Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2019”

Moscow, May 29—June 1, 2019

ASSESSING THEME ADHERENCE IN STUDENT THESIS

Tikhomirov M. M. (tikhomirov.mm@gmail.com),
Loukachevitch N. V. (louk_nat@mail.ru),
Dobrov B. V. (dobrov_bv@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

In this paper we study approaches to assessing the quality of student theses in pedagogics. We consider a specific subtask in thesis scoring of estimating its adherence to the thesis’s theme. The special document (theme header) comprising the theme, aim, object, tasks of the thesis is formed. The theme adherence is calculated as the similarity value between the theme header and thesis segments. For evaluation we order theses in the increased value of the calculated theme adherence and compare the ordering with expert grades using the average precision measure. The best configuration for theses ranking is based on the weighted averaged sum of word embeddings (word2vec) and keywords extracted from the theme header.

Key words: Thesis assesment, embeddings, cosine similarity, ontology

ОЦЕНКА СООТВЕТСТВИЯ ТЕМЫ И ТЕКСТА В СТУДЕНЧЕСКОЙ ВЫПУСКНОЙ РАБОТЕ

Тихомиров М. М. (tikhomirov.mm@gmail.com),
Лукашевич Н. В. (louk_nat@mail.ru),
Добров Б. В. (dobrov_bv@mail.ru)

МГУ им. М. В. Ломоносова, Москва, Россия

В данной статье изучаюся подходы к оценке качества студенческих выпускных работ по педагогике. Рассматривается подзадача определения соответствия текста теме работы. Для этого формируется специальный документ (тематический заголовок), содержащий тему, цель, объект, задачи работы. Соответствие теме рассчитывается как значение сходства между тематическим заголовком и сегментами работы. Для оценивания мы упорядочиваем работы по возрастанию значений соответствия теме и сравниваем полученный порядок с оценками экспертов, используя меру средней точности.

Ключевые слова: оценивание выпускных квалификационных работ, дистрибутивное представление слов, косинусная мера близости, онтология

1. Introduction

Currently, proper assessment of a student thesis (bachelor or magister) can be a very difficult task because of availability of various informational resources in Internet, which can be plagiarized by a student. It can seem that a thesis is well-done, but in fact the share of student own work is minimal. In order to detect borrowings in the thesis texts, so-called plagiarism detection systems (antiplagiat.ru, etxt.ru) have become widespread, which allow determining the percentage of borrowings either on the basis of their own source database or by analyzing the global search engine results (Yandex, Google) [7].

However, the quality requirements to a student thesis are not limited to plagiarism restrictions. The requirements include such important characteristics of a work as the theoretical and practical significance, elements of novelty in the work, knowledge of the modern literature on the research topic, consistency in the presentation of the material, the scientific style of presentation, and others. Checking these requirements could be automated in order to provide an expert with data on different characteristics of student works. The task of automated assessment can be compared to such a known task as automatic essay scoring [5], [6], [14], when student essays to specific topics should be assessed. But also, there are significant distinctions between thesis assessment and essay scoring tasks.

One of important characteristics of a student thesis is its relatedness to the thesis theme. The proclaimed theme is usually concretized in the following terms: aim of the study, object of the study, and the tasks of the work. It is possible to gather all these information into so-called theme header. It is usually supposed that a student should develop the theme and its details in the presented work. So, there is a subtask of thesis scoring to assess its adherence to the theme header. In the essay scoring, this subtask corresponds to the prompt relatedness subtask [6], [11].

In this paper, we study approaches to determining the relatedness between the theme header and student thesis in pedagogics. To evaluate the methods, we have the collection of 40 thousand student theses, 120 student theses among them have double expert scores. The aim of the assessment is as follows: if low relatedness is detected, then the problems should be visualized to experts and some penalties to the overall score for this work should be proposed by the system. We use several means for assessing relatedness including word embeddings and a thesaurus providing knowledge about domain term relations. As a thesaurus, we use Ontology on Natural Sciences and Technologies [4], where the pedagogics domain terms have been introduced.

We consider theme adherence as one of factors needed to be calculated for comprehensive assessment of student thesis. Also thesis fragments that found irrelevant to the thesis theme are considered as good candidates for plagiarism analysis.

2. Related Works

For assessing the quality of scientific papers, Osipov et al. [10] discuss such characteristics as the presence of the necessary sections (introduction, problem statement, list of references, etc.); scientific and non-scientific vocabularies; the presence of logical and semantic defects in the text of a scientific publication; selecting author's terms—new concepts defined by the authors of publications; highlighting the results presented in publications etc. Some authors study methods for the recognition of artificially generated scientific papers [2], [3], [8].

