Enrichment of information in multilingual Wikipedia based on quality analysis

Włodzimierz Lewoniewski[1]

Poznań University of Economics and Business,
Al. Niepodległości 10, 61-875 Poznań, Poland,
wlodzimierz.lewoniewski@ue.poznan.pl

Abstract. [*] Despite the fact that Wikipedia is one of the most popular sources of information in the world, it is often criticized for the poor quality of content. In this online encyclopaedia articles on the same topic can be created and edited independently in different languages. Some of this language versions can provide valuable information on a specific topics. Wikipedia articles may include infobox, which used to collect and present a subset of important information about its subject. This study presents method for quality assessment of Wikipedia articles and information contained in their infoboxes. Choosing the best language versions of a particular article will allow for enrichment of information in less developed version editions of particular articles.

Keywords: Wikipedia, article quality, infobox, DBpedia.

1 Introduction

Knowledge exchange is one of the key factors for success. Internet allows to interact and share global information. According to the Internet World Stats in March 2017, about half the world's population are Internet users[1]. Web 2.0 technologies allows users became producers of the online-content through collaborative platforms. Collaborative editing can be defined by its attributes: writing in a shared document, collaborative processes, data lineage, distributed teams, placeless document philosophy, flexible handling of content and layout [1]. At present everyone can contribute to common human knowledge on the Internet. One of the best examples of such online repositories are wiki websites where content can be created and changed from a web browser. The most popular wiki website is Wikipedia.

More than 15 years, Wikipedia exists as a general available encyclopaedia, where everyone can contribute to contributing content. Wikipedia also is one of the most successful examples of mass collaboration [2]. However, this free online-encyclopaedia does not fulfil all the attributes of a classical collaborative tool. For example, Wikipedia users do not work with separate documents, but with articles, which are integrated in a searchable knowledge base [3].

[*] This is a preprint version. The final publication is available at Springer via https://doi.org/10.1007/978-3-319-69023-0_19

[1] http://www.internetworldstats.com/stats.htm

According to the latest statistics, Wikipedia is fifth most popular website in the Internet[2]. For more than 15 years this free online encyclopedia as become more and more popular and important sources of knowledge throughout the world. Wikipedia contains over 44 million articles in about 300 language editions[3]. The biggest language version is English with over 5,4 million articles.

Despite the fact that an article in Wikipedia on the same topic can be presented in different languages, each of these versions can be created and edited by users separately. Consequently, this can often be observed differences between information quality in various languages of the same article. Naturally, to compare such versions it is often necessary for users to have knowledge in these languages.

Wikipedia pages about famous people, firms, products often appear as first in search results of Google, Bing, Yandex and other search engines. It is expected that visitors of Wikipedia and its editors are interested in the high quality of content contained in this online knowledge base. So presentation of information in different languages is particularly important for users who use search engines in their native (non-English) language. Also, some topics may be more popular in some countries and therefore more likely to find more information on same topic in relevant language versions (other than English). In addition, there are topics that are not described in English Wikipedia, despite the fact that the less developed language versions of Wikipedia have these informations [4, 5].

Wikipedia has a quality grading system for articles, but a specific language version may use its own standards and grades [6]. That means that each language community of Wikipedia can create own standards for the quality evaluation of articles.

Articles in Wikipedia can consist special tables which present shortly important information about subject. This table is usually placed at the top of the right side of the article, and has name „infobox". Information from these infoboxes also used to automatically enrich various public databases (such as DBpedia[4]). Just as in the case of articles, these infoboxes are often created and edited by users in each language separately.

This work try to answer the following main questions:

1. How to determine the quality of a Wikipedia articles?
2. How automatically enrich wiki pages with information (elements of the infoboxes) coming from the counterparts of this Wikipedia article in other languages?

In addition, auxiliary questions were formulated:

1. How to determine quality measures of a Wikipedia article?
2. How automatically evaluate the quality of a Wikipedia articles based on the selected quality measures related to timeliness, validity and completeness?
3. Can the quality of Wikipedia article help to evaluate the quality of infobox contained in it?
4. Is the quality of the infobox in particular language version of article dependent on the demand for the related content?

[2] http://www.alexa.com/siteinfo/wikipedia.org

[3] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[4] http://dbpedia.org

5. How to choose a better quality infobox parameters from different language versions?

2 Quality of Wikipedia articles

In each Wikipedia language editions there is system of grades for articles quality. Practical every language version has special mark for articles are considered to be the best. In English Wikipedia they are called „Featured Articles" (FA), in polish Wikipedia - „Artykuły na Medal". Such best articles should meet the specified quality criteria related to accuracy, neutrality, completeness and style. For example, FA articles content must be written with professional standard, neglects no major facts or details and places the subject in context, consist high-quality reliable sources, have lead section that summarizes the topic and prepares the reader for the detail in the subsequent sections and others [5]. There is also a mark for high-quality decent articles, not have met the criteria for FA - „Good articles" (GA) [6]. English Wikipedia have other marks for lower quality articles: B-class, C-class, Start, Stub. One of the important difference between high-quality grades (FA and GA) and lower ones is evaluation procedure. Articles can get or lose FA or GA grade after discussion and voting by Wikipedia users, which can be carried out within about one month from the date of nomination. In case of lower grades it is enough initiative of an individual user. It should be noted, there is also high-quality grade A-class, which can also be given without special voting procedure. However, A-class articles usually at the same time have FA or GA grade.

