Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction

Nina Khairova[1], Włodzimierz Lewoniewski[2], Krzysztof Węcel[2]

[1] National Technical University "Kharkiv Polytechnic Institute",
NTU "KhPI" 2, Kyrpychova str., 61002, Kharkiv, Ukraine,
khairova@kpi.kharkov.ua
[2] Poznań University of Economics and Business,
Al. Niepodległości 10, 61-875 Poznań, Poland

Abstract. * We present the method of estimating the quality of articles in Russian Wikipedia that is based on counting the number of facts in the article. For calculating the number of facts we use our logical-linguistic model of fact extraction. Basic mathematical means of the model are logical-algebraic equations of the finite predicates algebra. The model allows extracting of simple and complex types of facts in Russian sentences. We experimentally compare the effect of the density of these types of facts on the quality of articles in Russian Wikipedia. Better articles tend to have a higher density of facts.
Keywords: Russian Wikipedia, article quality, fact extraction, logical equations.

1 Introduction

Nowadays, in order to make correct financially significant economic decisions, a large amount of information and knowledge should be analyzed. Useful information can be found both in specialized economic sources and in Web-resources of general nature. In recent years Wikipedia has become one of the most important sources of knowledge throughout the world. In the ranking of the most popular websites this online encyclopedia with more than 44 million articles in almost 300 languages[1] occupies the 5th place in the world. Many articles of this multilingual encyclopedia contain information about the various types of products, e.g. cars, movies, video games, cell phones. Information in Wikipedia is also used to automatically enrich various public databases (such as DBpedia).

Russian-language edition of Wikipedia is one of the major language versions of the online encyclopedia. For instance, the largest language version, which is English Wikipedia, contains five million articles, while Russian Wikipedia contains one million articles.

The number of articles is continually rising, and authors of the articles may not have an official confirmation of their expertise in a given domain. Sometimes the authors are anonymous. Additionally, there is no process of obligatory expert reviewing of the

* This is a preprint version. The final publication is available at Springer via https://doi.org/10.1007/978-3-319-59336-4_3
[1] https://meta.wikimedia.org/wiki/List_of_Wikipedias

Wikipedia articles does not exist. All changes to the article immediately are visible on the site. Therefore, in order to provide the computer encyclopedia with qualitative information, which is reliable for making business decisions, its articles' quality must be evaluated.

Generally, the quality of the Wikipedia article is estimated manually in accordance with the Wikipedia policies, guidelines and community rules in a particular language version. Today, there exist techniques of automatic evaluation of the quality of articles that are mostly based on using different quantitative characteristics (article length, number of images, number of links and others). However, qualitative characteristics are rarely used to evaluate the quality of the Wikipedia articles. There are at least two reasons for that. First, text has relatively rich semantics [1]. Second, qualitative grammatical and stylistic characteristics of the text of the article depends on an article's language [2, 3].

We suggest to use qualitative characteristics of the density of simple and complex facts to automatically estimate the quality of articles in Russian Wikipedia. In order to identify a fact in a text, we developed the logical-linguistic model of fact extraction from Russian sentences.

2 Related work

Nowadays, there exist quite a lot of the approaches to measuring the quality of textual information [4, 5, 6]. Among other things in scientific works, various methods for automatic distinguishing of high-quality Wikipedia articles are written. Most of them use various quantitative features of the article as independent variables and the article quality class as a dependent one. Usually, the quality of Web content is assessed with such metrics as objectivity [7], content maturity and readability [8]. At the same time, current approaches to the automatic assessment of documentation quality are mostly based on statistical models or on some formalisation of grammar. For instance, Blumenstock [9] proposes to use word count as a simple metric for capturing quality indicator of Wikipedia articles, Lipka and Stein [2] exploit an article's character trigram distribution for the automatic assessment of information quality. Online service WikiRank[2] used different quantitative parameters of articles (text length, the number of images, references etc.) to calculate the so-called relative quality of the same article in various language versions of Wikipedia.