In the essay scoring, the most similar to our task is the task of prompt adherence that is assessing how the essay content corresponds to the announced essay topic [6], [11].

In [6] the Relatedness to Prompt feature is studied. The text of a essey fragment and the prompt (text of the essay question) must be related. If this relationship does not exist, this is perhaps evidence that the student has written an off-topic essay. The assessment was made for each sentence. The quality of the assessment was evaluated using double expert annotation for specific sentences. Most features proposed in this work are based on so-called Random indexing [13]. Random Indexing is a vector-based semantic representation system similar to Latent Semantic Analysis. In the current work, Random Indexing (RI) semantic space is trained on about 30 million words of newswire text. RI similarity to prompt for a sentence measures to what extent the sentence contains terms in the same semantic domain as compared to those found in the prompt. The SVM-classifier is trained on the calculated features and labeled data.

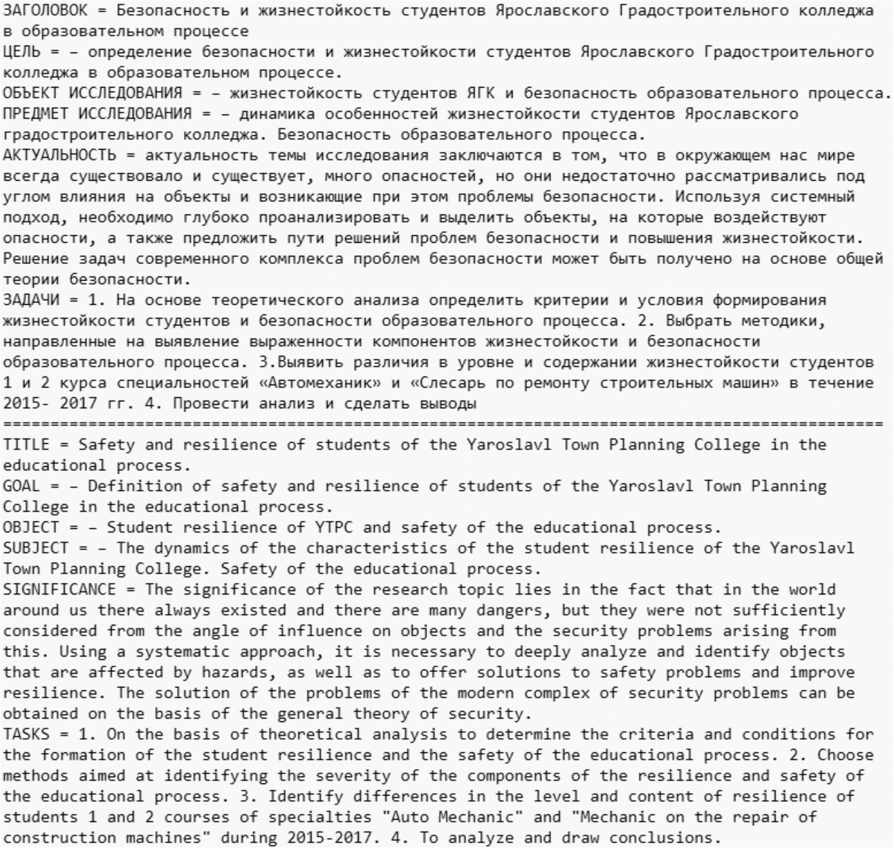
Persing and Ng [11] continue the study of the prompt relatedness in essay scoring using more diverse features. They try to predict the prompt relatedness for the whole essay, not for a single sentence. The predicted score ranges from one to four points at half-point intervals. 830 argumentative essays were annotated using a numerical score from one to four. Persing and Ng consider the task as a regression problem. Seven types of features were utilized in prompt-specific regressors based on linear SVM. Besides the random indexing features from the previous work, the authors used lemmatized unigram, bigram, and trigram similarity; thesis clarity keywords, which are the subdivision of the initial prompt to logical parts; LDA statistically generated topics.

3. Task, Data and Preprocessing

For experiments we usethe collection 40 thousand theses in pedagogics from various universities defended in 2017–2018 (further FullCollection). 120 theses from this collection have double scores from two experts belonging to different institutions (further AnnotatedCollection). This collection is new and the current study is the first one based on this collection.

The theses have similar structure. They include several parts: introduction, two-three chapters, sometimes recommendations, conclusion, appendices. In the introduction, a student introduces the theme of the thesis, the aim, the object, and the tasks of the work. The first chapter presents the survey of theoretical studies related to the theme of the work. The second and third chapters often describe practical experiments carried out by the student.

