Assisting cloze test making with a web application

Ayako HOSHINO and Hiroshi NAKAGAWA
University of Tokyo
{hoshino, nakagawa}@dl.itc.u-tokyo.ac.jp

Abstract: We present an internet-based system that helps teachers to make a cloze test on an online news article. The current system supports question making of grammar and vocabulary of English. The system works as an assistant to the teacher. It helps the user to choose an article, highlights grammar targets, suggests possible choices for the wrong alternatives and formats the questions in a printer-friendly way. We use the technique of NLP (Natural Language Processing) to provide suggestions for the wrong alternatives. The user interface is built on a web browser so the user can make a test easily (just by clicking on the text.) The user evaluation shows that about 80% of the questions made with this application are appropriate and the usefulness and the usability of the interface are satisfactory.

Introduction

Grammar and vocabulary are vital elements in SLA (Second Language Acquisition), therefore grammar and vocabulary testing is necessary for a teacher to assess students’ proficiency. Normally, those components are taught and assessed by non-native teachers. In such cases it’s hard to use new materials and bring in current topics, which would make the class more exciting and close to daily life, since it takes too much labor for them to develop new materials.

A cloze test is preferred in such situations. Cloze test in general refers to a format of questions where a/some word(s) are removed from a passage and the test takers are asked to figure out what the missing words are. The advantages of this format of question are easy to make and easy to mark. In this study, we especially focus on a multiple-choice fill-in-the-blank test, which is a subcategory of a cloze test. The questions we focus on are those on grammar and vocabulary, rather than reading comprehension and discourse structures.

Our approach

We explore the use of NLP (Natural Language Processing) for the purpose of cloze test making. Recent years have seen a large advance in technologies to analyze natural language, such as a sentence parser. A sentence parser analyzes a sentence in nested phrase structure. Also there are more available NLP resources such as machine-readable dictionary and thesaurus. Aside from that, there are more available online text and corpora, which is a large collection of texts, on the World Wide Web. We employ these technologies and resources for grammar and vocabulary test making. One of our motivations is to see if the state-of-the-art NLP is practical for an educational purpose.

The system is designed to semi-automatically make questions in a given input text. We use online news articles as a source text. The reasons for using online news are; first, news texts are properly written compared with other sources on the web. Secondly, their topics are suitable for English Education. Also, they are up-to-date so a quick question making benefits.

Our related studies attempt automated question generation but we find several problems in putting a totally automatic system in a practical use. The main reason is that the current NLP is not hundred percent accurate, and it is not appropriate to use it in a classroom if there can be errors. So we currently provide a semi-automatic way of making question with a view of fully automating the process. In principle, the system presented here works as an assistant for a teacher. The system helps the user to choose an article, highlights grammar targets, suggests possible choices for the wrong alternatives and formats the questions in a printer-friendly way.

The system assists the user to make certain types of questions, with an assumption that the alternatives of a good four-choice question should either be flat or symmetric. Flat alternatives mean a set of alternatives which differ
on single element (either grammar or vocabulary). For example, a set of alternatives where all alternatives have the same part-of-speech and the same inflection form (tense etc.) but are of different vocabulary items. Also, a set of alternatives that have all different inflection forms and have the same vocabulary item is flat. Symmetric alternatives are those which consist of two different elements, each from vocabulary and grammar. Below shows the examples for the alternative sets for each question type.

A sentence with blank: He finally managed [ ] the habit of smoking.

Example of flat vocabulary alternatives:
{to kick, to toss, to touch, to knit}

Example of flat grammar alternatives:
{to kick, kick, kicking, kicked}

Example of symmetric, grammar and vocabulary alternatives:
{to kick, kicking, to toss, tossing} (“to kick” is the right answer)

In our approach, the system provides candidates for wrong alternatives or distractors when the blank position is specified (this is one by the user’s click on the text) and let the user decide which ones to use as a distractor. The current version of the system provides seven choices for vocabulary items and three for grammar variations where available. The user can obtain an alternative set by selecting from these suggestions. An alternative set of symmetric type is obtained by picking one from each of vocabulary alternative suggestions and grammar alternative suggestions. The user is allowed to enter or edit the alternatives when they cannot find needed ones in the suggested alternatives.

The rest of the paper is organized as follows. In related studies, we discuss recent studies on question generation in the field of NLP. In the next section, we explain the core methods employed, namely the method for selecting candidates for the wrong alternatives. Then we present the outline of the system with the details of the two main components, which are preprocessing program and user interface. Then we present the results of user evaluation and conclusions.