However, it is obvious that quality of texts may depend not only on grammatical features of a document but also on its semantic characteristics [10]. The reason is that text informativeness mostly depends directly on semantics. Most applications that use semantic characteristics to assess the quality of textual documents are based on knowledge from ontologies such as WordNet [11]. In this approach, it is necessary to have explicitly expressed relationships such as meronymy and hypernymy between entities in the text. In [12, 13], it is proposed to use the number of facts and the factual density as features to identify high quality articles in English Wikipedia. Lex et al. consider

[2] http://wikirank.net

the fact in the form of a triplet with two entities and a relationship between them. Authors used the ReVerb Open Information Extraction framework to extract facts from the articles in English Wikipedia [14].

Today there exist a lot of different techniques for information extraction and, in particular, for facts extraction. The most of them are domain-specific or focus on a small number of relations in specific preselected domains [14]. More advanced Information Extraction systems use a domain-independent architecture and sentence analyzer. Nevertheless these systems demand either a large hand-tagged corpus to create a training set or knowledge from ontologies such as WordNet [15]. Anyway, every modern facts extraction system depends on the language of texts which are analyzed and the vast majority of these systems is focused on English, Spanish and German [16, 3].

In our study we consider densities of simple and complex facts as features to measure the quality of articles in Russian Wikipedia. In order to extract facts from Russian texts we propose to use the built logical-linguistic model.

3 Formal Model of Fact Extraction

In order to build a model we use logical-algebraic equations of the finite predicates algebra (FPA) [17], which can describe any finite and determined relations. These mathematical tools of the FPA have been successfully used for building different Artificial Intelligence and natural language models [18]. Basic predicates of the FPA are the predicates of recognition of the element a by the variable x_i:

$$x_a^i = \begin{cases} 1, \text{if } x_i = a \\ 0, \text{if } x_i \neq a \end{cases} \quad (1 \leq i \leq n), \tag{1}$$

where a is any of the elements of universe U. In our model, the universe U contains various elements of the language system: lexemes, morphemes, sentences, grammatical and semantic features of Russian words etc. We then introduce to the universe the subsets of grammatical and semantic features of words in Russian sentences $M = \{X, Y, Z\}$, where X is the finite subset of the characteristic of animacy, Y is the finite subset of semantic features of nouns and Z is the finite subset of morphological features that describe the grammatical cases of Russian nouns.

Let us write the grammatical cases of Russian nouns per the predicates of recognition of the element (1)

$$P(z) = z^{nom} \vee z^{gen} \vee z^{dat} \vee z^{acc} \vee z^{inc} \vee z^{loc}, \tag{2}$$

where nom, gen, dat, acc, ins, loc are nominative, genitive, dative, accusative, instrumental and prepositional cases of Russian nouns respectively. Similarly, we can write semantic features of the nouns that represent the participants of the sentences:

$$P(x) = z^{anim} \vee z^{inan}, \tag{3}$$

$$P(y) = y^{device} \vee y^{hum} \vee y^{tool} \vee y^{pc:hue} \vee y^{spacee} \vee y^{time:moment} \vee y^{time:period} \vee y^{s:loc} \vee y^{others}, \tag{4}$$

where $P(x)$ is predicate that describes the feature of animacy of the noun (index $anim$ means animate, index $inan$ means inanimate); $P(y)$ is predicate that describes others particular semantic feature of the noun ($device$, $tool$, $space$, $time : moment$, $time : period$, index hum means belonging to the semantic class "person", index $pc : hum$ means belonging to the semantic class "part of the body" and index $s : loc$ means belonging to the semantic class "destination"). Such choice of semantic categories is motivated by the necessity of the correct and complete description of the seven considered semantic roles. The labelling corresponds to semantic labelling in Russian National corpus.