To have more information about the thesis' theme, we gather the above-mentioned structural elements (the theme, aim, object, and tasks of the thesis) into a specific document called theme header. The theme header conveys the main idea and direction of the thesis. It is clear that all parts of the thesis should correspond to the theme header in some extent. In this paper we assess how the first chapter of the thesis, survey, is related to the theme header. We extract the theme header and chapters of the thesis using a specialized vocabulary and patterns. Figure 1 presents an example of a theme header.



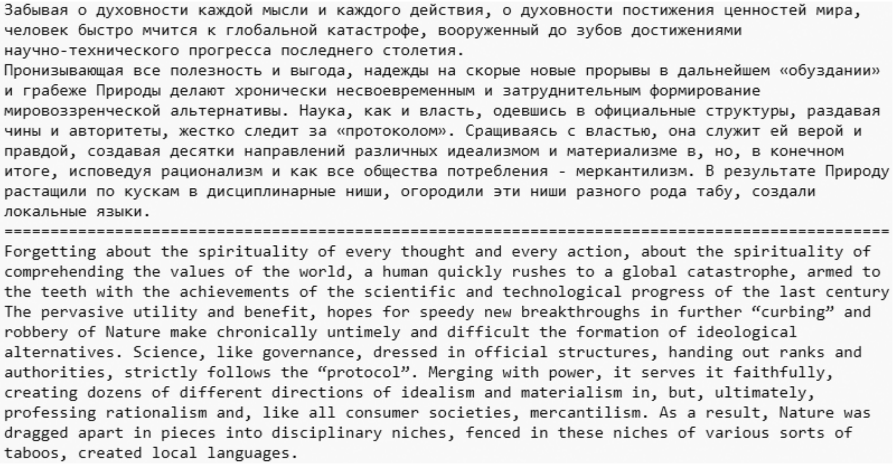
```
ЗАГОЛОВОК = Безопасность и жизнестойкость студентов Ярославского Градостроительного колледжа
в образовательном процессе
ЦЕЛЬ = – определение безопасности и жизнестойкости студентов Ярославского Градостроительного
колледжа в образовательном процессе.
ОБЪЕКТ ИССЛЕДОВАНИЯ = – жизнестойкость студентов ЯГК и безопасность образовательного процесса.
ПРЕДМЕТ ИССЛЕДОВАНИЯ = – динамика особенностей жизнестойкости студентов Ярославского
градостроительного колледжа. Безопасность образовательного процесса.
АКТУАЛЬНОСТЬ = актуальность темы исследования заключаются в том, что в окружающем нас мире
всегда существовало и существует, много опасностей, но они недостаточно рассматривались под
углом влияния на объекты и возникающие при этом проблемы безопасности. Используя системный
подход, необходимо глубоко проанализировать и выделить объекты, на которые воздействуют
опасности, а также предложить пути решений проблем безопасности и повышения жизнестойкости.
Решение задач современного комплекса проблем безопасности может быть получено на основе общей
теории безопасности.
ЗАДАЧИ = 1. На основе теоретического анализа определить критерии и условия формирования
жизнестойкости студентов и безопасности образовательного процесса. 2. Выбрать методики,
направленные на выявление выраженности компонентов жизнестойкости и безопасности
образовательного процесса. 3.Выявить различия в уровне и содержании жизнестойкости студентов
1 и 2 курса специальностей «Автомеханик» и «Слесарь по ремонту строительных машин» в течение
2015- 2017 гг. 4. Провести анализ и сделать выводы
===============================================================================================
TITLE = Safety and resilience of students of the Yaroslavl Town Planning College in the
educational process.
GOAL = – Definition of safety and resilience of students of the Yaroslavl Town Planning
College in the educational process.
OBJECT = – Student resilience of YTPC and safety of the educational process.
SUBJECT = – The dynamics of the characteristics of the student resilience of the Yaroslavl
Town Planning College. Safety of the educational process.
SIGNIFICANCE = The significance of the research topic lies in the fact that in the world
around us there always existed and there are many dangers, but they were not sufficiently
considered from the angle of influence on objects and the security problems arising from
this. Using a systematic approach, it is necessary to deeply analyze and identify objects
that are affected by hazards, as well as to offer solutions to safety problems and improve
resilience. The solution of the problems of the modern complex of security problems can be
obtained on the basis of the general theory of security.
TASKS = 1. On the basis of theoretical analysis to determine the criteria and conditions for
the formation of the student resilience and the safety of the educational process. 2. Choose
methods aimed at identifying the severity of the components of the resilience and safety of
the educational process. 3. Identify differences in the level and content of resilience of
students 1 and 2 courses of specialties "Auto Mechanic" and "Mechanic on the repair of
construction machines" during 2015-2017. 4. To analyze and draw conclusions.
```

