Estimation by Phrases for the Closeness of a Topical Text to the Semantic Pattern without Paraphrasing

Dmitry Mikhaylov[1,*] and Gennady Emelyanov[1]

[1] Yaroslav-the-Wise Novgorod State University, Velikii Novgorod, Russia
e-mail: Dmitry.Mikhaylov@novsu.ru

Abstract. In this paper, the numerical estimation method for the closeness of a topical text to the most rational linguistic variant (i.e., semantic pattern or sense standard) of description the corresponding knowledge fragment without paraphrasing, is offered. As the analyzed texts the abstracts of scientific articles together with their titles are considered. The base for estimation of the closeness of a text to the semantic pattern is the splitting of words of each of its phrase into classes by the value of the TF-IDF metric relative to the corpus pre-formed by an expert. The paper considers two variants of estimation: relatively to the article title and the phrase closest to the semantic pattern.

Keywords: intelligent data analysis, e-learning, natural-language expression of expert knowledge, human-computer interaction in education.

1 Introduction

Development of e-learning significantly increases the qualitative requirements for electronic training materials. The major requirement here may be formulated as the sorting of information sources by degree of reflection of the most significant concepts of the studied subject area at a maximal compactness and non-redundancy of narration. Ideally, the information sources form a hierarchy at a top level of which will be placed the start points for study.

Essentially close problem is the construction and verification of thematic models of major conferences with the finding of most relevant themes for a new participant [1]. Here the theme of a document is defined by its terms from the terminological dictionary of the conference. The significance value of the term is expressed via its entropy relatively to expert classification on a given level of hierarchy. A primary role here plays a revelation of a set of text units and their relations necessary and sufficient to represent a knowledge unit and satisfies the semantic pattern.

The current work considers the possibility of applying the estimation offered in the paper [2] for the closeness to a semantic pattern and based on the TF-IDF metric without paraphrasing the phrases of the analyzed text. Herewith as the analyzed texts, the abstracts of scientific articles together with their titles are considered. These parts of the articles reflect the main content of each paper and the most important results without unnecessary methodological details.

In: P. Sosnin, V. Maklaev, E. Sosnina (eds.): Proceedings of the IS-2019 Conference, Ulyanovsk, Russia, 24-27 September 2019, published at http://ceur-ws.org

2 The choice of estimating for the closeness to the pattern for phrases and their groups

Let D be a topical corpus of texts, selected by an expert. According to the definition, TF-IDF is the product of term frequency (TF) and inverse document frequency (IDF, [3]) and intended to reflect how important a word t_i is to a document $d \in D$. We have used the classic case of term frequency, it is the number of times that the word t_i occurs in a document d divided by the total number of words in d. The IDF metrics can be determined as $idf(t_i, D) = \log(|D|/|D_i|)$, where $|D_i \subset D|$ is the number of documents where the word t_i appears at least once (i.e. $tf(t_i, d) \neq 0$).

Let X be a descent-ordered sequence of TF-IDF values for words of the initial phrase relatively to a document d from the corpus D.

Let's split X into clusters $H_1, \ldots, H_r$ using the algorithm offered by us in [4] and close to FOREL class taxonomy algorithms [5]. Further in the current paper, concerning to clustering of phrases and documents, we'll have in mind this algorithm. As the mass center of cluster H_i the arithmetic mean of all $x_j \in H_i$ like in [4] is taken. Also, we note that for $\forall i \neq j \quad H_i \cap H_j = \varnothing$, and $H_1 \bullet H_2 \bullet \ldots \bullet H_r = X$. The rule to relate elements of X to the same cluster is identical to the one used in the paper [4].

Let $\text{first}(X)$ be the first, $\text{last}(X)$ be the last element of the X sequence, and $\text{mc}(X)$ be the center of mass of X.

Statement 1. The elements of X can be related to the same cluster, if

$$\begin{cases} |\text{mc}(X) - \text{first}(X)| < \dfrac{\text{mc}(X)}{4} \\ |\text{mc}(X) - \text{last}(X)| < \dfrac{\text{mc}(X)}{4} \end{cases}. \tag{1}$$

The choice of denominators of right-hand sides of inequalities in formula (1) was based on the assumption that the elements of the same cluster always have more similarities than differences. To estimate the affinity of some phrase to the semantic pattern the most important clusters obtained from the splitting of sequence X will be:

- the cluster H_1 (the terms from the source phrase which are the most unique for the analyzed text document);
- the "median" cluster $H_{r/2}$ which will host general vocabulary that ensures periphrases and synonymous terms;
- the cluster H_r to which the terms that prevail in the corpus are corresponded.

