Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2373–2380
Marseille, 11–16 May 2020

A Topic-Aligned Multilingual Corpus of Wikipedia Articles for Studying Information Asymmetry in Low Resource Languages

Dwaipayan Roy, Sumit Bhatia, Prateek Jain
GESIS - Cologne, IBM Research - Delhi, IIIT - Delhi
{dwaipayan.roy@gesis.org, sumitbhatia@in.ibm.com, prateek16068@iiitd.ac.in}

Abstract

Wikipedia is the largest web-based open encyclopedia covering more than three hundred languages. However, different language editions of Wikipedia differ significantly in terms of their information coverage. We present a systematic comparison of information coverage in English Wikipedia (most exhaustive) and Wikipedias in eight other widely spoken languages (Arabic, German, Hindi, Korean, Portuguese, Russian, Spanish and Turkish). We analyze the content present in the respective Wikipedias in terms of the coverage of topics as well as the depth of coverage of topics included in these Wikipedias. Our analysis quantifies and provides useful insights about the information gap that exists between different language editions of Wikipedia and offers a roadmap for the Information Retrieval (IR) community to bridge this gap.

Keywords: Wikipedia, Knowledge base, Information gap

1. Introduction

Wikipedia is the largest web-based encyclopedia covering more than 49.1 million articles spanning over three hundred languages[1]. The open nature and multi-lingual information present in Wikipedia makes it a valuable resource for various applications such as multi-lingual and cross-lingual information retrieval (Hieber and Riezler, 2015; Paramita et al., 2017; Potthast et al., 2008), question answering systems (Ferrández et al., 2007), entity linking (Zhang et al., 2018), creating parallel corpora for machine translation (Adafre and de Rijke, 2006), query performance prediction (Katz et al., 2014), cluster labeling (Carmel et al., 2009), and computing short-text similarity (Shirakawa et al., 2013).

However, there exists considerable information gap across different language editions of Wikipedia (Filatova, 2009) resulting in an *information divide* between users of different language Wikipedias. The English language Wikipedia is the largest among all the different language editions and contains more than 5.8 million articles. By some estimates, it is as much as 60 times larger than Encyclopedia Britannica, the next-largest English language Encyclopedia[2]. However, the information coverage in other language editions, even for widely spoken languages, is only a fraction of the content in English Wikipedia. For example, the Hindi edition of Wikipedia has just over 130,000 articles as of November 2019 , despite Hindi being the third most spoken language in the world[3]. While the issue of *quality* of Wikipedia content has been studied in detail (Stvilia et al., 2008; Wang and Li, 2019), there has been little work on analyzing the *quantity* of information in different Wikipedia that leads to this information divide and its subsequent impact on various downstream tasks.

Studying the Information Asymmetry is important: Different language editions of Wikipedia differ from each other with respect to the coverage of topics as well as the amount of information about overlapping topics. Even though the English language Wikipedia has the largest community of editors leading to the highest number of topics (articles) covered, many entities that are of interest to a specific country/region/community are often only present in the local language editions (e.g. articles on *Isaltino Morais*, a Portuguese politician and *Friedrich Wilhelm Ristenpart*, a German astronomer are present respectively in Portuguese and German Wikipedia but are missing from the English counterpart). Further, many topics that are majorly of local interest but are also known globally might be present in English as well as local language Wikipedia. In such cases, many facts present in the local language edition might not be present in the English edition, and vice versa. Due to different cultural backgrounds of Wikipedia editors, a bias in the selection of content to include based on local culture, and more intimate knowledge about local facts also leads to a difference in content present in different language editions of Wikipedias (Pfeil et al., 2006; Callahan and Herring, 2011; Hecht and Gergle, 2009).

Despite these differences, availability of open and free content in different language editions of Wikipedia makes it the *go-to* resource for users looking for information in their native languages. For example, while the monthly readership of English Wikipedia has remained mostly stable over the past two years, with slight variations each month, the monthly readership of Hindi Wikipedia has witnessed a steady increase and has more than doubled during the same period[4]. The total number of native Indian language internet users in India was about 234 million in 2016, and is estimated to grow to 536 million users by 2021[5]. Compare this with just 2% coverage of topics in Hindi Wikipedia when compared with English Wikipedia. Thus, there is going to be a tremendous *information scarcity* for this native language user base coming online in the next few years.