**Related Studies**

There are recently some attempts to automatically generate multiple-choice cloze questions. In principle they take input texts and generate questions by removing some words from a sentence. Although the method for providing the wrong alternatives varies from research to research, the main idea is to select similar words to the correct answer, as in one research stated; “Syntax, rather than football or sport, is more appropriate for a distractor of the correct answer semantics” (Mitkov 2003).

In selecting similar words, machine-readable dictionaries are used in most of the studies. Mitkov et al. Liu and Brown (Brown 2005) used WordNet [1], and Sumita and Kunichika (Kunichika 2002) used their in-house thesauri. Typically, they retrieve similar or related words (synonyms, antonyms, hypernyms, hyponyms, etc.) by consulting a machine-readable thesaurus and in their respective metrics select the most appropriate ones for the distractors. In Sumita’s study, they first retrieved similar words by consulting a thesaurus and then filters out the distractors that can result in a multiple right answers.

Liu et al., Sumita et al. and Brown et al. generate cloze questions on text, while Mitkov and Kunichika generated question sentences with conversion patterns that change declarative sentences into interrogatives. The type of question targeted in most of the studies is vocabulary type. Kunichika et al. report that their system is also capable to produce reading comprehension questions.

To our knowledge this is the first study that uses NLP techniques to generate grammar type questions. Especially the use of parse result makes this study a unique attempt. While other studies present automatic systems, we provide an interactive system which works as an assistance for a teacher. Except for those features, we adopted

---

[1] an online lexical reference system by Princeton University. (http://wordnet.princeton.edu/)
common methods in above studies: the format of the resulting questions is multiple-choice, fill-in-the-blank embedded in a given text. The resource we used for vocabulary distractors is WordNet.

**Methods**

This section explains the methods to select candidates for the distractors. We propose methods for selecting vocabulary distractors and grammar distractors. A combination of the two methods is used for the grammar-and-vocabulary type question.

**Vocabulary distractors**

In the method proposed by Coniam (Coniam 1997) and adapted by Brown et al. (Brown 1996), they select the words of close frequency for the distractors. We also use the frequency-based method. The main difference from the method in above researches is, while they select the distractors from the words appeared in the same passage (thus obtaining words which are related to the right answer), we select from the words appeared in the article plus their synonyms (thus obtaining words which are related but which do not necessarily appear in the same article). Also our system provides the same part of speech and inflection form as the right answer. We used Kilgarriff's word frequency list to obtain word frequencies.

**Grammar distractors**

Currently, the system has ten grammar targets as defined in Table 1 (Tab. 1). The selection of ten targets is based on a previous research (Tateno 2005). They the most frequent grammar targets reported in their study. For each grammar target we developed a pattern that matches a phrase which has specified phrase structure. Also, we developed three generation rules for each grammar target, which transform the matched phrase into possibly grammatically incorrect ones.

