A Study on the Automatic Selection of Candidate Sentences and Distractors

Itziar ALDABE\textsuperscript{a,1}, Montse MARITXALAR\textsuperscript{a}, Ruslan MITKOV\textsuperscript{b}

\textsuperscript{a} Computer Languages and Systems, University of the Basque Country, Spain
\textsuperscript{b} University of Wolverhampton

Abstract. The paper outlines a system for generating multiple-choice tests on the choice of English verbs. In particular, the selection of candidate sentences and distractors are discussed in detail. The system is an extension of a system which generates language tests for Basque.

Keywords. automatic question generation, distractor selection, Natural Language Processing, language resources

Introduction

In [1] we described a system which generates grammar exercises for Basque. The original system was extended to generate language drills focusing on verb use in English. This new system makes use of the following resources:

- Two corpora: a proportion of the British National Corpus (BNC), a 100 million word collection of texts of written and spoken language and a mystery/detective novel of 140,000 words.
- The Academic Word List (AWL): the list [2] is divided into mutually exclusive sublists and it contains 570 word families.
- Word frequency lists: a lemmatised frequency list of BNC described in [3] and the General Service List (GSL) [4].
- The Web 1T 5-gram dataset [5] which contains the counts of word sequences up to length five in a corpus derived from Web pages.

With a view to arriving at new heuristics, we first establish a baseline system which is manually evaluated by an expert EFL teacher; her opinion serves as the gold-standard. Then, an experiment is defined in order to improve the baseline system with the new heuristic rules. Finally, the result of the experiment is compared with the gold-standard.

As the system was first developed for Basque, its adaptation to English is an illustration of its ability for multilingual portability. The generation of items was made possible by parsing the BNC texts by Connexor's Machinese Syntax [6]. The output of the analysis is a syntax tree where chunks are not represented explicitly. To this end, we implemented a post-processing module for the English version which takes the output of the parser and produces chunks needed for the generation of test items.

\textsuperscript{1} Corresponding Author: Itziar Aldabe, IXA Research Group, Joxe Mari Korta Centre, Tolosa Etorbidea, 72. 20018 Donostia – San Sebastian, Spain; Email: itziar.aldabe@ehu.es
Unlike the original system for Basque which is an agglutinative language, no morphological generator was needed for English.

1. Sentence Selection

To select sentences to be presented to the teachers, two constraints were defined and set as the baseline. The system only took into account sentences with a particular length, while considering as candidate sentences those where the topic is part of the main clause.

To establish the gold-standard, the expert language teacher had to decide if the given sentence was appropriate to be the stem of the question. She accepted 52.73% of the questions generated by the baseline system.

**First experiment:** We first established the acceptance threshold for BNC and GSL lists, that is, the minimum number of words a sentence should have from the lists to be considered a candidate sentence and set it as the heuristic. The thresholds were established from both corpora before generating the questions.

**Second experiment:** In order to find the most frequent collocation of the verb, we extracted patterns of occurrences from the Web 1T 5-gram dataset. To obtain all occurrences of each verb, the system made use of three different patterns and then identified the occurrences of the patterns which shared the same words. The most common occurrences for the given verb tense and person were counted. Therefore, the heuristic considered a sentence as candidate if its three patterns had a minimum count.

<table>
<thead>
<tr>
<th></th>
<th>GSL</th>
<th>BNC</th>
<th>BNC_ALL</th>
<th>Web 1T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>&gt;= 0.6</td>
<td>&gt;= 0.65</td>
<td>&gt;= 0.75</td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.63</td>
<td>0.62</td>
<td>0.61</td>
<td>0.65</td>
</tr>
<tr>
<td>Recall</td>
<td>0.77</td>
<td>0.79</td>
<td>0.86</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Table 1. Precision and recall of the 1st and 2nd experiments

The results show that there is not a significant difference between lists and the patterns of occurrences in terms of precision. However, a more detailed study of the Web 1T dataset could yield new results and better recall.

2. Distractor Selection and Generation

One of the novel aspects of this system is its ability to automatically generate verbs as distractors; these were presented to the user in ranked order. The selected verbs are part of the AWL. The baseline system classified the candidate distractors on the basis of semantic similarity between verbs according to the distributional data (based on BNC). The distractors were compared for similarity employing the information radius (IR) measure [7]. Furthermore, all candidate distractors had to match the target verb in terms of transitiveness/intransitiveness, tense and person. The expert teacher marked 94.07% of the distractors which were part of the accepted questions, as valid.  

---

2 BNC takes into account the first 2,000 entries. BNC_ALL is the entire list.
Third experiment: In order to establish if there was any difference in terms of appropriateness when including the verbs from the previous sublists as candidate distractors, we divided the distractors evaluated by the teacher according to their sublists. The results show that the use of more data does not result in inferior performance. Thus, the use of the verbs from previous AWL sublists was set up as a parameter of the system.

Fourth experiment: In order to find out if the context could play a beneficial role in the selection of distractors, we obtained patterns for each verb. In our experiment, the lemmas of the contiguous words, their part-of-speech, and the combination of lemmas and their part-of-speech (POS) were used as variables of the patterns. This way, the heuristic discarded the candidate distractor in case the patterns match.

Fifth experiment: The Web 1T 5-gram dataset was used to obtain a language model which predicted the probability of occurrence of a word sequence. For that, we used the smoothing method proposed by [8] and we experimented with a 3-gram language model and a 5-gram language model. After analysing both language models with questions collected from preparatory tests of the CAE exam, a new heuristic rule was defined: a candidate distractor is confirmed as distractor if its probability is at least lower than a third of the correct 3-gram probability and is greater than zero.

The comparison between the gold-standard and the system performance showed that the patterns did not yield better results. After establishing the new heuristic in the fifth experiment, the precision was 94.30% and the recall 37.84%. There is not significant improvement; however the heuristic offers a way to select different distractors taking into account the context of the candidate sentence. Otherwise, the generated distractors would always be the same for each verb.

3. Conclusions

In this paper we outline a system which generates multiple-choice tests based on selecting the correct verb in an English sentence. The automatic generation of questions offers to teachers the opportunity to use ‘real-world’ texts in the form of sentences extracted from corpora. This system is an extension and modification of a system developed initially for Basque.

References