Automatic Generation of Lexical Exercises*

Alena Fenogenova[1] and Elizaveta Kuzmenko[2]

[1] National Research University Higher School of Economics, Moscow, Russia
alenush93@gmail.com,
[2] National Research University Higher School of Economics, Moscow, Russia
lizaku77@gmail.com

Abstract. We propose an approach towards automatic generation of lexical exercises for learners of English. The techniques and tools used for generation of five different exercise types are described. We provide examples and evaluate the quality of generated exercises. We also compare the exercises generated on the basis of two different corpora by conducting an experiment. In the experiment learners complete both automatically generated exercises and exercises from coursebooks, and the results reveal which corpus is better suited to the generation of exercises.

Keywords: language exercise generation, corpus, English as foreign language, lexical exercises

1 Introduction

Language lexical tests play an important role in the process of learning a foreign language. They help learners to increase lexical competence, to master new constructions, to put the recently learned words and collocations into real-world contexts, and thus expand their proficiency in academic English.

At the same time, language exercises are expensive to create manually. Apart from that, when instructors make up exercises, there is always a risk that the exercises will not sound authentic. Therefore, it may be more beneficial to generate exercises automatically. To add validity to the exercises, it may be necessary to produce them on the basis of a suitable corpus (chosen by the mode of language, genres, topics and specific characteristics of texts). The first issue our research deals with is contributing to the area of exercise generation by exploring new means by which the exercises can be created.

Moreover, in the present paper we elaborate on a particular type of exercise, omitted in the previous work. We produce lexical exercises designed for learning academic collocations. The importance of collocations in language learning is often underestimated, and while students practice grammatical constructions and enrich their vocabulary with

* The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2016 (grant 16-05-0057 Learner corpus REALEC: Lexicological observations) and supported within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program. The preliminary version of this paper was presented at the international conference on computational linguistics “Dialogue” in 2016.

single words, collocations are often simply learnt by heart or even go unnoticed [12]. We aim to facilitate the process of exercise generation for academic collocations and to test them among students.

Our tests are created on the data of the British National Corpus [14] and the British Academic Written English Corpus [2]. We carry out an experiment to compare the exercises generated on the basis of these two corpora with the ones taken from traditional coursebooks for English learners. Thus, we evaluate their quality and observe that their performance is on par with the exercises taken from traditional learner books.

1.1 Related work

The history of automatically generated exercises can be traced back to 1997, when Roger Levy introduced the notion of computer-assisted language learning. As stated in his work, "Computer Assisted Language learning (CALL) may be defined as the search for and study of application of the computer in language teaching and learning." [15]. Since then the automatic generation of learning content for exercises, presentations, teaching materials and courses has become a widely spread practice. Another established educational trend is blended learning [8], which satisfies the requirement for using online materials as well.

Various systems for automatic exercise generation have been developed in recent years. They differ in numerous aspects: 1) language supported, 2) types of exercises, 3) sources of creation, 4) the target aspect of learning (lexical or grammatical).

Most state-of-the-art approaches use various corpora as a source of data. There are a number of papers describing development of exercises from different corpora. In [17] the advantages and drawbacks of using corpora as a source of generation are outlined and possible improvements of such an approach are proposed. All in all, corpora are used as a general solution in automatic generation of various exercises for different languages [1] [4]. Apart from corpora, the research on exercise generation involves other sources such as ontologies and thesauruses [5] [10], especially for question generation tasks and vocabulary assessment of the students [9]. Another method uses the web as the data resource – there are a lot of systems that are designed on the basis of different word lists, dictionaries and datasets parsed from online web resources. Such resources are compiled together to produce exercises manually according to written rules [16]. A relatively novel approach is to apply statistical methods and machine learning in this task. [7] proposes a machine learning method for multiple-choice cloze questions – the system is able to select sentences from input student text based on Preference Learning [6], to estimate blanks by Conditional Random Fields [13] and to generate distractors based on statistical patterns of existing questions.

The task of automatically generated exercises demands using a wide range of NLP methods and techniques. The next section will discuss the methods we use in our system.