The estimation of the closeness of a separate phrase to the semantic pattern without paraphrasing the natural-language description of a corresponding knowledge unit is based on the following empirical consideration. First, the division of words into general vocabulary and terms here should be expressed as much as possible. Another

important aspect is that the words in clusters $H_1,\dots,H_r$, formed by the TF-IDF of words of the source phrase relative to a certain $d \in D$, should be distributed more or less evenly. But unlike the estimation from the paper [2], the number of resulted clusters must be close to three as much as possible at a maximum of TF-IDF values for words related to the cluster H_1. The latter requirement means the maximal relevance of term words in phrases of selected documents to the formed corpus. Essentially, the told above can be represented as the maximization of values

$$val_1 = -1 \big/ \log_{10}\left(\Sigma_{H_1}\right), \tag{2}$$

$$val_2 = 10^{-\sigma\left(\left|H_i,\, i=\{1,\, r/2\,,r\}\right|\right)}, \tag{3}$$

and, correspondingly,

$$val_3 = \left|H_1 \setminus H_{r/2} \setminus H_r\right| \big/ len(X). \tag{4}$$

The logarithm in the denominator of (2) is taken from the sum of TF-IDF values for words related to the cluster H_1 by the value of this metric relative to document $d \in D$ under consideration; $\sigma\left(\left|H_i,\, i=\{1,\, r/2\,,r\}\right|\right)$ in formula (3) is the root-mean-square deviation (RMSD) of number of elements in a cluster from $\left\{H_1, H_{r/2}, H_r\right\}$; $len(X)$ in the denominator of formula (4) is the length of X. In a case of $\Sigma_{H_1} = 0$, the value of val_1 is assumed to be zero. If the number of TF-IDF-clusters obtained is smaller than two, the values of $\left|H_{r/2}\right|$ and $\left|H_r\right|$ are assumed to be zero. In a case of only two TF-IDF-clusters obtained, the value of $\left|H_r\right|$ is assumed to be zero.

Documents $d \in D$ are sorted by the descending product of estimations (2), (3) and (4). As the numerical estimation of the closeness of an individual phrase to the pattern the greatest of the resulting values herewith is taken.

Let **Ts** be a group of phrases, first of which is the title of scientific article and others represent its abstract. In the current paper, two variants for estimation of the affinity of **Ts** to the semantic pattern are introduced. Both variants are equally assumed the minimum of RMSD for the value of affinity to the pattern for all $Ts_i \in \mathbf{Ts}$.

The first one essentially corresponds to the order of selection of articles with the analysis of title at first and assumes the maximal closeness to the standard for it, i.e.:

$$\mathrm{N}_1(\mathbf{Ts}, D) = \frac{\max\limits_{d \in D}\left[\, val_1(Ts_1, d) \cdot val_2(Ts_1, d) \cdot val_3(Ts_1, d)\right]}{\sigma\left(\max\limits_{d \in D}\left[val_1(Ts_i, d) \cdot val_2(Ts_i, d) \cdot val_3(Ts_i, d)\right], Ts_i \in \mathbf{Ts}\right) + 1}. \tag{5}$$

Note, that estimation (5) does not imply the sorting of phrases $Ts_i \in \mathbf{Ts}$ by affinity to the semantic pattern. Such a problem statement is the most adequate to requirement general accepted in scientific periodicals to reflect in the title the content of the arti-

cle. Nevertheless, the a priori assumption of maximal closeness to the standard exactly of the title of the article is not always performed in practice.

Taking into account the mentioned above, in the second variant the maximum of the found values of affinity to the standard for all phrases $Ts_i \in \mathbf{Ts}$ is used in the numerator of formula (5):

$$\mathrm{N}_2(\mathbf{Ts}, D) = \frac{\max\limits_{d \in D}\left[val_1(Ts_{\max}, d) \cdot val_2(Ts_{\max}, d) \cdot val_3(Ts_{\max}, d) \right]}{\sigma\left(\max\limits_{d \in D}\left[val_1(Ts_i, d) \cdot val_2(Ts_i, d) \cdot val_3(Ts_i, d) \right], Ts_i \in \mathbf{Ts} \right) + 1}, \qquad (6)$$

where $Ts_{\max} \in \mathbf{Ts}$ is the phrase for which the affinity to the sense standard is maximal. To prevent a possible division by zero, a one is added to the RMSD value in the denominator of each of the formulas (5) and (6).

