Exercise Maker: Automatic Language Exercise Generation

A. Malafeev
National Research University Higher School of Economics
Nizhny Novgorod, Russia
Problem Statement

Vocabulary and grammar exercises:

• widely used in teaching foreign languages
• expensive to create manually
• can be generated automatically

https://essentialthinking.wordpress.com/excerpts/kant/kant-answers-vocabulary/


Exercise Maker

• Arbitrary passages in English -> exercises
• Seven further customizable exercise types
• Freely available for non-commercial purposes
  • http://www.hse.ru/staff/aumalafeev#other
Overview

> Comparison with other systems
  - How Exercise Maker works
  - Evaluation
  - Conclusion and future work
Types of Systems

Exercise/test generation systems:

• Domain:
  • specific
  • multi-domain

• Degree of human participation:
  • fully automatic
  • semi-automatic (authoring tools)

Hot Potatoes (http://hotpot.uvic.ca/)
MaxAuthor (http://cali.arizona.edu/docs/wmaxa/)
Language Exercise Generation Systems

• 14 systems analyzed

• Differ in:
  • supported language(s)
  • type of input/output
  • exercise formats
  • external dependencies (e.g. corpora, taggers, parsers, WordNets, etc.)
  • availability
  • performance
Languages and Input/Output

Supported Languages
- English: 10
- Other: 2

Input Type
- Any Text: 4
- Other (corpora, grammars, etc.): 1

Output Type
- Text: 14
- Other (sentences, questions): 1
Exercise Types

- multiple choice: 7
- morphological transformation: 4
- open cloze: 3
- error correction: 2
- other: 8
Dependencies and Availability

• Only two self-contained systems (none for English)
• Only two are available (for English)
  • There are other ‘lightweight’ online tools, but with little or no formal description/evaluation
Exercise Maker vs Other Systems (1)

• Flexibility (arbitrary passages) -> motivation (Heilman et al., 2010)
• ‘Context-rich’ format like in Cambridge English certificate exams
  • FCE, CAE, CPE, and others
• Features some exercise types not supported by other systems:
  • filling in missing words (no gaps), word formation, and text fragments
  • commonly used in EFL, e.g. in Cambridge exams (word formation), Russian State Exam (text fragments)
Exercise Maker vs Other Systems (2)

• Tweakable difficulty – a very rare feature
• Fully self-contained, hence more easily extended to resource-poor languages
• Freely available
  • learning/teaching
  • research
  • comparison
Overview

+ Comparison with other systems
> How Exercise Maker works
- Evaluation
- Conclusion and future work
# Supported Exercise Formats (1)

<table>
<thead>
<tr>
<th>No.</th>
<th>Exercise</th>
<th>Exams</th>
<th>Description</th>
<th>Example(s)</th>
<th>Answer(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Word formation</td>
<td>FCE, CAE, CPE, RSE</td>
<td>Fill in blanks with derivatives of the words in parentheses.</td>
<td>...but the tools he had available were (6)____(sufficient) to do so.</td>
<td>insufficient</td>
</tr>
<tr>
<td>2</td>
<td>Error correction</td>
<td>BEC*</td>
<td>Correct spelling/lexical/grammar errors in the text.</td>
<td>Ralston had not informed nobody of his hiking plans &lt;...&gt; thus no one would searching for him &lt;...&gt; the dehydrated and deliriouse Ralston</td>
<td>had not informed anybody; no one would search/ be searching for him; delirious</td>
</tr>
<tr>
<td>3</td>
<td>Open cloze</td>
<td>FCE, CAE, CPE, BEC</td>
<td>Fill in blanks with suitable words (no candidate answers given). Sometimes, there are two or more correct answers.</td>
<td>When he ran (12)_____ of food and water on the fifth day...</td>
<td>out</td>
</tr>
<tr>
<td>4</td>
<td>Wordbank</td>
<td>none*</td>
<td>Fill in blanks with suitable words given a full list of answer choices (no distractors; each word is used only once).</td>
<td>(approximately, available, &lt;...&gt;, just, suspended) ...a (2)_____ boulder he was climbing down became dislodged...</td>
<td>suspended</td>
</tr>
</tbody>
</table>
**Supported Exercise Formats (2)**

<table>
<thead>
<tr>
<th>No.</th>
<th>Exercise</th>
<th>Exams</th>
<th>Description</th>
<th>Example(s)</th>
<th>Answer(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Missing words (articles or prepositions)</td>
<td>none</td>
<td>Insert prepositions (another subtype: articles) where appropriate.</td>
<td>Ralston had not informed anybody his hiking plans, thus no one would be searching him.</td>
<td>Ralston had not informed anybody of his hiking plans, thus no one would be searching for him.</td>
</tr>
<tr>
<td>6</td>
<td>Text fragments</td>
<td>RSE</td>
<td>Insert missing text fragments (all answer options are listed).</td>
<td>After three days of trying to lift (6) __________, the dehydrated and delirious Ralston</td>
<td>d) and break the boulder</td>
</tr>
<tr>
<td>7</td>
<td>Verb forms</td>
<td>RSE</td>
<td>Use the appropriate verb form to fill each of the gaps.</td>
<td>While he (1)_____(descend) a slot canyon, a suspended boulder...</td>
<td>was descending</td>
</tr>
</tbody>
</table>