2 Materials and methods

2.1 Exercises description

Current study is conducted within the research team project of the National Research University Higher School of Economics, and our team's task is to maintain the Russian Error-Annotated Learner English Corpus (REALEC) [11], investigate errors, assess the lexical level of student works, and produce recommendations for them. One of the most efficient ways for students to improve and transform their English writing skills and their proficiency in academic English is by learning collocations. For the purposes of the project special lexical exercises are needed to set up a system of developing lexical skills. Our exercises focus on the ACL (academic collocation lists) [3] which can help to increase students' lexical competence.

We generate 5 types of exercises:

1. **Match collocations**. Two columns with collocation parts in random order are offered to the learner. A student has to match the first part of the collocation with the second one from the given list.
2. **Multiple choice**. A student has to fill in the gaps in sentence. There are 4 choices of one of the part of the collocation are given. Only one answer is correct.
3. **Open cloze**. A learner has to fill in gaps with suitable whole collocation. No candidates for answers are given.
4. **Word bank**. A learner has to fill in the gaps with a suitable collocation. The full list of answer choices is given. No distractors are presented.
5. **Word formation**. A student has to fill in the gaps with derivatives of a part of a given collocation.

Our choices have been inspired by comparable tests from different English exams such as IELTS, FCA, etc. However, most of the exercises in these books are not lexical, but grammatical. At the same time, lexical exercises found in the tests do not focus on the collocation training. Therefore, we designed our exercises guided more by our project aims than the conventionally accepted exercises.

2.2 Data

For generating exercises we have used two different English corpora: BNC and BAWE.

1. **The British Academic Written English Corpus (BAWE)** [Alsop et Nesi, 2009] is an English corpus of academic written texts. The BAWE corpus contains about 6,700,000 tokens in 2761 assessed student writings, ranging in length from 500 words to 5000 words. Texts are evenly distributed across four broad disciplinary areas (Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences) and across four levels of difficulties, altogether thirty five disciplines are represented. The corpus is available online free of charge for academic purposes to researchers who agree to the conditions of use.

2. **British National Corpus (BNC)** [Leech, 1992] is a 100 million word collection of texts of written and spoken British English from the late twentieth century. Collection includes extracts from regional and national newspapers, specialist journals for all ages and interests, academic books and popular fiction, school and university essays and a lot of other texts. Corpus is available online and can be downloaded free of charge for academic purposes as well.

2.3 Methodology

The generation of all exercises is based on the corpora mentioned above and the Pearson's Academic collocation list. The whole system of exercise generation is written by means of scripts in the Python programming language.

Firstly, we take a list of collocations and generate for every item from the list its paradigm. That helps us to detect all possible variants of the collocations from the texts in the corpora. Secondly, our program looks for sentences in the corpus which contain the required collocation. It is possible to configure the length of the sentence and the context (plus/minus one or two sentences around). For specific types of exercise other techniques are implemented.

For the match a collocation exercise nothing but the collocation is needed. Word bank and open cloze exercises are generated only with corpora. An interesting type of exercise is word formation exercises. It presents a head form of the part of collocation to a student, for him to produce the real form. For this type of exercise we used the wiktionary list. The most sophisticated exercise type is multiple choice. In this case we need for generation not only texts with collocations, but also 3 distracting answers apart from the correct one. For this purpose we use a novel approach based on prediction-based word embedding models. The model was build with Word2vec, a tool based on neural networks, which computes vector representations of words from a big dataset, in our case from the BNC corpus. A word2vec model is trained to reconstruct linguistic contexts: the network gets a word and guesses the closest words that occurred in adjacent positions in the input text. Taking as input the word from collocation, the word2vec model returns us several candidates that are likely to occur in the same context as our word. There is a risk of getting completely interchangeable words, but finding the border between too close and too distant words, we find the ideal variants for multiple choice. In order to find candidates that are not interchangeable and at the same time can indeed replace the word in question, we form the list of semantic neighbors and take the words from the 5th place in this list and further. This border was defined empirically during the testing procedure.