Statement 2. The maximal final rank in the collection will be designated to the paper with the greatest value of estimation (5) related to the same cluster with the value of estimation (6) for this paper according to the condition of *Statement 1*.

Note. The correct applying of *Statement 2* assumes relating to the same cluster the value of estimation (5) for the article with the maximal final rank, and a maximal value of estimation (5) in the collection for paper selection. At the absence of an article meets this requirement, the maximal final rank will be designated to the article with the highest value of estimation (5) in analyzed collection.

3 Experimental research

To test the proposed estimations, as an expert-formed corpus D the variant from experiments in the paper [2] was involved. It was formed from the following editions:

- Taurida journal of computer science theory and mathematics (3 papers);
- Proceedings of International conferences “Intelligent Information Processing” IIP-8 and IIP-9 of the years 2010 and 2012 (2 papers);
- Proceedings of the 15th All-Russian Conference with International Participation on Mathematical Methods for Pattern Recognition (MMPR-15, 2011, 1 paper);
- Proceedings of the Conference MMPR-13 (2007, 2 papers);
- Proceedings of the Conference MMPR-16 (2013, 14 papers);
- Proceedings of the Conference IIP-10 (2014, 2 papers);
- a scientific report prepared by the first author of the current paper in 2003.

The scope of selected papers includes:

- mathematical methods for learning by precedents (K.V. Vorontsov, M.Yu. Khachay, E.V. Djukova, N.G. Zagoruiko, Yu.Yu. Dyulicheva, I.E. Genrikhov, A.A. Ivakhnenko);
- methods and models of pattern recognition and forecasting (V.V. Mottl, O.S. Seredin, A.I. Tatarchuk, P.A. Turkov, M.A. Suvorov, A.I. Maysuradze);

- intelligent processing of experimental information (S.D. Dvoenko, N.I. Borovykh);
- image processing, analysis, classification and recognition (A.L. Zhiznyakov, K.V. Zhukova, I.A. Reyer, D.M. Murashov, N.G. Fedotov, V.Yu. Martyanov, M.V. Kharinov).

Here the number of words in corpus documents is varied from 218 to 6298, and the number of phrases per document is varied between 9 and 587. Selection of articles was made from:

- proceedings of the conference IIP-9 (2012), section "Theory and Methods of Pattern Recognition and Classification" (14 articles);
- proceedings of the conference MMPR-14 (2009), section "Methods and Models of Pattern Recognition and Forecasting" (35 articles);
- proceedings of the conference MMPR-15, section "Theory and Methods of Pattern Recognition and Classification" (18 articles) and "Statistical Learning Theory" (10 articles).

The main criterion when choosing collections, as well as when selecting texts for corpus D, was the most complete and evident division of words of the analyzed texts into general vocabulary and terms.

The software implementation (in Python 2.7) of the offered solutions and experimental results are presented on the website of Yaroslav-the-Wise Novgorod State University at http://www.novsu.ru/file/1504831.

Taking into account the conclusions of [2] regarding the semantic context of terms, the evaluation of estimations (2)–(6) was made without consideration of prepositions and conjunctions. Text extraction from a PDF file was implemented using the functions of the *pdfinterp, converter, layout*, and *pdfpage* classes as part of the *PDFMiner* package [6]. For the correctness of formula recognition, as in [2], all formulas from the analyzed documents here were translated by an expert manually into a format close to that used in LaTeX. To select the boundaries of sentences in the text by punctuation marks, the method *sent_tokenize()* of the *tokenize* class from the open-source library NLTK [7] was used. Lemmatization of words was performed using the morphological analyzer *pymorphy2* [8]. If a word has more than one parsing variant when determining its initial form (lemma), to calculate the TF-IDF measure, the closest one issued by the *n*-gram tagger from the *nltk4russian* library [9] is taken.