Our contributions: We report initial results of our efforts

[1] https://stats.wikimedia.org/EN/TablesWikipediaZZ.html

[2] https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons

[3] https://en.wikipedia.org/wiki/Hindi_Wikipedia

[4] https://stats.wikimedia.org/v2/#/all-projects

[5] https://home.kpmg.com/in/en/home/insights/2017/04/indian-language-internet-users.html

to understand and quantify the information gap present in different language editions of Wikipedia. Using English Wikipedia as the baseline, we present a comparative analysis of information asymmetry in Wikipedias of eight other widely spoken languages (Arabic, German, Spanish, Hindi, Korean, Portuguese, Russian, and Turkish). Our analysis focuses on the different statistical properties as well as the content of overlapping topics present in these Wikipedias editions. Further, we also make available *a topic-aligned multi-lingual corpus of Wikipedia articles* (Section 3.). It is hoped that this corpus can be used to develop and test methods for comparing the information present about a topic in documents written in different languages. Specifically for Wikipedia, knowing missing information in different editions can be used to auto-populate the content using automated content generation tools such as WikiKreator (Banerjee and Mitra, 2015) or can be used by article recommender systems such as Wikipedia SuggestBot (Cosley et al., 2007) that recommends articles needing improvements to editors.

2. Related Work

Most of the efforts towards the construction of multilingual corpora are geared towards the task of statistical machine translation (SMT). The multi-lingual corpora thus produced are aligned at the article level (Kunchukuttan et al., 2018; Inoue et al., 2018) as well as at the level of sentences (Wolk and Marasek, 2015; Schwenk et al., 2019). Unlike these aligned corpora for tasks such as machine translation, our focus is on comparing the information content about a given topic in different language editions of Wikipedia. In the following section, we present a brief overview of works related to the coverage of concepts (or topics) in different Wikipedias, role of Wikipedia editors in leading to an information asymmetry and efforts to reduce these gaps in different Wikipedia editions.

Coverage of concepts in multi-lingual Wikipedia: One way to compare the different editions of Wikipedia is in terms of the coverage of concepts (or entities). Often, it is assumed that a single Wikipedia article corresponds to a unique concept (or topic or entity) and the overlap of these concepts is a useful measure to compare different Wikipedias. Hecht and Gergle (2010) found that about 74% of all the concepts present in Wikipedia are present in only one language edition. Filatova (2009) considered a set of 48 people in DUC 2004 biography generation task and studied how many language Wikipedias contained pages for these people and compared their length. Barrón-Cedeno et al. (2014) showed that language-independent similarity measures such as character n-grams, word-count ratio are effective in measuring the cross-lingual similarity of Wikipedia articles and found no statistically significant difference between language dependent models (translation, monolingual, etc.).

Role of Wikipedia editors: Since Wikipedia is a community-driven effort, quality and quantity of the content present in a specific language edition of Wikipedia are dependent on the community of Wikipedia editors for that language. Due to this, previous studies have also focussed on the behavior of Wikipedia editors and found that only about 15% of Wikipedia editors edit multiple editions of Wikipedia (Hale, 2014). Further, it is observed that users contribute more complex information in their primary language (Park et al., 2015).

Addressing the information asymmetry in different language editions of Wikipedia: A large body of work focused on addressing the information asymmetry between different language editions of Wikipedia. Balaraman et al. (2018), proposed the RECOIN system for measuring completeness of information about an entity using other similar entities as background information. The work by Wulczyn et al. (2016) described an algorithm that finds articles missing in a target language given a source language. The missing articles are ranked based on their expected future page views and are recommended to editors based on their interests. Bao et al. (2012) describe a system to present information about a concept present in multiple language Wikipedias to end-users. Adar et al. (2009) describe an automated system to align *infoboxes* about an entity across multiple language Wikipedias. The system can create new infoboxes or fill in missing information in already existing infoboxes in one language by using the information present in infoboxes about the same entity in different language Wikipedias.

3. Creating a Topic-Aligned Multilingual Collection of Wikipedia Articles

In this section, we describe our strategy of gathering topically-aligned articles from different language editions of Wikipedia. We also make our code available for processing and extracting Wikipedia data and replicating our analysis for other languages not considered in this study[6].

