L2 text Recommendation System for the Russian Language

Nikolay Babakov[1], Natalya Isupova[2], Anastasiya A. Bonch-Osmolovskaya[3]

Olga Eremina[4]

Abstract

Language learning is a complicated process that includes numerous actions aimed at the accumulation of a learner's real-life language experience. Reading is one of the most important language-learning processes, and searching for texts which will be appropriate for the current language level of a learner is quite a time-consuming task. We propose a system for automatic text recommendation for L2 learners of the Russian language, based on an evaluation of learners' language competence.

Though the system has been designed for the Russian language, its general principles can be applied to any other language.

Keywords: text recommendation, L2 texts reading.

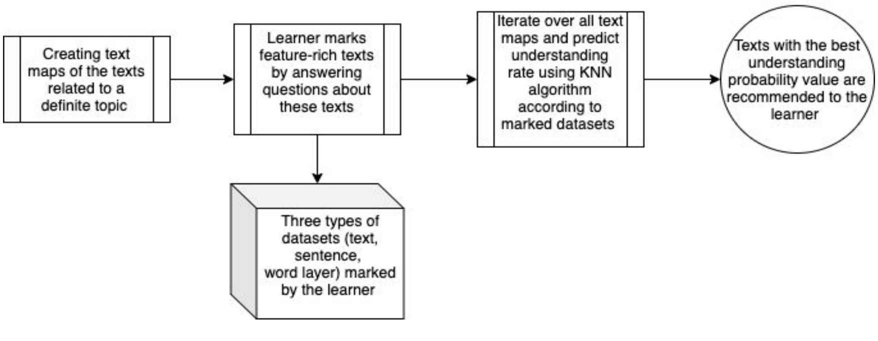
1. Introduction

So far, multiple L2 text recommendation approaches have been developed. The classical approach has been based on a learner's passing CEFR-like test and reading texts in of the corresponding CEFR level. This approach is considerably straightforward, therefore it does not take enough text complexity properties into account, and most words positioning in a definite CEFR level raise quite an arguable question.

At the same time, we have found some interesting approaches which used tests for further recommendations (Kurdi, M., 2018), but in that case, a hypothetical learner would be supposed to read text and to manually score such parameters as syntaxis, morphology and so on, which would actually be not so easy to evaluate.

We have tried to accumulate all experience of our colleagues and developed L2 text recommendation system which takes into consideration all nuances of text reading complexity and provides the most accurate information about user knowledge in each domain which effects further recommendations (see Figure 1).

Figure 1. System outline


Creating text maps of the texts related to a definite topic
Learner marks feature-rich texts by answering questions about these texts
Iterate over all text maps and predict understanding rate using KNN algorithm according to marked datasets
Texts with the best understanding probability value are recommended to the learner
Three types of datasets (text, sentence, word layer) marked by the learner

[1] Nikolay Babakov, Moscow, Russian Federation, bbkhse@gmail.com

[2] Natalya Isupova, Moscow, Russian Federation, nata_isupova@inbox.ru

[3] Anastasiya A. Bonch-Osmolovskaya, Moscow, Russian Federation, abonch@hse.ru

[4] Olga Eremina, Moscow, Russian Federation, oeremina@hse.ru

2. Method

2.1. Evaluation of learner's language competence

2.1.1. Textual annotation

To collect the information about a learner's language competence, we use 10-15 texts with a number of questions related to them. The texts are ranged either within all possible CEFR levels or within the level which the learner is supposed to have. There are several questions about each text learner should answer. The questions are designed in such a way that only one definite sentence in the text is enough for a correct answer to a given question.

The texts for both tests and further recommendations should be related to a specific topic. In this research, we have used the Lenta.ru news website. The structure of the website allowed us to have all texts marked with a definite topic.

The texts for a test are selected in accordance with the amount of high frequency-specific words each text contains. The ideal test should contain a minimum amount of texts with words that are the most typical for the topic inside which we will perform the recommendation.

Therethrough, when the user gives a correct or an incorrect answer to a question, he or she puts "+" or "-" mark to the corresponding group of words which shows whether these words and the sentence are understood or not. Such a marking is distributed within all layers of text features and then used for the collection of the three datasets which thoroughly demonstrate the learner's knowledge.

2.1.2. Learners Competence Datasets Collection

The next step is the collection dataset which will project the learner's knowledge into three domains as the text consists of words, sentences, and overall text features. All datasets' vectors have a target variable which shows whether the elements of a text in a corresponding layer are understood correctly or not.

Collocations Layer

We use the collocations layer to find out the group of words a learner has understood or not when answering the questions.

Each word or collocation inside is marked with 0 (for an incorrect answer) or 1 (for a correct answer) target variable.

Sentences Layer

Each sentence is represented as a vector, which includes a percentage of non-trivial morphological features such as specific parts of speech and other complicated language elements. The target variable for each sentence is 1/0 for correct or incorrect answer.

Text Layer

Each text is represented with the text complexity metrics such as LIX (Brown, J., Eskenazi, M., 2005), type-token ratio and averaged syntax ratios of sentence vectors. The target variable is within 0 and 1 which represents the percentage of correctly answered questions. For example, if there are ten questions for a text and a learner answered six of them correctly, the value for this text will be 0.6.