Let us introduce the system of the predicates $P_k(x, y, z)$ over Cartesian products $P(x) \times P(y) \times P(z)$:

$$P_k(x, y, z) = \gamma_k(x, y, z) \wedge P(x) \wedge P(y) \wedge P(z), \tag{5}$$

where the predicates $\gamma_k(x, y, z)$, $k \in [1, h]$ represent a complete set of semantic roles of the sentence participants of the facts that we consider, where h is the number of semantic roles of the facts in our model. We base on the assumption of Fillmore [19] that there is an action and participants of the action at the semantic level of a sentence. These are represented by a verb and nouns at the grammar level respectively. Every participant plays certain semantic role (a.k.a. deep case) in the action.

The predicate $\gamma_k(x, y, z)$ holds if the specific grammatical and semantic features of the noun in a Russian sentence define the specific semantic role of the sentence participant and the predicate is false otherwise. Therefore, the predicate excludes morphological and semantic features of the noun that are not inherent in the specific semantic role.

In our study, we consider the simple and the complex types of facts. The simple fact consists of the Subject and the Predicate[3]. In grammar, the simple fact is represented by the smallest grammatical clause. A typical clause is a group of words that includes a verb (or a verb phrase) and a noun (or a noun phrase) [20]. The complex fact apart from the Subject and the Predicate consists of the Object or others participants of an action. In grammar, the complex fact is represented by a sentence with a verb (or a verb phrase) and a few nouns or noun phrases.

We define the semantic role of the Subject of a fact via the predicate γ_1:

$$\gamma_1(x, y, z) = x^{anim} z^{nom} \vee x^{inan} z^{nom} (y^{device} \vee y^{tool} \vee y^{pc:hue}). \tag{6}$$

The predicate $\gamma_1(x, y, z)$ shows grammatical and semantic features of a noun in Russian sentence that denotes the Subject of a fact. We also explicitly distinguish the semantic role of Object of a fact via the predicate γ_2:

$$\gamma_2(x, y, z) = z^{acc} (x^{inan} \vee x^{anim}) \tag{7}$$

We also explicitly distinguish the semantic roles of other parts related to the fact via a set of the predicates $\{\gamma_3, ..., \gamma_7\}$. Grammatical and semantic characteristics of the beneficiary of an action is defined by the following predicate:

$$\gamma_3(x, y, z) = z^{dat} y^{hum} x^{anim} \tag{8}$$

The predicate γ_4 denotes semantic and grammatical features of the action tool or the action reason:

[3] We use 'Subject', 'Object' and 'Predicate' with the first upper-case letters to denote the element of a fact triplet Subject -> Predicate -> Object

$$\gamma_4(x,y,z) = z^{ins}x^{inan}(y^{tool} \vee y^{pc:hum} \vee y^{device}) \quad (9)$$

We distinguish the attributes of location, time and destination of the action via the predicates γ_5, γ_6,γ_7 respectively:

$$\gamma_5(x,y,z) = z^{loc}x^{inan}(y^{space} \vee y^{s:loc}) \quad (10)$$

$$\gamma_6(x,y,z) = x^{inan}(z^{acc}y^{time:moment} \vee z^{loc}y^{time:period}) \quad (11)$$

$$\gamma_7(x,y,z) = z^{acc}x^{inan}y^{space} \quad (12)$$

Based on the above predicates, we can define the simple fact and the complex fact as follows.

Definition 1. The *simple fact* in a Russian sentence is the smallest grammatical clause that includes a verb and a noun, where the semantic and grammatical features of the noun have to denote the Subject of the fact according to the equation (6).

Definition 2. The *complex fact* in Russian texts is a grammatical sentence that includes a verb and a few nouns. Among these nouns, one has to play the semantic role of the Subject (6), semantic and grammatical characteristics of the other nouns have to satisfy one or more equations (7-12).

Using some definitions from the recent works on measuring the quality of Web content [12, 13] we can also denote density of simple and complex facts in the Russian Wikipedia article.

Definition 3. The *density of simple facts* in the Russian Wikipedia article is defined as the number of simple facts divided by the number of words in the article.

Definition 4. The *density of complex facts* in the Russian Wikipedia article is defined as the number of complex facts divided by the number of words in the article.