3.1. Source Data

The number of active editions of Wikipedia in different languages is more than 290[7]. For our analysis, we selected *nine* languages, specifically English (**en**), Arabic (**ar**), German (**de**), Spanish (**es**), Hindi (**hi**), Korean (**ko**), Portuguese (**pt**), Russian (**ru**) and Turkish (**tr**)[8]. These specific languages were selected based on their popularity in terms of readership and edit activities in Wikipedia, and further for being among the most widely spoken languages across the globe.

We downloaded the article dumps of all the nine language editions released on 1st November, 2018 as a part of Wikimedia project[9]. Table 1 provides summary statistics of the article dumps for the nine selected languages. The English Wikipedia[10] is chosen as the benchmark (in terms of information coverage) because of its global popularity, highest scope in terms of number of articles, and for having the largest community of active editors and admins making it the most up to date and comprehensive snapshot of

[6] Available from: `https://github.com/dwaipayanroy/wiki-information-extractor`

[7] `https://en.wikipedia.org/wiki/List_of_Wikipedias`

[8] `https://lang.wikipedia.org`, `lang={en,ar,de,es,hi,ko,pt,ru,tr}`.

[9] `https://dumps.wikimedia.org/`

[10] `https://en.wikipedia.org`

Language	# speakers	# Wiki-articles	Ratio	# common articles	diff^{+}	diff^{-}	diff$^{\leftrightarrow}$
English (**en**)	1.268B	5,625,365	1.000	-	-	-	-
German (**de**)	131.6M	2,240,816	0.398	1,077,790 (48.1%)	391745	505201	180844
Russian (**ru**)	258.0M	1,506,914	0.268	764,954 (50.8%)	233798	406839	124317
Spanish (**es**)	537.9M	1,440,098	0.256	940,736 (65.3%)	178668	528360	233708
Portuguese (**pt**)	252.2M	1,010,539	0.180	695863 (68.9%)	61914	444735	189214
Arabic (**ar**)	274.0M	633,291	0.113	426253 (67.3%)	46379	284844	95030
Korean (**ko**)	79.4M	433,010	0.077	265552 (61.3%)	65334	169330	30888
Turkish (**tr**)	85.2M	319,191	0.057	223960 (70.2%)	25962	165584	32414
Hindi (**hi**)	637.3M	130,676	0.023	67642 (51.8%)	4710	55089	7843

Table 1: Overview of different language Wikipedias sorted in the decreasing order of article counts. Ratio represents the fraction of articles when compared with English Wikipedia. The common articles column presents, for each non-English Wikipedia articles, the number of articles having a corresponding English version. The *diff* columns respectively represents the number of common-articles (with English Wikipedia) having greater (diff^{+}), less (diff^{-}) and equal (diff$^{\leftrightarrow}$) document length.

information present in Wikipedia. The other selected languages, (Arabic, German, Hindi, Korean, Portuguese, Russian, Spanish and Turkish), are among the top 10 most spoken languages globally[11]. Further, Wikipedia editions for German, Russian and Spanish cover a fairly large number of topics (each having more than one million articles) and, these languages are relatively *resource-rich* in terms of availability of text processing tools such as parsers, stemmers, and translators. Hindi, Korean and Turkish, on the other hand, are chosen as a representative of low-resource languages. Note that despite having the fourth highest number of native speakers, Hindi Wikipedia suffers from poor topic coverage with only two percent overlapping articles with the English edition (Table 1). Considering English as our benchmark, the third column in Table 1 reports the fraction of topics (articles) present in the respective Wikipedia editions when compared with the English edition[12].

3.2. Tools

The Wikimedia project releases Wikipedia articles in XML formatted text dump. To extract textual information from the dump, we use WikiExtractor, an open source Wikipedia article parser[13]. To count the number of tokens, all the articles are processed using language specific stemmers as well as words are removed following list of stopwords provided by different analyzers as part of Lucene[14].