Table 1. Learner Knowledge Dataset Representation

Element id	**X**	**Target Variable**	**Comment**
COLLOCATIONS DATASET			
Word_1	word2vec(He)	1	Words from a correctly answered sentence
Word_2	word2vec(travels)	1	
Word_3	word2vec(alone)	1	
Word_4	word2vec(They)	0	Words from an incorrectly answered sentence
Word_5	word2vec(were)	0	
Word_6	word2vec(foolish)	0	
SENTENCES DATASET			
Sentence_1	Percentage of difficulty of a text under study, POS and mean syntax dependencies	1	Correct answer
Sentence_2		0	Incorrect answer
TEXTS DATASETS			
Text_1	LIX, TTR, average sentences properties	8/10	Percentage of correctly answered questions
Text_2		6/10	

2.2. Text Recommendation Process

2.2.1. Text Preprocessing

The text recommendation process requires the availability of all potentially recommended texts parsed to JSON-file.

The text is pre-processed by means of UDpipe. We need to extract three types of features and create a JSON text map which will include all mentioned features.

2.2.2. Recommendation Algorithm

The recommendation algorithm matches the vectors collected during competency evaluation with the text maps' vectors. KNN approach is applied to all feature vectors within all three feature layers to predict the understanding rate for each potentially recommended text.

The process of applying a recommendation algorithm to each potentially recommended text is illustrated in Table 2.

Table 2. Text Understanding Prediction Overview for All Three Layers

Dataset Type	**Prediction Object**	**Result**	**Prediction Result Meaning**	**Recommendation reference**
Words dataset	Word	W, +-val [0...1]	The value corresponds to the normalized tf-idf which is positive for correct understanding and negative for incorrect	Percentage of correctly understood words values
Sentences dataset	Sentence	S, [0...1]	0 for non-understanding 1 for understanding	Percentage of correctly understood sentences
Text dataset	Text	T, [0...1]	Percentage of correctly understood sentences	Percentage of correctly answered questions

As a result, for each analyzed text we have a vector with three numbers ranging from 0 to 1. The result can be illustrated as follows [0...1,0...1,0...1]

It is assumed that for a good understanding of the text, it is necessary to be familiar with 80% of the information or grammar rules. Therefore, the closer each calculated value to 0.8, the more likely the texts are recommended. We call this value Understanding Deviation and mark it as UD. The formula for calculating UD is shown below:

$$UD = \sqrt{\frac{(W - 0.8)^2 + (S - 0.8)^2 + (T - 0.8)^2}{3}} \qquad (1)$$

where
W - percentage of words in an examined text with correct understanding prediction
S - percentage of sentences in an examined text with correct understanding prediction
T - percentage of abstract text-related questions with correct answer prediction

Finally, when the analysis of all the text is finished, we have the UD corresponding to each text in a text database. The texts are sorted by value, and the three texts with the least standard deviation are recommended for a potential learner.

3. Model Performance Evaluation

We employ two different approaches to the evaluation of the recommended text understanding. The first one provides text-related questions, referring to several definite sentences in the text given. The second approach suggests asking the user to mark the sentences which include the biggest amount of any predefined features within some scale-like "I do not understand this sentence at all … I understand the general sense … I fully understand it". The choice of an approach depends on time availability of test designers. The first approach which implies questions to the recommended texts is more time consuming because the questions are supposed to be prepared manually.

After applying any of the named feedback models, we have the same knowledge datasets but they have different ways of applications. They are used for model performance evaluation.

The model is supposed to work properly if the examinee answers 80% of questions related to the recommended text or marks 80% of sentences from the text as “Fully understood”.

The model itself can be tested by providing learners with the recommended texts, collecting their answers and calculating standard deviation from 0.8.

Thus, the formula is the same as the previous one but we handle not predicted yet real values.

$$UD* = \sqrt{\frac{(W^* - 0.8)^2 + (S^* - 0.8)^2 + (T^* - 0.8)^2}{3}} \qquad (2)$$

W* - percentage of correctly understood words’ tf-idf normalized value in the recommended text
S* - percentage of correctly understood sentences in the recommended text
T* - percentage of correctly answered the text related questions

Obviously, the less UD* value we had, the better this system has worked. That is why the final metric of this is R-square-like (let’s name it UR-square which stands for Understanding Rate) calculated, using the following formula:

$$UR^2 = 1 - \sqrt{\frac{(W^* - 0.8)^2 + (S^* - 0.8)^2 + (T^* - 0.8)^2}{3}} \qquad (3)$$

For the time being, we have performed tests with 20 learners (each of them had three texts recommended for reading), and the UR-square achieved turned out to be 0.75.

4. Conclusion

In this paper, we have presented a new approach for L2 text recommendation according to learners' skills which are checked using the most natural way of text questions answering.
The next steps will be applying more tests with learners with different backgrounds, applying the approach to other languages and tuning the test parameters to reach a higher level of accuracy.

Reference List

Brown, J., Eskenazi, M. (2005) *Student, Text and Curriculum Modeling for Reader-Specific Document Retrieval.* Proceedings of the IASTED International Conference on Human-Computer Interaction, Phoenix

Kurdi, M. (2018). *A Reading Recommendation System for ESL Learners Based on Linguistic Features*. Proceedings of the Thirty-First International Florida Artificial Intelligence Research Society Conference, Melbourne, Florida