Implementation Details

• Written in Python 3, standard modules only
• Decision trees with manually written rules
  • algorithms vary depending on the exercise type
• A set of linguistic resources
  • compiled by the author manually and semi-automatically
Linguistic Resources (1)

1. 2274 and 10084 most common English word forms (including proper nouns)
   - based on a free film-subtitle-based frequency list
     ([https://invokeit.wordpress.com/frequency-word-lists/](https://invokeit.wordpress.com/frequency-word-lists/))

2. 11805 word forms used in the word formation exercise
   - heavily based on the BNC lists

3. Rules for making realistic spelling/lexical/grammar errors (795 words)
   - the spelling part is based on the Wikipedia list of common misspellings
   - the lexical and grammar error rules were compiled manually
Linguistic Resources (2)

4. Three ordered lists of 139 words each for generating open cloze tests
   • emulate specific Cambridge exam levels (FCE, CAE or CPE)
   • based on an empirical study of the mentioned exams

5. 91 adverbs used in the verb forms exercise

6. 13540 verb forms and an additional short list of auxiliary forms, both used in the verb forms exercise
   • automatically extracted from the Spelling Checker Oriented Word List (http://wordlist.sourceforge.net/) with some manual post-processing

7. A few manually written shorter lists of articles, conjunctions, prepositions, pronouns, etc.
Rules and Features

• Not merely looking up words in the lists; rules take into account:
  • word capitalization
  • spelling features
  • punctuation
  • word length
  • distance to other gaps
  • word context
  • sentence boundaries
  • other factors
Pre-processing

• Segmentation (very straightforward):
  • words (punctuation handled specially)
  • sentences
  • paragraphs

• Readability analysis
  • average sentence length and word frequency information
    • traditionally considered as most closely correlated with text readability (Klare, 1968; Chall, 1995)
  • source passages are ranked according to their complexity
Difficulty Adjustment

• Beyond complexity of source text
  • which is, of course, very important

• Exercise subtypes varying:
  • number of gaps in the exercise
  • target language material, i.e. the words in the text that are gapped (for the open cloze and verb forms exercises)
  • length of the gaps (for the fragments exercise)
Overview

+ Comparison with other systems
+ How Exercise Maker works

> Evaluation

- Conclusion and future work
Exercise Evaluation Notes

• Difficult to automate; evaluation may involve:
  • comparison with ‘gold standard’ exercises (but not straightforward)
  • students doing exercises
  • expert evaluation

• Classroom tests (since 2012)

• Exercise-specific evaluation
  • e.g. for open-cloze tests (Malafeev, 2014)
Evaluation of All Exercise Types

- Two independent TEFL experts (non-native speakers of English)
  - have taken no part in developing Exercise Maker
- Five abridged and simplified news articles

<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Date</th>
<th>Word count</th>
<th>Readability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Japanese government to play matchmaker</td>
<td>17th March, 2015</td>
<td>243</td>
<td>10.7</td>
</tr>
<tr>
<td>2</td>
<td>BBC Top Gear star punches producer</td>
<td>14th March 2015</td>
<td>237</td>
<td>7.1</td>
</tr>
<tr>
<td>3</td>
<td>Sportswear maker accused of sexism</td>
<td>11th March, 2015</td>
<td>231</td>
<td>12.1</td>
</tr>
<tr>
<td>4</td>
<td>China tops US at box office for first time</td>
<td>5th March, 2015</td>
<td>233</td>
<td>7.7</td>
</tr>
<tr>
<td>5</td>
<td>Cut music to an hour a day</td>
<td>2nd March, 2015</td>
<td>248</td>
<td>12.2</td>
</tr>
</tbody>
</table>

source: http://breakingnewsenglish.com
readability: http://www.editcentral.com/gwt1/EditCentral.html
Evaluation Details

• Generated 40 exercises, eight from each input text
  • two different subtypes of the missing words exercise: articles and prepositions

• Two kinds of expert assessment:
  • evaluate all gaps for validity (binary; precision only – standard practice)
  • assign to each exercise an overall score from 1 to 4, meaning:
    1 – the exercise cannot be used;
    2 – the exercise can be used only after making substantial alterations;
    3 – the exercise can be used, but it requires some minor alterations;
    4 – the exercise can be used as is.
    (score of 3 or 4 -> the exercise is ‘acceptable’)

