cognitive complexity and effectiveness of questions.

1 Introduction

Automatic item generation (AIG) techniques are beneficial to reading assessment development. First, AIG aids item writers to create items faster (Mitkov, Ha and Karamanis 2006), thereby reducing labor costs in test development. Second, while handcrafted items are subject to idiosyncratic interpretation of test specification, AIG can help to eliminate this undesirable feature and improve the consistency of item writing (Bormuth 1970). Third, AIG can boost the application of computerized adaptive reading assessment by accelerating construction of large-scale item banks (Ockey 2009). Furthermore, AIG can facilitate the development of individualized low-risk reading assessments like classroom quizzes and self-assessment.

Researchers have attempted to generate various forms of items. Usually, the stems of the items are either partial declarative sentences (i.e., cloze), or complete interrogative sentences (i.e., questions). This study focuses on items with question stems. These items require generation of grammatical sentences, presenting greater challenges in natural language processing (NLP) than cloze items. Specifically, we aim to generate short answer questions to which the answers may be a word, a phrase or a clause.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Questions can be classified into sixteen categories and three levels of depth according to the type of information that the questions ask (Graesser, Rus and Cai 2008). For example: questions on *who*, *what*, *when* and *where* belong to a shallow category called 'concept completion'; questions on *how much* and *how many* belong to an intermediate category called 'quantification', and questions on *why* and *how* belong to deep question categories like 'causal antecedent', 'goal orientation', or 'procedural'.

Additionally, questions can be classified by the cognitive skills they assess. These skills, from low to high, are knowledge, comprehension, application, analysis, evaluation and synthesis (Bloom *et al.* 1956). Intuitively, deeper questions are more likely to assess higher-order reading abilities or cognitive skills. However, the cognitive complexity of a question is determined not only by the question depth, but also other factors such as the exact wording of the question, the stimulus text and when the question is asked (Felker and Dapra 1975). For instance, given a sentence that states an explicit causal relationship (e.g., *it is dark because it is raining*), a deep question (e.g., *why is it dark*) that merely rearranges the words (excluding the answer) from the sentence, can be answered correctly by matching the surface words. As a result, the deep question is unlikely to require cognitive processing beyond the knowledge level. Contrastingly, this study uses 'question depth' and 'deep question' as described by Graesser *et al.* (2008), whilst using 'cognitive complexity' and 'comprehension' as described by Bloom *et al.* (1956).

This study presents an efficient and effective approach to automatically generating high-quality questions for reading comprehension assessment. Adopting LFG (Kaplan and Bresnan 1982) as the linguistic framework for question generation, the approach handles all three levels of question depth. The approach also includes paraphrasing to enhance the capability of the questions to assess reading comprehension, and uses sentence selection to improve the effectiveness of the questions. We experimentally evaluated the performance of our system in individual tasks (i.e., sentence selection, paraphrasing and question generation), and compared the automatically generated questions with those written by human experts. We also evaluated the performance of post-edited questions in reading comprehension assessment.

2 Literature review

This section reviews prior studies on text-based question generation, paraphrasing, and sentence selection.

2.1 Question generation

Research on automatic question generation from text (QGT) started in the 1970s (Wolfe 1976). However, no further studies on QGT were conducted until 30 years later when NLP techniques became mature enough (Kunichika *et al.* 2003, Mitkov and Ha 2003). The basic methodology of QGT is to substitute values, concepts or facts extracted from text for the variables of question templates. These question templates can be constructed based on various linguistic information. The depth of linguistic information is defined by the extent to which the information captures

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

long distance dependencies and underlying predicate argument structures (Baldwin *et al.* 2007). The linguistic information that has been utilized in the prior studies, from shallow to deep, includes surface words (Wolfe 1976), syntactic chunks (Mitkov and Ha 2003), syntactic relations (Kunichika *et al.* 2003, Gates 2008, Ali, Chali and Hasan 2010, Heilman and Smith 2010, Kalady, Elikkottil and Das 2010, Pal *et al.* 2010, Varga and Ha 2010, Bernhard *et al.* 2012, Curto, Mendes and Coheur 2012, Liu, Calvo and Rus 2012), semantic roles (Mannem, Prasad and Joshi 2010, Lindberg *et al.* 2013, Mazidi and Nielsen 2014), and deep semantic relations (Mostow and Chen 2009, Yao, Bouma and Zhang 2012, Olney, Graesser and Person 2012).

In general, QGT approaches that rely on shallower information require more specific templates. This is especially true for deep questions, whose question foci are too varied to describe by surface words or syntactic rules alone. Accordingly, semantics-based approaches have been embraced to enhance deep question generation (Piwek and Boyer 2012, Yao *et al.* 2012).

Most prior studies on QGT used self-defined criteria to evaluate the quality of individual questions. Nevertheless, there were exceptions in the first Question Generation Shared Task Evaluation Challenge (QG-STEC) (Rus *et al.* 2012). Four research teams participated in the task of sentence-based question generation for the first QG-STEC (Ali *et al.* 2010, Pal *et al.* 2010, Varga and Ha 2010, Yao and Zhang 2010). The questions were evaluated by relevance, question categories, syntactic correctness and fluency, ambiguity, and variety. Results showed that the syntax-based approach of Varga and Ha (2010) performed the best on all criteria except variety. However, when missing question types were penalized, the semantics-based approach of Yao and Zhang (2010) (MrsQG) was better on all criteria. Unlike most prior studies on question generation which manipulated linguistic structures by intuition, MrsQG transformed sentences based on Head-driven Phrase Structure Grammar (HPSG). This linguistic framework contributed to the good performance of MrsQG in relation to QG-STEC. Nevertheless, MrsQG required large memory footprint to run, and its processing speed was slow (Yao *et al.* 2012).

2.2 Paraphrasing

Paraphrasing can improve the cognitive complexity of questions. When a person reads, he or she constructs a mental meaning representation that is different from the surface representation of the text (Anderson 1972, Perfetti 1992). The goal of reading comprehension assessment is to examine whether or not a test-taker constructs a correct mental meaning representation during reading. As a result, making the surface words and the structure of a question different from its source sentence(s) ensures that the question cannot be answered correctly by matching the surface words, and increases the possibility that the question assesses reading comprehension.

To produce items for reading comprehension assessment, the content words (i.e., nouns, verbs and modifiers) in the source text should be replaced by their synonyms (Anderson 1972). Roid and Haladyna (1978) proved that items paraphrased in this manner were slightly more difficult than verbatim items. Similarly, Felker

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

and Dapra (1975) found that when people answered questions paraphrased by synonyms (compared to verbatim questions) during reading, they performed better on a subsequent application assessment which tested how well the test-takers could apply the rules learned from the reading. This meant that when answering the paraphrased questions, the participants triggered more cognitive activities (including comprehension) which were necessary for learning.

Few prior studies on QGT have considered paraphrasing. Kunichika *et al.* (2003) replaced words with their synonyms and antonyms, but failed to report how this was implemented. On the other hand, Bernhard *et al.* (2012) defined rules to vary question words, such as replacing *who* with *which person*, and to replace verbs with their nominalizations. These treatments enriched the form of questions, but did not introduce content words that seemed different enough from the original ones. Therefore, the paraphrasing by Bernhard *et al.* (2012) was unlikely to improve the cognitive complexity of questions.

2.3 Sentence selection

In order to generate items suitable for reading comprehension assessment, we should select appropriate source sentences. The essence of educational assessment is inferring target abilities from sample observations (Carroll 1968). In reading comprehension assessment, responses to questions that target the key information of a text provide good sample observations for assessing whether or not the respondents comprehend the entire text. As a result, selecting sentences that contain key information as the question source can improve the effectiveness of the items for reading comprehension assessment.

Sentence selection has been studied extensively in the field of extractive text summarization (Spärck Jones 2007). Therefore, prior studies of AIG used summarizing techniques to select sentences. For example, Agarwal and Mannem (2011) selected sentences that were informative and cloze-item-generatable based on surface features like the sentence position, sentence length and the number of superlative modifiers. Similarly, Becker, Basu and Vanderwende (2012) utilized a summarizing tool that scored sentences according to term frequency, whilst Adamson *et al.* (2013) compared the cosine similarity of word vectors, Latent Semantic Analysis (LSA), and Term Frequency-Inverse Document Frequency (TF-IDF) in selecting sentences for question generation.

In summary, all the aforementioned methods focused on selecting important sentences, and worked well for long stimulus texts. However, when the stimulus texts are short, which is common in low-intermediate level reading comprehension assessment, these methods may select sentences of similar meaning and result in redundancy.

3 Methodology

This section introduces our framework for generating short answer questions, and explains how to implement each module of the framework.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

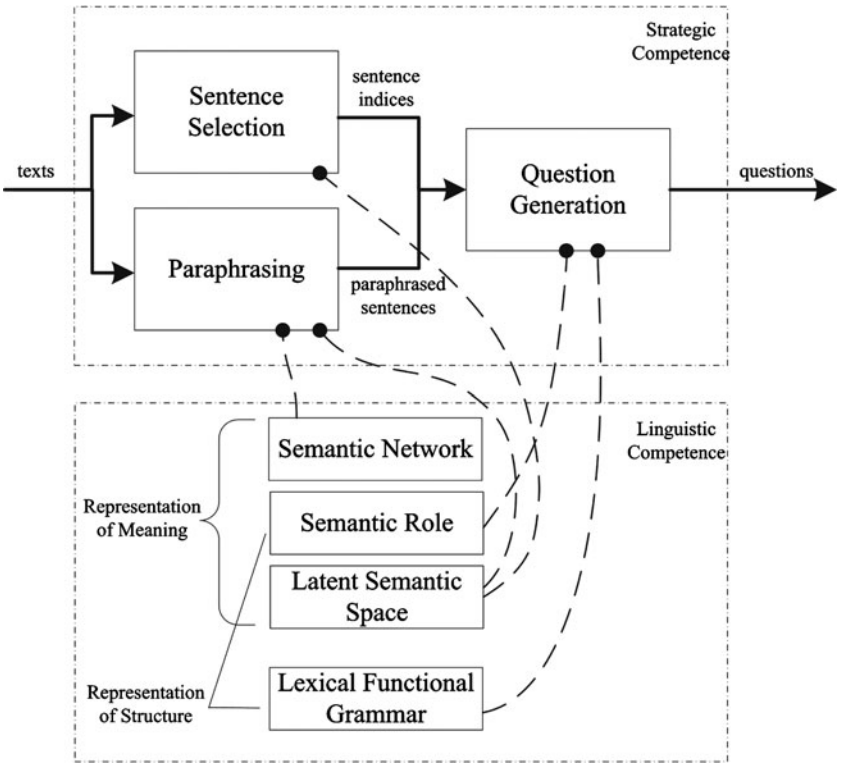


Fig. 1. Strategic and linguistic framework for automatic item generation.