Other language versions of Wikipedia may have own grading scheme. For German Wikipedia, which have over 2 million articles, there are only 2 higher-quality grades „Exzellente Artikel" and „Lesenswerte Artikel", which are equivalents for FA in GA in English Wikipedia. Russian Wikipedia have more developed grading scheme with 7 grades, but not all of them are equivalent for English grades. Figure 1 shows the differences between quality grades in particular Wikipedia language: English (EN), German (DE), French (FR), Russian (RU), Polish (PL), Ukrainian (UK), Belarussian (BE).

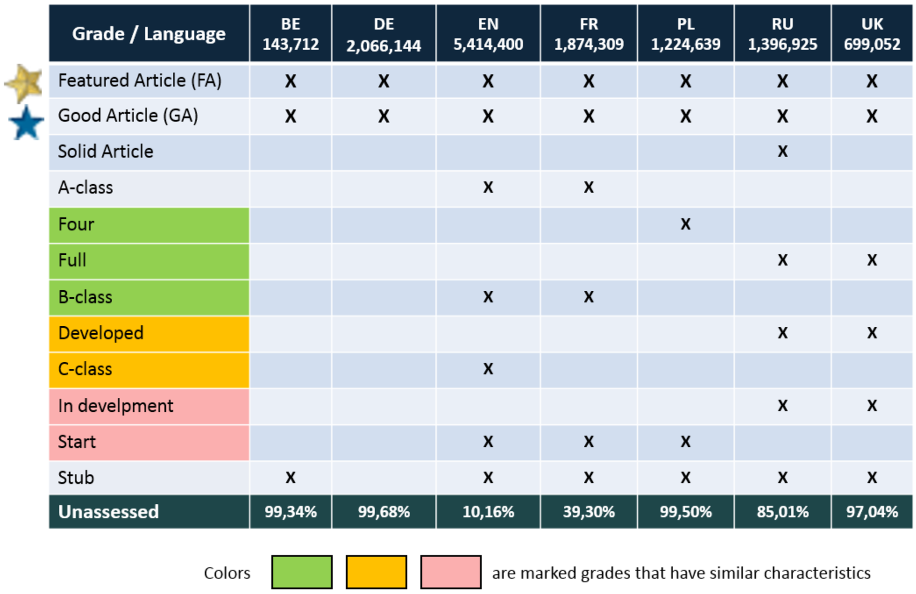
The great challenge is the large number of articles that do not have quality grade. Some language versions (such as BE, DE, PL) have over 99% unassessed articles.

Nowadays, there exist quite a lot of the studies that describes different methods for automatic quality prediction of Wikipedia articles. One of the first researches in this direction proposes to analyse volume of articles content [7]. Such simple metric as word count can help to assess quality of the Wikipedia articles [8]. The best articles use also more references and consits more sections [9, 6]. In addition it can be taken into account special templates which describes quality gaps such as credibility, writing style, structure, and other issues [10].

There are studies that use lungusitics features extracted from the articles texts to analyse the articles quality. Lipka [11] exploit an articles character trigram distribution for the automatic assessment of information quality. Another studies proposed to use

[5] https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

[6] https://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria



Grade / Language	BE 143,712	DE 2,066,144	EN 5,414,400	FR 1,874,309	PL 1,224,639	RU 1,396,925	UK 699,052
Featured Article (FA)	X	X	X	X	X	X	X
Good Article (GA)	X	X	X	X	X	X	X
Solid Article						X	
A-class			X	X			
Four					X		
Full						X	X
B-class			X	X			
Developed						X	X
C-class			X				
In develpment						X	X
Start			X	X	X		
Stub	X		X	X	X	X	X
Unassessed	99,34%	99,68%	10,16%	39,30%	99,50%	85,01%	97,04%

Colors 🟩 🟧 🟥 are marked grades that have similar characteristics

Fig. 1. Quality grading schemes in different language versions of Wikipedia. Source: own calculations.

the number of facts and the factual density as features to identify high quality articles in Wikipedia [12, 13], wherein Fact can have the form of a triplet with two entities and a relationship between them [14].

Assesment of the quality of Wikipedia articles can be based not only on content metrics. Other studies shows how characteristics related to contributors' reputations and edit network, article status, external factual support and other features can help in determining the quality of the article [15, 16].