The experimental results represented further in the tables confirm the rule of "good manners" of some periodicals on information science and computer engineering to display in the title the name of method, model, algorithm presented by paper, as well as the theoretical basis of the proposed solutions. For the collection *"MMPR-15, Statistical Learning Theory"* the maximums of estimations (5) and (6) took place relative to the same article, a similar result was reached for the collection *"MMPR-15, Theory and Methods of Pattern Recognition and Classification"*.

As can be seen from Tables 1 and 2, the values of estimations (5) and (6) for the mentioned articles are coincided. So, according to the condition of *Statement 2*, the papers *"Принцип максимизации зазора для монотонного классификатора ближайшего соседа" (The principle of gap maximization for nearest neighbor mon-*

otonic classifier) by *K.V. Vorontsov* and *G.A. Makhina*, and "*Полные решающие деревья в задачах классификации по прецедентам" (Complete decision trees in classification tasks by precedents)* by *I.E. Genrikhov* and *E.V. Djukova* will have a maximal final rank each in its collection.

Table 1. Articles with a maximum value of estimation (5) in collections

MMPR-15, Statistical Learning Theory	
Author(s)	*Vorontsov, K.V., Makhina, G.A*
Title of the article	*The principle of gap maximization for nearest neighbor monotonic classifier*
The maximum affinity to the sense standard for the title is achieved relative to the document	*Vorontsov, K.V. Combinatorial theory of overfitting: results, applications and open problems (Комбинаторная теория переобучения: результаты, приложения и открытые проблемы). In: MMPR-15 (2011)*
Value of estimation (5)	0.0711
Value of estimation (6)	0.0711
MMPR-15, Theory and Methods of Pattern Recognition and Classification	
Author(s)	*Genrikhov, I.E., Djukova, E.V.*
Title of the article	*Complete decision trees in classification tasks by precedents*
The maximum affinity to the sense standard for the title is achieved relative to the document	*Vorontsov, K.V. Combinatorial theory of overfitting: results, applications and open problems. In: MMPR-15 (2011)*
Value of estimation (5)	0.1194
Value of estimation (6)	0.1194
MMPR-14, Methods and Models of Pattern Recognition and Forecasting	
Author(s)	*Barinova, O.V., Vetrov, D.P.*
Title of the article	*Estimates of the generalization ability for boosting with a probabilistic entries (Оценки обобщающей способности бустинга с вероятностными входами)*
The maximum affinity to the sense standard for the title is achieved relative to the document	*Vorontsov, K.V. Combinatorial theory of overfitting: results, applications and open problems. In: MMPR-15 (2011)*
Value of estimation (5)	0.1295
Value of estimation (6)	0.1295
IIP-9, Theory and Methods of Pattern Recognition and Classification	
Author(s)	*Dvoenko, S.D., Pshenichny, D.O.*
Title of the article	*On negative eigenvalues removing from matrices of pairwise comparisons (Об устранении отрицательных собственных значений матриц парных сравнений)*

The maximum affinity to the sense standard for the title is achieved relative to the document	*Dvoenko, S.D., Pshenichny, D.O. Metrical correction of matrices of pairwise comparisons (Метрическая коррекция матриц парных сравнений). In: MMPR-16 (2013)*
Value of estimation (5)	0.0920
Value of estimation (6)	0.0920

The result obtained for the collection *"MMPR-14, Methods and Models of Pattern Recognition and Forecasting"* illustrates the case when an article with the greatest value of estimation (6) in the collection has the value of estimation (5) not relates to the same cluster with it. Indeed, for the article *"Selection of support object set for robust integral indicator construction" (Выбор опорного множества при построении устойчивых интегральных индикаторов)* by *D.I. Melnikov, V.V. Strijov, E.Yu. Andreeva and G. Edenharter* the values of estimations (5) and (6) equal, correspondingly, to 0.0129 and 0.1426, form two independent clusters according to the condition of *Statement 1*. By this virtue, the maximal final rank in the collection will be designated to the article by *O.V. Barinova and D.P. Vetrov* having the maximal value of estimation (5) relatively to the considering collection.

