Analyzing Wikipedia Article Networks in Different Languages

Ingvar Baranin, Eerik Sven Puudist, Karolin Rips
University of Tartu, Institute of Computer Science

***Abstract*—This article uses various methods from graph theory and network science to analyse the relations between Wikipedia articles in English and Estonian and observe cultural differences and values between language communities. The comparison between the English and Estonian Wikipedias offers a lens into how language and culture influence the selection and emphasis of topics. Through the analysis of article centrality, the most central and relevant topics within both communities are identified. Centrality measures help uncover core topics and focal points, highlighting divergences and convergences in cultural interests. Then link prediction is applied to find logical links between articles both within and across language borders with the aim of revealing potential missing connections and enhancing the overall knowledge network. The findings provide valuable insights into intercultural understanding and the localization of knowledge and help to understand the reasons of differences and similarities between English Wikipedia, which is one of the largest and most widely accessed editions, and Estonian Wikpedia, where articles are written by smaller Estonian-speaking community.**

***Index Terms*—Language-agnostic embeddings, link prediction, network analysis, text representations, wikipedia**

I. Introduction

Wikipedia, as a collaborative online encyclopedia, serves as a rich source of information across various languages. Each language edition of Wikipedia represents the knowledge and perspectives of its respective language community. Our project focuses on English and Estonian editions, which may be different in terms of topic coverage and content in the articles.

The English Wikipedia is one of the largest and most widely accessed editions, encompassing a vast range of topics from a global perspective. On the other hand, the Estonian Wikipedia caters specifically to the Estonian-speaking community, focusing on subjects of particular relevance and interest to that group.

The first aim of this project is to compare a subset of the English and Estonian Wikipedias as a means for observing differences in culture and values between the two language groups. In particular the aim is to uncover the most central articles in both languages to see which topics seem most relevant to the speakers of those languages.

Understanding the cultural differences and values between language groups is essential for fostering intercultural understanding and appreciating diverse perspectives. By comparing Wikipedia articles, which reflect the collective knowledge and interests of a language community, it is possible to gain insights into the topics that are most relevant and significant to each group. This knowledge contributes to bridging cultural gaps and promoting cross-cultural dialogue.

For this purpose, various centrality measures are used to find a good, yet computationally efficient way for determining the most central articles. Centrality measures offer a valuable lens through which we can assess the importance and relevance of articles within a network. The assumption is made that highly central articles are more likely to be cited and connected, indicating their significance in the knowledge structure. Analyzing centrality allows us to identify the core topics and interests within each language community and compare their relative prominence.

However, it is important to recognize that the absence of links between articles is not necessarily indicative of a lack of connection or relevance. Some links may be missing due to the randomness of article creation and editing. To address this, the second aim of this project is to perform link prediction throughout the datasets, testing various methods and parameters in order to achieve the best possible results. For this purpose, we are going to use GraphSAGE model with LASER embeddings, which are explained in more detail in methodology section. Calculating cosine similarity is used as baseline method.

Link prediction helps uncover hidden relationships and potential collaborations between topics. By leveraging link prediction, we can identify articles that are conceptually related but not explicitly connected, contributing to a more comprehensive and interconnected knowledge network.

By undertaking this project, we aim to gain a deeper understanding of the cultural differences and values expressed through Wikipedia articles in English and Estonian. Uncovering the most central articles and predicting missing links can provide insights into the core interests, historical narratives, and collective values within both language communities. Such knowledge contributes to fostering intercultural understanding, promoting inclusivity, and supporting the localization of knowledge platforms to better serve diverse communities.

We as native Estonian-speakers are interested in discovering differences and similarities between Estonian Wikipedia articles, which we have used a lot in our lives to gain quick information about various topics, and English Wikipedia articles, which are written by much larger language community members. We have noticed that for some topics that are very relevant in Estonia, but which the rest of the world knows nothing about, Estonian Wikipedia includes very detailed information, but in English Wikipedia, there may be only one general sentence about it. Similarly, some articles especially related to different concepts in science, are much more informative in English Wikipedia than in Estonian Wikipedia.

Since ready-made, publicly accessible and comparable networks describing the Estonian and English Wikipedias were not found, we scraped data from those Wikipedias and made two graphs of it, one for Estonian and the other for English.

Then we continued with performing descriptive analysis of our networks, determined the most influential articles in both languages and predicted missing links between articles.

The results of our analysis show that the most influential articles in both the English and Estonian samples are completely different and depend on the centrality measures that were used. At the same time, some centrality measures had highly correlated results, which means that it is reasonable to use centrality measures that need less computational resources for such tasks. In terms of link prediction, our approach with GraphSAGE increased the accuracy significantly compared to our baseline.