4 Experiments and Results

Our dataset includes about 31,000 present articles (December 2016) from the most popular domains from Russian Wikipedia. Table 1 shows the distributions of the analyzed articles according to domains of Russian Wikipedia. There is no generally accepted standard classification of articles quality in Wikipedia community. The classification schemes vary in language versions. For instance, Belarus version uses three quality classes, whereas German one uses only two classes. In Russian Wikipedia, there are seven quality classes that can show the "maturity" of an article. They are (in decreasing order): Featured, Good, Solid, Full, Developed, Developing and Stub.

According to previous studies [12, 21, 22], we distinguish two groups to evaluate the quality of the Russian Wikipedia articles. The first group includes Featured, Good, Solid classes and it is called **GoodEnough** group articles. The second group includes Full, Developed, Developing and Stub classes and is referred to as **NeedsWork** group. We consider that the articles in the first group are of higher quality than the articles in the other group. The main reason for such conclusion is the following. To receive any estimates from the first group of estimates, the article must be subjected to a complex procedure involving discussion and voting of the users of Wikipedia.

Using the capabilities of API Wikipedia we have created two corpora of plain articles texts of selected domains. The first corpus contains articles from Russian Wikipedia that are assigned to Featured, Good, Solid categories. The second corpus contains articles from Russian Wikipedia that are assigned to Full, Developed, Developing and Stub categories.

Table 1. The distributions of the analyzed articles according to domains of Russian Wikipedia.

Domain	All articles	NeedsWork articles				GoodEnough articles		
		Stub	Developing	Developed	Full	Solid	Good	Featured
Adm. division	28691	811	289	40	9	10	6	1
Album	14039	5153	760	212	46	75	109	35
Company	9343	318	385	100	18	15	14	3
Film	25148	92	157	53	8	73	46	27
Filmmaker	23155	468	251	66	11	42	34	4
Football player	28905	330	486	36	9	37	58	12
Human settlement	183411	1407	2153	135	24	22	22	7
Military person	32728	564	1237	298	48	412	55	6
Musician	19313	381	416	82	16	32	34	14
Officeholder	35969	1650	844	283	101	416	159	44
Person	40829	874	898	310	60	230	81	27
River	31008	166	103	22	6	19	7	2
Scientist	32327	1363	3337	421	46	176	74	40
Writer	16158	281	494	196	32	31	31	23

Before the application of our model of fact extraction, we apply the pymorphy2[4], the library for morphological analysis of the Russian language. Our algorithm uses the OpenCorpora dictionary[5].

In order to estimate the quality of articles in Russian Wikipedia based on our logical-linguistic model of fact extraction, we focus on two approaches. In the first approach, we determine the average densities of simple and complex facts in each category of each domain of our corpora.

The second approach is based on the hypothesis that subjectivity in an article has a large impact on the quality of Wikipedia text. For instance, according to a widely accepted standard that all editors English Wikipedia should normally follow, all content on Wikipedia must be written from a neutral point of view.

In the second approach, before we determine the average densities we excluded facts that may comprise some subjective assessment of the authors. In order to solve this problem, we have created the set of Russian verbs V that have certain semantic component of subjectivity. The set includes 120 speech verbs (such as *tell*, *recall*, *dictate* and others), 154 feelings verbs and 103 emotions verbs (such as *wish*, *rejoice*, *worry* and others). We designate these verbs as the mental verbs. If a simple or a complex fact includes Predicate that is represented by a verb from the set V, we exclude the fact from the number of facts in a calculation of density of facts. As a result of this procedure the number of simple facts was decreased by 7.37% and the number of complex facts was decreased by 6.86% in GoodEnough articles group. The number of simple facts was

[4] https://pymorphy2.readthedocs.io
[5] http://opencorpora.org

decreased by 10.5%, the number of complex facts was decreased by 9.46% in NeedsWork articles group. This supports our hypothesis that the higher quality Wikipedia articles the less subjective they are. The results of these studies are shown in the Table 2.

