Corpus-based Automatic Generation of Multiple Choice Questions for College English Tests

Lei Wang[1,2,a], Shujing Li[1,b], Houfeng Wang[2,c], Shiwen Yu[2,d]

[1]School of Foreign Languages, Peking University, Beijing 100871, China

[2]Key Lab of Computational Linguistics of Ministry of Education, Peking University, Beijing 100871, China

{[a]wangleics, [b]lishujing, [c]wanghf, [d]yusw}@pku.edu.cn

Keywords: corpus, multiple choice questions, automatic generation, College English tests

Abstract. At present multiple choice questions as a testing item are adopted widely in College English tests, and play a significant role in assessing students' English proficiency. However, in practice the design of multiple choice questions is also difficult in many aspects, especially with regard to the selection of the question that serves as its stem and the choices that need to be both interfering and exclusive. This paper introduces a method that can generate the stem question automatically based on an English-Chinese parallel corpus in which sentences that meet the designated criteria input by test question makers are extracted on the condition of complex considerations, such as sentence length, word choice and syntactic structure. Although the multiple choice questions generated by our system still need human involvement for correction and consideration, our work will surely provide valuable assistance to teachers who are designing multiple choice questions so that their efficiency will be improved significantly.

Introduction

In English tests of various levels, multiple-choice questions may be the most popular test items and thus they often account for a large proportion of the overall score of certain tests. They may appear in tests about vocabulary, grammar, listening, reading comprehension or cloze[1]. For instance, multiple-choice questions account for over 80% of the total marks in college entrance examinations in China, and over 85% in Band 4 and Band 6 of College English Tests(CET Band 4/6). In Test of English as Foreign Language (TOEFL), they amount to two-third of its total points. Among them, the type of blank-filling with one correct answer among four choices is also called standard multiple-choice questions and seen in almost all English language tests.

When designing multiple-choice questions, a testing item shall be established according to the designated teaching content at first, then the stem question will be chosen from various sources. Afterwards, the testing item is mixed up into inferring choices and all of them need to be marked by a series of alphabets, usually take the form of A), B), C) and D). The selection of a stem question is of great significance in making a successful multiple-choice question for it serves as a prerequisite of reaching the testing purpose of certain question. In principle, the selection of a stem question should be based on the characteristics of language communication, and should be complete in both structure and information and concise as well so as to let the test candidates fully understand the purpose of the testing item and fulfill the testing task successfully. On the contrary, if the stem question is either fragmentary, which lacks in effective information, or lengthy, which provides irrelevant information, it will affect both the candidate's speed of reading and efficiency of test-taking. The following is an example of a bad selection of stem question:

The Israeli troops _____ by launching a bombing attack on the Palestinian-controlled areas.

A) took a crack B) stick to C) took revenge D) followed up with

In the above example, those who do not know what is happening between Palestine and Israel may question why the correct answer is not A) because A) is totally alright semantically by meaning "the Israeli troops took a try by launching a bombing attack" whereas not necessarily "took revenge". The correct answer C) demands a knowledge that Palestinian militants often launch rockets to Israeli settlements, which incurs the revenge from Israeli troops. Unfortunately, the stem question does not provide information as such.

The Feasibility and Necessity of Computer-aided Design of Multiple-choice Questions

With the rapid development of computer technologies, computer aided testing, more technically Computer Adaptive Testing (CAT), has become an important method in the field of academic evaluation or assessment. The wide use of computer-aided testing method brings about the fast increase of testing subjects and types. Educators and researchers begin to construct various item banks to standardize and computerize their testing tasks. Nevertheless, for the huge amount of human labor and financial cost needed to achieve the goal, very few higher education institutions are able to afford a large-scale, applicable and well-maintained testing item bank for English language tests. Some have to outsource the service to some commercial organizations, such as the Education Testing Service (ETS) in the US. Take the author’s work for example, even though there are over forty teachers in the division responsible for teaching College English, a considerable amount of time and energy has to be spent to design testing items for the final exam of every semester, not mention those vocabulary and grammar quizzes during the semester. Given the large number of students (over 2,000 each semester) who take the courses of College English and the heavy teaching load, many faculty members regard the creation of testing items for the above-mentioned teaching tasks as an extra burden and expect that there should be a tool that can help them to reduce their pressure with this regard.