II. Related Work

Establishing connections and mapping specialized knowledge across disciplines is not a strictly new problem. For example, by utilizing their proposed method, Schwartz explored relationships between art and science by realizing a proof of concept [1]. Through creating a *knowledge universe* from Wikipedia articles, with article seeds of Pablo Picasso and Albert Einstein, Schwartz identified so-called *knowledge-dealers*: individuals who create notable bridges between the interconnected clusters of various disciplines. A comparison between the individuals identified Schwartz's quantitative work and an earlier, traditional historic research work exploring the lives of Picasso and Einstein, showed that there were 75 common elements - both persons and works - mentioned in the two studies, thus confirming the value of Schwartz's automatic, unsupervised method [1].

Continuing with the theme of connectedness, Sajadi's work went so far as to use the graph structure of Wikipedia for creating semantic vector representations of concepts, with the assumption that a concept can be represented by the neighborhood of the concept's Wikipedia article, including both incoming and outgoing links [2]. Sajadi's work makes use of both Natural Language Processing and the graph structure of Wikipedia to build vector representations, ultimately reporting superior quality of semantic relatedness compared to earlier approaches that utilized only one of these techniques [2], such as word2vec.

Roy, Bhatia and Jain analyzed the content present in English Wikipedia and Wikipedias in eight other widely spoken languages in terms of the coverage of topics as well as the depth of coverage of topics included in these Wikipedias [3]. They 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 [3]. In addition, significant differences were found among different Wikipedias even when comparing the content of overlapping articles [3].

Gabella's work showed that culture influences the architecture of first-link networks for Wikipedias in different languages [4]. He used betweenness centrality to find the most central articles and observed that for European languages, articles like Philosophy and Science are central, whereas Human and Earth dominate for East Asian languages [4].

Wang and Vinel benchmarked several existing graph neural network models on different datasets for link predictions and one of their datasets was based on Wikipedia article references [5]. They also used GraphSAGE model, which we are going to use in our project [5]. Though they did not get their best results with this model, GraphSAGE is especially useful for large graphs that grow over time and new articles are added to Wikipedia every day, so we are still interested in seeing how it performs with our data [6].

III. Dataset

At the time of writing, ready-made, publicly accessible and comparable networks describing the Estonian and English Wikipedias were not found. Due to this, the networks were scraped by our custom script utilizing pywikibot, a Python wrapper for interacting with the MediaWiki API [7]. Additionally, as the gathering of the entire Estonian and English Wikipedias would be encumbering both storage-wise and in terms of analysis, we focused on certain regions of both languages' networks.

This meant that we were compelled to set a certain article as the *seed* for scraping. It has been previously observed that by following the first link in any English Wikipedia article, it will quickly converge at the Philosophy article [8]. As this means the article should capture a wide range of topics in its relatively immediate neighborhood, we set the article as the heuristic starting point to a breadth-first search scraping algorithm. With this algorithm, we visited 100 nodes of the starting seed's neighborhood, while also focusing their subsequent neighbors, resulting in a dataset much bigger than just 100 nodes. For each node, we kept track of its connections, the article's title and contents, and its Estonian/English equivalent article, if it existed. To create content column, we extracted first 500 characters of article's main text, joined sentences with punctuation mark together and added punctuation mark in the end of the extracted content.

We repeated the scraping twice, once for each language, and retrieved two independent graph components, which can be connected by their interlanguage link equivalents. Various metrics of both networks can be see in Table I.

TABLE I
Descriptive analysis results for English and Estonian graph

	English graph	**Estonian graph**
Number of nodes	31817	7488
Number of edges	68744	12301
Average shortest path length	3.535	3.494
Diameter	4	4
Transitivity	0.01226	0.01388
Average clustering coefficient	0.275	0.124
Edge density	0.000136	0.000439
Average degree	4.321	3.286
Total number of triangles	480717	29997
Number of connected components	1	1

If we look at the number of nodes and edges of our graphs, we can see that English graph is more than 4 times bigger than Estonian graph. Since English Wikipedia had more than

6.3 million articles in April 2023 and an average of 600 new articles are added to English Wikipedia every day, we have less than 0.5% of the articles in our network [9]. Estonian Wikipedia is much smaller – as of May 2023, the edition has about 236, 000 articles, so our network contains about 3% of those articles [10]. This means that the coverage of Estonian Wikipedia is higher in our data and therefore our results may be more reliable with Estonian network.

Diameter for both graphs is 4 and an average shortest path length is also close to 4. This indicates that most of the paths between different nodes have length 3 or 4. Average degree is bigger for English graph, which means that nodes in English graph have on average more connections. English graph has also more local clusters and more triangles. Transitivity and edge density are bigger for Estonian graph, so it is more dense.