Table 2. Simple and complex facts density in Russian articles Wikipedia corpora (DSF - density of simple facts, DCF - density of complex facts)

Domain	with mental verbs				without mental verbs			
	GoodEnough		NeedsWork		GoodEnough		NeedsWork	
	DSF	***DCF***	***DSF***	***DCF***	***DSF***	***DCF***	***DSF***	***DCF***
Administrative division	0.043	0.041	0.040	0.034	0.041	0.039	0.040	0.033
Album	0.046	0.041	0.021	0.019	0.040	0.036	0.019	0.018
Company	0.040	0.038	0.033	0.031	0.038	0.036	0.032	0.030
Film	0.051	0.045	0.039	0.036	0.044	0.040	0.035	0.032
Filmmaker	0.046	0.043	0.022	0.021	0.042	0.040	0.021	0.020
Football player	0.051	0.048	0.038	0.036	0.048	0.045	0.037	0.035
Human settlement	0.043	0.040	0.033	0.031	0.041	0.038	0.031	0.029
Military person	0.033	0.031	0.028	0.027	0.032	0.030	0.027	0.026
Musician	0.043	0.039	0.028	0.026	0.038	0.035	0.026	0.025
Officeholder	0.043	0.040	0.031	0.029	0.039	0.036	0.029	0.028
Person	0.043	0.040	0.031	0.029	0.039	0.036	0.029	0.027
River	0.044	0.041	0.038	0.035	0.041	0.039	0.036	0.033
Scientist	0.030	0.028	0.024	0.023	0.027	0.026	0.022	0.021
Writer	0.039	0.036	0.029	0.027	0.035	0.032	0.027	0.025

Table 2 shows the dependence of the simple facts density and complex facts density from quality categories and domains of the articles. The table compares the results of the first and the second approaches. The table compares the results of two approaches. In the first approach, we calculate the average of densities of simple and complex facts in each category of each domain of our corpora. In the second approach, we carry out similar calculations, excluding facts with the so-called mental verbs.

Table 3 shows mean, standard deviation and median of simple and complex facts density in the articles of two corpora. Figure 1 shows four curves for the densities of simple and complex facts in different domains of two our corpora. The plain lines represent the densities of simple and complex facts in the GoodEnough group for higher quality articles. The dotted lines represent the densities of simple and complex facts in the NeedsWork group of articles.

We found that the densities of simple and complex facts in higher quality articles corpus are higher than the similar densities in the lower quality articles corpus for all domains. From this observation, we conclude that the densities of simple and complex facts, along with the article

Table 3. Mean, standard deviation and median (DSF - density of simple facts, DCF - density of complex facts)

Parameter	with mental verbs				without mental verbs			
	GoodEnough		NeedsWork		GoodEnough		NeedsWork	
	DSF	***DCF***	***DSF***	***DCF***	***DSF***	***DCF***	***DSF***	***DCF***
Mean	0.041	0.037	0.028	0.026	0.037	0.034	0.026	0.025
Median	0.042	0.038	0.027	0.026	0.038	0.035	0.025	0.024
Std. deviation	0.012	0.010	0.016	0.015	0.010	0.009	0.016	0.014