**Table 1: Ten grammar targets**

<table>
<thead>
<tr>
<th>Grammar target</th>
<th>Matching Rule</th>
<th>Generation rules for wrong alternatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) to + infinitive</td>
<td>(S (VP (TO to) (VP (VV ) ...)))</td>
<td>to do -&gt; do</td>
</tr>
<tr>
<td></td>
<td></td>
<td>to do -&gt; doing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>to do -&gt; done</td>
</tr>
<tr>
<td>2) passive voice</td>
<td>(S ((NP ..) (VP ( am/is/are/was/were) (VVN))))</td>
<td>NP is done -&gt; NP done</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NP is done -&gt; NP being done</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NP is done -&gt; NP do</td>
</tr>
<tr>
<td>3) S+V+O</td>
<td>(S ((NP ..) (VP V.. (NP ... )))</td>
<td>NP1V NP2 -&gt; NP1NP2V</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NP1V NP2 -&gt; V NP1 NP2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NP1V NP2 -&gt; NP2 V NP1</td>
</tr>
<tr>
<td>4) subject + be-verb + complement</td>
<td>(S (NP ...) (VP ( am/is/are/was/were) (NP ...)))</td>
<td>is, are -&gt; are, is; was, were -&gt; were, was; am -&gt; are</td>
</tr>
<tr>
<td></td>
<td></td>
<td>is, are, am -&gt; has, have, was, were -&gt; had</td>
</tr>
<tr>
<td></td>
<td></td>
<td>is, are, am, was, were -&gt; been</td>
</tr>
<tr>
<td>5) subordinate conjunction “that”</td>
<td>(S (NP) (SBAR (WHNP (WDT that) ... ) ) ... ))</td>
<td>that ... -&gt; that ... it</td>
</tr>
<tr>
<td></td>
<td></td>
<td>that NP VP -&gt; that VP NP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>that ... -&gt; those ...</td>
</tr>
<tr>
<td>6) gerund as a noun</td>
<td>(S ... (VP (VBG ... ))) ... }</td>
<td>doing -&gt; to do</td>
</tr>
<tr>
<td></td>
<td></td>
<td>doing -&gt; do</td>
</tr>
<tr>
<td></td>
<td></td>
<td>doing -&gt; done</td>
</tr>
<tr>
<td>7) relative clause</td>
<td>(S ... (SBAR (WHNP ...) (S ...) ... ) ... )</td>
<td>has/have done -&gt; did</td>
</tr>
<tr>
<td></td>
<td></td>
<td>has/have done -&gt; do</td>
</tr>
<tr>
<td></td>
<td></td>
<td>has/have done -&gt; had done</td>
</tr>
<tr>
<td>8) present perfective</td>
<td>(S ... (VP (AUX has/hae) ...) ... )</td>
<td>which/who/that do(es)... -&gt; which/who/that he does ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>which/who/that do(es)... -&gt; which/who/that it does ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>which/who/that do(es)... -&gt; he does ...</td>
</tr>
<tr>
<td>9) future tense with “will”</td>
<td>(S ... (VP (MD will) (VP ...) ... ) ... )</td>
<td>will do -&gt; do</td>
</tr>
<tr>
<td></td>
<td></td>
<td>will do -&gt; is doing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>will do -&gt; do</td>
</tr>
</tbody>
</table>
For example, pattern 1) matches the phrase “to go to the theater” in a sentence as “We are planning to go to the theater”. Then, the matched phrases are highlighted in the article. The right hand side of the arrows are generated phrases, which will be presented to the user as suggestions for grammar alternatives.

\[
\begin{align*}
to \text{ go to the theater} & \Rightarrow go \text{ to the theater} \\
& \Rightarrow going \text{ to the theater} \\
& \Rightarrow went \text{ to the theater}
\end{align*}
\]

The use of parse result is advantageous when the conversion to the distractors is achieved by phrase-based operations. For example, a sentence “I don’t know how the culprit has done that”, in which the phrases “the culprit”, “has done” consist of more than one word. Below shows the transformation when pattern 3) SVO is applied;

\[
\begin{align*}
\text{the culprit has done that} & \Rightarrow \text{the culprit that has done} \\
& \Rightarrow \text{that the culprit has done} \\
& \Rightarrow \text{that has done the culprit}
\end{align*}
\]

In this case the SVO pattern makes it possible to test on word ordering with a multiple-choice question. But the word order within phrase is retained, rather than the whole phrase is randomly messed up. Thus, the patterns can produce more likely mistakes.

**Grammar-and-vocabulary distractors**

A set of grammar-and-vocabulary distractors are generated on choosing a grammar distractor and a vocabulary distractor from the suggestions described above. The last distractor is determined by combining the selected phrase and word. The word is inflected in the same inflection form as the right answer. For example, when the user choose “going to the theater” as a grammar distractor for the phrase “to go to the theater” and the vocabulary distractor, say, “arrive” for “go”, then the following set of distractors is completed: \{to go to the theater, going to the theater, to arrive to the theater, arriving to the theater\}.

**System Outline**

The system is comprised of two main components: 1) a preprocess program and 2) a user interface. Figure 1. (Fig. 1) shows the overview of the system.

![Figure 1: The system overview](#)
Preprocess

The preprocess program extracts article text from an HTML document, analyzes sentences and add synonyms from a thesaurus. The output of preprocess program is added to the pool of preprocessed articles, and an index file which contains summarized information of all articles is updated. The preprocessing is totally automated. It takes about 10 minutes to process one article of about 500 words.

The input documents (in html) go through a series of subprocesses that are: 1) text extraction, 2) sentence splitting, 3) tagging and lemmatizing, 4) synonym lookup, 5) frequency annotation, 6) inflection generation, 7) grammar target mark up and grammar distractor generation and 8) selection of vocabulary distractor.