Dialogue 2015

A. Malafeev, National Research University Higher School of Economics
## Validity of Gaps (i.e. Precision)

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Total gaps</th>
<th>Expert 1, n</th>
<th>Expert 1, %</th>
<th>Expert 2, n</th>
<th>Expert 2, %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exercises</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Articles</td>
<td>110</td>
<td>109</td>
<td>99,09%</td>
<td>110</td>
<td>100,00%</td>
</tr>
<tr>
<td>Derivatives</td>
<td>60</td>
<td>52</td>
<td>86,67%</td>
<td>57</td>
<td>95,00%</td>
</tr>
<tr>
<td>Errors</td>
<td>131</td>
<td>131</td>
<td>100,00%</td>
<td>129</td>
<td>98,47%</td>
</tr>
<tr>
<td>Fragments</td>
<td>30</td>
<td>30</td>
<td>100,00%</td>
<td>30</td>
<td>100,00%</td>
</tr>
<tr>
<td>Open cloze</td>
<td>144</td>
<td>142</td>
<td>98,61%</td>
<td>144</td>
<td>100,00%</td>
</tr>
<tr>
<td>Prepositions</td>
<td>137</td>
<td>126</td>
<td>91,97%</td>
<td>131</td>
<td>95,62%</td>
</tr>
<tr>
<td>Verb forms</td>
<td>73</td>
<td>73</td>
<td>100,00%</td>
<td>69</td>
<td>94,52%</td>
</tr>
<tr>
<td>Wordbank</td>
<td>100</td>
<td>97</td>
<td>97,00%</td>
<td>100</td>
<td>100,00%</td>
</tr>
<tr>
<td><strong>Texts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>153</td>
<td>145</td>
<td>94,77%</td>
<td>151</td>
<td>98,69%</td>
</tr>
<tr>
<td>2</td>
<td>171</td>
<td>165</td>
<td>96,49%</td>
<td>168</td>
<td>98,25%</td>
</tr>
<tr>
<td>3</td>
<td>153</td>
<td>147</td>
<td>96,08%</td>
<td>146</td>
<td>95,42%</td>
</tr>
<tr>
<td>4</td>
<td>144</td>
<td>142</td>
<td>98,61%</td>
<td>142</td>
<td>98,61%</td>
</tr>
<tr>
<td>5</td>
<td>164</td>
<td>161</td>
<td>98,17%</td>
<td>163</td>
<td>99,39%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>785</td>
<td>760</td>
<td><strong>96,82%</strong></td>
<td>770</td>
<td><strong>98,09%</strong></td>
</tr>
</tbody>
</table>

**Notes:**
- **Micro** refers to the micro level of analysis.
- **Macro, exercises** refers to the macro level of analysis for exercises.
- **Macro, texts** refers to the macro level of analysis for texts.
Precision per Exercise Type

- Articles
- Derivatives
- Errors
- Fragments
- Open cloze
- Prepositions
- Verb forms
- Wordbank
Precision per Source Text

![Bar chart showing precision for five experts. Each bar represents the percentage of accuracy for each expert across different source text levels. The chart compares Expert 2 and Expert 1.]

- Expert 2:
  - Source Text 1: 94.00%
  - Source Text 2: 96.00%
  - Source Text 3: 98.00%
  - Source Text 4: 99.00%
  - Source Text 5: 100.00%

- Expert 1:
  - Source Text 1: 93.00%
  - Source Text 2: 95.00%
  - Source Text 3: 97.00%
  - Source Text 4: 98.00%
  - Source Text 5: 99.00%
Overall Quality

[Bar chart showing the distribution of ratings from two experts.]

- Expert 1: 4, 3, 2
- Expert 2: 3, 2, 1
Evaluation Summary

• Optimistic results
  • 97-98% gap precision

• Experts disagree w.r.t. acceptance
  • 90% (expert 1), 97.5% (expert 2)

• Disagree even more w.r.t. quality
  • score 4: 35% (expert 1), 77.5% (expert 2)

• Room for improvement: overall exercise quality
Overview

+ Comparison with other systems
+ How Exercise Maker works
+ Evaluation

> Conclusion and future work
To Recap

• Exercise Maker automatically generates a variety of lexical and grammatical exercises from arbitrary passages written in English
• Seven types of supported exercises can be further customized
  • adjusting difficulty to accommodate varying learner needs
• Source passages are ranked according to their readability
• Exercises are perceived by TEFL experts as quite useful
• Freely available; ongoing development
Future Work

• Support for other languages
  • Japanese version under development

• New exercise types
  • such as multiple choice

• Further improve exercise quality
  • especially overall (gap precision is good enough for most exercise types)
  • possibly with statistical methods and machine learning
Thank You!

amalafeev@yandex.ru
http://www.hse.ru/staff/amalafeev
References (1/3)


References (3/3)