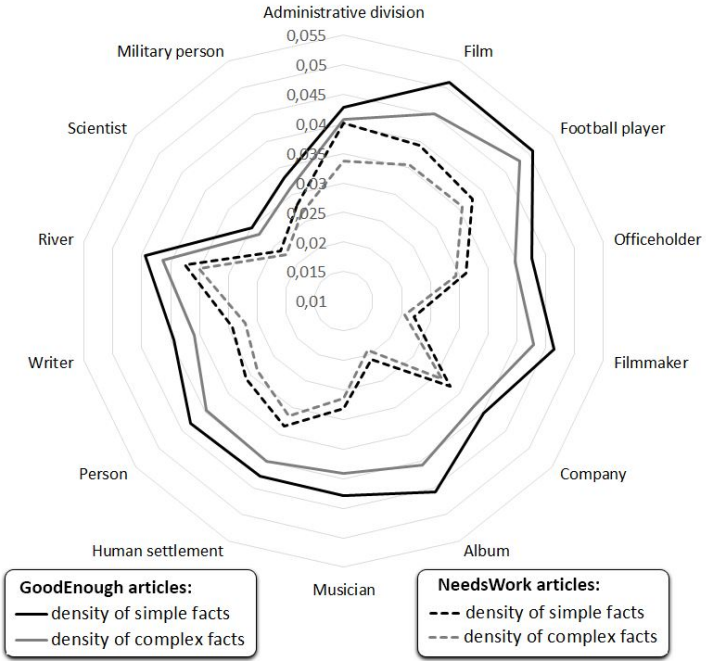

Fig. 1. Densities of simple and complex facts in different domains of the GoodEnough and NeedsWork groups of articles.

length, can be a good feature to separate higher quality articles of Russian Wikipedia from lower quality ones. Besides, data in Table 3 shows that standard deviations of complex facts distributions are less than standard deviations of simple facts distributions for all groups of articles. It means that values of densities for complex facts are closer to means than for simple facts. It should be noted also that the densities of simple and complex facts depend on a particular domain, though the ratio of the densities of simple and complex facts in the two corpora is retained.

Additionally, we can see that the density of complex facts is a more discriminative feature than the density of simple facts for distinction of higher quality articles of Russian Wikipedia.

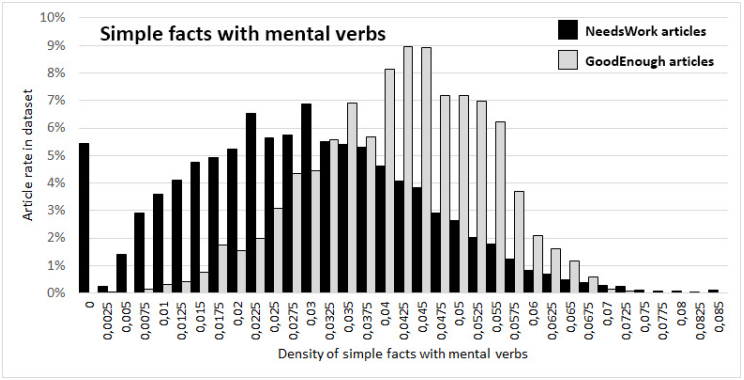

Fig. 2. Distributions articles of GoodEnough and NeedsWork groups according to the density of simple facts including mental verbs.

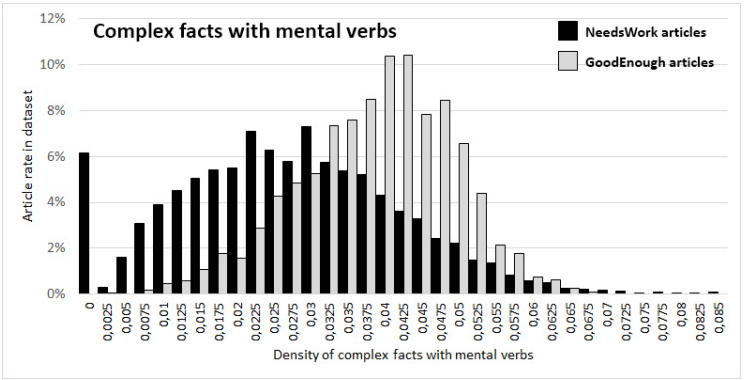

Fig. 3. Distributions articles of GoodEnough and NeedsWork groups according to the density of complex facts including mental verbs.

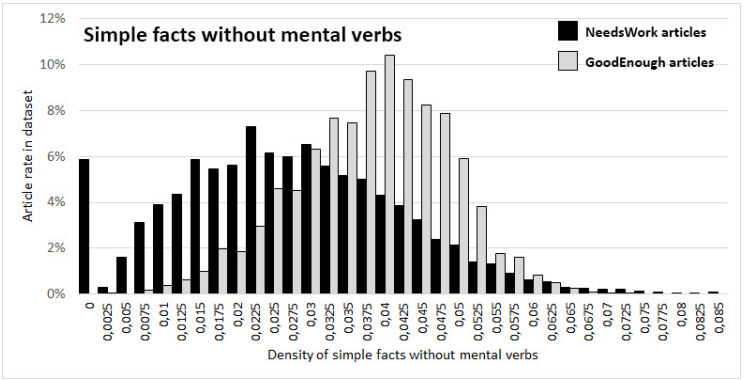

Fig. 4. Distributions of articles of GoodEnough and NeedsWork groups according to the density of simple facts excluding the mental verbs.

