CMC MSU Faculty of Computational Mathematics and Cybernetics
Lomonosov Moscow State University

MSU RCC

Assessing Theme Adherence in Student Thesis

Mikhail Tikhomirov

Natalia Loukachevitch

Boris Dobrov

Lomonosov Moscow State University

Significance

Thesis assessment task is a very difficult task that takes a lot of time for people.

Currently, in Russia the automated assessment of student theses exploits only so-called plagiarism detection systems.

Student thesis should also meet other requirements:

- Theoretical and practical significance
- Elements of novelty in the work
- Knowledge of the modern literature on the research topic
- Other

One of such important characteristics of a student thesis is its relatedness to the thesis theme – **theme adherence**.

Essay Scoring

Similar to the thesis assessment task, is the essay scoring task.

- An essay is, generally, a piece of writing that gives the author's own argument on some topic or question

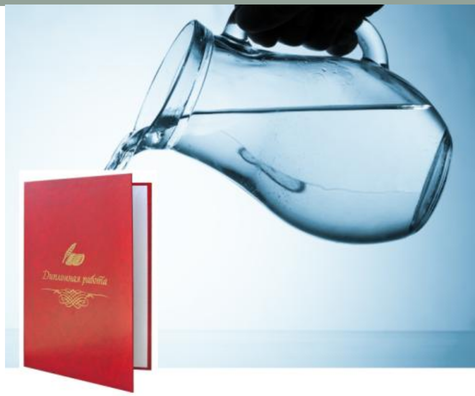
In this task there is prompt adherence (prompt relatedness) problem, which similar to theme adherence.

- The text of a essay fragment and the prompt (text of the essay question) must be related

There are some important differences between these two tasks:

- The essay is very short compared to the thesis
- The essay has a well-defined topic.
- For the essay and thesis different educational tasks are set

Problem



Theme:

“Safety and resilience of students of the Yaroslavl

Town Planning College in the educational process”

“Безопасность и жизнестойкость студентов Ярославского Градостроительного колледжа в образовательном процессе”

In text:

Inspired by the gifts of the outside world, fascinated by the successes in creating his artificial world, man forgot that within each of us there is an even more beautiful world with the treasures of the spirit. …

Увлекшись дарами внешнего мира, увлекшись успехами по созданию своего искусственного мира, человек забыл о том, что внутри каждого из нас имеется еще более прекрасный мир, располагающий сокровищами духа. …

Task

General:

- For any thesis to put in its correspondence a real number from range [0, 1] which reflects the degree of theme adherence

Concrete:

- Represent the thesis theme in an operable way
- Represent thesis text in an operable way
- Develop a similarity function that produces the necessary mapping

Dataset

40 thousand theses in pedagogics defended in 2017-2018 (further FullCollection).

120 theses each evaluated by two experts(further AnnotatedCollection).

We have scores for the thesis as a whole, but we assume that there is a strong correlation between the score as a whole and the theme adherence.

General Info About Students Theses

Average word length of thesis ~ 25000 words

Special structure

- Title page
- Table of contents
- Introduction
- …

Formally defined *goal, title, tasks, etc.*

Even among the experts, there are disagreement in the assessments: for half of the theses, the grades differ by at least one point.

Theme Header

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 = Resilience of YTPC students and safety of the educational process.
SUBJECT = The dynamics of the characteristics of the resilience of students 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, …
TASKS = 1. On the basis of theoretical analysis to determine the criteria and conditions for the formation … . 2. Choose methods aimed at identifying the severity of the components of the resilience …

Thesis Model

The thesis is presented as a sequence of small segments of the text (4-8 sentences), which can be considered mono-thematic.

$$thesis = (th, [seg_1, seg_2, \dots, seg_n]),$$

Where th *is theme_header,* seg_i is segment representation.

