Comparison of Algorithms for Automatic Terminology Extraction on Material of Educational Texts on Biology

Farida Ismaeva
Kazan Federal University
Kazan, Russian Federation
ismaevafarida@gmail.com

Elijah Tomin
Kazan Federal University
Text Analytics Laboratory
Kazan, Russian Federation
elijahtomin@gmail.com

Elvira Sharifullina
Kazan Federal University
Kazan, Russian Federation
ehvi-ehvi12@rambler.ru

***Abstract*—Two methods of automatic extraction of terms are considered in the research. An experiment on the basis of educational texts from biology textbooks, that are used in state schools of Russia, is carried out. The way of evaluation of the proposed methods for correspondence of analyzed parameters to validation values is offered.**

I. Introduction

Modern methods of natural language processing provide tools and methods for the text aggregation on a wide range of parameters. Text complexity according to the research [1] is analyzed: "…via linguo-mathematical, linguo-cognitive and machine learning methods". Text complexity identification is possible with text analysis on its lexical level. Thus, the level of terminological diversity, the volume and number of terms in the text are metrics that can be counted. The number of terms in a text can serve as a measure of their complexity. Therefore, in the field of pedagogy it is important to be able to distinguish terms in texts [2].

In the field of school education, the analysis of texts for the presence of terms can allow the teacher to make a decision on the appropriateness of using an educational text for a certain level of students' knowledge. Often, the use of materials that do not correspond to the cognitive abilities of students causes insufficient assimilation of educational material.

Although school textbooks often provide a list of terms used, an examination of them reveals that such lists are usually far from complete. And as far as we know, there are no dictionaries of terms for schoolchildren, at least for the Russian language. Thus, the task of automatic extraction of terms from educational texts aimed at schoolchildren is relevant.

II. State-of-the-Art

The task of extracting terms is one of the most difficult tasks in natural language processing. There are different approaches of term extraction. The classical approach is the statistical one, that is based on the frequency of occurrence of words and phrases in a text or corpus compared to their frequency in the language as a whole. Typically, this approach is used to extract 2-word and longer terminological word combinations [3]. Its obvious disadvantage is that long word combinations may occur too infrequently. Another approach is the use of templates - the constructions typical for terminological word combinations, which contain information about the parts of speech of the words included in the construction, the syntactic relations between them [4]. This approach is characterized by high labor intensity in compiling a set of templates.

Recently, the use of machine learning using deep learning neural networks has become predominant. This approach requires an extensive training set and specification of a set of features specific to the terms [5]. In the research [6] different neural network training methods for term extraction in Russian are compared. If the subject area of the text is clearly defined and there are large digital dictionaries for this area, it is possible to use them. In practice, hybrid methods with various additional tools are often used [7].

In the research of Braslavsky and Sokolov [8] the next approaches are considered: the usage of maximum length word-forms as term candidates via stop-words and text markup (MaxLen); the extension of terms that are large in length based on the original set of basic terms by means of Web search engines (k-factor); syntactic analysis of word chains based on word form meta-tagging (AOT).

The research of Amir et al. [9] is based on the complex application of methods (statistical method with corpus aggregation; meta-tagging usage; TF-IDF; classification by gradient descent; neural network model training, validation by measures of precision, recall and F-measure). In the research of Augustyniak et al. [10] extraction of specific terminology (Aspect Term Extraction) based on vector representations of words (word2vec), the use of meta-tagging of word vectors, preliminary BIO-tagging (inside-outside-beginning) of chains of candidate word forms into terms, training of a recurrent neural network on the material of BIO-markup are described.

The machine learning approach to extracting terms from text is based on the NER learning technology. NER itself stands for Named Entity Recognition. It is a subfield of Natural Language Processing (NLP) which involves identifying and classifying named entities in unstructured text, such as people, organizations, locations, dates, and numerical values. It is an important task in information extraction, document classification, and text analysis. By automatically identifying and categorizing named entities, NER can assist in extracting useful information from large volumes of data, such as news articles, customer feedback, or social media posts. NER has a wide range of applications, such as chatbots, recommendation systems, and search engines. NER can be used as a tool to extract terms if the latter are pre-defined.