3.3. Identifying Articles on Same Topic in Different Wikipedia Editions

Wikidata[15] is the community edited knowledge base hosted by the Wikimedia foundation. Each concept (or topic) in Wikidata is assigned a unique identifier (QID) that can be used to extract the facts in the Wikidata knowledge base about the concept. For each unique concept in Wikidata, links to other resources in sister projects of Wikimedia foundation (such as Wikipedia articles, Wikibooks, etc.) are also maintained. Thus, we can use this information to identify articles in different Wikipedia editions about the topic of interest and create a topic-wise parallel collection of articles in different languages by aligning the articles on the same topics (i.e. having the same QIDs).

The fourth column of Table 1 presents the number of topics in each non-English edition having a corresponding article in the English edition. This overlap is obtained using Wikidata information (to be described in Section 3.4. in details). Observe that on an average, only about 60% of the articles in non-English editions (particularly Arabic, Korean, Portuguese, Spanish and Turkish) have a corresponding article in the English edition despite the much higher coverage of topics in English Wikipedia. Surprisingly, topics having an article both in German and English Wikipedia editions is only 48% despite the fact that German edition has the largest number of articles after English (among the editions considered in this study) Wikipedia. This reflects the pre-eminence of German Wikipedia in terms of article coverage highlighting the latent gap of information coverage between the top editions of Wikipedia. Similar remark can also be drawn for Russian and Hindi Wikipedia in which, about 50% articles do not have a corresponding article in English. This highlights an interesting finding: despite the much higher coverage of topics in English Wikipedia (the largest edition among all the languages), all the non-English editions studied contain a significant number of articles about topics that are not present in the English edition. Such a stark difference hints at the orthogonality of local preference about certain topics that are either missed or not considered important to be added to the English Wikipedia.

Among the topics that have articles in both non-English and English editions, there can be a significant disparity in the coverage of information for the topic in different language versions. For a coarse assessment of this information variation, the difference in article length (for the same article in two different editions) is presented in the last three columns of Table 1. The number of common articles having a larger coverage in non-English editions than its English counterpart is presented by diff^{+} where as diff^{-} shows the count of articles having a comparatively broad coverage in the English edition. The diff$^{\leftrightarrow}$ column in Table 1 presents

[11] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[12] This is a simplifying assumption to compare topic coverage across different Wikipedias.

[13] https://github.com/attardi/wikiextractor

[14] https://lucene.apache.org/core/

[15] https://www.wikidata.org

the number of articles having similar coverage in both the non-English and corresponding English editions. Being the largest and widely covered editions among all, a comparatively greater number of articles have been seen to contain higher information coverage in English Wikipedia (column diff^{-}). Notably the comprehensiveness for a significant number of articles, especially in German, Russian, Spanish, Korean and Portuguese, are higher than the corresponding English counterparts (presented in column diff^{+}).

3.4. Topic Selection for the Dataset

For constructing the dataset with topic-wise parallel articles, we sampled a set of 2000 topics from each of the non-English Wikipedia editions (presented in Table 1) with the constraint that the topic must have the corresponding article in the English edition as well[16]. Note that some topics may have more information content in English language while for some topics, coverage in English Wikipedia may be less compared to other editions. Therefore, to ensure that our sampled topics capture all such possible variations, we sampled topics from the following three classes depending on the article length difference in the English and non-English editions of Wikipedia: (i) articles having higher information coverage in non-English than corresponding English (diff^{+}), (ii) articles having lower information coverage in non-English than corresponding English (diff^{-}), and (iii) articles having similar information coverage in both non-English and English editions (diff$^{\leftrightarrow}$). To have a diversified dataset with potential information gap, we concentrated primarily on the articles having noticeable length difference. Specifically, we selected most articles from category (i) and (ii) (approximately 40% from each). Figure 1 presents the range of article length difference of articles in the constructed dataset. The selected articles along with their coarse statistics can be found in the project directory[17].

The resulting dataset of the topically-aligned articles in different languages would enable a wide range of text and natural language processing related tasks such as machine translation, cross-lingual information retrieval, change detection, content analysis and comparison, etc. In next Section, we discuss some characteristics of the dataset and analysis of information asymmetry in Wikipedia using the developed dataset. The insights derived from the analysis can help in developing methods for identifying missing information in different Wikipedia editions, and recommending topics and content to Wikipedia editors to be included in respective Wikipedia editions.