3.1 Strategic and linguistic framework

Item writing is a type of language use which involves interaction among language knowledge, strategic competence, topical knowledge and the personal characteristics of a language user (Bachman and Palmer 1996). Our approach did not imitate the topical knowledge or the personal characteristics for three reasons. First, even though topical knowledge is crucial for inducing information not presented in the text, this knowledge was too complex to capture with current NLP techniques. Second, to prevent topical biases among test-takers with different educational backgrounds, test-designers usually restrict the stimulus texts of reading comprehension assessments to containing only general topics (Bachman and Palmer 1996). Therefore, specialized topical knowledge was unnecessary for our task. Third, eliminating personal characteristics during item generation aided in maintaining the consistency of items (Bormuth 1970).

As such, our automatic item generator had two competencies (denoted by the two dashed squares in Figure 1). The strategic competence related to how linguistic elements were implemented to generate items. This competence was composed of three independent modules: sentence selection, paraphrasing and question generation. The linguistic competence related to the knowledge of language. This competence represented the meaning of language with various linguistic models: at the word level, a semantic network, WordNet (Miller 1995), was used to exploit

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

manually-coded fine-grained relations between word senses; at the clause level, semantic roles were used to capture predicate-argument relations; and at the discourse level, latent semantic space (Dumais *et al.* 1988) was used to represent the meaning of large linguistic units. The structure of language was represented by LFG. Semantic roles were involved in constructing LFG structures. The LFG structures served as an efficient interface between syntax and semantics.

The following sections expand on two important components of the linguistic competence: the latent semantic space model and the LFG framework. Each component of the strategic competence will then be discussed in turn.

3.2 Latent semantic space

LSA is the classic approach to building latent semantic space (Dumais *et al.* 1988). In general, LSA constructs a word-context co-occurrence matrix first and reduces the dimensions of the matrix through Singular Value Decomposition (SVD). The resulting vector space is believed to represent latent semantic meaning (Landauer and Dumais 1997), whilst the cosine distance between the vectors is used to denote meaning similarity. However, LSA is unsuitable for online input, because the word-context matrix and SVD need to be constructed and implemented whenever there is a new set of input documents. Such processing can be time-consuming. Also, the exact meaning of each dimension of the latent semantic space changes across different inputs, which makes it impossible to compare contexts from different inputs. Furthermore, if the number of contexts and words is small, the resulting space may be meaningless.

We improved this method by combining it with the model of Hyperspace Analogue to Language (HAL) (Burgess, Livesay and Lund 1998). Instead of the word-context matrix, HAL constructs a weighted word-word co-occurrence matrix to obtain distributional representations of words. As such, we applied SVD to a word-word matrix, and then obtained the vector of a context (e.g., sentence) by calculating the geometric center of its word vectors. Thus, the co-occurrence information of a large external corpus was utilized, whilst the latent semantic representations of words were fixed and the construction of latent semantic space did not rely on the input. As a result, only simple computations were needed to calculate the vectors of context, which accelerated online processing.

3.3 Lexical functional grammar

Question generation involves transformation of syntactic structure. An influential idea about interrogative transformation came from Government and Binding (GB) theory (Chomsky 1977), which insists that a *wh*-question derives from a *wh*-movement on the deep structure of a canonical declarative sentence. However, GB theory has not been used to implement QGT directly. There are two reasons for that. First, GB theory is syntactocentric (Jackendoff 2010). That is, the theory postulates that syntax is central to the organization of all other linguistic aspects including morphology, phonology, semantics and pragmatics. This reduces the

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

cohesion of each linguistic aspect, making it difficult to develop and use modular resources. Second, the syntactic deep structure is difficult to parse. Computational implementation of GB theory is intractable. To date, no parser has successfully handled the transformational grammar (Sag and Wasow 2011).

To achieve efficiency, most prior studies on question generation substituted surface structure for the deep structure as the basis for transformation (Kunichika *et al.* 2003, Gates 2008, Ali *et al.* 2010, Heilman and Smith 2010, Kalady *et al.* 2010, Pal *et al.* 2010, Varga and Ha 2010, Bernhard *et al.* 2012, Curto *et al.* 2012, Liu *et al.* 2012). Unless very specific rules were defined, these syntax-based approaches were unable to produce flexible expression. Take grammatical voice as an example: given a passive sentence (e.g., *the cat was chased by a little boy*), a syntax-based approach may generate a passive question (e.g., *by whom was the cat chased*), rather than an active question (e.g., *who chased the cat*), which is more natural.

In contrast to GB theory, non-transformational grammars such as HPSG assume that each linguistic aspect is an independent generative component, and that the linguistic aspects interact with each other through interface rules (Jackendoff 2010). Such linguistic architecture enables unification-based parsing in accordance with features and constraints from multiple linguistic aspects, reducing ambiguity and increasing efficiency. Nevertheless, parsing sentences following non-transformational grammars exactly remain computationally expensive. For example, the question generation approach based on HPSG (Yao *et al.* 2012) is considered to be slow.

As such, this study uses LFG (Kaplan and Bresnan 1982) to strike a balance between the flexibility and the computational complexity of structural manipulation. LFG analyzes syntax in two structures: the constituent structure (c-structure) and the functional structure (f-structure). The former is a phrase structure storing the information of syntactic categories as well as the dominance and precedence relations among the phrasal components. The latter is a typed feature structure containing functional features such as subject, predicate and object, and abstract grammatical features such as tense, aspect, number and case. The f-structure offers an interface between semantics and syntax. A series of constraints, i.e., the well-formedness conditions, are supposed to be satisfied during construction of the f-structure. However, checking these conditions can be computationally expensive. As such, this study introduced some heuristic rules to reduce the computational complexity of constructing the f-structure:

(1) The functional features of the f-structure were filled by sentence components according to their semantic roles. Following the theory of binary partition of semantic roles (Dowty 1991), we mapped proto-actors to subjects, and proto-patients to objects (Section 3.6.1 explains the specific mapping rules). Since each functional feature can further expand into an f-structure, predicate-semantic relations in different clauses of a sentence were stored in an embedded f-structure.
(2) The abstract grammatical features in the f-structure were decided according to the morphological information that passed along the hierarchy of the c-structure through endocentric projection. That is, the head of each phrase

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

structure was identified with heuristic rules; the phrase structures then inherited the grammatical information from their heads. When the words of a phrase structure were filled into a function of the f-structure, their grammatical information assisted in making decisions about the abstract grammatical features of the particular f-structure. At this stage, checking of grammatical well-formedness was omitted. Instead, we decided the value of each abstract grammatical feature by examining relevant functional features, and adopting the marked grammatical value (see Section 3.6.1 for an illustration). This relaxation of constraints assumed that the input text was grammatical, and compromised the ability of LFG to reject ungrammatical input. Nevertheless, the relaxation reduced the computational cost and the influence of errors from morphological analyzers.

Moreover, we added to the f-structure a semantic feature which recorded the supersenses (Ciaramita and Altun 2006) of the head words. A supersense refers to the basic logical category of a word. For example, the supersense of *woman* is Person. The supersense feature was used to decide *wh*-words during question generation (see Section 3.6).

Thus, a semantics-driven approach to question generation can be established. Different functions in the f-structure can be selected as the question focus, whilst sentence structures can be manipulated under the constraints of abstract grammatical features.

3.4 Sentence selection

The goal of sentence selection was to identify a set of sentences that represented the content of the input text. This study operationalized the representativeness as consisting of three dimensions: keyness, completeness and independence.

3.4.1 Keyness

As important ideas are usually emphasized redundantly in a text, we defined the keyness of a sentence as:

$$d = \frac{\sum_{i=1}^{n} w_i}{n}, \tag{1}$$

where n is the number of words (excluding stop words) in a sentence; w_i denotes the number of times the ith word occurs in the text. Equation (1) assumes that a sentence containing denser frequent words is likely to contain more of the important meaning of the text.

3.4.2 Completeness

From the perspective of content distribution, a discourse can be deconstructed into several suprasentential units. A suprasentential unit is the aggregation of sentences that convey relatively independent sub-meaning (Hřebček 1993). To test whether test-takers fully grasp a text, all the sub-meanings should be examined. Normally, a

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

paragraph can be viewed as a suprasentential unit defined by its author. To ensure that the selected sentences spread over different paragraphs, we set the number of sentences that should be selected from each paragraph as:

$$n_i = \frac{N n_{si}}{\sum_{j=1}^{p} n_{sj}}, \tag{2}$$

where N is the number of items set by test-designers; n_i denotes the quota of selected sentences assigned to the ith paragraph ($i = 1, 2, \ldots, p$) which has n_{si} sentences.

3.4.3 Independence

Redundancy is almost inevitable in any discourse. Therefore, following the criteria of keyness and completeness alone, sentences of similar meaning may be chosen. As the number of items in a reading assessment is limited, the selected sentences should cover different aspects of text content. This means that the meaning of sentences should be as independent from each other as possible.