In a number of cases, including in biomedicine, the term extraction problem and the NER problem are very close and are solved by a common method [11]. The research of Bruches and Batura [12] is based on the dictionary approach imposed on the corpus, its automatic markup, followed by training of the NER-model and the model quality measures usage (precision, recall, F-measure). The research of Starostin, Bocharov and Alexeeva [13] as a method of term recognition uses fact extraction track to identify the related NER-entities.

III. Data and Methods

For term extraction we tested two approaches: the dictionary approach and the NER-approach. The dictionary approach is based on superimposing a dictionary of terms of a certain field of knowledge on the text. The algorithm [14] is based on a six-step analysis of n-grams of words in term candidates with of word n-grams in the dictionary. Term candidates are pre-lemmatized. The result of the algorithm is an html-document with highlighted terms.

The six-step analysis of the text is justified in view of the specifics of the statistical significance of the n-grams of the terminological dictionary [14] presented in Table I, formed on the basis of the terms of Gilyarov's biological encyclopedic dictionary.

TABLE I. Statistics of n-grams in the Dictionary of Biology.

N-grams	1	2	3	4	5	6	7	8
N of terms	21552	58781	6887	2150	168	50	8	3

This feature of the dictionary allows us not to analyze the text at the level of 7 and 8-word n-grams. The dictionary approach is briefly described as follows: the word form of each term is processed and the initial word form (lemma) is returned. The lemmatized list of terms is combined with the initial list, a dictionary (data structure in Python) is created, where the key is the lemma of the term, and the value is a terminological unit, designed with the preservation of all affixations presented in the encyclopedic dictionary, note: клетка организм_ – клетка организмА (body cell), эукариотнЫЙ клетка – эукаритнАЯ клетка (eukaryotic cell). As Russian is a language that is ineffective to preprocess using stemming, we use lemmatization to save some lexical features.

The text processing algorithm for lemmatization uses pipeline technology (pipeline). The input is text. The text is segmented into paragraphs, the set of paragraphs is formed into a tuple with data. Further, the elements of the tuple for processing are fed to the pipeline through a cyclic construction. The pipeline architecture contains such text processing components as: tokenizer, part-of-speech tag determiner, lemmatizer, syntactic parser, and named entity recognizer. For our purposes, only the tokenizer, the part-of-speech tag qualifier, and the lemmatizer are used. So, a paragraph of text then contains linguistic features; the text is divided into words and symbols (tokens), each token contains information about the part of speech of the word form and its initial form.

The elements of the tuple, containing linguistic features, are then fed to the input of six cyclic constructions. The first construction takes the first six-word forms and, with a step of one token, collects sequences of word form lemmas for checking with the keys of the terminological dictionary. Provided that a match is found, the ordinal numbers of the text tokens defined as terms are appended into the preformed nested tuple. This allows us not to re-apply to words and phrases already defined as a term. 5 further subsequent cyclic constructions are guided by the same principle. The difference is in the selection of word forms in descending order from 6-component terms to single-word terms. The pipeline of the dictionary approach algorithm is presented in Fig. 1.

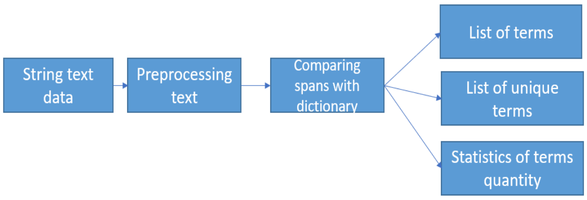
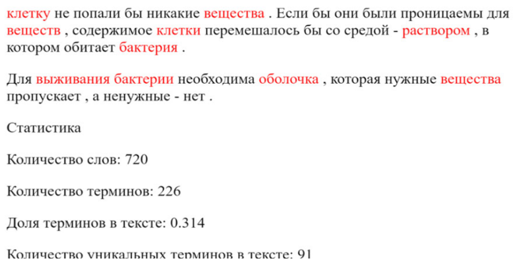


Fig. 1. Dictionary approach pipleline

Upon completion of these six cycles, the filled tuple with lemmatized terms is processed. The tuple is superimposed on a dictionary of terms and lemmas are replaced by the usual forms of words, which are given as the titles of dictionary entries, note: эукариотнЫЙ клетка – эукариотнАЯ клетка (eukaryotic cell). After that, a new dictionary is formed, the key of which is the term, and the value is the occurrence of the term in the text. So, the statistics of the frequency of terms is displayed.