Table 2. Articles with a maximum value of estimation (6) in collections

MMPR-15, Statistical Learning Theory	
Author(s)	*Vorontsov, K.V., Makhina, G.A.*
Title of the article	*The principle of gap maximization for nearest neighbor monotonic classifier*
Phrase closest to the standard	*The principle of gap maximization for nearest neighbor monotonic classifier*
The maximum affinity to the sense standard for the phrase is achieved relative to the document	*Vorontsov, K.V. Combinatorial theory of overfitting: results, applications and open problems. In: MMPR-15 (2011)*
Value of estimation (6)	0.0711
Value of estimation (5)	0.0711
MMPR-15, Theory and Methods of Pattern Recognition and Classification	
Author(s)	*Genrikhov, I.E., Djukova, E.V.*
Title of the article	*Complete decision trees in classification tasks by precedents*
Phrase closest to the standard	*Complete decision trees in classification tasks by precedents*
The maximum affinity to the sense standard for the phrase is achieved relative to the document	*Vorontsov, K.V. Combinatorial theory of overfitting: results, applications and open problems. In: MMPR-15 (2011)*
Value of estimation (6)	0.1194
Value of estimation (5)	0.1194
MMPR-14, Methods and Models of Pattern Recognition and Forecasting	

Author(s)	*Melnikov, D.I., Strijov, V.V. Andreeva, E.Yu. and Edenharter, G.*
Title of the article	*Selection of support object set for robust integral indicator construction*
Phrase closest to the standard	*Objects are described in linear scales (Объекты описаны в линейных шкалах)*
The maximum affinity to the sense standard for the phrase is achieved relative to the document	*Abramov, V.I., Seredin, O.S., Sulimova, V.V., Mottl, V.V. Equivalence of kernel functions and linear-space representations of arbitrary real-world objects (Эквивалентность потенциальных функций и линейных пространств представления объектов произвольной природы). In: IIP-8 (2010)*
Value of estimation (6)	0.1426
Value of estimation (5)	0.0129
IIP-9, Theory and Methods of Pattern Recognition and Classification	
Author(s)	*Zhivotovskiy, N.K., Vorontsov, K.V.*
Title of the article	*The exactness criteria of combinatorial generalization bounds*
Phrase closest to the standard	*Combinatorial theory of overfitting gives exact estimations of overfitting probability for some non-trivial sets of classification algorithms (Комбинаторная теория переобучения даёт точные оценки вероятности переобучения для некоторых нетривиальных семейств алгоритмов классификации)*
The maximum affinity to the sense standard for the phrase is achieved relative to the document	*Vorontsov, K.V. Combinatorial theory of overfitting: results, applications and open problems. In: MMPR-15 (2011)*
Value of estimation (6)	0.1336
Value of estimation (5)	0.0600

Table 3. Articles with a maximal final rank in collections

MMPR-15, Statistical Learning Theory	
Author(s)	*Vorontsov, K.V., Makhina, G.A.*
Title of the article	*The principle of gap maximization for nearest neighbor monotonic classifier*
MMPR-15, Theory and Methods of Pattern Recognition and Classification	
Author(s)	*Genrikhov, I.E., Djukova, E.V.*
Title of the article	*Complete decision trees in classification tasks by precedents*
MMPR-14, Methods and Models of Pattern Recognition and Forecasting	
Author(s)	*Barinova, O.V., Vetrov, D.P.*
Title of the article	*Estimates of the generalization ability for boosting with a probabilistic entries*

IIP-9, Theory and Methods of Pattern Recognition and Classification	
Author(s)	*Dvoenko, S.D., Pshenichny, D.O.*
Title of the article	*On negative eigenvalues removing from matrices of pairwise comparisons*

A similar situation also takes place for the collection *"IIP-9, Theory and Methods of Pattern Recognition and Classification"*. Here the maximal value of estimation (6) equal to 0.1336 will be belonged to the article *"Критерии точности комбинаторных оценок обобщающей способности" (The exactness criteria of combinatorial generalization bounds)* by *N.K. Zhivotovskiy and K.V. Vorontsov*. The value of estimation (5) here is equal to 0.0600 and related to the same cluster with the maximal value equal to 0.0920 for this estimation in collection, but not lies in the same cluster with the value of estimation (6) for this article. Therefore the maximal final rank obtains the article by *S.D. Dvoenko and D.O. Pshenichny* having the greatest value of estimation (5) in the considered collection.