This study used the reciprocal of meaning similarity to measure meaning distance. As mentioned earlier, context meaning was represented by latent semantic space vectors. The meaning distance between two sentence vectors $\vec{s_1}$ and $\vec{s_2}$ was calculated as:

$$I(\vec{s_1}, \vec{s_2}) = \frac{1}{sim(\vec{s_1}, \vec{s_2})} = \frac{1}{\cos(\vec{s_1}, \vec{s_2})} = \frac{\|\vec{s_1}\| \, \|\vec{s_2}\|}{\vec{s_1}\vec{s_2}}. \tag{3}$$

A threshold was then set for deciding whether the meanings of two sentences were truly independent. If the meaning distance between two sentences was lower than the threshold, these sentences could not be selected at the same time.

Based on the aforementioned criteria, our sentence selection method contained three steps:

(1) Given an input text, the sentences were ranked by their keyness.
(2) The quota of selected sentences for each paragraph was calculated.
(3) Starting from the sentence of highest keyness, we selected sentences under the following constraints: a candidate sentence was rejected if its paragraph had already met the quota, or if its meaning was similar to that of a sentence which had been selected.

3.5 Paraphrasing

This study implemented two kinds of paraphrasing: thesaurus-based synonym substitution and pronoun substitution.

3.5.1 Synonym substitution

Retrieving the synonym of a polyseme requires identification of the specific sense of the polyseme. However, the accuracy of current techniques for fine-grained word sense disambiguation (WSD) is far from satisfactory. The highest precision achieved on SemEval-2007 data, at the time of our study, was 59 per cent (Navigli and

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Ponzetto 2012). We adopted the extended Lesk algorithm (Banerjee and Pedersen 2002), a light knowledge-based approach to WSD. The basic idea was to find a sense whose gloss and the glosses of the concepts related to this sense (e.g., hypernym, holonym, etc.) were most similar to the context of the target word. The latent semantic space model was used again to calculate the similarity. As the accuracy of the algorithm was only 34.6 per cent on Senseval-2 data for fine-grained WSD, we improved the precision of synonym substitution by:

(1) Constructing multiword units for WSD. A multiword unit tends to have fewer senses than its components. For example, while *get* has up to thirty-six senses in WordNet, *get rid of* has only three. Constructing multiword units restricted the candidate senses and helped to increase the precision of WSD.
(2) Setting a threshold on sense frequency. In second language reading comprehension assessment, which mainly targets the low-intermediate proficiency level, the words of a stimulus text usually take the most frequent senses. Therefore, a frequent sense is even more likely to be correct than a rare sense. As WordNet ranks the senses of a word according to their frequencies in a reference corpus, we set a fixed threshold on the sense numbers to restrict sense frequencies. If WSD decided that a word was being used with a sense number higher than the threshold, no synonym substitution was made.
(3) Setting a lexicon boundary for candidate synonyms. As a language assessment should not include words beyond the proficiency level of its target test-takers, a synonym can be rejected if it is not within the lexical range of the assessment. This lexical range, as well as the aforementioned threshold of sense frequency, should be decided by the proficiency level of the target assessment. Section 4.1.2 describes the setting of the lexical range and sense frequency threshold in our experiments.

When a word had several acceptable synonyms, we selected one of them randomly. Each sentence was paraphrased only once, with all the words that had acceptable synonyms being paraphrased.

3.5.2 Pronoun substitution

When a sentence is split from its context and transformed into a question directly, ambiguity may occur if the sentence has pronouns whose antecedents are located in other sentences. To reduce ambiguity, some pronouns should be replaced by their antecedents. Specifically, pronoun substitution recombines the linguistic elements of a text to generate a paraphrase.

In this study, we used emPronoun (Charniak and Elsner 2009) to retrieve the antecedents of personal, possessive and reflexive pronouns. A pronoun was then replaced by its antecedent if none of the following conditions was present:

(1) The antecedent of the pronoun was in the same clause;
(2) Another pronoun of the same antecedent was present before the target pronoun in the sentence.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Table 1. *PropBank semantic role tag set*

Argument			
A0	Proto-agent	A1	Proto-patient
A2-A5	Other arguments		
Adjunct			
AM-LOC	Location	AM-CAU	Cause
AM-TMP	Time	AM-EXT	Extent
AM-PNC	Purpose	AM-DIS	Discourse connectives
AM-MNR	Manner	AM-ADV	General-purpose
AM-NEG	Negation marker	AM-MOD	Modal verb
AM-DIR	Direction		

3.6 Question generation

This module transformed declarative sentences into questions. For each sentence, the module constructed a c-structure and an f-structure first. The module then selected a question focus on the f-structure, and chose a *wh*-word. A question was then generated according to the f-structure.

This section explains the question generation process in detail with the following example sentence: *the study shows that after school begins children are far more influenced by parents.*

3.6.1 Constructing c-structures and f-structures

First, we extracted syntactic and semantic information with existing parsers:

(1) A c-structure was extracted by a probabilistic context-free grammar (PCFG) parser (Charniak 2001).
(2) Morphological information was obtained by a finite-state morphological analyzer (Minnen, Carroll and Pearce 2001).
(3) Semantic roles were assigned by a semantic role labeler (Punyakanok, Roth and Yih 2008) trained on PropBank (Palmer, Gildea and Kingsbury 2005). Table 1 lists the semantic role tag set. A0-A5 are arguments governed by predicates. A0 and A1 correspond with proto-agent and proto-patient respectively. Other semantic roles are adjuncts which are optional in predicate-argument structures. These adjuncts modify the meaning of predicts, such as specifying where an action denoted by a predicate occurs (AM-LOC). Table 2 shows the semantic roles assigned to the components of the example sentence. There are three sets of semantic roles. A sentence segment may serve different semantic roles in different predicate-argument structures.
(4) Supersenses were assigned by a supersense tagger (Ciaramita and Altun 2006) following the BBN entity tag set (Weischedel and Brunstein 2005). The full tag

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Table 2. *Supersenses and semantic roles for the example sentence*

Sentence element	Supersense	Semantic roles		
		1	2	3
The		A0		
study		A0		
shows		V		
that		A1		
after		|		AM-TMP
school	Organization	|	A1	|
begins		|	V	AM-TMP
children	Person	|		A1
are		|		
far		|		AM-MNR
more		|		AM-MNR
influenced		|		V
by		|		A0
parents	Person	A1		A0

set contains 105 tags, which are grouped into 24 categories including Person, Location and Organization. For the sake of brevity, this paper refers to the tags by the names of their categories. Table 2 shows the supersenses assigned to the words of the example sentence.

The semantic roles were then mapped to the f-structure according to heuristic rules. This study treated the numbers in the argument tags as an 'agent-like degree' indicator. The subject function of the f-structure, SUBJ, was mapped with the lowest numbered argument, or the most agent-like argument. For example, when A0 and A1 were present, we mapped A0 to SUBJ, and A1 to OBJ. But if A0 was absent while A1 existed, then A1 was mapped to SUBJ. Moreover, when A2, which usually served as a second object, was present in addition to A0 and A1, we mapped A2 to the oblique function OBL_θ if A2 began with a preposition, or to the secondary object function OBJ_θ otherwise. Adjuncts, as they were not governed by predicates, were mapped to loose adjunct functions of ADJ's. For example, AM-TMP was mapped to the adjunct function $\mathrm{ADJ}_{\mathrm{TMP}}$.

Next, the abstract grammatical information of each f-structure was decided according to the morphological information of relevant functions. For instance, there are two cues for deciding the number of the subject *the study* in the example sentence: one is the number of the head *study*, and the other is the number indicated by the predicate *shows* ('show + singular person + present tense'). For predicates, singularity is marked, whereas for noun phrases, plurality is marked (by '-s', '-es' or irregular suffixes). We specified that predicates were checked first when deciding the feature of number (NUM). As *shows* is marked with a singular suffix '-s', the grammatical feature NUM is determined to be singular (SG). Figure 2 shows the f-structure constructed for the example sentence.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

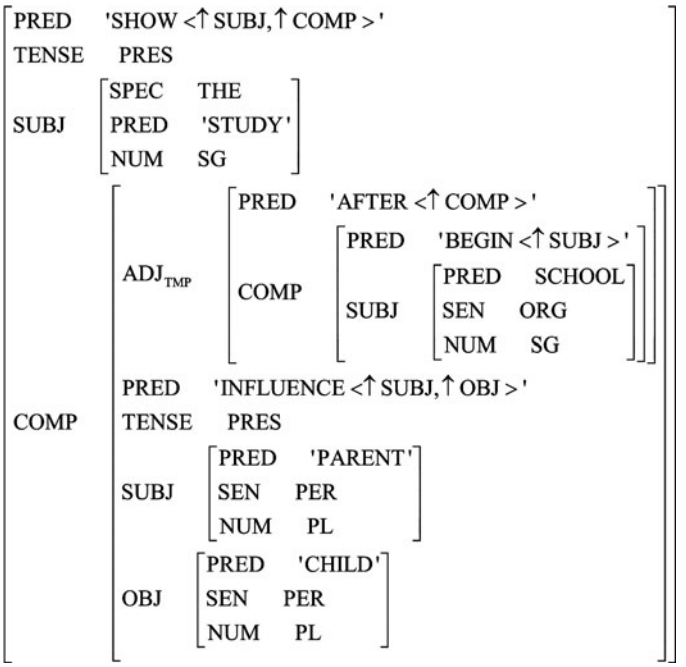


Fig. 2. Example f-structure.

3.6.2 Picking question foci and deciding wh-words

Other than predicates, any function in the f-structure can be picked as a question focus. This means that several questions can be generated from one source sentence. There has been research into what question focus is desirable for gap-filling items (Becker *et al.* 2012). Results showed that good candidates for question foci span all kinds of semantic roles. Our approach did not consider selection of the best question focus. Rather, the approach produced several questions for each selected sentence, providing users (i.e., human item writers) with opportunities to select and modify the questions.