Many of these studies solve the problem of evaluating articles as a classification task - all grades are divided into two groups: Complete and Incomplete [14, 9, 6, 17, 13]. Complete group consist FA and GA grades. The remaining lower-quality grades are included in Incomplete group. So, various measures that describes the Wikipedia articles are independent variables, quality group - binary dependent variable [9, 6, 18, 17]. Studies have shown that there differences between quality models of particular language versions of Wikipedia using same set of independent variables [6, 18, 17]. Most commonly for these tasks researches used data mining algorithms, and in particular Random Forest, which showed the highest precision in the classification [9, 6, 18, 17].

Using a binary measure for quality assessment of the articles typically give a high precision in classification models (over 95% in different versions of Wikipedia), but this approach has some disadvantages and limitations. If articles belong to the same group (eg Incomplete), then it is not possible to compare their quality between each other.

In some works quality in quality models is also considered a categorial variable [9, 6, 19]. However, the precision in such models is much lower than with the use of a stochastic dependent variable - about 60%. In addition, in this approach, the comparison of the quality of the articles in different languages will be challenging task due to the differences in grade systems in different language versions of Wikipedia.

It may be more useful to use articles quality grade as continuous measure. For example, online service WikiRank[7] used some quality metrics of articles (such as text length, number of images, references, headers, etc.) to calculate the so-called relative quality of the same article in different language versions of Wikipedia on a scale from 0 to 100 [6].

Metrics of wiki pages analysis are extracted in different ways. One of the most important elements of wiki pages are references. Most research usually focuses on counting number of references used in Wikipedia articles and uses that number to create other (derivative) indicators (eg ref./length). Promising is the direction to study the similarity of references between different language versions of a wiki page on a specific topic.

Like in other websites with user-generated content, quality of Wikipedia articles can be determined through an evaluation of the following measures: Completeness, Validity, Timeliness and others [20]. Quality measure - appropriately selected set of metrics. Metric - quantitative value, calculated on the basis of the certain rules. For example completeness consist number of headers, test length, number of images and other.

3 Quality of infoboxes

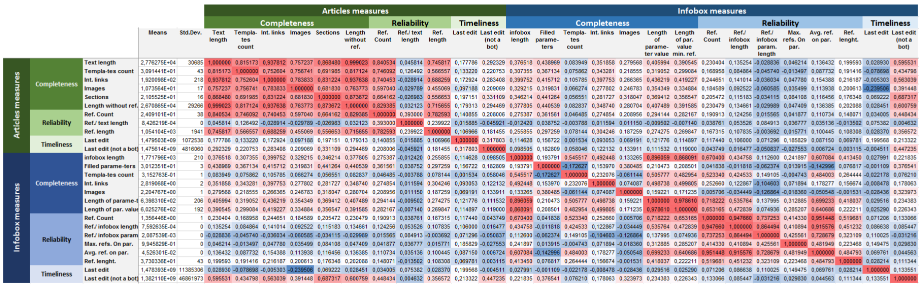
In fact, the infobox is a template that contains list of items "parameter = value". Depending on the topic, the infobox can contain a certain set of possible parameters. Data from the infoboxes can be used not only for receiving main facts about the subject by the reader of Wikipedia, but also to enrich other popular databases such as DBpedia. For this reason, it is particularly important to verify these informations, which are provided by users.

Quality of infoboxes is much less developed topic than articles quality in researches. Existing studies often explore the quality of databases created from infoboxes. Good example of such databases is DBpedia, which additionaly contains many links to other datasets in the LOD cloud such as Freebase, OpenCycand others [21]. Using comprehensive set of generic Data Quality Test Patterns it is possible to reveal a substantial amount of data quality issues [22]. Using special methods it is possible to analyze, the consistency, syntactic validity, conciseness and semantic accuracy of data contained in DBpedia [23]. Analysis of data quality in this semantic knowledge base is also possible without using of ontology [24]. There are also studies related to data fusion from different language versions of DBpedia [25]. However, most of the works does not take into account the various aspects of the quality of the infoboxes and the wiki pages from which the data was extracted.

One of the approaches propose to define relevant metrics and respective scoring functions for specific data quality assessment task[26]. In Linked Open Data (LOD) there are more than 50 quality metrics related to accessibility, intrinsic, trust, dynamicity and contextual dimension coategories [27]. For example, contextual category includes:

[7] `http://wikirank.net`

completeness, amount-of-data, relevancy. Due to the fact that DBpedia is one of the biggest representatives of the LOD, quality of Wikipedia infoboxes can be determined through an evaluation of the following measures: completeness, validity, timeliness and others. The completeness can be connected with the analysis of the filled parameters in infobox [27]. Validity of infoboxes may include an analysis of references, including their similarity between different language versions of Wikipedia [28]. Preliminary analyzes have shown that articles that have been assessed by Wikipedia users for high quality do not always have the best quality information in the infobox in a given language version. Further experiments have shown a correlation between some quality metrics of articles and infoboxes. Figure 2 shows correlation matrix of quality measures of infoboxes and articles in selected language.