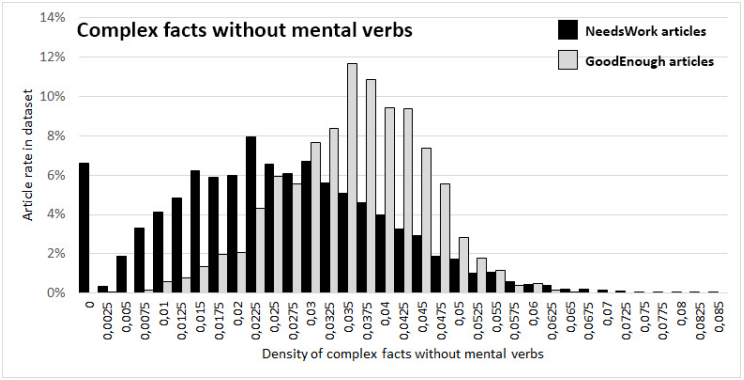

Fig. 5. Distributions of articles of GoodEnough and NeedsWork groups according to the density of complex facts excluding the mental verbs.

Figures 2-5 show the distributions of the articles of both corpora according to the densities of simple and complex facts. Figure 2 represents the distribution of the articles of two groups according to the density of simple facts including mental verbs. Figure 3 represents the similar distribution according to the density of complex facts including mental verbs. Analogously, figures 4 and 5 show the distributions of the articles of both corpora according to the densities of simple and and complex facts respectvely, excluding the mental verbs. The separation of distributions calculated with the so-called mental verbs and the one without mental verbs helps in understanding the impact of neutral point of view on quality of the article.

Since the numbers of articles in GoodEnough and NeedsWork groups are different, we normalise them by representing the article rate in the respective corpus. We can see that the articles from GoodEnough corpus have relatively higher densities of simple and complex facts than the articles from NeeedWork corpus. Additionally, we found that the distribution of the articles according to the density of complex facts (figures 3 and 5) is more demonstrative than the distribution of the articles according to the density of simple facts (figures 2 and 4).

5 Conclusions and Directions for Future Work

In this paper we leveraged the semantic categories of the densities of simple and complex facts to determine the quality of Wikipedia articles. In order to calculate number of simple and complex facts we proposed to use our logical-linguistic model of fact extraction from Russian texts.

The performed experiment showed that density of simple facts and density of complex facts, which were selected as a result of the model application, indeed characterise the quality level of articles in Russian Wikipedia. Additionally, elimination of facts with the so-called mental verbs allowed us to better distinguish the quality of articles, as the reduction rate of density of facts was higher in articles of lower quality. However, the influence of the facts whose Predicates have a flavour of subjectivity on the quality of Wikipedia articles requires further study.

The results of the paper can increase precision of the quality classification of articles in Russian Wikipedia. The obtained features, along with others, can be used in supervised learning algorithms that have shown their effectiveness in other studies related to the automatic evaluation of the Wikipedia articles quality. One should note that regarding distinction of higher quality articles of Russian Wikipedia, the density of complex facts is a more discriminative feature than the

density of simple facts. Furthermore, the density of facts (complex or simple) excluding the mental verbs is a more discriminative feature than the density of facts (complex or simple) including the mental verbs.

We suggest that trends and dependencies between qualitative characteristics of the density of simple and complex facts and the quality of Wikipedia articles also cover other languages. However, we should develop the specific logical-linguistic model of fact extraction for every language [3].