Figure 1: Theme header for thesis “Safety and resilience of students of the Yaroslavl Town Planning College in the educational process”

The similarity between the theme header in a thesis and the prompt in essay writing can be seen. But the difference between two tasks: relatedness of a thesis to its theme header and an essay to the prompt is also significant:

- The theme elements are written by a student and can be poorly worded but the prompt in essay writing is formulated by professionals,
- There can be many essays for a given prompt. Therefore some methods can be specially tuned for a specific prompt [1]. The student thesis’s theme header is unique,
- The survey chapters are much longer than essays. In our data, they contain 250 sentences on average. Also, the theme headers are in most cases much longer than usual essay prompts,
- The survey chapters are never pervasive or argumentative in contrast to essays.

In the current study, we do not have any manual annotation of the theme relatedness of the first chapters. For evaluation, we use the overall score given to a thesis by two professional experts according to 2–5 scale, where 5 is the maximum grade in the scale. We suppose that low-scored theses should also have some problems in its surveys. As an example of “bad” segments for the same thesis, the theme header of which was presented, see Figure 2.



```
Забывая о духовности каждой мысли и каждого действия, о духовности постижения ценностей мира,
человек быстро мчится к глобальной катастрофе, вооруженный до зубов достижениями
научно-технического прогресса последнего столетия.
Пронизывающая все полезность и выгода, надежды на скорые новые прорывы в дальнейшем «обуздании»
и грабеже Природы делают хронически несвоевременным и затруднительным формирование
мировоззренческой альтернативы. Наука, как и власть, одевшись в официальные структуры, раздавая
чины и авторитеты, жестко следит за «протоколом». Сращиваясь с властью, она служит ей верой и
правдой, создавая десятки направлений различных идеализмом и материализме в, но, в конечном
итоге, исповедуя рационализм и как все общества потребления - меркантилизм. В результате Природу
растащили по кускам в дисциплинарные ниши, огородили эти ниши разного рода табу, создали
локальные языки.
================================================================================================
Forgetting about the spirituality of every thought and every action, about the spirituality of
comprehending the values of the world, a human quickly rushes to a global catastrophe, armed to
the teeth with the achievements of the scientific and technological progress of the last century
The pervasive utility and benefit, hopes for speedy new breakthroughs in further “curbing” and
robbery of Nature make chronically untimely and difficult the formation of ideological
alternatives. Science, like governance, dressed in official structures, handing out ranks and
authorities, strictly follows the “protocol”. Merging with power, it serves it faithfully,
creating dozens of different directions of idealism and materialism in, but, ultimately,
professing rationalism and, like all consumer societies, mercantilism. As a result, Nature was
dragged apart in pieces into disciplinary niches, fenced in these niches of various sorts of
taboos, created local languages.
```

Figure 2: “Bad” segment for thesis “Safety and resilience of students of the Yaroslavl Town Planning College in the educational process”

We order all the theses according to the increase of the automatic scores of theme relatedness and evaluate methods of theme relatedness calculation using Average precision of 2-grade in the first positions of the created ranking. Table 1 shows the distribution of grades in 120 theses. Table 2 shows the deviation between grades of two experts. We calculate the Average precision measures according to the minimal grades of the theses. Thus, a thesis should have at least one 2 grade to be considered as a correct result in the beginning of the calculated rating.

Table 1: Distribution of marks in student theses

Mark	Number of theses with minimal mark	Number of theses with maximal mark
2	42	8
3	42	33
4	28	51
5	8	28

Table 2: Deviations between thesis marks

Points of difference	Number of works
0	49
1	51
2	16
3	4

As preprocessing, we use the procedure of segmenting the whole chapter to thematic fragments. Then we calculate similarity between specific segments and the theme header. The overall score of the theme relatedness of a chapter is based on averaging segment scores of this chapter.

4. Segmentation of Thesis Chapter to Thematic Fragments

The thematic segmentation module should break up a long sequence of the text into segments so that the sentences in one segment are thematically similar, and the boundaries of the segments signal a violation of connectivity between blocks of text. This procedure is based on the TopicTiling algorithm [12]. For splitting the text into segments, a procedure for assessing connectivity violations has been implemented, which is applied to each sentence. Based on the values of this metric, selection of separating sentences is carried out, which become the beginnings of segments. The procedure for assessing connectivity violation is as follows.