Basic html tools are used to visualize the operation of the algorithm. The analyzed text goes into the html document, the word forms of the terms are highlighted. At the end, text statistics is displayed. “Количество слов” parameter stands for the number of words, “Количество терминов” parameter stands for the number of terms, Количество уникальных терминов в тексте” parameter stands for the number of unique terms and parameter “Доля терминов в тексте” stands for proportion of terms in the text. An example of the visualization of the algorithm is shown in Fig. 2.



клетку не попали бы никакие вещества . Если бы они были проницаемы для веществ , содержимое клетки перемешалось бы со средой - раствором , в котором обитает бактерия .

Для выживания бактерии необходима оболочка , которая нужные вещества пропускает , а ненужные - нет .

Статистика

Количество слов: 720

Количество терминов: 226

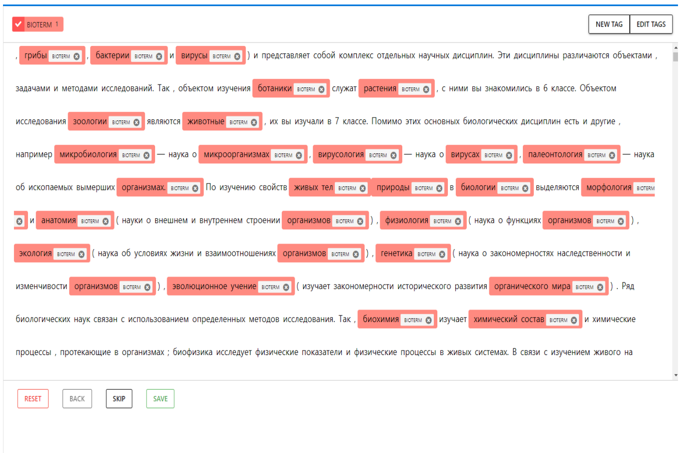
Доля терминов в тексте: 0.314

Количество уникальных терминов в тексте: 91

Fig. 2. Visualization of dictionary approach algorithm’s result

For the implementation of NER-approach, the following stages of work were carried out: collection of a corpus of educational texts in biology; corpus cleaning; span labelling and model training.

At the first stage, 10 biology textbooks included in the State Textbook list were collected. The second stage included work with optical text analyzers for extracting texts, subsequent cleaning and their restoration. At the third stage, using the NER annotator program, the marking of words and phrases (spans) was implemented. An example of markup is shown in Fig. 3.

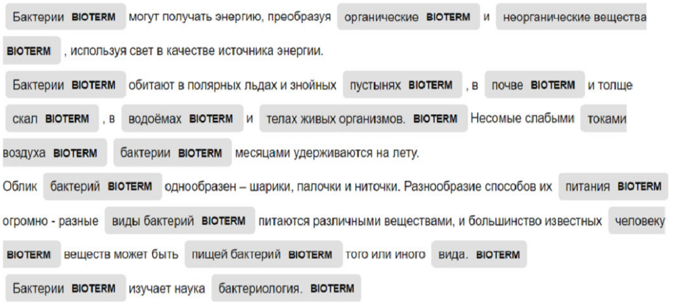


, грибы BIOTERM , бактерии BIOTERM и вирусы BIOTERM) и представляет собой комплекс отдельных научных дисциплин. Эти дисциплины различаются объектами ,
задачами и методами исследований. Так , объектом изучения ботаники BIOTERM служат растения BIOTERM , с ними вы знакомились в 6 классе. Объектом
исследования зоологии BIOTERM являются животные BIOTERM , их вы изучали в 7 классе. Помимо этих основных биологических дисциплин есть и другие ,
например микробиология BIOTERM — наука о микроорганизмах BIOTERM , вирусология BIOTERM — наука о вирусах BIOTERM , палеонтология BIOTERM — наука
об ископаемых вымерших организмах. BIOTERM По изучению свойств живых тел BIOTERM природы BIOTERM в биологии BIOTERM выделяются морфология BIOTERM
и анатомия BIOTERM (науки о внешнем и внутреннем строении организмов BIOTERM) , физиология BIOTERM (наука о функциях организмов BIOTERM) ,
экология BIOTERM (наука об условиях жизни и взаимоотношениях организмов BIOTERM) , генетика BIOTERM (наука о закономерностях наследственности и
изменчивости организмов BIOTERM) , эволюционное учение BIOTERM (изучает закономерности исторического развития органического мира BIOTERM) . Ряд
биологических наук связан с использованием определенных методов исследования. Так , биохимия BIOTERM изучает химический состав BIOTERM и химические
процессы , протекающие в организмах ; биофизика исследует физические показатели и физические процессы в живых системах. В связи с изучением живого на

