A.V. Belyi, O.A. Mitrofanova, N.A. Dubinina

DISTRIBUTIVE SEMANTIC MODELS IN LANGUAGE LEARNING: AUTOMATIC GENERATION OF LEXICAL-GRAMMATICAL TESTS FOR RUSSIAN AS A FOREIGN LANGUAGE[1]

Abstract. The report discusses the results of experiments on automatic generation of lexical-grammatical tests in Russian as a foreign language for gap-filling with multiple choice on authentic texts. The proposed method is based on the use of distributional semantic models for Russian and can be expanded to other languages. The toolkit developed in course of the study was evaluated in terms of the quality of resulting tests, evaluation procedure proved their correctness and practical suitability.

Keywords. TORFL, automatic generation of language exercises, gap-filling, lexical substitutions, distributional semantics

1. Introduction

It is well known that language tests aimed at analyzing students' language competence play a significant role in mastering a foreign language, such tests are considered as an effective means of control and are involved in all stages of language learning. Our attention is focused on lexical-grammatical tests, which are used in many manuals and language exams (for example, IELTS, DELE, etc.). However, manual composition of such tests is a particularly complex procedure. We propose a solution to this problem by automation of the process, namely, by the development of a toolkit for generation of lexical-grammatical tasks based on the text supplied to its input and lexical substitution of text slots from predictive embedding models.

[1] Research is performed with partial support of SPbSU Scientific Project № 75254082 «Modeling the communicative behavior of residents of a Russian metropolis in the socio-speech and pragmatic aspects with the use of artificial intelligence methods», RSF grant № 21-78-10148 «Modeling the meaning of a word in individual linguistic consciousness based on distributive semantics».

2. Related work

The problem of automatic generation of training tasks in various subject areas is a matter of discussion both in Russia and abroad. A detailed review of task generation methods and existing software solutions is given in [Balashova et al. 2016; Malafeev 2015]. However, there is currently no available platform for automating the process of task construction, which includes algorithms for selecting target words (TW) and relevant distractors. In both cases, the complexity of the text and/or language level should be taken into account. We consider task generation for students of Russian at levels from A1/A2 CEFR to B2 CEFR. Our choice is grounded on the fact that reaching levels B1 – B2, students learn basic vocabulary and grammar. B2 lexical minimum is about 5,000 lexical units [Andryushina et al. 2015] including semantically ambiguous lexemes. It is at the B2 level that students reach a «plateau» in language learning, which increases the need for practice and control of language competence.

3. Toolkit architecture

Task structure. In TORFL lexical-grammatical tests for gap-filling with multiple choice TWs are excluded from the contexts and are replaced by gaps. The student is asked to choose an answer within a set of options, only one of which is correct. Selection of distractors for TWs may be confined to the problem of lexical substitution [Arefyev et al. 2020]. We exemplify the output of our toolkit by the following task generated for B2 level.

Когда однажды я сказала своей подруге из Германии, что мы пойдём на __________(1) в метро, она очень __________(2) , ведь метро – это обычный __________(3) транспорта и всё! После экскурсии она уже не думала так. Она поняла, что московское метро – это музей под землёй, а многие станции – настоящие __________(4) искусства. При строительстве дизайнеры __________(5) очень дорогие материалы : мрамор, гранит, золото и серебро. В метро можно увидеть __________(6) скульптуры, мозаики, фрески, керамику.