4. Studying Information Asymmetry in Different Wikipedias

Different language editions of Wikipedia differ considerably in terms of coverage of topics as well as information contained in overlapping topics. The reason for this asymmetric information is rooted in the fact that different sets of editors generate the content in different language editions. Wikipedia editions that have a higher number of active editors generally have more up to date and recent information. Further, only a small fraction of Wikipedia editors edit multiple editions of Wikipedia (Hale, 2014), and multi-lingual users make most of their contributions in their primary language (Park et al., 2015). These different editors come with their own biases (Pfeil et al., 2006; Callahan and Herring, 2011) and different knowledge (Hecht and Gergle, 2009) about the topics they are editing leading to very different content about the same topic in different editions. For example, consider the Wikipedia pages about Rabindranath Tagore (1913 Nobel laureate in literature from the state of Bengal, India) in English and Bengali (native language in Bengal). Although English is the dominating edition in Wikipedia with a significantly larger number of editors, some of the essential details about Tagore's family and early life are not present in English but are included in the Bengali page.

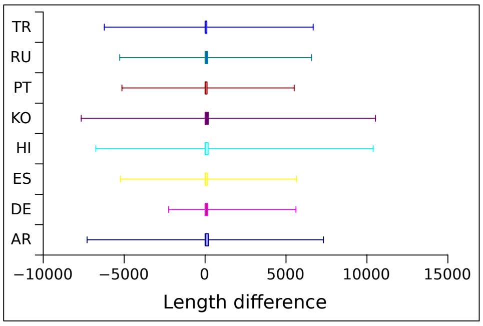

Figure 1: Variation in length difference of the selected articles where horizontal axis indicates the difference in length of the non-English and corresponding English articles.

Given the scale of Wikipedia, manual comparison of content present in different language Wikipedias is not feasible. Therefore, in order to study the differences across different Wikipedias, we resort to multiple indirect ways of comparing the content in different languages that can provide an estimate of the information gap that exists between different Wikipedia editions. Table 1 provides a very high level view of the differences that exist between different Wikipedias in terms of the number of articles present in them. For a fine-grained analysis, we use the parallel corpus constructed (discussed in Section 3.4.) to study the differences that exist between articles about same topics in different Wikipedia editions. We first report the results of our analysis based on statistical comparisons of different properties of the non-English Wikipedia editions using English as the baseline version (Section 4.1.). Next, we describe the results of comparing content in English Wikipedia and content in non-English editions (translated to English) for the articles in parallel corpus (Section 4.2.).

4.1. Statistical Comparison

4.1.1. Article Length

A general advice given to the Wikipedia editors is to be precise and less redundant while writing an article[18]. As-

[16] Similar dataset with recent Wikipedia dumps can also be constructed for other languages by using our code.

[17] http://tiny.cc/wikipedia-parallel-corpus

[18] https://en.wikipedia.org/wiki/Wikipedia:Writing_better_articles

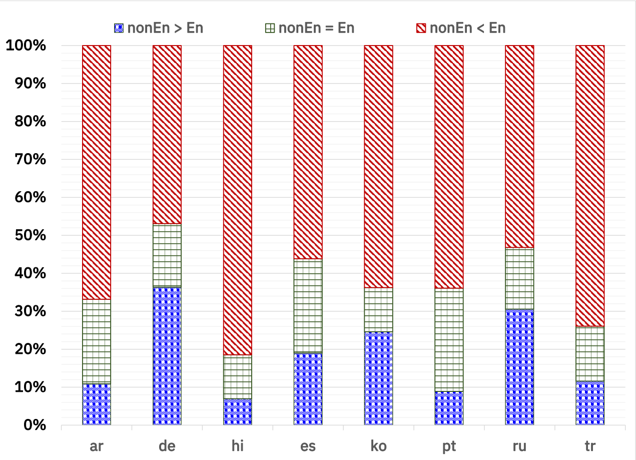

Figure 2: Per-article difference in article length for each of the language pairs.