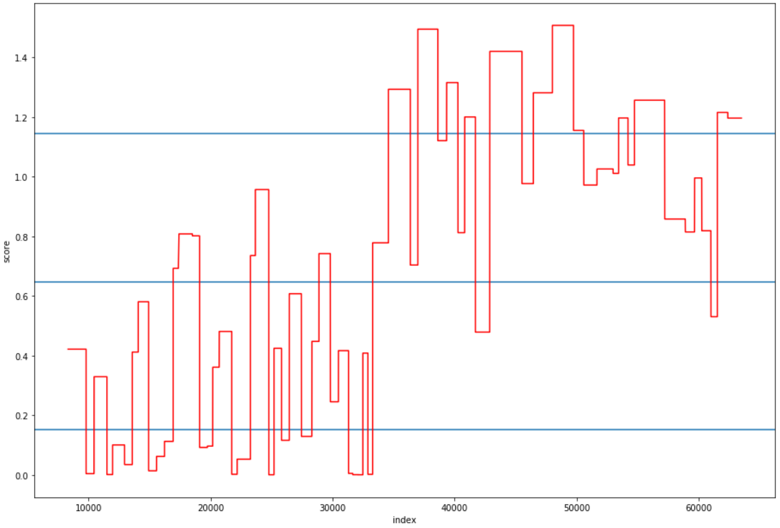
For each thesis are calculated:

$$adherence\_vec = [sim(th, seg_1), \dots, sim(th, seg_n)]$$

Where sim is some similarity function.

$adherence\_vec \rightarrow adherence\_score$?

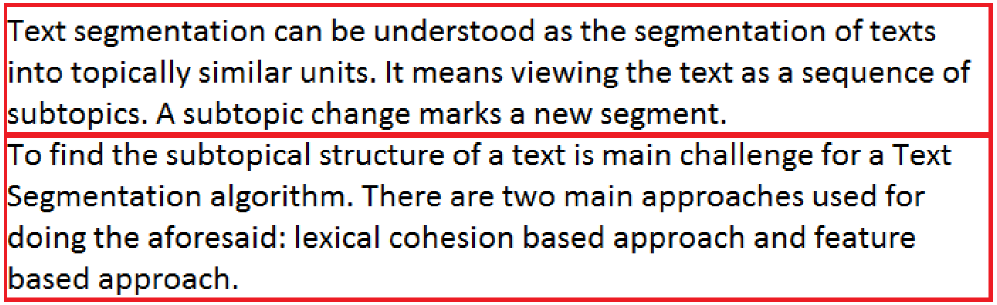
- Mean
- Mean among the worst 20% of segments



Segmentation

Based on TopicTiling[*] algorithm.

For each sentence, two blocks of *k* sentences are considered, “left” and “right”.



Text segmentation can be understood as the segmentation of texts into topically similar units. It means viewing the text as a sequence of subtopics. A subtopic change marks a new segment.

To find the subtopical structure of a text is main challenge for a Text Segmentation algorithm. There are two main approaches used for doing the aforesaid: lexical cohesion based approach and feature based approach.

Left block

← Sentence candidate

Right block

Similarity between left and right block is calculated, forming coherence score.