After a question focus was picked, the *wh*-word was chosen according to the function of the question focus and the supersense of the head word. The major conditions for deciding *wh*-words are displayed in Figure 4. We also defined a supplementary condition: *who* was chosen if the head word was a personal pronoun. Figure 3 shows the change of the f-structure after picking the COMP of the predicate (*shows*) as the question focus: the function becomes QFOCUS, with a *wh*-word *what* assigned to it.

3.6.3 Generating questions based on f-structures

In general, this stage involved extracting values from the functions of the f-structure, inflecting them according to the types of the functions and abstract grammatical features, and pushing the inflected values into a question array. A question was then generated by connecting the array elements.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

$$\begin{bmatrix} \text{PRED} & \text{'SHOW } <\uparrow \text{SUBJ}, \uparrow \text{COMP} >\text{'} \\ \text{TENSE} & \text{PRES} \\ \text{SUBJ} & \begin{bmatrix} \text{SPEC} & \text{THE} \\ \text{PRED} & \text{'STUDY'} \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \text{QFOCUS} & \text{WHAT} \end{bmatrix}$$

Fig. 3. Example of interrogatively transformed f-structure.

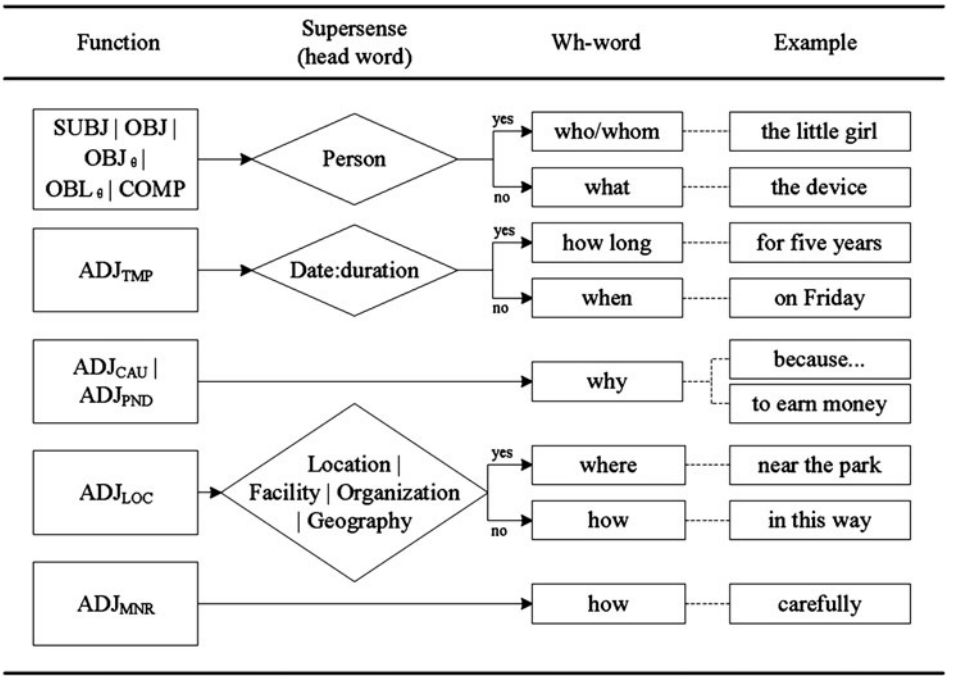


Fig. 4. Major conditions for decision of *wh*-words.

The values were pushed in the following order:

(1) QFOCUS, i.e., the *wh*-word. If the QFOCUS was in the position of SUBJ, Steps 2 and 3 were skipped.
(2) The first auxiliary. The values of TENSE and MODALITY were checked first. If the TENSE was 'Perfect' or 'Future', or the MODALITY had a value (e.g., 'can' or 'may' etc.), an inflected 'have', 'will', or the respective modal verb was pushed as the auxiliary. Then, the value of ASPECT was checked. If no auxiliary had been pushed, an inflected 'is' or 'do' was pushed. In case of multiple auxiliaries (e.g., 'will be doing'), only the first auxiliary is pushed at this step, while the remaining ones are pushed at step 4.
(3) SUBJ. The value of SUBJ was turned to nominative case if it was an objective case pronoun.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

(4) The remaining auxiliaries, if they existed.
(5) The remaining functions except ADJ_{TMP}, ADJ_{LOC}, ADJ_{DIS} (ADJ_{DIS} was left out during question generation). They were pushed into the question array following their original order in the source sentence. The value of PRED was inflected according to abstract grammatical information, and the value of OBJ was turned to objective case if it was a nominative case pronoun.
(6) ADJ_{LOC} and ADJ_{TMP}, if they existed.

Following the aforementioned procedure, the approach generates a question from the f-structure of Figure 3: *what does the study show?* By choosing different question foci, the approach generates five more questions as follows:

When do parents far more influence children?
How do parents influence children after school begins?
Who far more influences children after school begins?
Whom do parents far more influence?
What shows that after school begins children are far more influenced by parents?

4 Experiment and evaluation

Accordingly, we built a system called ASIG (**A**utomatic **S**hort Answer **I**tem **G**eneration System) in Perl. The following section presents our evaluation of the system.

4.1 Evaluation method

Since our approach consisted of three aspects (i.e., sentence selection, paraphrasing and question generation), we first fed reading materials to each module directly to evaluate the independent performance of each module. We then compared the short answer items generated by the whole system with those written by human experts with regard to each aspect. Finally, we compared the performance of post-edited items and the items written by human experts in a reading comprehension assessment.

4.1.1 Input material

We selected reading materials from the College English Test Band 4 (CET4) as input. Administered by the Ministry of Education (MoE) of China, CET4 is developed by testing experts in China. Therefore, its original test items provided a reliable reference for our study. CET4 mainly assesses low-intermediate English proficiency, which corresponds to the B1 level in the Common European Framework of Reference (CEFR) for languages (Ishikawa 2013).

There have been two test formats for the reading comprehension section of CET4. One follows a stimulus text with multiple choice items, the other with short answer items. The first format has been predominant. However, because our goal was to

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

compare the short answer items generated by the system and humans, we chose only the reading materials in the second test format.

There were seven passages in the second test format in CET4 from 1989 to 2006. Each of them contained 250 to 400 words, and was followed by 5-8 short answer items. Some of the items took a cloze form, and therefore could not be compared with the questions generated by ASIG directly. Nevertheless, we took the source sentences of the cloze items into account when comparing the source sentences selected by humans and ASIG. This was because all the source sentences for a passage were selected as a whole.

4.1.2 System setting

We set the system to choose four sentences from each input text, because the input texts were shorter than 400 words and the key ideas of these texts could be represented by 3-4 sentences. To obtain word vectors, we first built a word-word matrix from a corpus containing all the reading passages in CET4 during 1989-2010. The original elements of the matrix were frequency counts of co-occurrence. The matrix was then normalized by TF-IDF method (i.e., the value of element (i, j) was divided by the sum of all values in row i and the number of rows where the value of column j was non-zero). Following this, the matrix was reduced to a 100-dimensional latent semantic space by SVD. The specific constraints during paraphrasing were set as follows: the sense frequency threshold was the top-two most frequent sense in WordNet; and the lexicon boundary was the vocabulary list of the CET4 Test Syllabus.

4.1.3 Procedure

Experiment 1: evaluation of sentence selection module. We evaluated the degree of keyness, completeness and independence of the selected sentences. The baseline was obtained by dividing each input passage into four segments of nearly the same number of sentences, and selecting the first sentence of each segment. Two college English teachers who had at least five years of item writing experience jointly evaluated the selected sentences. First, they collaborated to choose a list of key sentences and suprasentential units for each passage. Here, we assumed there was an optimal choice of the key sentences and suprasentential units. While it may not apply to all reading materials, the assumption was reasonable for the reading texts of CET4, whose structure was clear as a result of adaptation to low-intermediate level language assessment. The selected sentences were then compared to the list. For keyness, a sentence scored one if it was on the key sentence list. For independence, a sentence scored one unless there was another sentence, whose keyness was higher, selected from the same suprasentential unit. The completeness of the selected sentences of a text was scored by subtracting the number of unrepresented suprasentential units from the total number of selected sentences (i.e., four). The information about which algorithm was involved in selecting each set of the sentences was unknown by the evaluators.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Experiment 2: evaluation of paraphrasing module. We evaluated the precision and coverage of paraphrasing. All the generated paraphrases were presented to a graduate student in applied linguistics, who then decided whether a paraphrase was correct in context. Following this, the precision of paraphrasing was calculated as the proportion of correct paraphrases, and the coverage of paraphrasing was measured by the number of paraphrases per sentence.

Experiment 3: evaluation of question generation module. We evaluated the efficiency, productivity, variety and quality of question generation. Two of the top performing question generation systems were used for comparison: one was the syntax-based QG system built by Heilman and Smith (2010) (hereafter referred to as SynQG), whilst the other was the semantics-based MrsQG built by Yao *et al.* (2012). The depth of the linguistic information used by ASIG was somewhere in between. These systems and ASIG were intended to generate questions of similarly great variety. SynQG and MrsQG each has a module that ranks and filters the questions. While these modules can reduce deficient questions (e.g., ungrammatical questions), they obscure the productivity of the systems and their ability to identify question foci. To investigate these aspects of system performance, we disabled the ranking and filtering modules.[1] In the experiment, the systems were run in the same hardware setting with the same input. Their processing time, number of output questions and distributions of question categories were recorded and extracted. Four graduate students in applied linguistics rated the quality of the questions according to a 9-point evaluation framework proposed by Heilman and Smith (2010) (see Table 7 in section 4.2.3). First, three raters independently classified each question according to its most severe deficiency. Second, each question was assigned the category winning at least two out of the three votes or, if the raters all disagreed, a category determined by the fourth rater. The questions from all systems were mixed and the information about which system generated them was unknown by the raters.