Language pair	Title	Length Diff.
ar-en	Harem Messaăd	−4908 5702
de-en	Račak massacre List of Supernatural characters	−36983 48532
es-en	Mahatma Gandhi Rafael Nadal	−9889 12447
hi-en	World War I Drought	−13598 5277
ko-en	Oil sands Cronus	−10238 9402
pt-en	Mahatma Gandhi Pregnancy	−9663 6835
ru-en	Laurence Olivier The Grand Inquisitor	−6856 5271
tr-en	Bahrain Iğdır	−7071 4146

Table 2: Some parallel articles in both **en** and **non-en** edition with extensive difference in information coverage. Last column in each row indicates difference in article length between the **non-en** and **en** editions.

suming that practice, the basic unit to measure the containing information in an article can be approximated to be proportional to the size.[19] Approximating the amount of information with the raw count of tokens in an article can be rudimentary but reflective of the amount of information. In Figure 2, the count of per-article (stopword removed and stemmed) length difference with English Wikipedia for each of the selected non-English editions are presented. From the figure, it is evidently seen that the length of English articles are in general longer (shown with red shades for each languages) than the corresponding non-English counterparts. This is as per our expectation as English Wikipedia has the highest coverage in terms for information. Notably, we can also observe from the blue shades of Figure 2 that there are a number of documents in non-English editions having larger content than their English counterpart. Specifically, for German (**de**) and Russian (**ru**), more than 30% overlapping articles have greater coverage in terms of information than their corresponding English (en) edition. The extreme cases for some of the articles where the difference in article length in non-English and English Wikipedia is maximum, are presented in Table 2. To realise the persistent gap in information between a pair of articles on the same topic, let us consider the articles for *Oil sands* in English and Korean Wikipedia[20]. As presented in Table 2, there is an extensive difference in the length of the respective articles. Besides that, the news regarding *KNOC accquring stake of Canadian Black Gold Block* was missing in the English article which was covered in its Korean counterpart. In another example, consider the pair of article on the devastating fire that broke out at *Notre-Dame de Paris* in 2019[21]. Though it was an eminent structural assets and was a famous place of attraction for people all around the globe, the information coverage in the French edition contains significantly detailed report on the incidence and the present improvements regarding reconstruction. Of course the local reports (in French) regarding the incidence have enriched the information contained in the French edition. Specifically, a *conspiracy theory* has been around in the local news suspecting a connection with the fire at the Al-Aqsa Mosque in Jerusalem that took place around the same time. In connection with the conspiracy theory, a former member of *National Front (France)* tried to set fire in a mosque in Bayonne, France.[22] Although relevant to the topic of the article, these information are only referenced in the local language edition of Wikipedia (French) but missing in the overall popular editions (such as English or German).

4.1.2. Coverage of must-have articles

ar	de	es	en	hi	ko	pt	ru	tr
2197.17	3287.42	2503.42	4683.86	739.41	2649.46	2147.29	2916.22	1159.14

Table 3: Average length of the must-have articles in different editions of Wikipedia.

With the growing number of Wikipedia projects in different languages, a list of articles has been presented containing the minimum amount of basic, useful information that every editions must contain[23]. In Table 3, the average article length of these must-have articles are presented for each of the selected editions. As an example must-have article, consider the different editions of *Mahatma Gandhi* in different Wikipedia editions under consideration. The length of the articles in the nine of the selected editions (En, Ar, De, Es, Hi, Ko, Pt, Ru and Tr) are respectively

[19] However, the size of an article may depend on the rules followed while constructing a sentences (e.g. use of active or passive voice), and different languages have there own rule. Also, the presence of stopwords can unnecessarily participate in the article length calculation.

[20] Accessible from `https://www.wikidata.org/wiki/Q297322`

[21] `https://www.wikidata.org/wiki/Q63167656`

[22] `https://fr.wikipedia.org/wiki/Incendie_de_Notre-Dame_de_Paris#cite_ref-363`

[23] `https://meta.wikimedia.org/wiki/List_of_articles_every_Wikipedia_should_have`

$2531, 12237, 8375, 2338, 5842, 2875, 2481, 1241,$ and 7067. The corresponding number of references for the same article (Mahatma Gandhi) in each of the selected languages are also quite different (367, 23 16 and 89).