Fig. 3. NER-annotator interface

After markup, the annotated corpus is formatted into a json file as a dictionary with nested data structures, which contains the name of the training entity tag and the boundaries of its segment (span) in the training example. A total of 12249 training examples were marked up.

So, the NER-approach is based on machine learning. A marked-up corpus [15] with spans (symbolic segments) in the training example of the corpus is fed to the input of the algorithm. The corpus markup is done by the NER annotator program [16]. After the corpus is formed, the program generates a json file with the BIOTERM tag. The result of the algorithm [17] is a NER model. After model is created, it is possible to highlight terms of an input text. The DisplaCy render in the SpaCy library allows us to cope with it. An example of visualization is shown in Fig. 4.



Бактерии BIOTERM могут получать энергию, преобразуя органические BIOTERM и неорганические вещества
BIOTERM , используя свет в качестве источника энергии.
Бактерии BIOTERM обитают в полярных льдах и знойных пустынях BIOTERM , в почве BIOTERM и толще
скал BIOTERM , в водоёмах BIOTERM и телах живых организмов. BIOTERM Несомые слабыми токами
воздуха BIOTERM бактерии BIOTERM месяцами удерживаются на лету.
Облик бактерий BIOTERM однообразен – шарики, палочки и ниточки. Разнообразие способов их питания BIOTERM
огромно - разные виды бактерий BIOTERM питаются различными веществами, и большинство известных человеку
BIOTERM веществ может быть пищей бактерий BIOTERM того или иного вида. BIOTERM
Бактерии BIOTERM изучает наука бактериология. BIOTERM

Fig. 4. Visualization of NER-approach model implementation

The list of identified terms is displayed in a tuple, the element of the tuple is a term. Using the SpaCy pipeline, terms are lemmatized, data is displayed on the total number of terms, the number of unique terms, the proportion of terms in the text, as well as a list of terms by frequency and alphabet.

The learning algorithm is based on tok2vec, which uses a recurrent neural network, which contains 64 hidden layers and a volume for one training period – 1000 training examples.

Such parameters are necessary to improve the accuracy of the term recognizer. 55 training periods were conducted. 10 textbooks were used to train the model, including: 5th grade (5 textbooks), 5th-6th grade (3 textbooks), 7th grade (1 textbook), 7-8 grade (1 textbook). The corpus consists of 255081 word-forms. To validate the model we use precision, recall and F-measure metrics [11]. They are calculated as follows:

$$Precision = \frac{TP}{TP+FP}; \quad Recall = \frac{TP}{TP+FN};$$

$$F1 = 2 \times \frac{Precision \ \times Recall}{(Precision + Recall)} \qquad (1)$$

where *TP* is True Positive predictions, *FP* – False Positive predictions, *FN* – False Negative predictions.

Validation of the NER model was carried out on 5 biology school-oriented texts [15] that are not contained in the training dataset. The results of model validation are in Table II.

TABLE II. VALIDATION OF THE NER MODEL.

№	Grade	TP	FP	FN	Precision	Recall	F1
1	6	278	6	36	0.97	0.88	0.91
2	7	200	13	24	0.93	0.89	0.91
3	8	169	5	15	0.97	0.87	0.91
4	9	94	20	16	0.82	0.85	0.84
5	10	159	6	40	0.96	0.79	0.87

Table II shows metrics of precision, recall and F-measure. There is a separation into two categories, where on texts up to grade 9, the F-measure is 91%. On texts from grades 9 to 10, the F-measure is 87%. There is the next tendency: the quality of the model's work gets worse as well as the level of a text gets higher. This trend, we suppose, is caused by the nature of the training dataset, which contained more text data from 6-8 grade level textbooks than 9-10 grade level ones. Overall, we used standard pipeline for SpaCy library to build a NER-model. The pipeline scheme we used to train RNN is presented in Fig. 5.