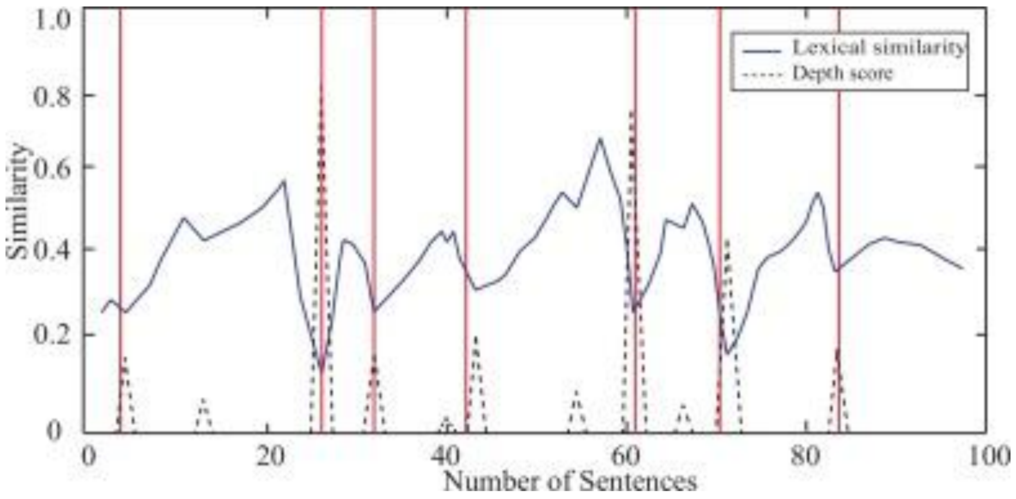
$$coh\_score = Sim(L_{S_i}, R_{S_i})$$

Riedl M., Biemann C. TopicTiling: a text segmentation algorithm based on LDA

Segmentation - 2

Based on *coherence score*, *depth score* is calculated.

$$depth\_score = 0.5 * (high_{left} + high_{right} - 2 * coh\_score(L_{S_i}, R_{S_i}))$$



Segment separators are selected based on *depth scores*.

Segment Representation Models

TF-IDF

- Each segment – sparse vector, calculated based on TF-IDF

SegWord2Vec

- Using Word2Vec for forming vector for segment, by weighted averaging of word vectors.

Topic Modeling

- Build Topic Model for theses and use it to obtain vectors for segments
- In our work we used LDA

Additional features

Theme:

“Safety and resilience of students of the Yaroslavl Town Planning College in the educational process”

“Безопасность и жизнестойкость студентов Ярославского Градостроительного колледжа в образовательном процессе”

Keywords – additional multiplayer for top k words in theme header

- Parameters: keywords count, keywords multiplier
- Example: resilience, safety, town-planning, Yaroslavl, college, YTPC, student

EmbedExp – theme header extension by similar words by word2vec

- Example: hardiness, Maddy, tough, stories, coping, freshman, security, safe, highschool, scholar

Additional features - 2

Used Ontology on Natural Science and Technology [*].

- Concepts of pedagogy were added
- Allow us to account for synonyms and multiword expressions
- Example: *DEAF AND HARD OF HEARING EDUCATION* -> deaf education, deaf teaching, education of the deaf, teaching of the deaf (translation from Russian).

Concepts - select *concepts* from the text using the ontology

- Parameters: alpha - the degree of influence on the similarity function, relative to the model in words
- Example: urban planning, safety, college, student, system approach

Dobrov, B.V., Loukachevitch, N.V.: Development of linguistic ontology on natural sciences and technology.

Evaluation

Of the two expert grades in the AnnotatedCollection, the worst was chosen as a gold.

How can we understand that the system is well compute theme adherence?

Hypothesis: “Bad” theses have poor theme adherence.

We can set the ranking task.

Goal: Get the «worst» theses at the top of the search results.