4.2. Comparison of Translated Content

To perform a solidified comparison between a pair of articles, the containing information of the respective articles have to be analyzed in parallel to check for information overlap and mismatch. As the articles in separate editions of Wikipedia are in different languages, to perform human assessments, the assessors need to have knowledge about both the languages, which is costly as well as time consuming. To avoid performing human judgment, a way of estimating the difference is to unify the language by translating one of the articles in the other language (similar to (Filatova, 2009)) and compare the content. However, as the articles in different languages are not aligned in parallel, like the data set used for statistical machine translations (SMT), the content of a pair of articles cannot be compared in the similar way as SMT tasks. Due to the absence of sentence alignment, BLEU (Papineni et al., 2002) cannot be used be used to compare the content after translation. As an alternative, ROUGE (Lin, 2004) is a recall oriented similarity measurement for automatic translation and summarization that computes similarity on the basis of overlap of unigram and bigram, as well as, longest common subsequent based statistics. In case of Wikipedia, as English is the dominating language with significantly larger coverage, it can be used as the reference while comparing the articles. However, there are articles in non-English with less information coverage than the corresponding non-English article. Yet, as the focus is on computing the difference in information between articles, ROUGE can be used to measure the disparity although it will not be able to highlight the article with information abundance or scarcity.

The study reported in Section 4.1. analyses the content of the articles on the basis of the statistical counts. In this section, we explore the fine-grained analyse to see the information gap. To compare the content in a pair of articles in different languages, one needs to have the knowledge about both the languages. For automatic comparison, the articles can be translated to the same reference. As discussed above, ROUGE metric can be used for this study to compare a pair of article after translation. For our analysis, we have used the IBM Watson Language translator[24] and considered English as the reference. That is, the translator is called to translate each non-English articles in English. The overall comparison of the translated text with the actual English article is performed using ROUGE (considering English as reference).

The average ROUGE scores (we only reported overlapping unigram and bigram metric in this report as LCS based metrics are not suitable to apply in this scenario) for all the 2000 articles in each of the languages are presented in Table 4. For each non-English languages, the standard metrics, i.e. precision, recall and F1 score is given. A high ROUGE score indicates that the coverage between a pair of articles, in general, is symmetrical which is seen for the articles in De, Es, Pt and Tr editions. On an average, Hi, Ko and Ar articles are seen to be disproportional with the corresponding English articles. This disproportional behaviour is especially seen predominating for the Hindi Wikipedia.

Lang	Metric	Recall	Precision	F-Score
ar	ROUGE-1	0.0805	0.0756	0.0589
	ROUGE-2	0.0209	0.0188	0.0158
de	ROUGE-1	0.2543	0.1408	0.1441
	ROUGE-2	0.0962	0.0478	0.0508
es	ROUGE-1	0.2298	0.1347	0.1322
	ROUGE-2	0.0842	0.0488	0.0477
hi	ROUGE-1	0.0420	0.0302	0.0237
	ROUGE-2	0.0085	0.0085	0.0065
ko	ROUGE-1	0.0609	0.0949	0.0564
	ROUGE-2	0.0106	0.0146	0.0098
pt	ROUGE-1	0.2201	0.1620	0.1301
	ROUGE-2	0.0840	0.0586	0.0482
ru	ROUGE-1	0.0906	0.0666	0.0613
	ROUGE-2	0.0182	0.0145	0.0128
tr	ROUGE-1	0.2073	0.1811	0.1283
	ROUGE-2	0.0734	0.0598	0.0433

Table 4: ROUGE scores for non-En articles, considering En as reference. Translations have been performed using the Watson's language translator.

5. Conclusion and Future Work

We reported a thorough comparison results of our efforts towards a systematic study of the information gap that exists across different language editions of Wikipedia. Taking the English Wikipedia as the baseline, we compared eight different editions of Wikipedia in terms of different statistical properties and content of overlapping topics. As suspected, we found that despite being the largest Wikipedia and having the largest number of editors, articles in English Wikipedia often miss out on many important details that are present in other Wikipedia editions. Further, we found that almost 50% of the articles present in the non-English Wikipedias studied in this paper did not have a corresponding article in English Wikipedia. Significant differences were found among different Wikipedias even when comparing the content of overlapping articles. Surprisingly, on analyzing articles of topics that have been identified as *must-have* by Wikipedia community, the general trend of the information gap remained intact.