Experiment 4: evaluation of the whole system. We compared the items generated by ASIG against the original items of CET4. First, the two evaluators for Experiment 1 collaborated again to identify the source sentences of the original CET4 items according to the following principles: the source sentences should contain answers to the items; more than one sentence may be identified as the source of an item, and a sentence may be identified as the source of several items. The overlaps between the source sentences and the sentences selected by ASIG were then calculated. Second, the four evaluators for Experiment 3 rated the quality of the paraphrased questions on the 9-point scale again. Third, one of the four evaluators qualitatively compared the paraphrased questions with the original CET4 items.

Experiment 5: evaluation of post-edited items. We post-edited the items generated by ASIG, and evaluated the correlation between these items and the original CET4

[1] MrsQG has another ranking module which chooses the best question generated from a semantic structure. This module was not disabled in our experiment.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

items in a reading comprehension assessment. We also compared the difficulty and discrimination power of the two sets of items. The assessment paper contained six passages. The passage with the lowest number of original question items was removed. Each passage was tested by four original CET4 questions and four post-edited questions. The questions were prepared in the following way: first, we randomly selected four questions from the original CET4 items for each passage. For each of the three passages with only three original question items, we supplemented a question by randomly selecting and adapting an original cloze item to the question form. Second, the questions generated by ASIG were post-edited by a graduate student in applied linguistics who had no item writing experience. The post-editor was presented with the sentences and questions selected and generated by ASIG. She then chose and modified a question for each selected sentence. The goal of modification was to make the questions appropriate for assessing reading comprehension whilst keeping the modification to a minimum level. Thus, twenty-four questions were created. Among these questions, 25 per cent received no modification;[2] 33 per cent involved insertion, deletion and reordering in no more than two positions, and maintained the basic grammatical structures (e.g., *whom will employers exclusively look at* was modified into *whom will employers of English teachers look for*); 42 per cent involved modification beyond those defined in the previous group.

Due to the overlap of the source sentences, the post-edited items and the original CET4 items for the same passages should not be presented simultaneously. Therefore, our assessment had two sections; each section presented either the post-edited items or the original CET4 items for each passage. To control for the effect of testing order, we designed two versions of test papers: version A allocated the post-edited items to the odd-numbered passages, and the original CET4 items to the even-numbered passages; version B was the opposite. Half of the test-takers took version A for the first section and version B for the second section, whereas the other half took the papers in reverse order. The first section lasted 75 minutes, after which the test papers were collected and the test papers of the second section were handed out. The second section lasted 45 minutes. Forty-two college students in China participated in the in-class assessment.

After the assessment, the post-editor scored the answers. The possible scores for an item were 0, 0.5, 1, 1.5 and 2. First, we summed up the total scores obtained by each test-taker on the post-edited items and the original CET4 items. We then calculated the correlation between the two sets of scores. Second, we analyzed the parameters of the items based on Classical Test Theory (Ebel 1972). The difficulty index of an item was calculated by:

$$Diff = \frac{S}{NM}, \tag{4}$$

[2] This percentage was lower than the percentage of acceptable questions reported in section 4.2.4 (54.1 per cent) for two reasons. First, the post-editor happened to select more questions with deficiencies. Second, even though an acceptable question did not have any of the deficiencies defined in the 9-point evaluation framework in Experiment 3 and 4, it may still suffer from insufficient paraphrases, which is discussed in section 4.3

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Table 3. *Quality of selected sentences*[3]

	Keyness		Completeness		Independence	
Text	Baseline	ASIG	Baseline	ASIG	Baseline	ASIG
CET4021	2	4	2	3	2	3
CET4031	3	3	2	4	2	4
CET4032	2	4	2	3	2	3
CET4951	1	2	3	3	3	3
CET4971	1	2	3	4	4	4
CET4991	3	4	3	4	3	3
CET4992	1	4	2	4	3	4
Total	13	23	17	25	19	24

where S is the sum of the scores obtained by all test-takers on the item; N denotes the total number of test-takers, and M represents the maximum individual score for the item. Next, we calculated the total test scores of each test-taker, and then classified the top 27 per cent of the test-takers as the upper group, and the bottom 27 per cent the lower group. The discrimination power was then calculated by:

$$Disc = \frac{S_u - S_l}{N_g M}, \tag{5}$$

where S_u and S_l are the total scores obtained by the upper group and the lower group on the item respectively; N_g denotes the number of test-takers in each group (i.e., eleven), and M represents the maximum individual score for the item.

4.2 Experimental results

This section presents the experimental results and performance analyses.

4.2.1 Performance of sentence selection module

The number of key sentences pre-identified by the two evaluators for each input text ranged from seven to ten, whereas the number of suprasentential units was three or four. Table 3 shows the scores for the sentences selected by different algorithms. Even though no inter-rater reliability could be reported for these scores, our evaluation method ensured that there was no bias towards either algorithm.

As Table 3 shows, ASIG scored higher than the baseline in almost every dimension on all input texts. The greatest difference lay in keyness, where in total ASIG scored twenty-three (out of twenty-eight) and the baseline only thirteen. This demonstrated the validity of selecting important information based on word frequency density. The baseline performed better on completeness and independence than on keyness,

[3] The maximum score on each dimension is four for each text, and twenty-eight for all the texts.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Table 4. *Precision and coverage of synonym paraphrasing*

Text	# of paraphrases	Precision (per cent)	Coverage (# of paraphrases per sentence)
CET4021	30	86.7	2.5
CET4031	51	88.2	2.8
CET4032	41	78.0	2.0
CET4951	34	79.4	1.8
CET4971	41	68.3	2.6
CET4991	41	85.4	3.4
CET4992	40	82.5	2.5
Total	278	81.3	2.4

because it ensured that the sentences were selected from different parts of the text. Nevertheless, ASIG still scored higher than the baseline on completeness and independence. This proved the effectiveness of selecting sentences according to paragraph length and the constraint on the similarity in meaning[4] between the selected sentences.

Meanwhile, we calculated the average length of the sentences selected by each algorithm and the average length of all input sentences. The sentences selected by ASIG were longer than the average of input sentences (21.1 versus 18.2 words per sentence, $p > 0.05$). The baseline sentences were also longer (19.4 words per sentence, $p > 0.05$). Although the greater length of ASIG's sentences was insignificant, it illustrated that ASIG did not necessarily select short summary sentences.

4.2.2 Performance of synonym paraphrasing module

Table 4 shows the precision and coverage of paraphrasing for each input text. As this section was evaluated by only one evaluator, no inter-rater agreement could be reported.

The average precision of synonym paraphrasing was 81.3 per cent, which was much higher than the precision of the extended Lesk WSD method (Banerjee and Pedersen 2002). This showed the effectiveness of sense constraints. Even though restricting senses reduced the number of paraphrases, Table 4 demonstrates that on average a sentence still had 2.4 content words or phrases replaced by their synonyms.

The wrong synonym paraphrases were mainly caused by errors in WSD and differences between synonyms. An example of the latter was: *what is most important (*of import) for the decision?* The original word *important*, an adjective that can be

[4] Recently developed neural-network semantic space models built with word2vec (Mikolov *et al.* 2013) and GloVe (Pennington, Socher and Manning 2014) performed better in many similarity-related tasks than LSA. In the future it would be interesting to investigate whether using the newer models or adapting the hyper-parameters of these models to LSA (Levy, Goldberg and Dagan 2015) can achieve better results on sentence selection and WSD for AIG.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Table 5. *Processing speed of question generation*

System	MrsQG	SynQG	ASIG
Processing time (min.)	451.47	2.75	7.30
# of questions	483 (399)[5]	325	459
Processing speed (# of questions per sec.)	0.018 (0.015)	1.97	1.05

either predicative or attributive, was used as a predicative. However, the word was replaced by a synonym phrase *of import*, which can only serve as a postnominal attributive. Such errors indicated that to improve the precision of thesaurus-based paraphrasing, subtle differences between the usages of synonyms should be resolved.

4.2.3 Performance of question generation module

Table 5 shows the processing speed of ASIG (1.05 questions per second), which was lower than that of the syntax-based SynQG (1.97 questions per second), but much higher than that of the semantics-based MrsQG (0.018 questions per second). This demonstrated that the processing speed decreased with the depth of the linguistic information utilized.

Meanwhile, Table 5 depicts that the semantics-based approaches generated many more questions than the syntax-based one. ASIG and MrsQG generated more than 450 questions, while SynQG generated only 325 questions. The reasons for the production of more sentences by MrsQG and ASIG were different. MrsQG over-generated questions radically: given an input sentence, it produced several parse structures (Yao *et al.* 2012), as opposed to a single parse structure that SynQG and ASIG produced. Based on some of the parse structures, MrsQG generated similar questions that differed from each other only in the inflection or position of a word. When these nearly identical questions were removed, only 399 questions were left. Still, some of the remaining questions targeted the same question foci and source sentences. These questions differed from one another either in grammatical voice, the presence or absence of certain auxiliary clauses, or the order of several words. Such variances reflected the flexibility of the natural language generation mechanism of MrsQG.

On the other hand, ASIG generated more sentences by recognizing more question foci. First, SynQG and MrsQG simplified sentences before generating questions, and discarded some clauses that contained potential question foci during the simplification. In contrast, ASIG built embedded f-structures to store functional relations in all clauses, and posed questions on any non-predicate function. In Case 1, for example, SynQG generated no questions for the bracketed subordinate clause, whereas ASIG generated the *who* and *whom* questions. Second, the functions of ASIG were mapped by semantic roles. The semantic role labeler recognized more

[5] The parenthesized numbers were achieved when nearly identical questions were filtered out.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Table 6. *Question category distribution*

Question depth	*wh*-word	MrsQG	SynQG	ASIG
Shallow	What	427	215	232
	Which	5	—	—
	Who	4	82	109
	Whom	2	—	16
	Whose	—	3	—
	When	5	15	44
	Where	4	7	5
Intermediate	How many	11	3	—
	How long	—	—	1
	How much	2	—	—
Deep	Why	11	—	7
	How	1	—	45
Uncertain	11	0	0	

sentence components of various syntactic structures and surface realizations than using hand-crafted rules alone. In Case 1, neither SynQG nor MrsQG generated any question concerning time, while ASIG recognized *until recently* as a temporal adjunct, and generated the *when* question for it.