For each sentence, its “left” and “right” contexts with the length of w sentences are considered (sentences having the length of less than k words are ignored). For each sentence from the context, its vector representation is calculated using word2vec [9]. The weighted average of the word embeddings based on idf multiplier is calculated. Using cosine similarity, the similarity of all pairwise combinations of sentences between the “left” and “right” context is calculated. Based on these values, the coherence score is calculated by averaging all the similarities—*coherence_score*.

$$coherence\_score = \frac{1}{w} \sum_{v_l} \sum_{v_r} similarity(v_l, v_r) \tag{1}$$

After each sentence has its *coherence_score* calculated, in the second pass for each sentence its *depth_score* value is calculated using the following formula.

$$depth\_score(s_i) = 0.5 * (top\_left + top\_right - 2 * coherence\_score(s_i)) \tag{2}$$

Where *top_left* is the peak value of *coherence_score* to the left of the sentence, in non-descending order, *top_right* calculated by analogy.

The mean value and variance are calculated for the *depth_score* vector, on the basis of which the threshold values are chosen for the selection of candidate-sentences. If the *depth_score* of sentence is above the threshold and no separating sentences were selected in the neighborhood of several sentences, then this sentence is chosen as the separator. sentences are considered in the order of their *depth_score* values. Thus, after completing the procedure described above, the source text is broken up into segments.

5. Methods of Assessing Theme Adherence

5.1. Preprocessing

The text of the thesis was pre-processed as follows:

- The whole text was lemmatised and stop words were removed. The frequency characteristics of the words were calculated and the idf values were obtained;
- In addition, some words were removed from the text of the theme header based on a part of speech: verbs, adjectives and functional parts of speech.
- The procedure of extracting the concepts was carried out using the ontology [4]. Thus for each sentence there is a list of concepts contained in it. For concepts, idf was also counted;
- The word2vec model was trained on full collection of the theses, which contains 40 thousand documents. The parameters are: CBOW, vector length is 150, window size is 10. The training was conducted using the python gensim package;[1]

As an additional source of information about the domain, we use the ontology on natural sciences and technologies [4], which comprises terms of scientific fields (mathematics, physics, chemistry, geology, astronomy etc.) and terms of technological domains (oil and gas, power stations, cosmic technologies, aircrafts, etc.). Currently, about 6,500 terms (including term variants) were added to OENT to describe the pedagogics domain.

The main unit of the ontology is a concept, which has a unique name, the set of text entries, which express this concept in the text, and concept relations. For example, the concept *DEAF AND HARD OF HEARING EDUCATION* has the following text entries: *deaf education, deaf teaching, education of the deaf, teaching of the deaf* (translation from Russian).

5.2. Features of Theme Adherence Assessment

The chapter under analysis (further document) is presented in the form of N segments S and compared with the theme header H. Two baseline models were implemented to accomplish this task.

Baseline 1 (Tf-Idf). In this baseline model, the theme header and segments are represented as sparse vectors, where each element of the vector corresponds to a word from the collection's vocabulary. The values of vector elements are calculated as tf-idf. Based on the cosine similarity between the theme header and the S_i segment vectors, the *adherence_score* with the theme is formed.

Baseline 2 (SegWord2Vec). Each segment, including the theme header, is converted into a vector using word2vec embeddings. This operation is performed by weighted averaging of the word vectors with idf as weights, as used in the segmentation procedure. The following features were implemented and combined with the baselines:

[1] https://radimrehurek.com/gensim/index.html

NoNorm: Disabling normalization for the theme header vector. We introduce this feature, because the normalization gives lower *adherence_score*'s to larger and richer theme headers.

Concepts: Adding an additional vector for the theme header and segments, which is formed by analogy with the word vector, but based on the concepts of the ontology founded in the text of the segment/header. The concepts allow accounting for synonyms and multi-word expressions. In this case, the thematic adherence is formed as a weighted sum of the vectors similarities.

$$adherence\_score = \alpha * sim_{word} + (1 - \alpha) * sim_{conc} \quad (3)$$

Keywords: For the theme header, the most significant k_w words and k_c concepts are determined according to tf-idf. The set of keywords includes words and concepts whose tf-idf weights exceed the threshold. This threshold is calculated as 0.2 * average value to the 3 (2 for concepts) most significant (by tf-idf) words (or concepts). The weights of the keywords are multiplied by additional factors w_w and w_c, respectively. Keywords for thesis "Safety and resilience of students of the Yaroslavl Town Planning College in the educational process" are as follow:

- *words*—resilience (1.00), safety (0.47), town-planning (0.42), Yaroslavl (0.30), college (0.21), ytpc (0.17), student (0.17);
- *concepts*—urban planning (1.00), safety (0.63), college (0.57), student (0.20), system approach (0.17);

EmbedExp: There is an expansion of keywords for the theme header, by adding most similar words to them using word2vec embeddings. To do this, for each of the k_w keywords, the n_{w2v} closest words are calculated using the word2vec representation. Those words that are present in the whole thesis are added to the theme header vector. In order to calculate the weight of new words in the vector, the following formula is used:

$$weight_{wordext} = sim(word_{raw}, word_{ext}) * tf(word_{raw}) * idf(word_{ext}) \quad (4)$$

Where $word_{raw}$—the keyword on which the expansion is made, $word_{ext}$—new word. In addition, the weights of the new words are multiplied by the factor of w_{ext}. The set of expanded words for the thesis "Safety and resilience of students of the Yaroslavl Town Planning College in the educational process", includes:

- *new words*—hardiness (0.88), Maddy (0.76), tough (0.66), stories (0.62), coping (0.54), freshman (0.48), security (0.44), safe (0.41), highschool (0.14), scholar (0.13);

Regardless of the specific configuration, each segment and theme header are represented by a vector (or vectors) and *adherence_score* is calculated as cosine similarity between corresponding vectors. We calculate *adherence_score* for the whole chapter in two ways:

- *mean*: The average value of *adherence_score*'s of all segments;
- *mean_worse*: The average value of *adherence_score*'s among the worst 20% of segments;

The *mean_worse* variant corresponds to the hypothesis that a thesis is characterized by its worst fragments. Further, theses are ranked in ascending order of their *adherence_score* values.

6. Evaluation and Results

As mentioned earlier, we have 2 expert grades for each thesis. The minimum score (2) is chosen as the reference value, assuming that at least one expert could find serious problems in the theses. It was also previously shown that in the reference collection there are 42 works with the minimum grade of 2.

The evaluation methodology is proposed as follows: the algorithm for each thesis forms the values of *adherence_score* (*mean* and *mean_worse*), on the basis of which the reference collection is ranked so that the "worst" thesis was "above". For evaluation we use average precision measure.

$$average\_precision(n) = \frac{\sum_{k=1}^{n} P(k) * rel(k)}{n} \tag{5}$$

Where $P(k)$—precision at k, $rel(k)$ is equal to 1 if k-th element of the list is relevant, otherwise 0.

We calculate *average_precision(25)* measure in 25th position (20% of the collection). The thesis is considered as relevant if it has grade 2 for at least one expert. In addition, the mean values of the *average_precision(n)* for positions from 1 to 25 were calculated.

The results of the evaluation of the configurations can be seen in Table 3. It also presents the result of random ordering (averaging 25000 random permutations). *NoNorm* did not give any significant improvements for any configurations. The results of average precision measures are also shown in the Figures 3 and 4.

Table 3: Evaluation results on 120 reference theses

	av_prec(25) by mean_worse	av_prec(25) by mean	mean av_prec(25) by mean_worse	mean av_prec(25) by mean
Random	0.15	0.15	0.19	0.19
Tf-Idf	0.15	0.16	0.11	0.12
Tf-Idf + Concepts	0.15	0.22	0.22	0.23
Tf-Idf + Keywords	0.23	0.21	0.16	0.29
Tf-Idf + Concepts + Keywords	0.20	0.27	0.26	0.43
Tf-Idf + Keywords + EmbedExp	0.21	0.22	0.19	0.18
Tf-Idf + Concepts + Keywords + EmbedExp	0.20	0.28	0.25	0.41
SegWord2Vec	0.28	0.22	0.37	0.18
SegWord2Vec + Concepts	0.28	0.27	0.37	0.40
SegWord2Vec + Keywords	0.30	0.19	0.47	0.36
SegWord2Vec + Concepts + Keywords	0.29	0.29	0.46	0.49

For the best configuration (accounting for both *mean* and *mean_worse*) , the optimal parameters were as follows:

- *mean*: *SegWord2Vec* + *Concepts* + *Keywords*:
  $\alpha = 0.25, k_w = 12, w_w = 1.25, k_c = 5, w_c = 1.$
- *mean_worse*: *SegWord2Vec* + *Concepts* + *Keywords*:
  $\alpha = 0.75, k_w = 12, w_w = 1.25, k_c = 3, w_c = 5.$
  Where α—the parameter which controls the participation of concepts,
  k_w—number of word keywords, w_w—multiplier of word keywords,
  k_c—number of concept keywords, w_w—multiplier of concept keywords.