(1) (a) гонку (b) экскурсию (c) экспедицию (d) прогулку

(2) (a) рассердилась (b) испугалась (c) обрадовалась (d) удивилась

(3) (a) вид (b) тон (c) пейзаж (d) размер

(4) (a) произведения (b) стихотворения (c) сочинения (d) изобретения

(5) (a) употребили (b) объединили (c) исправили (d) использовали

(6) (a) необычные (b) оригинальные (c) интересные (d) сложные

Corpus for task generation. Following [Jingjing 2005], we chose the following sources for corpus development: children's and adolescent fiction from M. Moshkov's Library, adapted texts for students of Russian as a foreign language from sites about teaching and learning Russian as a foreign language, original and adapted texts from the RuAdapt corpus [Dmitrieva et al. 2021], school and university textbooks on the Russian language and literature, RFL textbooks. A sample of about 50 million tokens was collected from these resources. To determine text complexity, we used readability metrics (Flesch index, Flesch-Kincaid Readability Test, SMOG, Automated Readability Index, (Gunning) Fog, etc.) [Solovyev et al. 2018] as well as modern readability assessment tools, in particular, *Textometr*[2] [Laposhina et al. 2021]. Text selection was performed on the assumption that the native speaker reaches a level equivalent to B2 – C1 by about 8 – 10 years of schooling, or by the age of 15 – 16 [Dubinina et al. 2021]. Then, morphosyntactic annotation and corpus filtering were carried out: texts were converted to lower case, stop words and non-text elements were removed, POS tagging, lemmatization, and chunking were performed using *spacy* library. Thus, the volume of the corpus reached 30 million tokens.

Training a Word2Vec Model. As the task of test generation implies prediction of lexical substitutes for TWs based on the corpus data, we worked out the procedure of model selection combining fundamental ideas of distributional semantics and machine learning techniques

[2] URL: https://textometr.ru/

implemented in predictive embedding models. We trained a static Word2Vec model [Mikolov 2013] on the corpus-in question. The model is used to select distractors (clusters of similar substitutes of TWs) [Kuzmenko et al. 2016; Perez et al. 2017], and in the given experimental settings CBOW (continuous bag of words) architecture is more efficient. Semantic similarity of two words is determined by the cosine distance between their vectors. The model was trained until convergence with the minimum value of the loss function, training parameters being vector dimension *d*=300 and context window width *w*=10.

Algorithm for selecting contexts for gap-filling. In order to detect text segments with TWs, it is necessary to weigh sentences, in this case extractive summation algorithms are applicable [Pilán et al. 2017]. To meet the need we use extractive summarization based on the BERT neural network model in *bert-extractive-summarizer*[3] library. Most language proficiency tests offer 10, 15 or 20 assignments, text size varying from 10 to 50 sentences or from 150 to 1200 words. In our case, the user can select an arbitrary text to generate tasks. At the stage of preprocessing the required number of sentences with TWs is selected from the text according to their weighs.

Algorithm for choosing a TW in a sentence. When choosing a TW, we follow [Agarwal et al. 2011]. TWs must be contextually restorable and meet a number of requirements. At the moment, the algorithm considers all tokens as TWs except for stop words, named entities, functional parts of speech, and numerals. Further TWs are chosen at random, although preference is given to words with a large number of dependents. The choice of several TWs in one assignment is not excluded, however, TWs must not be adjacent and/or be syntactically linked. Obviously, TWs are unique for the entire text: the same word (and/or its forms) cannot serve as a TW or as a distractor.

Algorithm for distractor selection. We use predictions of the CBOW Word2Vec model to generate distractors for TWs. First, a TW

[3] URL: https://pypi.org/project/bert-extractive-summarizer/

is lemmatized, and then sent as a request to the model. Potential distractors are selected from the list of model predictions, which is filtered in accordance with the lexical minimum chosen by the user. Distractors are expected to be of the same part of speech as the TW, to have approximately the same length, and to be absent in the text. Cosine similarity between distractor vectors and TW vectors should be below a predetermined limit, and lemmata of distractors and TWs are not orthographically close, the given condition is ensured by filtering candidates by means of Levenshtein distance. The set of distractor lemmata is processed by *pymorphy2*[4] morphological tagger to generate forms with the grammatical characteristics corresponding to those of TWs. Finally, a limited set of distractors is selected at random.

Software implementation. We have created a web toolkit[5] *russian-task-generator* on HuggingFace platform. The toolkit implements the proposed task generation method. The data for generating tasks is selected from a user's text, which is uploaded directly into the field of the toolkit or as a text file. For a single text a user can generate tasks of varying complexity, depending on the required L2 lexical minimum. The settings allow to choose a wide range of distractors (from 2 to 9), taking into account that some of them may be incorrect. Our web application supports several input versions: «for students»: a file with tasks and hidden correct answers; «for teachers»: a file with tasks and highlighted correct answers, a file with keys; output tasks on the screen; displaying an online test.