Case 1

Source: [*Until recently*] *parents had been warned by educators* [*who asked them not to educate their children*]. (CET4021)

ASIG: *Who asked parents not to educate their children?*

Whom did parents not educate?

When had educators who asked parents not to educate their children warned parents?

Table 6 outlines the distribution of question categories. While *what* questions were the majority for each system, the distribution of question categories for ASIG was more balanced, as it had the highest proportion of *who* and *when* questions, and a larger number of deep questions (mainly *how* questions) than MrsQG. In contrast, even though MrsQG was intended to generate the most varied question categories, its distribution of question categories was highly skewed to *what* questions. Evaluation of question quality (Table 7) showed that this was due to its failure in choosing the correct *wh*-words for the question foci of common nouns. Based on name entity recognizers, MrsQG chose *wh*-words only for proper names. However, common nouns should not be neglected when choosing *wh*-words, because they are common components of question foci in second language reading comprehension assessment.

As for the evaluation of question quality, the inter-rater agreement among the first three raters, over all the sentences they evaluated, was Fleiss's $\kappa = 0.68$, which meant that there was 'substantial agreement' (Landis and Koch 1977). Table 7 shows the final evaluation results on question quality.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Table 7. *Quality of questions generated by three systems respectively*

Deficiency	Example	Percentage		
		MrsQG	SynQG	ASIG
Ungrammaticality	*What do many adult females select whether?*	25.1	20.0	6.3
Unintelligibility	*How do attitudes instruct English?*	10.8	2.5	4.1
Vagueness	*Where is the easiest place to do this?*	24.4	15.1	13.5
Wrong *wh*-word	*What faces a difficult decision?*	12.0	5.0	4.6
Format problem	*What did cities as time passed begin to do?*	4.8	4.3	2.0
Obvious answer	*Who educated their kids? (Parents)*	4.1	1.5	1.5
Missing an answer	*How did hotel rooms (Germans) handle to make a net income?*	1.9	6.2	9.6
Other error	*Who saw how much I (she) could have gained at home?*	1.2	3.4	2.2
Acceptable	*What has replaced the power of the school?*	19.5	42.2	56.2

In general, the quality of ASIG's questions was the highest. They had the highest rate of acceptable questions (56.2 per cent), and the lowest rates of severe deficiencies including ungrammaticality, vagueness, wrong *wh*-words and format problems. In particular, the percentage of ungrammatical questions from ASIG (6.3 per cent) was less than 1/4 of that for MrsQG (25.1 per cent) and 1/3 of that for SynQG (20.0 per cent). Investigation into the causes of the ungrammaticality illustrated that while all the systems suffered from some errors from syntactic and/or semantic parsers, the LFG-based transformation mechanism enabled ASIG to make fewer errors during question generation:

(1) ASIG recognized sentence components based on their semantic roles, which was more accurate than using syntactic rules. For example, in Case 2, SynQG mistook the bracketed phrase *of course* in the source sentence as a compositional noun phrase and generated an ungrammatical question for it. In contrast, ASIG correctly recognized *of course* as a discourse marker (ADJ_{DIS}), and did not make such errors.

Case 2

Source: [*Of course*], *children shouldn't be pushed to read by their parents.*

SynQG: *What should children not be pushed to read by their parents of?*

(2) ASIG transformed syntactic structures with fewer errors. ASIG involved substitution of QFOCUS only on the f-structure, and generated a question string by connecting the surface words extracted from various functions. In contrast, SynQG made more mistakes during complicated transformation of syntactic tree structure.

(3) ASIG tackled grammaticality during generation in a systematic way. When the value of a function was extracted, it was morphologically adjusted according

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

to its function and the abstract grammatical features in the f-structure. SynQG also defined rules to ensure grammaticality during generation, but they did not consider functional information. On the other hand, MrsQG first overgenerated questions for a given semantic structure, and then used a ranking and filtering module to select the most grammatically sound question. The accuracy of this module was only 64.28 per cent (Yao *et al.* 2012). As a result, MrsQG performed worst in terms of grammaticality.

Although vagueness was the major deficiency that the questions of ASIG suffered (13.5 per cent), its incidence rate was lower than those of SynQG (15.1 per cent) and MrsQG (24.4 per cent). MrsQG performed worst due to a lack of pronoun resolver. Some vagueness problems of ASIG and SynQG were due to the mistakes by their pronoun resolvers. For ASIG, altogether there were ninety-eight instances of pronoun substitution, among which 80.6 per cent of the pronouns were resolved correctly. Of these correctly-resolved pronouns, 94.9 per cent were replaced correctly by the heuristic rules in section 3.5.2. Nevertheless, the pronoun resolver of ASIG, emPronoun (Charniak and Elsner 2009), cannot resolve pronouns whose antecedents require world knowledge to identify. All three systems suffered from failure in capturing context information when generating questions for relatively short clauses. This was proved by the fact that the average length of the vague questions from each system (MrsQG: 4.9 words, ASIG: 6.4 words, SynQG: 7.8 words) was much lower than the average length of all the questions from the same system (MrsQG 8.9 words, ASIG: 8.9 words, SynQG: 11.0 words). The average question length of each semantics-based approach was shorter than that of the syntax-based approach. This meant that while the semantics-based approaches were able to generate more questions that targeted local semantic relations in short clauses, their questions were more likely to suffer from vagueness when the context issue was ignored. Case 3 illustrates a vague question generated by ASIG, where the modifiers in the parentheses were needed to restrict the objects or events:

Case 3

Source: *In 1896… The rules barred athletes who had ever received a $50 prize or an athletic scholarship or who had spent four weeks in a training camp.* (CET4032)

ASIG: *How long had athletes* (who were barred by the rules at the end of the nineteenth century at least) *spent in a training camp.*

The only category where ASIG performed worst was missing an answer (9.6 per cent). Errors from pronoun resolution were one of the major causes, as the wrong antecedents created new propositional meaning that did not exist in the input text. For instance, the question that missed an answer in Table 7 was generated by replacing a pronoun *they*, whose antecedent was *Germans*, with a wrong noun phrase of *hotel rooms*. The other major cause was *how* questions, of which ASIG generated forty-five, whereas other systems did not (SynQG) or generated only one (MrsQG). ASIG generated *how* questions for general-purpose adjuncts. This kind of adjunct covered various adverbial phrases and clauses, whereas *how* may not be a suitable *wh*-word for all of the cases. Case 4 illustrates such an error. The ambiguous *how*

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

question could be interpreted as questioning the manner in which *the game remained amateur*, or the mechanism behind this phenomenon. In either case, the intended question focus (i.e., the bracketed concessive clause) was not the answer. To solve this problem, more specific conditions should be considered when choosing *how* as a *wh*-word.

Case 4

Source: [AM-ADV *Even though the winners became national heroes*], *the games remained amateur for centuries.* (CET4032)
ASIG: *How did the games remain amateur for centuries?*

The unintelligibility rate of ASIG was between those of MrsQG and SynQG. A close examination showed that this deficiency was mainly caused by parser errors. As semantic parsing was less accurate than syntactic parsing, the semantics-based approaches produced more unintelligible questions than the syntax-based one.

4.2.4 Performance of the whole system

When the modules functioned together, the system presented 4.1 candidate questions on average for each selected source sentence.

We first compared the automatically-selected sentences with those selected by human testing experts (see Appendix for an example of the selection). Evaluation showed that among the twenty-eight sentences selected by ASIG, only twelve were also selected by the CET4 human testing experts. The result was only slightly better than the baseline sentence selection algorithm (ten overlaps). A close examination showed that this was caused by differences in the sentence selection criteria. Some of the original items of CET4, especially the cloze ones, focused on content details. In one of the two input texts upon which ASIG performed worst (CET4032), five out of the eight original items were in the cloze form, and only three out of the seven source sentences were in the key sentence list. In contrast, the sentences selected by ASIG were mainly key sentences (Table 3). In another text (CET4991), two original items asked the meanings of rare phrases; ASIG did not select the source sentences of these items due to their low word frequency density.

The quality of questions declined slightly when paraphrasing was involved. The rate of acceptable questions dropped from 56.2 per cent to 54.1 per cent as a result of an increase in the proportion of ungrammatical and unintelligible questions caused by wrong paraphrases. The drop of acceptable questions was much less than the product of the error rate of paraphrasing multiplied by the acceptable rate of non-paraphrased questions (10.5 per cent). There were two reasons. First, the system paraphrased texts before generating questions; some erroneous semantic relations caused by wrong paraphrasing were neglected during semantic role labeling. Second, some inappropriate paraphrases were concealed when they were selected as question foci.

Qualitative analysis showed that the original items produced by humans involved structural changes and flexible recombination of words from different parts of a text. Answering these questions requires inference. In contrast, even though the items

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Table 8. *Difficulty indices and discrimination power of items*

	Post-edited ASIG		Original CET4	
Item parameter	mean	stand. dev.	mean	stand. dev
Difficulty index	0.54	0.15	0.52	0.20
Discrimination power	0.36	0.20	0.40	0.18

produced by ASIG used some words and phrases that were different from those in the source clauses, they were still similar to the source clauses in terms of length and structure. The only components of the questions that came from sentences other than the sources were the antecedents of pronouns. As a result, the items were mainly intended to assess comprehension of single sentences. They targeted the ability to retrieve information right there in the stimulus texts, in contrast to those that require inference, world knowledge, or an opinion (Raphael 1982).