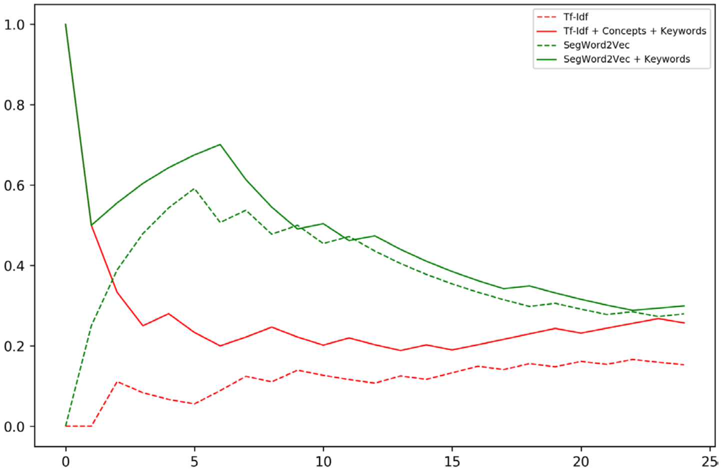


Figure 3: Average precisions graphs for base and best configurations for av_precision by mean_worse

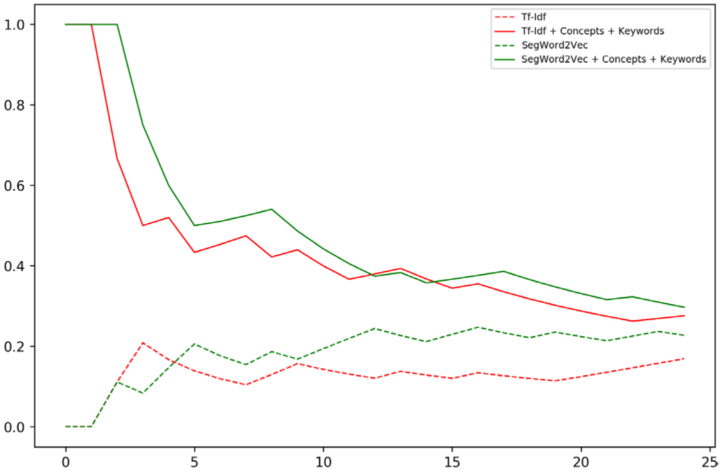


Figure 4: Average precisions graphs for base and best configurations for av_precision by mean

Based on the results, we can conclude the following:

- The configurations based on *SegWord2Vec* are better than those based on *Tf-Idf*.
- Ranking based on *mean_worse* (worst segments) at least not worse then *mean adherence_score* and better corresponds to the task of finding the worst thesis.
- Use of *Keywords* always leads to better results.
- Use of *Concepts* is very useful for *Tf-Idf* and sometimes useful for *SegWord2Vec*

7. Error analysis

For the analysis of the system, we consider two configurations: *Tf-Idf* и *SegWord2Vec + Concepts + Keywords*.

Tf-Idf.The first 5 theses have the following scores and grades:

1) 0.0 : 4;
2) 0.001 : 3;
3) 0.001 : 2;
4) 0.001 : 3;
5) 0.001 : 4.

The scores of the worst theses are very close to each other and are practically zero. This means that the algorithm did not reveal similarities between the worst segments and theme header.

The first thesis with the grade 4 has score 0.0. Among the 13 worst segments the similarity to the theme header is zero. The thesis itself has the name “Differentiated approach in improving the physical fitness of students in 6th grade”, but in all worst segments author is talking about the biological characteristics of children and their development. The amount of specific biological information seems excessive. In addition, all the worst segments are plagiarized, and the text was deliberately distorted (every 3–4 words were simply removed from the text), which prevented the anti-plagiarism systems from detecting plagiarism.

The second thesis with grade 4 also has a low score of 0.001. Among the 10 worst segments, there are also no exact matches of words with the theme header and this is due to the fact that they also deviate from the main theme of the work. The theme of the thesis is “The implementation of a systematic approach to teaching biology in primary school” but in all the worst segments there is a serious bias in the philosophical direction, to questions of knowledge.

SegWord2Vec + Concepts + Keywords. The first 5 theses have the following scores and grades:

1) −0.099 : 2;
2) −0.098 : 3;
3) −0.076 : 2;
4) −0.056 : 2;
5) −0.037 : 2.