4. Experiments on the generation of lexical-grammatical tasks for Russian as a foreign language

To assess the quality of generated tasks, two experiments were carried out with similar settings, differing in the target audience and the data requested from respondents. Experimental tasks were created on the basis of texts from the «Reading» sections of the TORFL manuals.

[4] URL: https://pymorphy2.readthedocs.io/en/stable/

[5] URL: https://huggingface.co/spaces/a-v-bely/russian-task-generator

For each text L2 level for selecting lexical minimums was determined with the help of *Textometr* toolkit. Following the procedure, worked out in our study, tests with six response options were obtained: the correct answer and five distractors. This is due to the need to evaluate the largest possible number of tasks and distractors at the lowest cost: the standard number of distractors in tasks of this type are commonly reduced to three lemmata.

In the first experiment, the participants were native speakers of the Russian language, specialized in linguistic studies (except for TORFL). Respondents were asked to take tests, indicating the degree of confidence in the chosen answer with a scale from 1 to 3, where 1 is «not confident at all» and 3 is «absolutely confident». The questionnaire was distributed through the GoogleForms platform and corresponds to the recommendations for conducting psycholinguistic experiments.

In the second experiment, professional assessment was carried out by five expert teachers of Russian as a foreign language, including Language Testing Center specialists, St. Petersburg State University. Experts evaluated the tasks for suitability in terms of complexity and goals of testing, the unambiguity of the solution, and the relevance of distractors.

Analysis of results provided in the first experiment (57 responses) is presented in Table 1. Those terms which were selected by the absolute majority of respondents «fully confident» in their choice were considered as relevant distractors. The values of recall, precision and accuracy were calculated in a modified form: $recall = CCR/(CCR{+}UEA)$; $precision = CCR/(CCR{+}UCA)$; $accuracy = (CCR{+}CEA)/TN$, where CCR − the number of confidently correct answers, CEA − the number of confident-erroneous answers, UEA − the number of unconfident erroneous answers, UCA − the number of unconfident correct answers, TN is the total number of responses.

In the first experiment, the best results were shown for levels A1−B1, proving the central position of B1 in the system of levels. For level B2, task development requires additional procedures concerning text peculiarities (a high proportion of polysemic words, emotional expression, etc.). The small number of confidently erroneous answers is significant in this experiment, as it demonstrates a low degree of ambiguity within sets of distractors. At the same time, their complexity was preserved, the number of uncertain answers being close to 25%.

Table 1. Evaluation of experimental results.

Mertics	**L2 level**				**Average**
	A1	A2	B1	B2	
Recall	0,95	0,99	1,00	0,80	0,90
Precision	0,91	0,77	0,95	0,73	0,81
Accuracy	0,87	0,77	0,95	0,63	0,75

The results of the second experiment inspired us to introduce improvements aimed at greater correlation of tasks with the students' competence. We made significant observations on the properties of generated distractors. First of all, there are groups of irrelevant distractors in the tasks, e.g. synonyms close to TWs (*решительный*/*энергичный*, *идеология*/*мораль*, etc.), obsolete words from fiction texts (*орда*, *барыш*, etc.), cohyponyms (*ожерелье*/*кольцо*, etc.), single-root words. Secondly, irrelevant distractors are found in about a third of tasks, while their number does not exceed one or two (out of six), which allows us to use our method taking into account manual selection of relevant distractors. Thus, the authors come to the conclusion that the developed method for generating tasks is suitable for work and application in the educational process.

5. Conclusion

We proposed a method for automatic generation of lexical and grammatical gap-filling tasks with multiple choice. The method is currently adapted for teaching Russian as a foreign language and is based on distributional semantic models, namely, CBOW Word2Vec model architecture. To train the model under consideration, we assembled a research corpus that emulates the language experience of students of Russian as a foreign language. To test the quality of the generated tasks, we developed a web application and conducted two experiments with native Russian language speakers and experienced TORFL experts. Their results show high degree of correctness of the tasks and low ambiguity of distractors. The experts made conclusions concerning about usability of the web application, practical suitability of the proposed method and the possibility of its adaptation for teaching other foreign languages.