Research in this field has come into being with the application of computer in the education domain and that leads to the birth of Automatic Item Generation (AIG)[2] in computer-aided testing technologies. At present, AIG is able to generate testing items for various subjects and applicable for language testing in particular, with an emphasis on the generation of multiple-choice questions. In a sense, it is also difficult because in the very process an appropriate stem question and three interfering choices need to be generated[3]. Currently, the application of AIG in language testing focuses mainly on generating items in terms of grammar and vocabulary[4]. As for generating items involved in semantics or pragmatics, which requires more complicated technologies of computational linguistics and tends to incur controversies themselves, cannot be put into practical use in a short run.

The characteristics of College English tests enable the automatic generation of multiple choice questions via certain human aid possible. First, for the limitation of test content and time, this part will not exceed the range of vocabulary stipulated by the syllabus issued by the Ministry of Education. Since The English level of new undergraduates are usually low, the test items will only consist of single words or simple phrases. Second, big student population, huge number of papers and heavy marking workload also make it both feasible and wise for educators of college English abide by the same standard and scale in the tests or quizzes to ensure credibility and validity. Test designers need to work out accurate standards and useful tool for purpose of making the assessment process as quantitative and operative as possible.

Principles and Methodology of System Design

In order to realize the automatic generation of multiple-choice questions, we need to separate our research into two branches – the generation of stem questions and the generation of interfering choices. To generate stem questions, we shall extract qualified sentences from a certain source and then find the appropriate position to form the blank that corresponds to the item to be tested1. Now there are usually two channels to find stem questions: corpora[5] or documents downloaded from web pages[6].

[1] In our system, since the word or the phrase to be tested is input by the test designer, the goal turns to be positioning the word or the phrase in the stem sentence successfully.

It provides great freedom and facility to get documents from the Internet via search engine nowadays and these documents can serve as the sources for us to select sentences as stem questions. However, for problems like copyright, character coding etc., we have to make complicated legitimacy check to the sentences extracted in the process of which those grammatically or syntactically wrong sentences have to be eliminated. In addition, if needed we have to unify character coding and transfer the documents into the format that meets our standards. In this process, professional knowledge and human labor are required. Therefore, in our system we obtain stem questions from corpora, which have the advantages of being unified, formalized and reliable. The tags in the corpora enable the user to get easy access to linguistic knowledge which will improve the quality and efficiency of designing problems for tests of English language.

In our research we also use a parallel corpus from the Institute of Computational Linguistics at Peking University (ICL/PKU). Our corpus is consisted of over 300,000 pairs of aligned English/Chinese sentences with a total number of over 3.2 million English words and over 3 million Chinese characters. The reason why we use a parallel corpus is that we expect a Chinese equivalent will be provided in the meantime to make a better judgment when the teacher is exacting a stem question from the corpus. The following in Figure 1 is the sentence pairs from the corpus.

756097 I made that horn sound like it never had before; I made it cry for all the miles and years that separated them.

我把那号吹出从来没有过的声音,我让它为他们分离的那些年月,为他们相隔的那千万里路而哭泣。

756098 Finally, I take a boat over to the island where he lived. It was an old cabin-shack, really-down by the water.

最后打到了一条船到他住的那个岛上去,那是在水边的一间旧屋子,说实在的就是个棚子。

756099 So I don't work anymore.

所以我就不再工作了。

756100 It made the boy sad to see the old man come in each day with his skiff empty and he always went down to help him carry either the coiled lines or the gaff and harpoon and the sail that was furled around the mast.