Here are the examples of generated exercises (answers are given within the # # signs):

– **Word formation**
The children are producing their own spelling dictionary which is #freely# available to the whole class. (free)
The provision of private medicine both within and without the NHS has remained a #controversial# issue. (controversy)

– **Multiple choice**
  The relative #status# and esteem accorded to husband and wife will be roughly equal. Usually each will have some paid employment outside the home and each may have his or her career.
  Choices: status, autonomy, competence, identity
– **Match collocation**
– **Open cloze** *Environmental and #climatic conditions# have combined with agricultural techniques to produce in Japan exceptionally high yields. This approach she adopted in all her #subsequent work# thereby introducing a revolutionary style of attack on problems of algebra.*
  The problems we face when generating the exercises are as follows:
  - inability to induce the right answer from the provided context;
  - general complexity of exercises due to the authenticity of texts;
  - the possibility of having more than one right answer.
– **Word bank**
  Choices: renewed interest, dynamic system, stress level, unrelated topic
  - *The black feminist movement again threw the crisis of African-American masculinity and gender relations into relief, and so inspired _____ in men's studies by the late 1980s.*
  - *They could see that nature was not static and unchanging, but that it was a _____ that ever changing.*
  - *For example, excessive noise can raise _____ and also gives the impression of a lack of privacy.*
  - *Science is a very broad field comprising of many varied and seemingly _____ from, zoology to astronomy and geology to medicine.*

We conducted an experiment to evaluate how well students cope with our exercises. In the next section the experiment and its outcomes will be described.

3 Evaluation of the exercises

In our experiment we aimed to compare the exercises generated by our program with the exercises manually compiled by language instructors. To do this, we have made three sets of exercises:

– exercises generated on the BNC data;
– exercises generated on the BAWE data;
– exercises taken from English coursebooks.

We have tested only three types of exercise: word formation, multiple choice and word bank. Match collocation and open cloze exercises are more appropriate for learning, when student has a topic or list of collocations to be learned. In our experiment we do not prepare student for the tasks, thus, these exercises are not included in experiment.

We have prepared three set of exercises (BNC, BAWE and from coursebooks), each one containing three different types of exercises. Each type consisted of 4 pieces, the reason for it is not to make experiment time consuming and respectively to make student

tired. All together 36 items were given to 22 students of the program "Fundamental and Applied Linguistics" at National Research University Higher School of Economics. Students did not know which exercises are manually compiled and which are generated. Table 1 shows the percentage of correct answers per person in each set and the number of maximum score of students in each variant. Table 2 represents the percentage of correct answers per student in each set of each type of the exercises.

Table 1. Results of passing the exercises for all types in each set.

Score	Coursebooks	BAWE	BNC
Correct answers, %	86,7	84.5	65.5
Max score, %	100	100	91.6

Table 2. Results of passing the exercises for every type of the exercise.

Score	Word formation	Multiple choice	Word bank
Coursebooks score, %	86,36	89,70	84,09
BNC score, %	75,02	60,22	61,36
BAWE score, %	81,81	78,4	93,18

As we can see, the original exercises taken from coursebooks seem to be more appropriate for learners. At the same time, the exercises generated with the BAWE corpus are very promising and comparable to the manually compiled exercises. The percentage of correct answers in the BAWE set is still high and in the word bank exercise it is even higher than in the coursebooks exercises. The worst performance was found in the tests generated with the BNC corpus, but still the participants answered correctly more than half of the exercises.

Moreover, for multiple choice exercises we can see the distribution of answers (see an example in Figure 1) for each exercise and thus understand which choices are too easy, which ones are not appropriate at all, catch the cases when two choices are equally chosen by the students and thus can be considered interchangeable.

Among all three sets the distribution of right answers (Table 3) shows that in coursebooks the correct variant is very likely to be selected, and it may mean that the exercises are too easy for students. Exercises based on the BNC corpus are, on the contrary, too hard or ambiguous for learners. In further experiments we will carry out an accurate analysis of these data and find the appropriate border between the similarity of choices.

Our experiment shows that the quality of generated exercises is heavily dependent on the corpora used for their creation. At the same time, the computational approach

8. Small developing countries in particular tend to rely heavily on a narrow _____ of primary commodities for their export earnings.

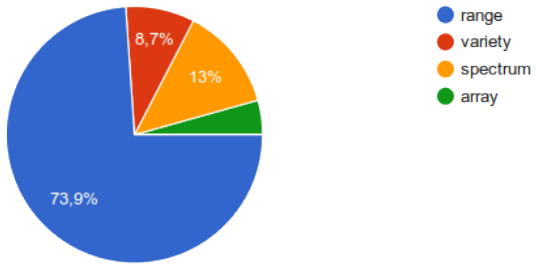

Fig. 1. Example of multiple choice result distribution.