					Articles measures											Infobox measures														
					Completeness						Reliability			Timeliness		Completeness							Reliability						Timeliness	
			Means	Std.Dev.	Text length	Templa-tes count	Int. links	Images	Sections	Length without ref.	Ref. Count	Ref./ text length	Ref. length	Last edit	Last edit (not a bot)	infobox length	Filled parame-ters	Templa-tes count	Int. links	Images	Length of parame-ter value	Length of par. value min. ref.	Ref. Count	Ref./ infobox length	Ref./ infobox param. length	Max. refs. On par.	Avg. ref. on par.	Ref. lenght.	Last edit	Last edit (not a bot)
Articles measures	Completeness	Text length	2,776275E+04	30685	1,000000	0,815173	0,937812	0,757237	0,868480	0,999023	0,840534	0,045814	0,745817	0,177786	0,292329	0,376518	0,438969	0,083949	0,351858	0,279568	0,405994	0,390545	0,230404	0,135254	-0,028836	0,046214	0,136432	0,199593	0,028930	0,595531
		Templa-tes count	3,091441E+01	43	0,815173	1,000000	0,752604	0,756741	0,691985	0,817124	0,746092	0,126492	0,566557	0,133220	0,220753	0,307355	0,367134	0,075862	0,343281	0,218555	0,319052	0,299084	0,168958	0,084864	-0,045740	-0,013497	0,087732	0,191416	-0,078698	0,434798
		Int. links	1,920098E+02	218	0,937812	0,752604	1,000000	0,783833	0,831224	0,937638	0,740453	-0,028914	0,688259	0,172924	0,283408	0,399752	0,415712	0,105785	0,397753	0,266365	0,436219	0,419227	0,244651	0,141014	-0,036034	0,047780	0,154388	0,216187	-0,005303	0,563039
		Images	1,073564E+01	17	0,757237	0,756741	0,783833	1,000000	0,681830	0,763773	0,597040	-0,029789	0,455069	0,097188	0,209069	0,329215	0,319831	0,066274	0,277802	0,246783	0,354349	0,334884	0,184589	0,092522	-0,060585	0,035499	0,113938	0,200613	-0,239506	0,391448
		Sections	2,105525E+01	16	0,868480	0,691985	0,831224	0,681830	1,000000	0,873672	0,664162	-0,026983	0,556653	0,197151	0,331109	0,346214	0,441264	0,056551	0,281727	0,318047	0,369412	0,356547	0,205472	0,115183	-0,034115	0,084108	0,116456	0,176348	0,069222	0,687317
		Length without ref.	2,670865E+04	29266	0,999023	0,817124	0,937638	0,763773	0,873672	1,000000	0,829385	0,032123	0,715655	0,179313	0,294469	0,377805	0,440539	0,082837	0,348740	0,280704	0,407489	0,391585	0,230479	0,134661	-0,029989	0,047409	0,136385	0,202088	0,028451	0,600759
	Reliability	Ref. Count	2,409101E+01	38	0,840534	0,746092	0,740453	0,597040	0,664162	0,829385	1,000000	0,393000	0,782593	0,140855	0,208006	0,275387	0,361561	0,046485	0,274854	0,208956	0,294144	0,282167	0,190913	0,124256	0,015565	0,041877	0,110734	0,148071	0,034005	0,448434
		Ref./ text length	8,426219E-04	0	0,045814	0,126492	-0,028914	-0,029789	-0,026983	0,032123	0,393000	1,000000	0,239922	0,015885	-0,045921	-0,012420	0,038752	-0,003788	0,011594	0,011150	-0,009502	-0,007140	0,038761	0,053526	0,084913	0,036777	0,035136	-0,013502	0,075382	0,004632
		Ref. length	1,054104E+03	1941	0,745817	0,566557	0,688259	0,455069	0,556653	0,715655	0,782593	0,239922	1,000000	0,106966	0,181455	0,255855	0,297259	0,078144	0,304246	0,187259	0,274275	0,269847	0,167315	0,107835	-0,003692	0,015771	0,100445	0,108308	0,028370	0,356572
	Timeliness	Last edit	1,479503E+09	1072538	0,177786	0,133220	0,172924	0,097188	0,197151	0,179313	0,140855	0,015885	0,106966	1,000000	0,317803	0,114628	0,156722	0,001534	0,093053	0,069191	0,121776	0,114897	0,117440	0,106000	0,071296	0,185829	0,087150	0,069781	0,199568	0,213322
		Last edit (not a bot)	1,475614E+09	4816060	0,292329	0,220753	0,283408	0,209069	0,331109	0,294469	0,208006	-0,045921	0,181455	0,317803	1,000000	0,098505	0,182809	0,058046	0,122132	0,133911	0,111532	0,119000	0,043749	0,016477	-0,050837	-0,027553	0,006724	0,003115	-0,004511	0,447235
Infobox measures	Completeness	infobox length	1,771796E+03	210	0,376518	0,307355	0,399752	0,329215	0,346214	0,377805	0,275387	-0,012420	0,255855	0,114628	0,098505	1,000000	0,193791	0,545517	0,492448	0,133265	0,896059	0,868091	0,670400	0,434758	0,112600	0,241897	0,607084	0,413450	0,027991	0,221835