Since the title and phrases of the article abstract (by definition) represent a certain single semantic image, it is entirely acceptable to swap with each other the estimations (5) and (6) in *Statement 2*. In considered examples for both collections by *MMPR-15* conference, the maximal final ranks herewith will be designated to the same articles. For the collection *"MMPR-14, Methods and Models of Pattern Recognition and Forecasting"* the maximal rank here the article by *O.V. Barinova and D.P. Vetrov* obtains again. Indeed, the maximal value of estimation (6) in this collection will be for the article by *D.I. Melnikov, V.V. Strijov, E.Yu. Andreeva and G. Edenharter*. But as we showed earlier, the values of estimations (5) and (6) for this article are related to different clusters. Therefore, according to the condition of *Statement 2*, the maximal final rank obtains the article having among the remaining articles (except the article mentioned above) the maximal value of estimation (6) relating to the same cluster with the value of estimation (5) for itself, i.e. the article by *O.V. Barinova and D.P. Vetrov*.

The single exclusion in the considered series of experiments will be the result for collection *"IIP-9, Theory and Methods of Pattern Recognition and Classification"*. As in the previous example, the maximal final rank in the collection may be designated to the article by *S.D. Dvoenko and D.O. Pshenichny* as having the maximal value of estimation (6) which relates to the same cluster with the value of estimation (5) for this paper. But the value of estimation (6) for it does not relate to the same cluster with the maximal value of this estimation in the collection. So, the maximal final rank together with the maximal value of estimation (6) in this collection here obtains the article by *N.K. Zhivotovskiy and K.V. Vorontsov*.

It should be noted that both estimations, like previously proposed in [2], depend essentially on the selection of a subject-oriented corpus D by the expert. Nevertheless, the presented results confirm the hypothesis relative to the semantic load of title for scientific paper on information science and computer engineering. For disputable cases similar to the shown in the previous paragraph, depending on the subject area it's possible to give preference to the requirement of relating to the cluster of maximal value of either estimation (5) or estimation (6).

4 Conclusion

The main result of this paper is the *proposed method for estimating the closeness of a text to the semantic pattern relative to a topical text corpus.*

The effectiveness of the proposed method can be estimated by splitting of texts in the collection into clusters by the value of used estimation for the closeness to a pattern and the ratio of the number of texts assigned to the cluster of the highest evaluation values to the total number of texts in the collection. So, on the material of collections mentioned in Tables 1–3, we have at least a threefold reduction in the number of documents that should be read first when studying a given subject area.

Taking into account the evaluated degree of division of its words into general vocabulary and terms, when a phrase is referred to as a "representative of the pattern", it is also of interest to reveal key combinations from words with the greatest TF-IDF values. At disputable cases, the presence of key combinations in abstracts and titles can be a basis for designating the final rank to the article. To identify the key combination of words herewith it is necessary to enter into consideration the interpretation of TF-IDF metrics which would estimate the number of simultaneous presence of all words from analyzed combination in the phrases of separate document.

The work was supported by the RFBR (project no. 19-01-00006).

References

1. Kuzmin, A.A., Aduenko, A.A., Strijov, V.V.: Thematic Classification Using Expert Model for Major Conference Abstracts (in Russian). Informational Technologies **6** (214), 22–26 (2014)
2. Emelyanov, G.M., Mikhailov, D.V., Kozlov, A.P.: Relevance of a Set of Topical Texts to a Knowledge Unit and the Estimation of the Closeness of Linguistic Forms of Its Expression to a Semantic Pattern. Pattern Recognition and Image Analysis **28** (4), 771–782 (2018)
3. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation **60** (5), 493–502 (2004)
4. Emelyanov, G.M., Mikhaylov, D.V., Kozlov, A.P.: Formation of the representation of topical knowledge units in the problem of their estimation on the basis of open tests (in Russian). Machine learning and data analysis **1** (8), 1089–1106 (2014)
5. Zagoruiko, N.G.: Applied Methods of Data and Knowledge Analysis (in Russian). Institute of Mathematics SD RAS, Novosibirsk (1999)
6. PDFMiner – Python PDF parser and analyzer, https://euske.github.io/pdfminer/
7. Natural Language Toolkit, http://www.nltk.org
8. Korobov, M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: 4th International Conference on Analysis of Images, Social Networks and Texts, pp. 320–332. Springer (2015)
9. Moskvina, A., Orlova, D., Panicheva, P., Mitrofanova, O.: Development of the Core for Syntactic Parser for Russian based on NLTK libraries (in Russian). In: Computational Linguistics and Digital Ontologies: Proceedings of the XIX International Joint Conference on Internet and Modern Society (IMS 2016), pp. 44–54. St. Petersburg (2016)