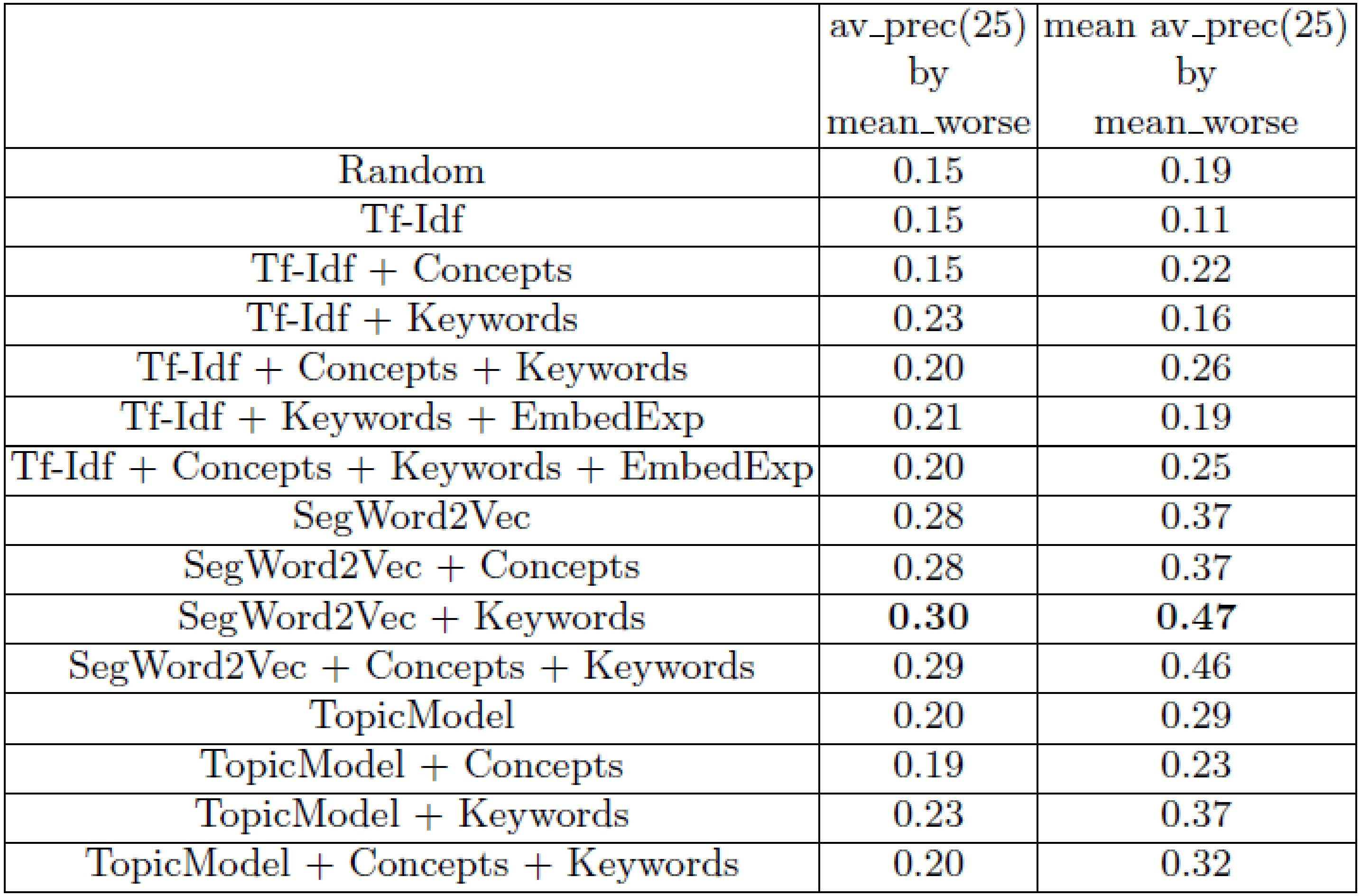
Metrics

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

When using ***average precision,***

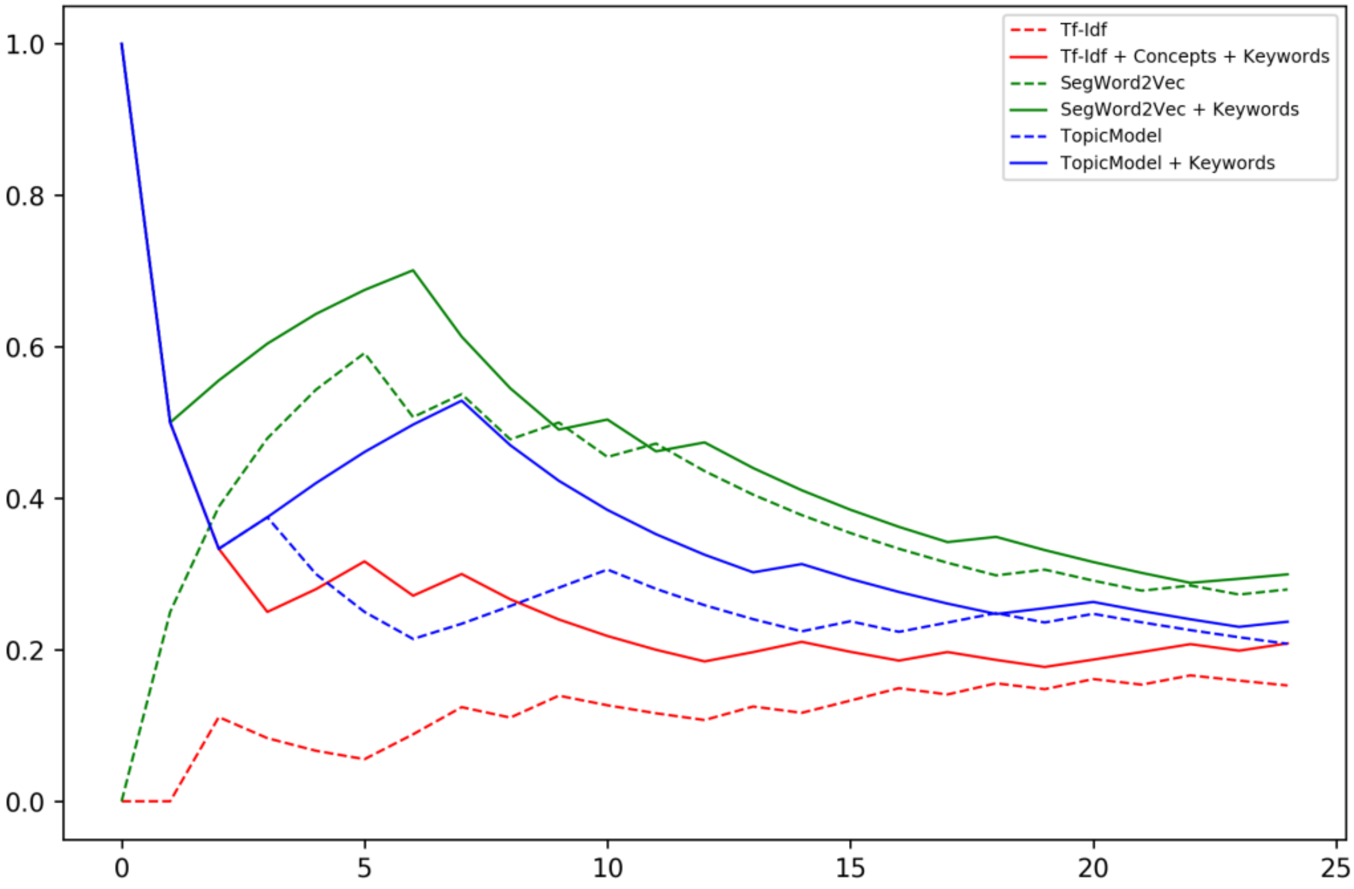
$$rel(k) = \begin{cases} 1; if\ grade\ at\ k\ is\ 2 \\ 0; \qquad\qquad otherwise \end{cases}$$

Results



	av_prec(25) by mean_worse	mean av_prec(25) by mean_worse
Random	0.15	0.19
Tf-Idf	0.15	0.11
Tf-Idf + Concepts	0.15	0.22
Tf-Idf + Keywords	0.23	0.16
Tf-Idf + Concepts + Keywords	0.20	0.26
Tf-Idf + Keywords + EmbedExp	0.21	0.19
Tf-Idf + Concepts + Keywords + EmbedExp	0.20	0.25
SegWord2Vec	0.28	0.37
SegWord2Vec + Concepts	0.28	0.37
SegWord2Vec + Keywords	**0.30**	**0.47**
SegWord2Vec + Concepts + Keywords	0.29	0.46
TopicModel	0.20	0.29
TopicModel + Concepts	0.19	0.23
TopicModel + Keywords	0.23	0.37
TopicModel + Concepts + Keywords	0.20	0.32