孩子看见老人每天回来时船总是空的，感到很难受，他总是走下岸去，帮老人拿卷起的钓索，或者鱼钩和鱼叉，还有绕在桅杆上的帆。

……

Figure 1. The sentence pairs in the parallel corpus

With natural language processing technology, we apply our criteria to the sentences extracted from the corpus and leave blanks in them to form the stem questions. Because our research aims to College English tests, the testing points[2] should not exceed the scope stipulated by the Requirements of College English Teaching[7] and the glossary for Band 4 of College English Test. The actual process is as the following: The teacher inputs the testing point into the system and then the system searches the corpus for qualified stem questions according to certain algorithm and return the questions to the

[2] They are often new words or phrases that appear in the textbooks of college English.

teacher for his/her selection. The method for generating interfering choices is mainly dependent on the thesaurus (including three lexicons) built for the purpose of our research. Due to the fact that most multiple choice questions in the quizzes and tests of college English teaching are modeled by the test questions in CET Band 4/6, we build our first lexicon based on the choices extracted from the original test questions of CET Band 4/6 and those practice tests we have gathered. The advantages of using these ready-made choices are: 1) They are designed elaborately by experts in this field and have a fairly high credibility. 2) They are both comparable, exclusive and meet the criteria generally required by multiple choice questions. Part of the lexicon can be seen in the following Table 1.

Table 1. The lexicon of choices from Band 4/6 original/practice test questions

ID	**Entry**			
1	**A)**[3] fainted	**B)** faded	**C)** paled	**D)** grew
2	**A)** feeble	**B)** extinct	**C)**extinguished	**D)** intricate
3	**A)** advocate	**B)** demonstrate	**C)** exhibit	**D)** reveal
4	**A)** In spite of	**B)** But for	**C)** Because of	**D)** As for
…	**…**	**…**	**…**	**…**

The second lexicon we have built includes those words that are formally similar. The reason of using such a lexicon is that it is a common technique of creating choices that look alike in order to confuse test candidates. By doing so, we define the formally similar words as: 1) They begin or end with the same three letters at least in the same order. 2) Other than 1), additional identical letters in the words will add their weight in being chosen into a group of formally similar words. Part of the lexicon of formally similar words can be seen in the following Table 2.

Table 2. The lexicon of formally similar words

ID	**Entry**			
1	**A)** offensive	**B)** attractive	**C)** decisive	**D)**conservative
2	**A)** affirmed	**B)** informed	**C)**conformed	**D)** confirmed
3	**A)** assured	**B)** ensured	**C)** secured	**D)** insured
4	**A)** award	**B)** forward	**C)** reward	**D)** toward
…	**…**	**…**	**…**	**…**

The last lexicon we have built is used when appropriate choices cannot be found in the other two lexicons or the test designer is not satisfied with what he or she has got and wants to further look into other options. The entries of this lexicon are mainly consisted of synonyms, i.e. words that have similar or close meanings. Words as such will help the test designer compare them with the test point and therefore select those that can serve as interfering choices with subtler differences. Part of the lexicon of synonyms can be seen in the following Table 3.

[3] In the real lexicon there are not A), B), C) or D) proceeding the words. Here they are marked to better illustrate their functions of being choices of a test question.

Table 3. The lexicon of synonyms

Index	**Explanation**
department, store, shop	department-作商店解时，是美国英语，通常写为 department store。 store-在美国指出售同一类商品的小型商店。在英国用复数形式表示百货商店。 shop-指规模较小，出售同一类商品的店铺。
depend, rely	depend-侧重指因自身能力不足或缺乏自信心而依靠他人或物给予帮助或支持。 rely-通常包含着以前的经验证明对方是能依赖的意思。
descend, drop, fall, sink	descend-通常指沿斜线或斜坡下降。 drop-指物体从一定高度落下。fall 与 drop 同义，指突然或猛烈地降落，但 fall 也可指任何下落，同高度或形式无关。 sink-指在空气或水中垂直下降、下沉。
…	…

You may notice that in the groups of synonyms the number of words in each group is not equal. The purpose of this lexicon is to help the teacher for reference when making decisions about interfering choices rather than provide a ready group of four choices to use immediately. Therefore the teacher can make his or her decision when considering synonyms as choices for the test question.

System Design and Experiment

We use C# in VS2008 to develop the interface and the corpus is stored in Microsoft Access database for the program to apply. The natural language processing technique is mainly string matching, therefore we choose the KMP[8] algorithm which has the advantage of a faster matching speech and lower time complexity. Compared with traditional method of plain string matching with a movement of the pattern string (simply plus one step), KMP can dynamically adjust the movement of pattern string.