References

1. Andryushina N. P., Afanasyeva I. N., Bitekhtina G. A., Klobukova L. P., Yatsenko I. I. (2015) *Lexical minimum in Russian as a foreign language. Second certification level. General language proficiency*. 5th ed., St. Petersburg: (electronic edition). (In Russian)

2. Balashova I. Yu., Volynskaya K. I., Makarychev P. P. (2016) Methods and tools for generating test items from natural language texts. *Modeli, sistemy, seti v ekonomike, tekhnike, prirode i obshchestve.* №1(17). P. 195–202. (In Russian)

3. Balykhina T. M. (2004) *Fundamentals of the theory of tests and the practice of testing (in the aspect of Russian as a foreign language)*: Textbook. M.: Russian language. Courses. 240 p. (In Russian)

4. Blinova O. V., Tarasov N. A. (2021) The complexity of Russian legal texts: assessment methods and language data. *Trudy mezhdunarodnoy konferentsii «Korpusnaya lingvistika-2021»*. SPb.: Izdatel'stvo Skifiya-print. P. 175–182. (In Russian)

5. Dubinina N. A., Ptyushkin D. V. (2021) Levels of testing in Russian as a foreign language in terms of age specificity of schoolchildren. *Rusistika*. Vol.19. № 2. P. 222–234. (In Russian)

6. Laposhina A. N., Lebedeva M. Yu. (2021) Textometer: an online tool for determining the level of complexity of a text in Russian as a foreign language. *Rusistika*. Vol.19. № 3. P. 331–345. (In Russian)

7. Malafeev A. Yu. (2015) Method of automatic creation of lexico-grammatical exercises in wordbank cloze format. *Inostrannyye yazyki v vysshey shkole.* №2(33). P. 88–95. (In Russian)

8. Jingjing L. (2017) The system of principles for the selection of educational texts for the formation of intercultural competence of foreign students-philologists (level B2). *Vestnik Tomskogo gosudarstvennogo pedagogicheskogo universiteta.* №7(184): 128–133. (In Russian)

9. Agarwal M., Mannem P. (2011) Automatic Gap-fill Question Generation from Text Books. In: *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*. Portland, Oregon. P. 56–64.

10. Arefyev N., Sheludko B., Podolskiy A.V., Panchenko A. (2020) Always Keep your Target in Mind: Studying Semantics and Improving Performance of Neural Lexical Substitution. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona. P. 1242–1255.

11. Dmitrieva A., Tiedemann J. (2021) Creating an Aligned Russian Text Simplification Dataset from Language Learner Data. *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Stroudsburg, 2021. P. 73–79.

12. Kuzmenko E., Fenogenova A. (2016) Automatic generation of lexical exercises. *CLLS 2016. Computational Linguistics and Language Science. Proceedings of the Workshop on Computational Linguistics and Language Science*. Aachen: CEUR Workshop Proceedings. Vol. 1886. P. 20–27.

13. Malafeev A. (2014) Automatic Generation of Text-Based Open Cloze Exercises. In: *Communications in Computer and Information Science, Ignatov D., Khachay M., Panchenko A., Konstantinova N., Yavorsky R. (eds) Analysis of Images, Social Networks and Texts AIST 2014*. Springer. Vol. 436. P. 140–151.

14. Malafeev A. (2015) Exercise Maker: Automatic Language Exercise Generation. In: *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue» (2015).* Moscow: RSUH. Issue 14(21). P. 441–452.

15. Mikolov T., Chen K., Corrado G.S., Dean J. (2013) Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations ICLR*.

16. Pérez N., Cuadros M. (2017) Multilingual CALL Framework for Automatic Language Exercise Generation from Free Text. *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics.* Valencia. P. 45–52.

17. Pilán I., Volodina E., Borin L. (2017) Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues.* Vol. 57(3). P. 67–91.

18. Solovyev V., Ivanov V., Solnyshkina M. (2018) Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of Intelligent and Fuzzy Systems*. Vol.34. Issue 5. P. 3049–3058.

Belyi Andrei Vladimirovich
Saint-Petersburg State University (Russia).
E-mail: a.v.belij@yandex.ru

Mitrofanova Olga Alexandrovna
Saint-Petersburg State University (Russia).
E-mail: o.mitrofanova@spbu.ru

Dubinina Nadezhda Alexandrovna
Saint-Petersburg State University (Russia).
E-mail: dnadine@yandex.ru