The spread of values here is much larger, which suggests that the model better separates different theses, among other things, it is clearly seen that the estimates are better grouped (this is also evident on the Figure 3).

We looked at the content of the worst theses and the text of the worst segments is poorly relevant to a given topic. But at the same time there were situations that large segments sometimes get low scores even if they contained some amount of relevant information. This is due to the fact that the averaging of word2vec vectors works worse on large texts. In addition, in comparison with *Tf-Idf*, it became more difficult to interpret, why exactly one segment is worse than the other.

As a result, we can draw the following conclusions from the error analysis:

- *Tf-Idf* badly detects links between related segments, but the text of which use different words. This leads to the fact that among the worst segments are those that do not seem to relate directly to the topic, but at the same time have some consistency with it.
- *Tf-Idf* poorly separates bad theses from each other.
- *SegWord2Vec + Concepts + Keywords* on the other hand, organizes theses well and, on average, highlights really bad segments, in which there is no useful information for thesis, but at the same time, the interpretability suffers a little.
- *SegWord2Vec + Concepts + Keywords* also sometimes puts very low weights on large segments, but at the same time in which there is some amount of relevant information.

8. Conclusion

In this paper we studied approaches to assessing the quality of student theses in pedagogics. We considered a specific subtask in thesis scoring of estimating its adherence to the thesis's theme. The special document (theme header) comprising the theme, aim, object, tasks of the thesis is formed. The theme adherence is calculated as the similarity value between the theme header and thesis segments.

For evaluation we ordered theses in the increased value of the calculated theme adherence and compared the ordering with expert grades using the average precision measure. The best configuration for theses ranking is based on the weighted averaged sum of word embeddings (word2vec) and keywords extracted from the theme header.

9. Acknowledgements

This article contains the results of the project "Developing new methods for analyzing large text data using linguistic and ontological resources, machine learning methods and neural networks", carried out as part of the Competence Center program of the National Technology Initiative "Center for Big Data Storage and Analysis", supported by the Ministry of Science and Higher Education of the Russian Federation under the Contract of Moscow State University with the Fund for Support of Projects of the National Technology Initiative No. 13/1251/2018 dated December 11, 2018.

References

1. *Attali, Y., Burstein, J.:* Automated essay scoring with e-rater v. 2. The Journal of Technology, Learning and Assessment. 4, 3, (2006).
2. *Avros, R., Volkovich, Z.:* Detection of computer-generated papers using one-class svm and cluster approaches. In: International conference on machine learning and data mining in pattern recognition. pp. 42–55 Springer (2018).
3. *Bakhteev, O. et al.:* About one method of detecting artificial and unscientific texts in an extensive collection of documents. Electronic Libraries. 20, 5, 298–304 (2017).
4. *Dobrov, B. V., Loukachevitch, N. V.:* Development of linguistic ontology on natural sciences and technology. In: LREC. pp. 1077–1082 (2006).
5. *Foltz, P. W. et al.:* Automated essay scoring: Applications to educational technology. In: EdMedia+ innovate learning. pp. 939–944 Association for the Advancement of Computing in Education (AACE) (1999).
6. *Higgins, D. et al.:* Evaluating multiple aspects of coherence in student essays. In: Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: HLT-naacl 2004. (2004).
7. *Khritankov, A. S. et al.:* Discovering text reuse in large collections of documents: A study of theses in history sciences. In: 2015 artificial intelligence and natural language and information extraction, social media and web search fruct conference (ainl-ismw fruct). pp. 26–32 IEEE (2015).
8. *Labbé, C., Labbé, D.:* Duplicate and fake publications in the scientific literature: How many scigen papers in computer science? Scientometrics. 94, 1, 379–396 (2013).
9. *Mikolov, T. et al.:* Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013).
10. *Osipov, G. et al.:* Technologies for semantic analysis of scientific publications. In: 2012 6th ieee international conference intelligent systems. pp. 058–062 IEEE (2012).
11. *Persing, I., Ng, V.:* Modeling prompt adherence in student essays. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 1534–1543 (2014).
12. *Riedl, M., Biemann, C.:* TopicTiling: A text segmentation algorithm based on lda. In: Proceedings of acl 2012 student research workshop. pp. 37–42 Association for Computational Linguistics (2012).
13. *Sahlgren, M.:* Vector-based semantic analysis: Representing word meanings based on random labels. In: In essli workshop on semantic knowledge acquistion and categorization. Citeseer (2001).
14. *Taghipour, K., Ng, H. T.:* A neural approach to automated essay scoring. In: Proceedings of the 2016 conference on empirical methods in natural language processing. pp. 1882–1891 (2016).