User interface

The user interface is implemented as a web application consisting of three screens, which are: 1) Article selection 2) Test making 3) Confirm and print. The screenshots of 1) and 2) are shown (Fig. 1) (Fig. 2).

![Figure 2: Article selection screen](image)

![Figure 3: Question making screen](image)

Article selection

On the first screen, the system helps the user to choose an article on which she is going to make questions. It shows the list of articles it has preprocessed. An article is shown as a list item with its news source, date, title and number of words. The user can see more information on the article, such as its lead (the first 10 words of the article), vocabulary level (Kincaid score) and grammar targets included and their times of appearance, of an article by putting the cursor on a list item.

The user can narrow down this list of articles with conditions on grammar targets and/or vocabulary levels. The system shows a menu so the user can specify conditions by clicking on the checkboxes or selecting from combo boxes. When the conditions are specified, the system returns a list of (the only and all) articles that meet the conditions. The user can choose an article with the grammar targets she wants to test and in a desired difficulty level. For example the user can choose an article which contains sentences in perfect present tense and within a difficulty level of 5~6 in Kincaid.

Test making

When the article is selected, the screen moves on to test making page. The system assists the user’s making blanks on the passages and deciding the wrong alternatives. Article text is shown on the left side and the question making area is on the right side. In the article text, the words that have distractor suggestions are clickable and shown in black and the words that have no suggestions (proper nouns, articles, etc.) unclickable and shown in gray. When the user clicks on a word or a highlighted phrase in the article, the system replaces the word/phrase with a blank and
shows it on question making pane as the right answer, along with three input forms. Simultaneously, the suggested alternatives are shown on question making pane, which fill the forms when clicked.

Menu on question making pane shows the grammar targets contained in the article. When a target is clicked, the corresponding phrases in the article are highlighted (black turns red and gray turns orange). When a particular phrase is selected by the user’s click on the article text, the suggestions for grammar distractors appears on question making pane. On selecting one grammar distractor, suggestions for vocabulary distractors appear. If the user chooses the second distractor from vocabulary distractor suggestions, then the distractor fields are automatically filled with a set of symmetric grammar vocabulary distractors.

Confirm and Print

When the questions are completed and the user clicked the submit button, the system shows the resulting questions in a test format. On moving to this page from question making page, the blanks containing phrases shrink where appropriate. In the example “We are planning [to go] to the theater.”, the phrase “to go to the theater” is retrieved on test making page. On the next screen, the blank shrinks so it contains only “to go”, since the rest of the phrase “to the theater” is the same in all alternatives. On “print” button clicked, it provides forms for the date, the name of test takers and the answer boxes.

User Evaluation

We have conducted a user test on the system. The purposes of the user test are: 1) evaluating in appropriateness questions they make with this application 2) evaluating the prototype version of the system in its usefulness and its usability, 3) getting implications about how to improve the system. The participants were teachers who are or have been teaching English. The number of the users participated was ten (all Japanese, teaching undergraduate classes).

The articles we used were eight from Student times, eight from CDLP, 10 from BBC Learning English\(^2\). Student Times\(^3\) and CDLP\(^4\). Length of each article is about 150-300 words. Difficulty levels are: Student Times 10~, CDLP 5, BBC Learning English 6~10 in Kincaid score.

Method of Evaluation

A user test was conducted with a task and questionnaires (pre-task and post-task). The pre-task questionnaire was on participant’s background as an English teacher. The participants are asked to describe a particular class that she teaches or has been teaching most recently. Then, as a task, the participants are asked to make a test for that class and evaluate each question. We asked them to make questions at a difficulty level which is as appropriate as possible for the class. This instruction is added to make the participants to evaluate the questions with a more realistic viewpoint. The post-questionnaire was on usefulness and usability and an overall evaluation on this system.

Background of participants

In pre-task questionnaire, we asked their 1) Experience as a teacher in years, 2) the level of the class 3) Familiarity with Internet browsers. Table 2 (Tab. 2) shows the summary of pre-task questionnaire.