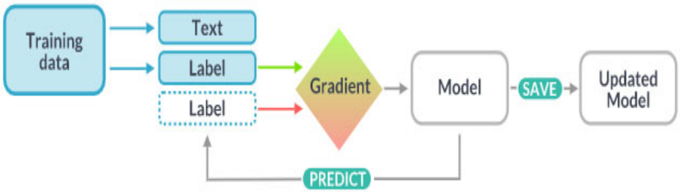


Fig. 5. NER-model pipeline scheme

The comparative analysis of the approaches is based on the degree of deviation from the values of validation indicators for a number of parameters. These parameters are: the number of terms in the text, the proportion of terms in the text, the number of unique terms in the text.

The first parameter is calculated by counting all found n-grams in the analyzed text. The second parameter summarizes the n-gram tokens and divides them into the word tokens of the entire text. The third parameter is calculated by counting the terminological variety (the term identified for the first time later in the text is not considered). The dictionary-based algorithm already contains tools for calculating these parameters.

An additional algorithm was written to calculate the parameters when applying the machine learning approach [18]. The data for comparative analysis on three parameters contain validation values for texts, the values of the dictionary approach and the approach with NER. In total, 15 educational texts [15] on biology for the secondary level of education are analyzed.

IV. Results

To validate both approaches, we have generated variational statistics on three parameters (the proportion of terms, the number of terms, the number of unique terms). With the help of proven methods, it is possible to derive three main parameters for terminological content:

1) the number of terms by volume of text (total number of terms);

2) the proportion of terms in the text (word forms of terms / all word forms);

3) the number of unique terms (the term is counted once if it is met in the analyzed text).

To implement the calculation of the first parameter, the length of the tuple with lemmatized terms is considered. The calculation of the second parameter involves counting all word forms stripped of punctuation marks. Information about the number of word forms of terms is stored in a nested tuple to prohibit access to certain terms.

Private word forms of terms and all word forms are rounded to 3 decimal places. This metric shows the amount of terminological apparatus that the text contains. To count unique terms, a dictionary of the frequency of found terms is aggregated. To do this, the sum of its keys is calculated.

The validation general set contains 15 observations (texts) for analysis, covering the educational texts of biology textbooks for secondary school (grades 5-9). Descriptive statistics on the parameter number of terms per text is presented in Table III.

TABLE III. Number of terms.

Approach	N texts	Mean	Min.	Max.	Std. Dev.
Validation	15	136.4	86	189	26.9
Dictionary approach	15	184.7	109	226	36.9
NER	15	128.9	81	159	22.4

According to the parameter “number of terms”, NER provided data close to authentic. The average value from the validation general set is 8 terms, in contrast to the dictionary approach (48 terms). The minimum value of NER is also closer to validation set with inaccuracy of 5 terms, with a dictionary approach the inaccuracy is 23 terms. According to the maximum value, both the dictionary approach and the NER are approximately equidistant from the value of validation statistics (37 and 30 terms, respectively).

Despite this, it cannot be said that both sets of observations differ greatly from the validation one, since the standard deviation in NER differs from the validation one by 4 units, and with the dictionary approach – by 12 units. This means that the values of NER are closer to the values of the validation set. A comparative graph of approaches for the number of terms parameter is shown in Fig. 6.

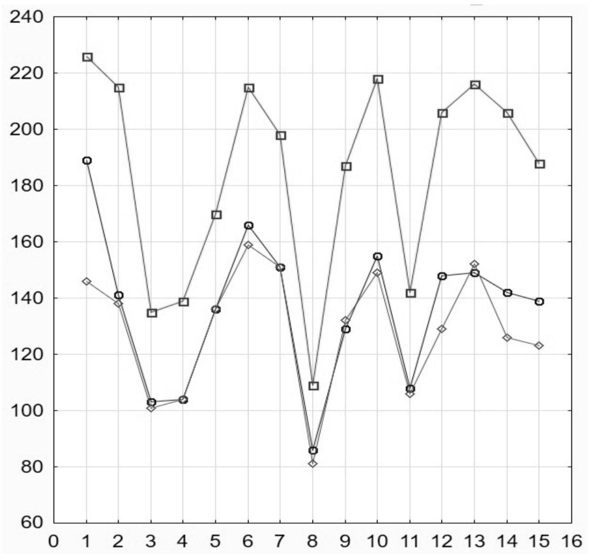


Fig. 6. Comparative graph for the parameter “number of terms”