Results - 2


Tf-Idf
Tf-Idf + Concepts + Keywords
SegWord2Vec
SegWord2Vec + Keywords
TopicModel
TopicModel + Keywords
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25

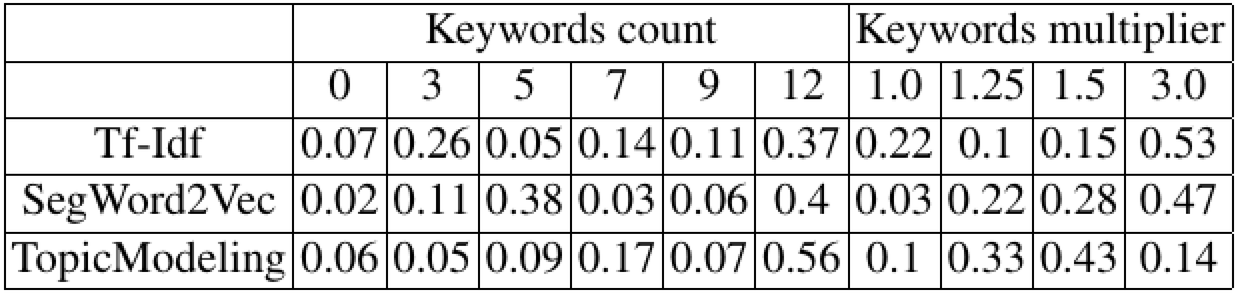
Hyperparameter Analysis

- Some values of the target parameter are fixed
- The remaining parameters are selected randomly
- By performing this procedure *k* times, there are a total of *k * p* configurations, where *p* - number of fixed values
- How often does one parameter beat others?

According to the described scheme, the influence of the majority of features was analyzed. For each configuration *k = 100* runs were performed.

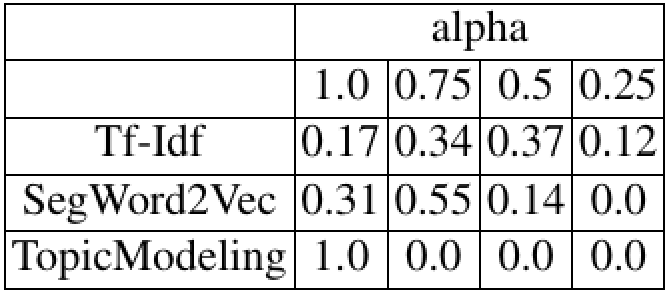
Hyperparameter Analysis Results

Keywords feature



	Keywords count						Keywords multiplier			
	0	3	5	7	9	12	1.0	1.25	1.5	3.0
Tf-Idf	0.07	0.26	0.05	0.14	0.11	0.37	0.22	0.1	0.15	0.53
SegWord2Vec	0.02	0.11	0.38	0.03	0.06	0.4	0.03	0.22	0.28	0.47
TopicModeling	0.06	0.05	0.09	0.17	0.07	0.56	0.1	0.33	0.43	0.14

Concept feature



	alpha			
	1.0	0.75	0.5	0.25
Tf-Idf	0.17	0.34	0.37	0.12
SegWord2Vec	0.31	0.55	0.14	0.0
TopicModeling	1.0	0.0	0.0	0.0

In addition, it was evaluated which segment representation model is better for this task: *Tf-Idf*: 0.0, *SegWord2Vec*: 0.9, *TopicModeling*: 0.1.

Conclusions

We have explored a new task: *Assessing Theme Adherence in Student Thesis*. The following results were obtained:

- *SegWord2Vec* segment representation model is better, than *Tf-Idf* and *Topic Modeling*
- Use of *Keywords* leads to better results
- Use of *Concepts* is very useful for *Tf-Idf* especially in combination with *Keywords*, but totally useless for *Topic Modeling*
- Use of *EmbedExp* (theme header extension using Word2Vec) is useful for *Tf-Idf*, but works not so good as *Concepts*

Questions?