		Filled parame-ters	3,012351E+01	3	0,438969	0,367134	0,415712	0,319831	0,441264	0,440539	0,361561	0,038752	0,297259	0,156722	0,182809	0,193791	1,000000	-0,172627	0,153970	0,380485	0,210473	0,208501	0,041838	-0,011818	-0,062374	0,013915	-0,142996	0,076817	-0,001109	0,376541
		Templa-tes count	3,152763E-01	1	0,083949	0,075862	0,105785	0,066274	0,056551	0,082837	0,046485	-0,003788	0,078144	0,001534	0,058046	0,545517	-0,172627	1,000000	0,232076	-0,061144	0,505777	0,482954	0,523340	0,424533	0,149105	-0,004743	0,484003	0,264444	-0,022178	0,076210
		Int. links	2,819068E+00	2	0,351858	0,343281	0,397753	0,277802	0,281727	0,348740	0,274854	0,011594	0,304246	0,093053	0,122132	0,492448	0,153970	0,232076	1,000000	0,074087	0,498738	0,499805	0,252660	0,122867	-0,104603	0,071894	0,178277	0,156674	-0,008478	0,178063
		Images	2,204767E+00	1	0,279568	0,218555	0,266365	0,246783	0,318047	0,280704	0,208956	0,011150	0,187259	0,069191	0,133911	0,133265	0,380485	-0,061144	0,074087	1,000000	0,159221	0,171235	0,005706	-0,034449	-0,126864	-0,018360	-0,050548	-0,001531	-0,028436	0,323973
		Length of parame-t	6,398310E+02	206	0,405994	0,319052	0,436219	0,354349	0,369412	0,407489	0,294144	-0,009502	0,274275	0,121776	0,111532	0,896059	0,210473	0,505777	0,498738	0,159221	1,000000	0,978610	0,718222	0,535764	0,137995	0,312885	0,699233	0,418037	0,029516	0,234383
		Length of par. value	6,025276E+02	192	0,390545	0,299084	0,419227	0,334884	0,356547	0,391585	0,282167	-0,007140	0,269847	0,114897	0,119000	0,868091	0,208501	0,482954	0,499805	0,171235	0,978610	1,000000	0,653165	0,472839	0,074936	0,285207	0,640686	0,222211	0,025290	0,226343
	Reliability	Ref. Count	1,356446E+00	1	0,230404	0,168958	0,244651	0,184589	0,205472	0,230479	0,190913	0,038761	0,167315	0,117440	0,043749	0,670400	0,041838	0,523340	0,252660	0,005706	0,718222	0,653165	1,000000	0,947660	0,737253	0,414330	0,951448	0,519681	0,071206	0,133066
		Ref./ infobox length	7,592635E-04	0	0,135254	0,084864	0,141014	0,092522	0,115183	0,134661	0,124256	0,053526	0,107835	0,106000	0,016477	0,434758	-0,011818	0,424533	0,122867	-0,034449	0,535764	0,472839	0,947660	1,000000	0,864494	0,410894	0,915576	0,451232	0,086638	0,085447
		Ref./ infobox param	2,087539E-03	0	-0,028836	-0,045740	-0,036034	-0,060585	-0,034115	-0,029989	0,015565	0,084913	-0,003692	0,071296	-0,050837	0,112600	-0,062374	0,149105	-0,104603	-0,126864	0,137995	0,074936	0,737253	0,864494	1,000000	0,425561	0,728679	0,323109	0,110025	-0,031216
		Max. refs. On par.	9,945829E-01	0	0,046214	-0,013497	0,047780	0,035499	0,084108	0,047409	0,041877	0,036777	0,015771	0,185829	-0,027553	0,241897	0,013915	-0,004743	0,071894	-0,018360	0,312885	0,285207	0,414330	0,410894	0,425561	1,000000	0,481949	0,223468	0,149475	0,029830
		Avg. ref. on par.	4,526301E-02	0	0,136432	0,087732	0,154388	0,113938	0,116456	0,136385	0,110734	0,035136	0,100445	0,087150	0,006724	0,607084	-0,142996	0,484003	0,178277	-0,050548	0,699233	0,640686	0,951448	0,915576	0,728679	0,481949	1,000000	0,484793	0,069761	0,044563
		Ref. lenght.	3,730336E+01	43	0,199593	0,191416	0,216187	0,200613	0,176348	0,202088	0,148071	-0,013502	0,108308	0,069781	0,003115	0,413450	0,076817	0,264444	0,156674	-0,001531	0,418037	0,222211	0,519681	0,451232	0,323109	0,223468	0,484793	1,000000	0,028214	0,111344
	Timeliness	Last edit	1,478393E+09	11385306	0,028930	-0,078698	-0,005303	-0,239506	0,069222	0,028451	0,034005	0,075382	0,028370	0,199568	-0,004511	0,027991	-0,001109	-0,022178	-0,008478	-0,028436	0,029516	0,025290	0,071206	0,086638	0,110025	0,149475	0,069761	0,028214	1,000000	0,133551
		Last edit (not a bot)	1,382110E+09	46861973	0,595531	0,434798	0,563039	0,391448	0,687317	0,600759	0,448434	0,004632	0,356572	0,213322	0,447235	0,221835	0,376541	0,076210	0,178063	0,323973	0,234383	0,226343	0,133066	0,085447	-0,031216	0,029830	0,044563	0,111344	0,133551	1,000000