where ⊖ is validation plot, ⊟ – dictionary approach plot, ◊ – NER plot, vertical axis – number of terms, horizontal axis – number of texts. As shown in the figure, according to the parameter “number of terms”, the values of NER are the closest to validation set. The dictionary approach slightly overestimates the values for this parameter, but, in general, both approaches approximately equally reflect the pattern of real data.

Descriptive statistics on the parameter “proportion of terms” is presented in Table IV.

TABLE IV. Proportion of terms.

Approach	N texts	Mean	Min.	Max.	Std. Dev.
Validation	15	0.371	0.27	0.52	0.07
Dictionary approach	15	0.331	0.28	0.43	0.04
NER	15	0.416	0.29	0.53	0.07

On average, the real value of terms proportion is 37%. Both in the dictionary approach and with NER, the deviation is 4%. The minimum value of the terms proportion in the dictionary approach is closest to validation set. The deviation from validation set in the NER is 2 percent. The maximum value of the NER term proportion is the closest to the validation set.

The deviation is 1%, in contrast to the dictionary approach, the deviation is 9%. According to the distribution of observations, NER is closest to validation (deviation of 0.01 units), the dictionary approach has a deviation of 0.02 units. A comparative graph of approaches for the parameter “proportion of terms” is shown in Fig. 7.

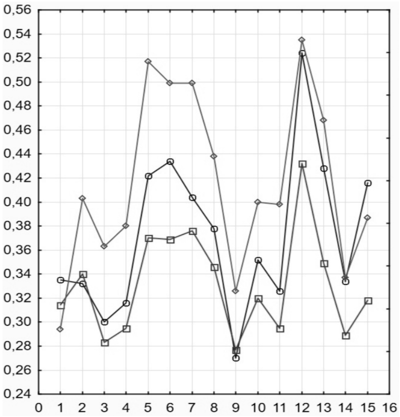


Fig. 7. Comparative graph for the parameter "proportion of terms"

where ⊖ is validation plot, ⊟ – dictionary approach plot, ◊ – NER plot, vertical axis – proportion of terms, horizontal axis – number of texts. According to the set of measurements, there is a slight overestimation of the NER values relative to the validation set. The dictionary approach slightly underestimates the indicators regarding validation set. In general, both approaches repeat the change in the values of the validation set. Descriptive statistics on the parameter "number of unique terms" is presented in Table V.

TABLE V. NUMBER OF UNIQUE TERMS.

Approach	N texts	Mean	Min.	Max.	Std. Dev.
Validation	15	83.3	53	106	17.8
Dictionary approach	15	72.3	49	91	13.8
NER	15	91.5	62	120	17.5

According to the parameter "number of unique terms", the average value of validation set is 83 terms. The deviation according to the dictionary approach is 11 terms, the deviation according to the NER is 9 terms. According to the minimum value from validation (53 terms), the deviation from the dictionary approach is 4 terms, for NER – 9 terms. According to the maximum validation value (106 terms), the inaccuracy of the dictionary approach is 16 terms, for NER – 14 terms. The standard deviation for validation is 17 units, the set of observations for NER is the closest to validation set in terms of uniformity (17 units), in contrast to the dictionary approach – 13 units. A comparative graph of approaches for the parameter "number of unique terms" is shown in Fig. 8.

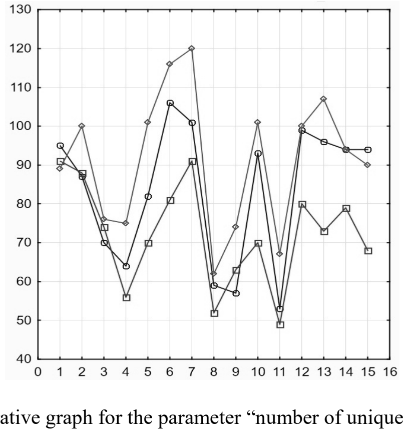


Fig. 8. Comparative graph for the parameter "number of unique terms"

where ⊖ is validation plot, ⊟ – dictionary approach plot, ◊ – NER plot, vertical axis – number of unique terms, horizontal axis – number of texts. According to the set of measurements, there is a slight overestimation of the NER values relative to the validation set.