Table 3. Distribution of multiple choice answers.

Score	Coursebooks	BAWE	BNC
Correct answers chosen, %	90,2	80,1	62,5

towards the creation of exercises still can be adopted in any learning environment. In our experiment exercises generated with the BAWE corpus were not significantly worse than the manually compiled exercises while they are more flexible and easier to generate.

4 Conclusion

In our work we presented an approach towards automatic generation of lexical exercises. The data for exercises was taken from the British National Corpus and the British Academic Writing Corpus. Exercises of several types were developed: multiple choice, match, word formation, open cloze and word bank.

We evaluated the quality of the generated exercises by comparing them to the exercises from coursebooks for learners of English compiled manually by language instructors. Three types of exercises was evaluated, namely, multiple choice, word bank and word formation. The comparison included how well language learners perform while completing original and generated exercises. This experiment showed that the quality of exercises can differ depending on the corpus used for their generation. However, the automatically generated exercises are found to be comparable in quality to the ones published in coursebooks.

The usage of generated exercises in the classroom raises many issues, which can be resolved in subsequent work. In particular, we should elaborate on other types of exercises and experiment with English corpora to find out which data is more suitable for learners. Also we intend to improve the quality of exercises by identifying and eliminating ambiguous questions.

Acknowledgments

We express our gratitude to Olga Vinogradova for commenting on our generated exercises and helping us with the organization of the experiment.

References

1. Aldabe I. et al. Arikiturri: an automatic question generator based on corpora and nlp techniques. Intelligent Tutoring Systems. Springer Berlin Heidelberg, 584–594 (2006)
2. Alsop, S., Nesi, H.: Issues in the development of the British Academic Written English (BAWE) corpus. Corpora, 4(1), 71–83 (2009)
3. Ackermann K., Chen Y. H.: Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. Journal of English for Academic Purposes,12, 4 235–247 (2013)
4. Bick E.: Live use of Corpus data and Corpus annotation tools in CALL: Some new developments in VISL. H. Holmboe (red.), Nordic Language Technology, Årbog for Nordisk Sprogteknologisk Forskningsprogram, 171–186 (2000)
5. Brown J. C., Frishkoff G. A., Eskenazi M.: Automatic question generation for vocabulary assessment Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 819–826 (2005)

6. Fürnkranz, J., Hüllermeier, E.: Preference learning, Springer US, 789–795 (2011)
7. Goto T. et al.: Automatic generation system of multiple-choice cloze questions and its evaluation Knowledge Management and E-Learning: An International Journal (KM&EL), 210–224 (2010)
8. Graham C. R.: Blended learning systems CJ Bonk and CR Graham, The handbook of blended learning: Global perspectives, local designs. Pfeiffer, (2006).
9. Heilman M., Eskenazi M.: Application of automatic thesaurus extraction for computer generation of vocabulary questions SLaTE, 65–68 (2007)
10. Knoop S., Wilske S. : WordGap - Automatic generation of gap-filling vocabulary exercises for mobile learning Proceedings of Second Workshop NLP Computer-Assisted Language Learning at NODALIDA, 39–47 (2013)
11. Kuzmenko E., Kutuzov A. : Russian error-annotated learner english corpus: a tool for computer-assisted language learning Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University
12. Meunier, F., Granger, S.: hraseology in foreign language learning and teaching. John Benjamins Publishing (2008)
13. Lafferty, J., McCallum, A., and Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
14. Leech, G.: 100 million words of English: the British National Corpus (BNC). Language Research, 28(1), 1–13 (1992)
15. Levy M.: Computer-assisted language learning: Context and conceptualization. Oxford University Press (1997)
16. Malafeev A. Y. Exercise Maker: Automatic Language Exercise Generation, in: Computational Linguistics and Intellectual Technologies. International Conference “Dialogue”, Issue 14(21), Russian State University for the Humanitie, 441–452 (2015)
17. Wilson E.: The automatic generation of CALL exercises from general corpora Teaching and language corpora, 1–23 (1997)