Although the analysis reported here has helped in revealing the information gap that exists in Wikipedia, it does not provide a unified measure to represent the information gap between different Wikipedia editions or to compare two different language versions of an article. An interesting area to pursue would be to define a metric to quantify the differences that exist between different Wikipedia editions. That metric can then be used to identify most information-rich articles about a topic from different editions. Further, in ad-

[24] www.ibm.com/watson/services/language-translator/

dition to comparing the content of two articles in terms of phrase overlap, it would be interesting to quantify the differences in terms of *facts* that are present in different language version of an article. Such a structured metric can help in recommending additions/modification to the editors of the articles in different languages.

6. Bibliographical References

Adafre, S. F. and de Rijke, M. (2006). Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*.

Adar, E., Skinner, M., and Weld, D. S. (2009). Information arbitrage across multi-lingual wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 94–103. ACM.

Balaraman, V., Razniewski, S., and Nutt, W. (2018). Recoin: Relative completeness in wikidata. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 1787–1792. International World Wide Web Conferences Steering Committee.

Banerjee, S. and Mitra, P. (2015). Wikikreator: Improving wikipedia stubs automatically. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 867–877.

Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., and Gergle, D. (2012). Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1075–1084. ACM.

Barrón-Cedeno, A., Paramita, M. L., Clough, P., and Rosso, P. (2014). A comparison of approaches for measuring cross-lingual similarity of wikipedia articles. In *European Conference on Information Retrieval*, pages 424–429. Springer.

Callahan, E. S. and Herring, S. C. (2011). Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915.

Carmel, D., Roitman, H., and Zwerdling, N. (2009). Enhancing cluster labeling using wikipedia. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 139–146, New York, NY, USA. ACM.

Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. (2007). Suggestbot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 32–41. ACM.

Ferrández, S., Toral, A., Ferrández, O., Ferrández, A., and Munoz, R. (2007). Applying wikipedia's multilingual knowledge to cross–lingual question answering. In *International Conference on Application of Natural Language to Information Systems*, pages 352–363. Springer.

Filatova, E. (2009). Directions for exploiting asymmetries in multilingual wikipedia. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*, pages 30–37. Association for Computational Linguistics.

Hale, S. A. (2014). Multilinguals and wikipedia editing. In *Proceedings of the 2014 ACM conference on Web science*, pages 99–108. ACM.

Hecht, B. and Gergle, D. (2009). Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on Communities and technologies*, pages 11–20. ACM.

Hecht, B. and Gergle, D. (2010). The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 291–300. ACM.

Hieber, F. and Riezler, S. (2015). Bag-of-words forced decoding for cross-lingual information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1182.

Inoue, G., Habash, N., Matsumoto, Y., and Aoyama, H. (2018). A parallel corpus of Arabic-Japanese news articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Katz, G., Shtock, A., Kurland, O., Shapira, B., and Rokach, L. (2014). Wikipedia-based query performance prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '14, pages 1235–1238, New York, NY, USA. ACM.

Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2018). The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Paramita, M. L., Clough, P. D., and Gaizauskas, R. J. (2017). Using section headings to compute cross-lingual similarity of wikipedia articles. In *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*, pages 633–639.

Park, S., Kim, S., Hale, S. A., Kim, S., Byun, J., and Oh, A. (2015). Multilingual wikipedia: Editors of primary language contribute to more complex articles. In *Ninth International AAAI Conference on Web and Social Media*.

Pfeil, U., Zaphiris, P., and Ang, C. S. (2006). Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113.

Potthast, M., Stein, B., and Anderka, M. (2008). A wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, pages 522–530.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

Shirakawa, M., Nakayama, K., Hara, T., and Nishio, S. (2013). Probabilistic semantic similarity measurements for noisy short texts using wikipedia entities. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 903–908, New York, NY, USA. ACM.

Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. (2008). Information quality work organization in wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001.

Wang, P. and Li, X. (2019). Assessing the quality of information on wikipedia: A deep-learning approach. *Journal of the Association for Information Science and Technology*, 0(0).

Wolk, K. and Marasek, K. (2015). Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. *CoRR*, abs/1509.08881.

Wulczyn, E., West, R., Zia, L., and Leskovec, J. (2016). Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 975–985. International World Wide Web Conferences Steering Committee.

Zhang, B., Lin, Y., Pan, X., Lu, D., May, J., Knight, K., and Ji, H. (2018). Elisa-edl: A cross-lingual entity extraction, linking and localization system. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 41–45.