The dictionary approach slightly underestimates the indicators regarding validation. The underestimation of values is most clearly observed in the last 4 measurements. In general, both approaches repeat the change in the values of the validation set. To check the deviation of the values, we calculated percentage for each observation for each analyzed parameter and derived a new data set [8]. The average values are shown in Table VI.

TABLE VI. DEVIATION FROM VALIDATION OBSERVATIONS.

Approach	N Terms	Terms Proportion	N Unique Terms
Dictionary approach	34.45%	10.53%	14.21%
NER	5.50%	15.17%	12.34%

According to the parameters "proportion of terms" and "number of unique terms", both approaches show deviation of 5 and 2%, respectively. Strong differences are present in the analysis of the parameter "number of terms". NER for this parameter represents the most approximate values, in contrast to the dictionary approach (the difference is 30%). This difference is expressed in the specifics of the algorithm based on the dictionary approach.

We have identified typical errors of the algorithm, which lead to such inaccuracy:

1) sensitivity to the work of the lemmatizer (the precision of any lemmatizer cannot cover absolutely all words of the language) – highly specialized terms may not be captured by the algorithm, since the dictionary contains word forms in the initial forms of words, note: камбиЕМ instead of камбИЙ (cambium), цитоплазМЕ instead of цитоплазМА (cytoplasm);

2) homonyms, note: каркас (frame), класс (category), тип (type), форма (form), пол (sex) etc.;

3) the absence of words or phrases in the dictionary of terms, which is the reason for capturing two or more terms that are part of one, note: межклеточное пространство (intercellular space), клетка-железа (gland-cell) etc.

V. CONCLUSION

This article presents applicability of two methods for extracting terms from school textbooks. They are applied to a collection of 10 biology textbooks for secondary schools in Russia. The dictionary method is rarely used nowadays, but in this case its use is justified, since we have a large encyclopedic dictionary of biology terms with more than 90,000 terminological units. The dictionary used is the largest dictionary of biological terms in Russian, covering all levels of biological terminology, including those from school textbooks. All other dictionaries are much smaller, so when applying the dictionary approach to extracting terms from biology texts, its use is the only option. The standard SpaCy and tok2vec tools were used to implement machine learning, allowing both verification of our results and easy application of our approach to other subject areas. In the machine learning approach, the neural network was trained on texts of the type from which the terms are supposed to be extracted in the future.

The two methods are compared on three parameters. The parameter “number of terms” and its variant “proportion of terms” provide a practical visual representation of the quality of the method, as can be seen in diagrams 6-8. Counting at the level of unique terms was previously used in the article [19] specifically on biological topics.

Thus, the approbation and comparison of algorithms for term extraction revealed that both approaches are equally applicable to the extraction of terminology, since the trends of approximation of data values repeat the trends of the graph of the validation set.

The dictionary approach has the following characteristics:

- advantages: scalability (it is possible to connect a terminological dictionary of another field of knowledge), flexibility (the ability to remove homonyms and homonymous phrases from the dictionary), as well as the ability to upload data (lists of terms by alphabet and frequency).

- disadvantages: it does not solve the problem of homonymy, interterms, general scientific terms and words with polysemy are also captured by the algorithm, it is sensitive to the work of the lemmatizer and typos.

Named Entity Recognition approach has the following characteristics:

- advantages: accuracy (word forms and phrases that are present in the corpus are captured), the problem of homonymy is partially removed (thanks to tok2vec), there is no need to work with a dictionary, speed (text analysis for the presence of terms occurs in less than a second).

- disadvantages: the conversion of a term into a dictionary form is possible only through the processing of a dictionary approach, the accuracy of training depends on the volume of training examples (manual marking of the corpus requires a significant time resource).

The methods that are presented in the article can be used to solve the applied problem of assessing the complexity of educational texts for schoolchildren. The dictionary approach may be transferred into other termfields in school discourse as well as Named Entity Recognition approach, particularly if we apply verified termlists. The main result of the study is the equal applicability of both methods for extracting terminology from school textbooks, provided, of course, by a large common dictionary on the subject area.