Fig. 2. Correlation matrix of articles and infoboxes quality measures in selected language topic. Source: own calculations.

An assessment of the quality of the infoboxes for each topic will allow you to select those language versions where the data has the best quality. In consequence that can help to improve the quality and enrich other Wikipedia language versions.

4 Comparing and enrichment of information in Wikipedia

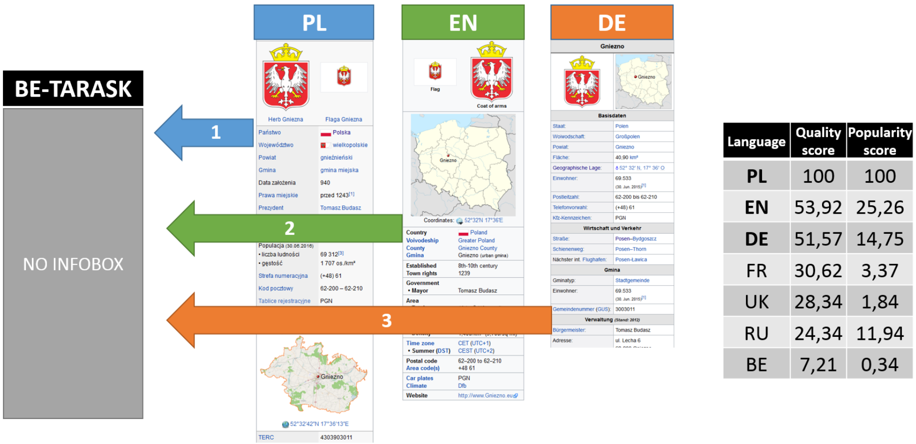
An analysis of current approaches in assessing the quality of information in wiki sites shows that further research to develop new methods are required. The results obtained by using such methods may allow to more accurate assessment of the quality of information in wiki sites in different languages and thus help to improve their quality.

One of the good examples of such researches is Sieve framework, which is used to increase completeness, conciseness and consistency of Linked Oped Data[26]. However, in case of Wikipedia infoboxes it is necessary to take into account additional quality dimensions for more objective analysis [27].

In situations where none of the considered language versions have an infobox, special tools can be used to extract the necessary facts from the text of the article [29] with the best quality. In addition, other approaches for gathering knowledge from semi- and unstructured content can be used [30].

It is also possible that the described subject in Wikipedia has different infoboxes in the examined language versions. This is connected to the fact that Wikipedia communities tend to structure the articles and infoboxes in different ways. In this case cross-language can be exploited cross-language links to represent each infobox with parameters extracted from the corresponding articles [31].

The future researches will address the issue of evaluating the quality of information contained in wiki pages by developing an authoritative method for comparing and enriching information in multilingual wiki services based on their quality analysis.



Language	Quality score	Popularity score
PL	100	100
EN	53,92	25,26
DE	51,57	14,75
FR	30,62	3,37
UK	28,34	1,84
RU	24,34	11,94
BE	7,21	0,34

Fig. 3. Scheme of information enrichment of Wikipedia infobox based on quality and popularity assessment of other language versions on an example of a Gniezno city. Source: own calculations.