Additionally, in our logical-linguistic model, we consider grammatical and semantic features of words in Russian sentences. However, consideration of the semantic characteristics in the conducted experimental research is limited. This is due to using the pymorphy2 library, which uses a limited number of tags. In the future, we aim to consider influence of all semantic features of Russian words on the result of the implementation of the logical-linguistic model of fact extraction in more details.

References

1. Anderka, M.: Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. Phd, Bauhaus-Universitaet Weimar Germany (2013)
2. Lipka, N., Stein, B.: Identifying Featured Articles in Wikipedia: Writing Style Matters. Proceedings of the 19th International Conference on World Wide Web (2010) (2010) 1147–1148
3. Khairova, N.F., Petrasova, S., Gautam, A.P.S. In: The Logical-Linguistic Model of Fact Extraction from English Texts. Springer International Publishing, Cham (2016) 625–635
4. Arthur, J.D., Stevens, K.T.: Document quality indicators: A framework for assessing documentation adequacy. Journal of Software Maintenance: Research and Practice **4**(3) (1992) 129–142
5. Knight, S.a., Burn, J.: Developing a framework for assessing information quality on the world wide web. InformingSciJ **8** (2005) 159–172
6. Shpak, O., Löwe, W., Wingkvist, A., Ericsson, M.: A method to test the information quality of technical documentation on websites. In: 2014 14th International Conference on Quality Software. (Oct 2014) 296–304
7. Lex, E., Juffinger, A., Granitzer, M.: Objectivity classification in online media. In: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia. HT '10, New York, NY, USA, ACM (2010) 293–294
8. Weber, N., Schoefegger, K., Bimrose, J., Ley, T., Lindstaedt, S., Brown, A., Barnes, S.A. In: Knowledge Maturing in the Semantic MediaWiki: A Design Study in Career Guidance. Springer Berlin Heidelberg, Berlin, Heidelberg (2009) 700–705
9. Blumenstock, J.E.: Size matters: word count as a measure of quality on wikipedia. In: WWW. (2008) 1095–1096
10. Wingkvist, A., Ericsson, M., Löwe, W.: Making sense of technical information quality — a software-based approach measuring the quality of technical data depends on developing models from which metrics can be extracted and analyzed. using an open source tool the authors describe one approach to this. (2012)
11. Fellbaum, C.: Wordnet: An electronic lexical database, edited by christiane fellbaum (1998)
12. Lex, E., Voelske, M., Errecalde, M., Ferretti, E., Cagnina, L., Horn, C., Stein, B., Granitzer, M.: Measuring the quality of web content using factual information. Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality - WebQuality '12 (2012) 7

13. Horn, C., Zhila, A., Gelbukh, A., Kern, R., Lex, E.: Using factual density to measure informativeness of web documents. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16. Number 085, Linköping University Electronic Press (2013) 227–238
14. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Communications of the ACM **51**(12) (2008) 68–74
15. Eugene, A., Luis, G.: Extracting relations from large plain-text collections. Proc. ACM **2000** (2000)
16. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2011) 1535–1545
17. Bondarenko, M., Shabanov-Kushnarenko, J. In: The intelligence theory. Kharkiv: "SMIT" (2007) 576
18. Petrasova, S., Khairova, N.: Automatic identification of collocation similarity. In: 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT). (Sept 2015) 136–138
19. Fillmore, C.J.: The case for case, dins. In Bach, E., Harms, R., eds.: Universals in Linguistic Theory. Holt, Rinehart, and Winston (1968)
20. Osborne, T., Gross, T.: Constructions are catenae: Construction grammar meets dependency grammar. Cognitive Linguistics **23**(1) (2012)
21. Węcel, K., Lewoniewski, W.: Modelling the Quality of Attributes in Wikipedia Infoboxes. In Abramowicz, W., ed.: Business Information Systems Workshops. Volume 228 of Lecture Notes in Business Information Processing. Springer International Publishing (2015) 308–320
22. Lewoniewski, W., Węcel, K., Abramowicz, W.: Quality and importance of wikipedia articles in different languages. In: Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016, Proceedings. Springer International Publishing, Cham (2016) 613–624