<table>
<thead>
<tr>
<th>Teaching experience</th>
<th>Have made questions of this type</th>
<th>Level of the class</th>
<th>Familiarity with Internet browser</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-10 years: 2</td>
<td>Yes : 7</td>
<td>High : 0</td>
<td>Yes : 3</td>
</tr>
<tr>
<td>10- years: 3</td>
<td>No : 3</td>
<td>Relatively high : 2</td>
<td>Somewhat : 5</td>
</tr>
<tr>
<td>no answer: 5</td>
<td></td>
<td>Intermediate : 4</td>
<td>No : 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relatively low : 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low : 1</td>
<td></td>
</tr>
</tbody>
</table>

\(^2\) BBC Learning English [http://www.bbc.co.uk/worldservice/learningenglish/](http://www.bbc.co.uk/worldservice/learningenglish/)
\(^3\) Student Times [http://www.japantimes.co.jp/shukan-st/st-news.htm](http://www.japantimes.co.jp/shukan-st/st-news.htm)
\(^4\) CDLP(California Distance Learning Project) [http://www.cdlponline.org/](http://www.cdlponline.org/)
Task details

The task took about 30 minutes to one hour. After having read the user’s manual, the participants were asked to make three questions in two articles. Then the participants evaluated the resulting questions in three-graded rating. When working on the first article, they were given instructions in person. On the second article, they worked mostly on their own.

The user’s manual introduced three types of questions. They chose articles using grammar and vocabulary conditions. We asked the participants to make one question from each in a set order, namely 1) flat vocabulary 2) flat grammar 3) symmetric vocabulary and grammar. In total, 60 questions are collected from 10 participants.

Evaluation Result

Table 3 (Tab. 3) shows the number of questions rated as “Appropriate”, “So so” and “inappropriate”. The numbers in the brackets indicates the number of questions whose one or more distractors are edited or entered by the participant. 12 out of 60 questions were edited or entered by the participants, the rest of the set of suggestions are considered acceptable on question making screen.

<table>
<thead>
<tr>
<th>Total number of questions</th>
<th>Vocabulary Question (number of questions)</th>
<th>Grammar Question (number of questions)</th>
<th>Grammar and Vocabulary Question (number of questions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriate</td>
<td>48(10)</td>
<td>18(4)</td>
<td>17(3)</td>
</tr>
<tr>
<td>So so</td>
<td>7(1)</td>
<td>1 (1)</td>
<td>5</td>
</tr>
<tr>
<td>Inappropriate</td>
<td>5(1)</td>
<td>1 (1)</td>
<td>2</td>
</tr>
</tbody>
</table>

The result shows that the most of the questions made with this application are of appropriate quality, with 80% of total, 79% of unedited questions rated as “appropriate”. This is a very promising result with regard to the fact that from 1/3 to a half of manually made questions are usually discarded after pilot testing (Brown 96).

The participants did not provide any reason for “so so” ratings, the reasons for “inappropriate” include too obvious answer, multiple right answers and a too long blank caused by a sentence split error. There are more so-so questions in grammar and grammar-and-vocabulary questions, which could be attributed to the difficulty in determining the blank of right range.

All participants completed the task in reasonable time despite of the fact that some of them were not familiar with Internet browsers. The result of the post-task questionnaire (Tab. 4) also confirms the conclusion that the current version of the interface is easy to use even for general users.

Table 4: Usefulness and usability of each functionality
The result shows this application is useful and usable; with all functionality receive “good” from more than half participants. In overall evaluation, all participants gave rating four in five-graded rating. The grammar target highlight and suggestions received lower scores by some participants. A major complaint about the grammar targets was about its length. The system assigns a grammar target to a phrase, which is often longer than the range that English teachers need. Another problem was its response time on the browser. The highlight feature was not fully optimized in the tested version, resulting in a 1–2 second delay since the button clicked.

Conclusions

We have presented an assisting system for cloze or multiple-choice fill-in-the-blank questions. As opposed to related studies, we built an assistance system, which allows the teachers to choose the distractors from the suggestions or enter them by themselves. The result was 80% appropriate questions in total, 79% appropriate unedited questions, which is very promising even for automatic system.

Usefulness of each function was approved by most of the participants. Similarly, usability was also satisfactory in the tested version. Still, further improvements were possible, and some of the improvements are already added to the system. Future work is putting the improved version on a more proper evaluation such as distractor efficiency analysis.

References


Tateno, Junko, Sano Hiroshi, Aizawa Hiroshi, Nakamura Takehiro and Morita Yasuo, (2005), Producing English Educational Materials form the BNC and releasing them on the Web, IEICE Technical report, TL2005-18–26, pp. 7-12, Tokyo, JAPAN.