The developed method will then be evaluated on the basis of actual data from seven language versions of Wikipedia: English (EN), German (DE), French (FR), Russian (RU), Polish (PL), Ukrainian (UK), Belarusian BE). One of the important metrics, which will be taken to the account is popularity of article, which contain the analyzed infobox. We can expect more relevant and verified information in articles, where the infobox is regularly reviewed by the bigger number of users. Figure 3 shows the general scheme of enrichment of information of infoboxes from the most popular language versions with the best quality score to Belarusian Wikipedia with classical orthography (BE-Tarask). In case when the information from the best language version is insufficient, other parameters will be transferred from other versions with high quality and popularity score. Before transferring values of particular parameters, information will be compared to other language versions, but versions with higher quality will have higher influence (weight) on decision making process on selecting the right value.

Volume for creating new Wikipedia articles based on other language versions can be assessed from the table 1, which presents the numbers of overlapping articles across language versions of Wikipedia. Despite the fact that the English version of Wikipedia is the largest, it can also be enriched by other language versions.

Table 1. Number of overlapping articles across language versions of Wikipedia. Source: own calculations in May, 2017.

	BE	DE	EN	FR	PL	RU	UK
BE	**143 105**	66 673	74 765	67 097	73 012	95 871	79 536
DE	66 673	**2 058 152**	1 045 390	704 217	489 443	492 078	275 593
EN	74 765	1 045 390	**5 405 997**	1 186 508	756 724	723 090	380 539
FR	67 097	704 217	1 186 508	**1 867 289**	550 315	519 651	313 706
PL	73 012	489 443	756 724	550 315	**1 220 272**	429 944	297 109
RU	95 871	492 078	723 090	519 651	429 944	**1 392 818**	401 051
UK	79 536	275 593	380 539	313 706	297 109	401 051	**694 670**

While the table 1 contains information on the coverage of articles between pairs of languages, the figure 4 shows the coverage between the triples of selected Wikipedia language versions using a Venn diagram.

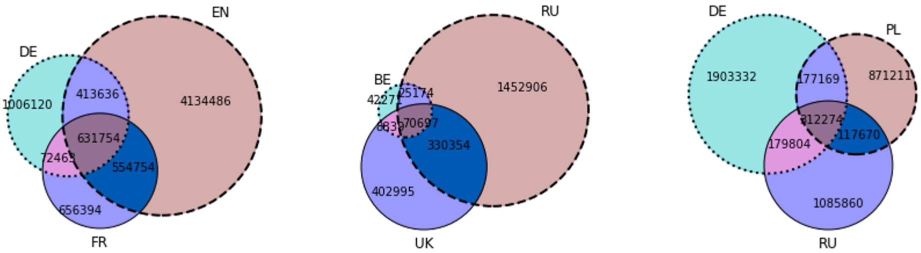


Fig. 4. Overlaps of articles between selected language versions of Wikipedia. Source: own calculations in May, 2017.

5 Discussion and Future Works

The purpose of the proposed research is to develop a method of comparing and enriching information in multilingual wiki services based on their quality analysis on the example of Wikipedia. The proposed method differs from the approaches used so far in several respects. Firstly, in the past work, quality analysis was carried out mainly in one language version - mostly for English Wikipedia. Some metrics that can be considered when building article quality models are dependent on the language in which these articles are written. This includes also linguistic measures. Secondly, there is no study that would automatically assess the quality of the Wikipedia article selected in various language versions. It is related also on differences in evaluation systems used in each language version of Wikipedia. Thirdly, the current works focused mainly on the quality of the whole article, not on the particular elements of it - such as infobox. Preliminary studies show that not always the article with the highest grade among other languages also has infobox with the best data quality in a given language version.

In addition, most research uses a set of metrics to build quality models of Wikipedia articles. The selection of some of these metrics depends on the language, some on the data source, some on the extraction method. An additional factor is the development of wiki technology, which gives you the ability to extract new metrics. This means that extracting and combining multiple metrics based on literature and own experiments may allow a more objective and comprehensive approach to the analysis of the quality of Wikipedia articles in different languages.

Another issue is the constant updating and creating of new wiki pages, on the basis of which Wikipedia article quality models are built. The time factor is important not only because of the varying number of articles, but also because of the continuous changing of the rules of articles assessment by the Wikipedia community in every language version. As a result, articles that have previously been rated with the highest grade for a certain time may no longer meet the criteria and lose their featured status.

Initial experiments showed, that in language sensitive topics, quality of information in infoboxes are high. Typically, such articles are popular in their local language versions. So, measurement of popularity can help in assessment infoboxes quality. It is also connected with the fact that some part of users may notice outdated or incorrect information. If an article is popular in this language - then this corrects can happen faster.

In addition, auxiliary targets are defined that contribute to the main objective:

1. Develop a method for automatically evaluating wiki pages in different languages using appropriate metrics
2. Developing a method for comparing the quality of infobox and wiki page quality
3. Developing a method for identifying high quality infobox elements from wiki pages in different language versions
4. Developing a method of enriching infoboxes between wiki language versions using semantic representation of elements of these infoboxes.
5. Developing a method for creating a new page in a specific language with selected high-quality infobox elements from other language versions of the wiki.