Acknowledgment

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (“PRIORITY-2030”), Strategic Project №5.

References

[1] M.I. Solnyshkina, V.D. Solovyev, E.V. Gafiyatova, and E.V. Martynova, “Text Complexity as Interdisciplinary Problem”, *Cognitive Linguistics Issues.*, vol.1, Jan. 2022, pp. 18-39.

[2] V.D. Solovyev, M.I. Solnyshkina, and D.S. McNamara, “Computational Linguistics and Discourse Complexology: Paradigms and Research Methods”, *Russian Journal of Linguistics.*, vol.26, Oct. 2022., pp. 275-316.

[3] V. Stoykova, and R. Stankovic, “Using Query Expansion for Cross-Lingual Mathematical Terminology Extraction”, *in Proc. Computer Science On-line Conf.*, May. 2018, pp. 154-164.

[4] W. Sha, B. Hua, and S. Linqi, “A Pattern and Pos Auto-learning Method for Terminology Extraction from Scientific Text”, *Data and Information Management.*, vol.5, Jul. 2021, pp. 329-335.

[5] F.Z.R. Saraiva, T.L.C. da Silva, and J.A.F. de Macêdo, “Aspect Term Extraction Using Deep Learning Model with Minimal Feature Engineering”, *in Proc. International Conference on Advanced Information Systems Engineering.*, vol.12127, Jun. 2020, pp. 185-198.

[6] Ya.Yu. Dementeva, E.P. Bruches, and T.V. Batura, “Terms Extraction from Texts of Scientific Papers”, *Software & Systems.*, vol.35, Sep. 2022 pp. 689-697.

[7] R. Stanković, C. Krstev, I. Obradović, B. Lazić, and A. Trtovac. “Rule-based Automatic Multiword Term Extraction and Lemmatization”, *in Proc. 10th International Conference on Language Resources and Evaluation.*, vol.10, May. 2016, pp. 507-514.

[8] I.P. Braslavskiy, and A.E. Sokolov, “Comparison of Five Methods for Variable Length Term Extraction”, *in Proc. Dialogue Conf.*, May. 2008, pp. 67-73.

[9] H. Amir et al., “TermEval 2020: TALN-LS2N System for Automatic Term Extraction”, *In Proc. 6th International Workshop on Computational Terminology Conf.*, May. 2020, pp. 95-100.

[10] L. Augustyniak et al., “Comprehensive Analysis of Aspect Term Extraction Methods Using Various Text Embeddings”, *Computer Speech & Language.*, vol.69, Sep. 2021, pp. 1-27.

[11] R. Ramachandran, and K. Arutchelvan, “Named Entity Recognition on Bio-Medical Literature Documents Using Hybrid Based Approach”, *Journal of Ambient Intelligence and Humanized Computing.*, vol.2, Mar. 2021, pp. 1-10.

[12] E.P. Bruches, and T.V. Batura, “Method for Automatic Term Extraction from Scientific Articles Based on Weak Supervision”, *NGU Journal.*, vol.2, Oct. 2021, pp. 5-16.

[13] A.S. Starostin, V.V. Bocharov, and S.V. Alexeeva, “Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian”, *FactRuEval.*, vol.1, Jun. 2016, pp. 1-19.

[14] Training Algorithm, Web: https://github.com/amrrs/custom-ner-withspacy3/blob/main/Custom_NER_with_Spacy3.ipynb.

[15] Corpus for NER, Marked up Data on the Biology Corpus, Educational Texts for Comparative Analysis, Educational Texts for Validating NER Model Validation, Data with Percentage Deviations, Web: https://github.com/Ertiops/Data-for-Research_TermExtraction-Approaches.

[16] NER annotator, Web: https://github.com/Ertiops/Dictionary-Approach-Algorithm.

[17] NER-Model, Web: https://github.com/Ertiops/NER-algorithm.

[18] Data for comparative analysis, https://github.com/Ertiops/Data-for-Research_TermExtraction-Approaches.

[19] L. Shi, and F. Campagne, “Building a protein name dictionary from full text: a machine learning term extraction approach”, *BMC Bioinformatics.*, vol.6, Apr. 2005, pp. 1-13.