4.2.5 Test performance of post-edited items

There was a significant strong positive correlation between the scores obtained by individual test-takers on the post-edited items and the original CET4 items ($r = 0.821, p < 0.001$). This meant that the performance of a test-taker on the post-edited items could predict his or her performance on the original CET4 reading comprehension assessment. In other words, the post-edited items can test reading comprehension as well as the items written by the testing experts of CET4.

Table 8 shows that the average difficulty index of the post-edited ASIG questions was slightly higher than that of the original CET4 items (0.54 versus 0.52), which meant that the post-edited items were slightly easier than the latter. Nevertheless, this difference was insignificant ($t = 0.38, p > 0.05$). The difficulty indices of both sets of items were close to the recommended value of 0.5, which meant that both sets of items were suitable for the proficiency level of the test-takers.

The average discrimination power of the post-edited ASIG questions was slightly lower than that of the original CET4 items (0.36 versus 0.40). According to the classification framework of Ebel (1972), the average discrimination power of the post-edited ASIG questions was 'good', while that of the original CET4 items was 'excellent'. Nevertheless, this difference was also insignificant ($t = -0.68, p > 0.05$). Furthermore, the discrimination power of all items were positive, which meant that all of the items were useful in discriminating between the test-takers of higher and lower reading proficiency.

There were no significant differences in the difficulty indices ($F = 1.45, p > 0.05$) or discrimination power ($F = 0.18, p > 0.05$) of the post-edited questions across different degrees of human modification, even though the difficulty index slightly decreased with the modification degree, whereas the discrimination power of the

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

items without modification was slightly higher than that of the items with minor modification but slightly lower than that of the items with substantial modification.

Meanwhile, among the forty-two questions, four post-edited questions and nine CET4 questions were based on content details rather than key ideas. Nevertheless, there were no significant differences in the item parameters of the questions derived from different types of source sentences, except for the discrimination power of the post-edited questions, which was higher when the questions were based on details rather than key sentences (marginal significant at $F = 6.49, p = 0.018$). However, it should be cautioned that this marginally significant difference was observed on the small and unevenly distributed data (four detail-based questions versus twenty key-idea-based questions). Contrastingly, the discrimination power of CET4 questions was insignificantly higher when the questions were based on key sentences.

4.3 Discussion

The results show that our approach can select a set of non-redundant sentences that contain all the key ideas of a passage. Our sentence selection criteria differ from those used in developing CET4 reading comprehension assessment. If we aim to assess comprehension of the details of a text, different criteria for sentence selection should be set.

Paraphrasing at the word and phrase level can be accurate despite the low precision of fine-grained word sense disambiguation. This can be achieved by restricting senses during synonym substitution according to the prevalence of frequent senses in low-intermediate level language assessment. Paraphrasing stimulus texts before question generation also helps to limit the influence of paraphrasing errors. However, the restrictions on senses reduce the number of paraphrases in an item, which could jeopardize the ability of the item to test reading comprehension. To improve the coverage of paraphrasing, it may be helpful to use larger databases such as augmented Wordnets (Snow, Jurafsky and Ng 2006) and Paraphrase Database (PPDB) (Ganitkevitch, Van Durme and Callison-Burch 2013).[6]

The LFG-based question generation approach can generate high-quality questions. The superiority stems from the systematic utilization of both syntactic and semantic information. Based on the functions and abstract grammatical features in the f-structure, rules can be defined to ensure grammaticality during question generation. Meanwhile, the f-structure is built on semantic roles, which helps to recognize different sentence components more accurately than using syntactic rules. From a cognitive perspective, question asking is motivated by meaning rather than form. Therefore, generating questions based on the semantic roles is a better simulation of the mental mechanism of questioners. From an informational perspective, the semantic roles bring more information, as they embody what the semantic role labeler learned from a large human-annotated training data set. The semantic roles are predicted based on combinations of various parse features (Punyakanok

[6] We would like to thank an anonymous reviewer for suggesting the databases.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

et al. 2008), and provide more precise information for the identification of question foci than hand-crafted syntactic rules.

Furthermore, our question generation approach is productive and can generate varied questions. This is due to the integration of semantic roles and supersenses in the f-structure. The shallow semantic information enabled us to construct a flexible question template. The question foci are not restricted to proper names, whilst questions can be posed on adjunct functions whose forms are too varied to recognize by surface words or syntactic rules.

Moreover, the question generation approach is efficient. While it utilizes semantic information to present questions of higher quality, its running time is comparable to the syntax-based approach and is much less than the deep semantics-based approach. This shows that dropping the checking of grammatical well-formedness during construction of the f-structure can effectively reduce the computational cost during parsing.

The automatically generated items are far from perfect. More than 40 per cent of the items suffer from some deficiencies, among which vagueness is the most prevalent. This is mainly caused by the inability of the sentence-based approach to incorporate context information. Apart from pronoun resolution, there is a lack of flexible recombination of the components of different sentences. Also, the structure of the items is similar to that of their source sentences due to an absence of structural paraphrase. Most of the other errors in paraphrasing and interrogative transformation are caused by the deficiencies of the basic natural language processing techniques.

Although the automatically generated items are imperfect, they can facilitate human item writing. After limited post-editing by an inexperienced item writer, these items can assess reading comprehension, and have difficulty and discrimination power similar to those of the items created by human experts.

5 Conclusion and future work

This study presents an approach to automatically generating short answer questions for reading comprehension. The approach includes question generation, paraphrasing and sentence selection. Our experiments demonstrated that adopting LFG as the linguistic framework aided in improving the efficiency, productivity, variety and quality of question generation. It was also proven that the performance of paraphrasing and sentence selection can be enhanced by imposing constraints according to the sense distribution and discourse features of stimulus texts in second language reading comprehension assessment. The system can help inexperienced item writers to generate items that assess reading comprehension.

Our results revealed challenges for automatic question generation in relation to reading comprehension assessment. Unlike cloze items, which can be generated more easily by deleting words from original sentences, and can test reading comprehension and inter-sentential inference when they are inserted during reading (Mostow *et al.* 2004, Mostow and Jang 2012), questions generated to test reading comprehension require more sophisticated natural language processing that can integrate context

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

information and realize flexible paraphrasing. In future, this may be addressed by introducing information from discourse parsing, designing a better question generation framework that can combine information from several sentences, and exploring the use of structural paraphrasing techniques.

Acknowledgements

The majority of this work was performed when the first author was a master's student in the School of International Studies at Zhejiang University. The work was sponsored by the China National Social Science Foundation project 15BYY086 as well as Cambridge Commonwealth, European and International Trust.

References

Adamson, D., Bhartiya, D., Gujral, B., Kedia, R., Singh, A., and Rosé, C. P. 2013. Automatically generating discussion questions. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education*, Memphis, TN, USA. Berlin: Springer, pp. 81–90.

Agarwal, M., and Mannem, P. 2011. Automatic gap-fill question generation from text books. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, Portland, OR, USA. Stroudsburg, PA: ACL, pp. 56–64.

Ali, H., Chali, Y., and Hasan, S. A. 2010. Automation of question generation from sentences. In *Proceedings of QG2010: the 3rd Workshop on Question Generation*, Pittsburgh, PA, USA. Pittsburgh, PA: questiongeneration.org, pp. 58–67.

Anderson, R. C. 1972. How to construct achievement tests to assess comprehension. *Review of Educational Research* **42**(2): 145–70.

Bachman, L. F., and Palmer, A. S. 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Baldwin, T., Dras, M., Hockenmaier, J., King, T. H., and van Noord, G. 2007. The impact of deep linguistic processing on parsing technology. In *Proceedings of the 10th International Conference on Parsing Technologies*, Prague, Czech Republic. Stroudsburg, PA: ACL, pp. 36–8.

Banerjee, S., and Pedersen, T. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*. Berlin: Springer, pp. 136–45.

Becker, L., Basu, S., and Vanderwende, L. 2012. Mind the gap: learning to choose gaps for question generation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada. Stroudsburg, PA: ACL, pp. 742–51.

Bernhard, D., De Viron, L., Moriceau, V. E. R., and Tannier, X. 2012. Question generation for French: collating parsers and paraphrasing questions. *Dialogue & Discourse* **3**(2): 43–74.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. 1956. *Taxonomy of Educational Objectives: Handbook I: Cognitive Domain*. New York, NY: David McKay.

Bormuth, J. R. 1970. *On the Theory of Achievement Test Items: with an Appendix on the Linguistic Bases of the Theory of Writing Items*. Chicago: University of Chicago Press.

Burgess, C., Livesay, K., and Lund, K. 1998. Explorations in context space: words, sentences, discourse. *Discourse Processes* **25**(2-3): 211–57.

Carroll, J. B. 1968. The psychology of language testing. In A. Davies (ed.), *Language Testing Symposium: a Psycholinguistic Perspective*. London: Oxford University Press.

Charniak, E. 2001. Immediate-head parsing for language models. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France. Stroudsburg, PA: ACL, pp. 124–31.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Charniak, E., and Elsner, M. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece. Stroudsburg, PA: ACL, pp. 148–56.

Chomsky, N. 1977. On *wh*-movement. In W. Culicover and Akmajan (eds.), *Formal Syntax*, pp. 91–132. New York, NY: Academic Press.

Ciaramita, M., and Altun, Y. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia. Stroudsburg, PA: ACL, pp. 594–602.

Curto, S., Mendes, A. C., and Coheur, L. 2012. Question generation based on lexico-syntactic patterns learned from the web. *Dialogue & Discourse* **3**(2): 147–75.

Dowty, D. 1991. Thematic proto-roles and argument selection. *Language* **67**(3): 547–619.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Washington, DC, USA. New York, NY: ACM, pp. 281–5.

Ebel, R. L. 1972. *Essentials of Educational Measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Felker, D. B., and Dapra, R. A. 1975. Effects of question type and question placement on problem-solving ability from prose material. *Journal of Educational Psychology* **67**(3): 380–4.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. 2013. PPDB: the paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, GA, USA. Stroudsburg, PA: ACL, pp. 758–64.