References

1. Hodel-Widmer, T.B., Dittrich, K.R.: Concept and prototype of a collaborative business process environment for document processing. Data & Knowledge Engineering **52**(1) (2005) 61–120
2. Oeberst, A., Cress, U., Back, M., Nestler, S.: Individual versus collaborative information processing: The case of biases in wikipedia. In: Mass Collaboration and Education. Springer (2016) 165–185
3. Staub, T., Hodel, T.: Wikipedia vs. academia: An investigation into the role of the internet in education, with a special focus on wikipedia. Universal Journal of Educational Research **4**(2) (2016) 349–354
4. Callahan, E.S., Herring, S.C.: Cultural bias in wikipedia content on famous persons. Journal of the American society for information science and technology **62**(10) (2011) 1899–1915
5. Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., Gergle, D.: Omnipedia: bridging the wikipedia language gap. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM (2012) 1075–1084

6. 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
7. Stvilia, B., Twidale, M.B., Smith, L.C., Gasser, L.: Assessing information quality of a community-based encyclopedia. Proc. ICIQ (2005) 442–454
8. Blumenstock, J.E.: Size matters: word count as a measure of quality on wikipedia. In: WWW. (2008) 1095–1096
9. Warncke-wang, M., Cosley, D., Riedl, J.: Tell Me More : An Actionable Quality Model for Wikipedia. In: WikiSym 2013. (2013) 1–10
10. Anderka, M.: Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. Phd, Bauhaus-Universitaet Weimar Germany (2013)
11. 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
12. 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
13. Khairova, N., Lewoniewski, W., Węcel, K. In: Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction. Springer International Publishing, Cham (2017) 28–40
14. 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
15. Wu, G., Harrigan, M., Cunningham, P.: Characterizing wikipedia pages using edit network motif profiles. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents, ACM (2011) 45–52
16. Velázquez, C.G., Cagnina, L.C., Errecalde, M.L.: On the feasibility of external factual support as wikipedia's quality metric. Procesamiento del Lenguaje Natural **58** (2017) 93–100
17. 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
18. Lewoniewski, W., Węcel, K., Abramowicz, W.: Analiza porównawcza modeli jakości informacji w narodowych wersjach Wikipedii. In Porębska-Miąc, T., ed.: Systemy Wspomagania Organizacji SWO 2015. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach (2015) 133–154
19. Dang, Q.V., Ignat, C.L.: Quality assessment of wikipedia articles without feature engineering. In: Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on, IEEE (2016) 27–30
20. Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P.: A general multiview framework for assessing the quality of collaboratively created content on web 2.0. Journal of the Association for Information Science and Technology **68**(2) (2017) 286–308
21. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. Semantic Web (Preprint) (2016) 1–53
22. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.: Test-driven evaluation of linked data quality. In: Proceedings of the 23rd international conference on World Wide Web, ACM (2014) 747–758
23. Mihindukulasooriya, N., Rico, M., García-Castro, R., Gómez-Pérez, A.: An analysis of the quality issues of the properties available in the spanish dbpedia. In: Conference of the Spanish Association for Artificial Intelligence, Springer (2015) 198–209

24. Jang, S., Megawati, M., Choi, J., Yi, M.: Semi-automatic quality assessment of linked data without requiring ontology. In: NLP-DBPEDIA@ ISWC. (2015) 45–55
25. Tacchini, E., Schultz, A., Bizer, C.: Experiments with wikipedia cross-language data fusion. In: Workshop on Scripting and Development. (2009)
26. Mendes, P.N., Mühleisen, H., Bizer, C.: Sieve: Linked data quality assessment and fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops. EDBT-ICDT '12, New York, NY, USA, ACM (2012) 116–123
27. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web **7**(1) (2016) 63–93
28. Lewoniewski, W., Węcel, K., Abramowicz, W.: Analysis of references across wikipedia languages. In: Information and Software Technologies: 23nd International Conference, ICIST 2017, Druskininkai, Lithuania, October 12-14, 2017, Proceedings. (in press)
29. Lange, D., Böhm, C., Naumann, F.: Extracting structured information from wikipedia articles to populate infoboxes. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. CIKM '10, New York, NY, USA, ACM (2010) 1661–1664
30. Schmidt, R., Möhring, M., Härting, R.C., Zimmermann, A., Heitmann, J., Blum, F. In: Leveraging Textual Information for Improving Decision-Making in the Business Process Lifecycle. Springer International Publishing, Cham (2015) 563–574
31. Palmero Aprosio, A., Giuliano, C., Lavelli, A. In: Automatic Expansion of DBpedia Exploiting Wikipedia Cross-Language Information. Springer Berlin Heidelberg, Berlin, Heidelberg (2013) 397–411