We distribute our system to 5 teachers of college English to make a trial and obtain their feedback by survey. In our questionnaire we evaluate our system by these criteria: accuracy(to measure whether the stem question is chosen and the blank formed correctly), validity(to measure whether the stem question is matched with the choices generated), response(to measure whether the system provides in-time response for inquiry and generation), exclusive(to measure whether the test point input is the only correct answer among the four choices), convenience(to measure whether the system saves the test designer’s time and energy). For each question in the survey, we provide a range of scaling from 1 (not satisfactory) to 5 (completely satisfactory) and the result is in Figure 2.

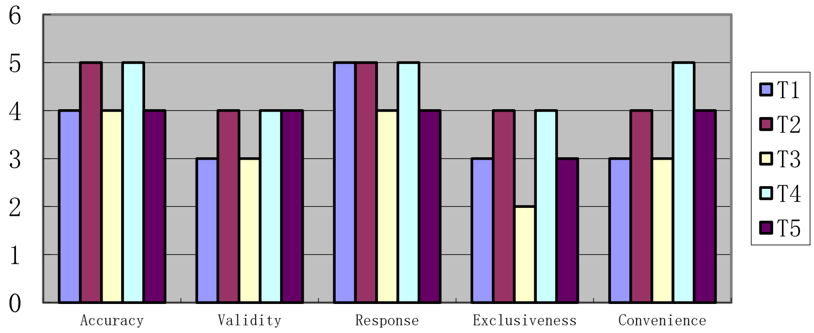

Figure 2. The result of survey on test designers using our system

We will notice that most teachers agree that the system performs well on accuracy, response and convenience, which indicates that it does save them time for reviewing related documents or browsing web pages for materials and provide real assistance in designing multiple choice questions for tests of college English.

Conclusion

For the popular application of computers in language teaching and learning, the design, application, statistics and analysis of English multiple choice questions become more scientific and systematic. Teachers are able to be relieved from the heavy load of making quizzes and tests and the time gained can be used on perfecting curricula and syllabus, organizing classes, etc. We hope that our research on the automatic generation of multiple choice questions for English language tests will have a broad prospect in the field of computer-aided language testing, therefore improve the quality of College English teaching and students' satisfaction.

Acknowledgement

Our work is supported by National High Technology Research and Development Program of China (863 Program) (No. 2015AA015402) and the Visiting Scholar Project by China Scholarship Council.

References

[1] Gao, S., Zhu, H., Liu, J., Song, Y.: The Historical Evolution and Future Prospect of English Multiple Choice Questions, Journal of Northwest University(Philosophy and Social Sciences Edition), pp. 156-159, 2009

[2] Deane, P., Sheehan, K.: Automatic Item Generation Via Frame Semantics, Education Testing Service: http://www.ets.org/ research/dload/ncme03-deane.pdf, 2003, 7(39-4): 43-48

[3] Jin, W., How to Design Multiple Choice Questions: A Perspective of English Language Testing： http://www.cnki.net/kcms/detail/42.1617.G4.20130131.1839.094.html

[4] Brown, J. C., Frishkoff, G. A., and Eskenazi, M. Automatic Question Generation for Vocabulary Assessment. http://wenku.baidu.com/view/ddc4dd1cfc4ffe473368abb6.html

[5] Aldabe, I., M. Lopez de Lacalle, Maritxalar, M., Martinez, E., ArikIturri, L.: An Automatic Question Generator Based on Corpora and NLP Techniques, Proceedings of the Eight International Conference on Intelligent Tutoring Systems (ITS'06), 2006, 6: 584-594

[6] Yang, D., The Construction of Computerized Corpus for Foreign Language Tests, Journal of Yunnan Normal University(Philosophy and Social Sciences Edition), 2005, 7(22-4): 142-144

[7] Ministry of Education. Requirements of College English Teaching, Shanghai Foreign Language Education Press, 2007

[8] Knuth, D.; Morris, J. H., jr; Pratt, V.. Fast pattern matching in strings. SIAM Journal on Computing 6 (2): 323-350, 1977