Gates, D. M. 2008. Automatically generating reading comprehension look-back strategy: questions from expository texts. Technical Report CMU-LTI-08-011. School of Computer Science, Carnegie-Mellon University.

Graesser, A., Rus, V. and Cai, Z. 2008. Question classification schemes. *Proceedings of the NSF-sponsored Workshop on the Question Generation Shared Task and Evaluation Challenge*, Arlington, VA, USA. Retrieved on April 5, 2013: http://www.cs.memphis.edu/~vrus/questiongeneration/16-GraesserEtAl-QG08.pdf.

Heilman, M., and Smith, N. A. 2010. Good question! statistical ranking for question generation. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Los Angeles, CA, USA. Stroudsburg, PA: ACL, pp. 609–17.

Hřebček, L. 1993. Text as a construct of aggregations. *Contributions to Quantitative Linguistics: Proceedings of the 1st International Conference on Quantitative Linguistics*, Trier, German. Dordrecht: Springer, pp. 33–9.

Ishikawa, S. 2013. ICNALE: the international corpus network of Asian learners of English. Retrieved on November 21, 2014: http://language.sakura.ne.jp/icnale/about.html.

Jackendoff, R. 2010. Your theory of language evolution depends on your theory of language. In R. K. Larson, V. Déprez, and H. Yamakido (eds.), *The Evolution of Human Languages: Biolinguistic Perspectives*. Cambridge: Cambridge University Press, pp. 63–72.

Kalady, S., Elikkottil, A., and Das, R. 2010. Natural language question generation using syntax and keywords. In *Proceedings of QG2010: the 3rd Workshop on Question Generation*, Pittsburgh, PA, USA. Pittsburgh, PA: questiongeneration.org, pp. 1–10.

Kaplan, R. M., and Bresnan, J. 1982. Lexical-functional grammar: a formal system for grammatical representation. In M. Dalrymple, R. M. Kaplan, J. Maxwell III, and A. Zaenen (eds.), *Formal Issues in Lexical-Functional Grammar*. Stanford, CA: CSLI, pp. 29–130.

Kunichika, H., Katayama, T., Hirashima, T., and Takeuchi, A. 2003. Automated question generation methods for intelligent English learning systems and its evaluation. In

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

Proceedings of the International Conference on Computers in Education, Hong Kong, China. Hong Kong: AACE, pp. 2–5.

Landauer, T. K., and Dumais, S. T. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* **104**(2): 211–40.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* **33**(1): 159–74.

Levy, O., Goldberg, Y., and Dagan, I. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* **3**: 211–25.

Lindberg, F., Popowich, D., Winne, J., and Nesbit, P. 2013. Generating natural language questions to support learning on-line. In *Proceedings of the 14th European Workshop on Natural Language Generation*, Sofia, Bulgaria. Stroudsburg, PA: ACL, pp. 105–14.

Liu, M., Calvo, R. A., and Rus, V. 2012. G-Asks: an intelligent automatic question generation system for academic writing support. *Dialogue & Discourse* **3**(2): 101–24.

Mannem, P., Prasad, R., and Joshi, A. 2010. Question generation from paragraphs at UPenn: QGSTEC system description. In *Proceedings of QG2010: the 3rd Workshop on Question Generation*, Pittsburgh, PA, USA. Pittsburgh, PA: questiongeneration.org, pp. 84–91.

Mazidi, K., and Nielsen, R. D. 2014. Linguistic considerations in automatic question generation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, MD, USA. Stroudsburg, PA: ACL, pp. 321–6.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, Stateline, Nevada, San Diego, CA: NIPS Foudation, pp. 3111–9.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM* **38**(11): 39–41.

Minnen, G., Carroll, J., and Pearce, D. 2001. Applied morphological processing of English. *Natural Language Engineering* **7**(03): 207–23.

Mitkov, R., and Ha, L. A. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing*, Edmonton, Canada. Stroudsburg, PA: ACL, vol. 2, pp. 17–22.

Mitkov, R., Ha, L. A., and Karamanis, N. 2006. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering* **12**(2): 177–94.

Mostow, J., Beck, J., Bey, J., Cuneo, A., Sison, J., Tobin, B., Valeri, J. *et al.* 2004. Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology Instruction Cognition and Learning* **2**: 97–134.

Mostow, J., and Chen, W. 2009. Generating instruction automatically for the reading strategy of self-questioning. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, Brighton, UK. IOS Press, pp. 465–72.

Mostow, J., and Jang, H. 2012. Generating diagnostic multiple choice comprehension cloze questions. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP*, Montréal, Canada. Stroudsburg, PA: ACL, pp. 136–46.

Navigli, R., and Ponzetto, S. P. 2012. Multilingual WSD with just a few lines of code: the BabelNet API. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, South Korea. Stroudsburg, PA: ACL, pp. 67–72.

Ockey, G. J. 2009. Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal* **93**: 836–47.

Olney, A. M., Graesser, A. C., and Person, N. K. 2012. Question generation from concept maps. *Dialogue & Discourse* **3**(2): 75–99.

Pal, S., Mondal, T., Pakray, P., Das, D., and Bandyopadhyay, S. 2010. QGSTEC system description–JUQGG: a rule based approach, *Proceedings of QG2010: the 3rd Workshop*

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

on Question Generation, Pittsburgh, PA, USA. Pittsburgh, PA: questiongeneration.org, pp. 76–9.

Palmer, M., Gildea, D., and Kingsbury, P. 2005. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics* **31**(1): 71–106.

Pennington, J., Socher, R., and Manning, C. D. 2014. GloVe: global vectors for word representation. In *Proceedings of the Empirical Methods in Natural Language Processing*, Doha, Qatar. Stroudsburg, PA: ACL, vol. 12, 1532–43.

Perfetti, C. A. 1992. The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, and R. Treiman (ed.), *Reading Acquisition*, pp. 145–74. Hillsdale, NJ: Lawrence Erlbaum Associates.

Piwek, P., and Boyer, K. E. 2012. Varieties of question generation: introduction to this special issue. *Dialogue & Discourse* **3**(2): 1–9.

Punyakanok, V., Roth, D., and Yih, W.-T. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics* **34**(2): 257–87.

Raphael, T. E. 1982. Question-answering strategies for children. *The Reading Teacher* **36**(2): pp. 186–90.

Roid, G., and Haladyna, T. M. 1978. A comparison of objective-based and modified-Bormuth item writing techniques. *Educational and Psychological Measurement* **38**(1): 19–28.

Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., and Moldovan, C. 2012. A detailed account of the first question generation shared task evaluation challenge. *Dialogue & Discourse* **3**(2): 177–204.

Sag, I. A., and Wasow, T. 2011. Performance-compatible competence grammar. In R. D. Borsley, and K. Borjars (eds.), *Non-Transformational Syntax: Formal and Explicit Models of Grammar*. Oxford: Wiley-Blackwell, pp. 359–77.

Snow, R., Jurafsky, D., and Ng, A. Y. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia. Stroudsburg, PA: ACL, pp. 801–8.

Spärck Jones, K. 2007. Automatic summarising: the state of the art. *Information Processing & Management* **43**(6): 1449–81.

Varga, A., and Ha, L. A. 2010. WLV: a question generation system for the QGSTEC 2010 task B. In *Proceedings of QG2010: the 3rd Workshop on Question Generation*, Pittsburgh, PA, USA. Pittsburgh, PA: questiongeneration.org, pp. 80–3.

Weischedel, R., and Brunstein, A. 2005. BBN pronoun coreference and entity type corpus. Technical Report LDC2005T33. Linguistic Data Consortium.

Wolfe, J. H. 1976. Automatic question generation from text - an aid to independent study. *SIGCUE Outlook* **10**(SI): 104–12.

Yao, X., Bouma, G., and Zhang, Y. 2012. Semantics-based question generation and implementation. *Dialogue & Discourse* **3**(2): 11–42.

Yao, X., and Zhang, Y. 2010. Question generation with minimal recursion semantics. In *Proceedings of QG2010: the 3rd Workshop on Question Generation*, Pittsburgh, PA, USA. Pittsburgh, PA: questiongeneration.org, pp. 68–75.

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press

A Appendix: an example of sentence selection

This table illustrates the source sentences selected by the human item writers of CET4, the baseline algorithm and the sentence selection module of ASIG.

Table A1. *Sentences selected for passage CET4992*

#	Sentence	State of selection		
		Human	Baseline	ASIG
1	For many women choosing whether to work or not to work outside their home is a luxury.		√	
2	They must work to survive.			
3	Others face a hard decision.			
4	Perhaps the easiest choice has to do with economics.	√		√
5	One husband said, 'Marge and I decided after careful consideration that for her to go back to work at this moment was an extravagance we couldn't afford.'	√	√	
6	With two preschool children, it soon became clear in their figuring that with babysitters, transportation, and increased taxes, rather than having more money, they might actually end up with less.	√		
7	Economic factors are usually the first to be considered, but they are not the most important.	√		
8	The most important aspects of the decision have to do with the emotional needs of each member of the family.	√		√
9	It is in this area that husbands and wives find themselves having to face many confusing and conflicting feelings.		√	
10	There are many women who find that homemaking is boring or who feel imprisoned if they have to stay home with a young child or several children.			
11	On the other hand, there are women who think that homemaking gives them the deepest satisfaction.	√		√
12	From my own experience, I would like to suggest that sometimes the decision to go back to work is made in too much haste.			
13	There are few decisions that I now regret more.		√	
14	I wasn't mature enough to see how much I could have gained at home.	√		
15	I regret my impatience to get on with my career.			√
16	I wish I had allowed myself the luxury of watching the world through my little girl's eyes.			

https://doi.org/10.1017/S1351324915000455 Published online by Cambridge University Press