IV. Methodology

This chapter covers the methodologies used for both graph analysis and link prediction, where the former focuses on network analysis and the latter discusses the Natural Language Processing background and procedure which generates detailed node attributes used for subsequent link prediction.

A. *Determining the most influential articles*

The aim of that part of the experiment is twofold. Firstly, to find In order to see which articles are the most central in both languages, a variety of centrality measures were calculated and their correlation with one another was examined.

In particular, the following centrality measures were used:

1) degree centrality
2) betweenness centrality
3) eigenvector centrality
4) closeness centrality
5) pagerank

B. *Link Prediction*

In order to predict whether two Wikipedia articles are linked, we calculate node embeddings with a GraphSAGE model, which are then passed to a link prediction layer as randomly sampled pairs. Both of these tasks depend on their corresponding StellarGraph implementation [11].

However, simply passing in the graph structure and Wikipedia article ID pairs as input data to the model would considerably limit the model's ability to learn underlying patterns of what connects two articles. While our dataset includes summaries of each article, neural networks cannot directly compute non-linear transformations on text. Therefore, in the interest of improved modelling, we employ Language-Agnostic Sentence Representation (LASER) [12] models to encode the article summaries into 1024-element vectors. LASER models, as the name may suggest, can transform semantically similar texts close together in latent vector space regardless of the language used in the texts. Figure 1 displays an example of this for two texts in three languages, as shown by Google researchers.

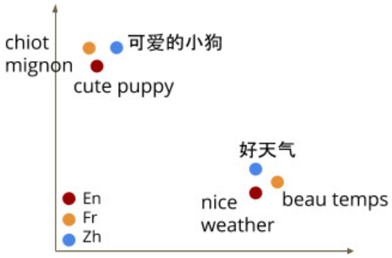

Fig. 1. An example of how LASER may cluster texts ”cute puppy” and ”nice weather” in English, French and Chinese in a dimensionally reduced vector space [13].

There are obvious limitations to the models, such as the inability to encode rare languages due to the lack of training data, but our task is unaffected, as both Estonian and English is covered by this method.

As the original LASER implementation by Facebook provides only bash scripts for encoding, our approach makes use of a LASER Python wrapper [14] which provides a simpler set of encoding functions.

Next, we add these embeddings to their corresponding graph nodes and sample an equal amount of positive (linked articles) and negative (non-linked) edges between Wikipedia articles using GraphSage's EdgeSplitter class for train-test sampling. The node data of these edges are fed to a two-layered GraphSAGE model which combines the embeddings with the homogeneous graph structure to calculate node embeddings. By training a binary classifier on pairs of node embeddings, we ultimately report our results on a test set separated from the training data in the very beginning.

Further, we provide a simple baseline in comparison to the GraphSAGE approach, which helps to understand whether the graph neural network benefits from better performance. The baseline approach uses the text embeddings of the same article page pairs used for the GraphSAGE approach and calculates their cosine similarities. By finding the best threshold for separating the cosine similarity distributions of the linked and non-linked pairs, we find and report the link prediction results when no graph structure is manipulated for classification.

V. Results

In this section, we report the results for both aforementioned explorations.

A. *Comparing different centrality measures*

Figure 2 shows the time complexity of computing various centrality measures for the English Wikipedia sample. The differences in compute speed are extremely prominent ranging from 0.005 seconds for degree centrality and 0.02 seconds for pagerank up to 140 for betweenness centrality.

Figure 3 shows the distributions and correlations of those centrality measures for the same English Wikipedia sample. As seen from the figure, degree, pagerank, and betweenness have a very strong positive correlation. This means that in

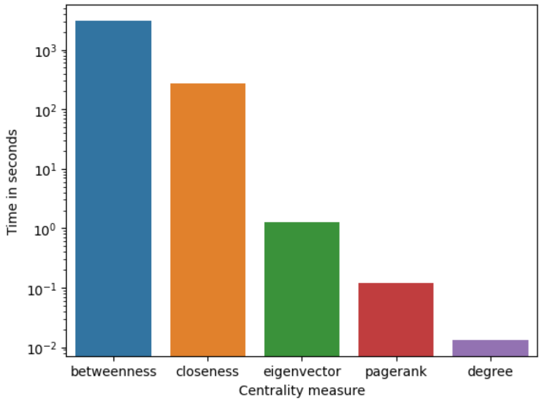

Fig. 2. Comparison of time spent on computing centrality measures for the English Wikipedia sample (notice logartimic scale)

case of a larger Wikipedia sample computing only degree and pagerank while skipping betweenness will bring a significant saving in computational resource consumption without causing much loss in insights gained.

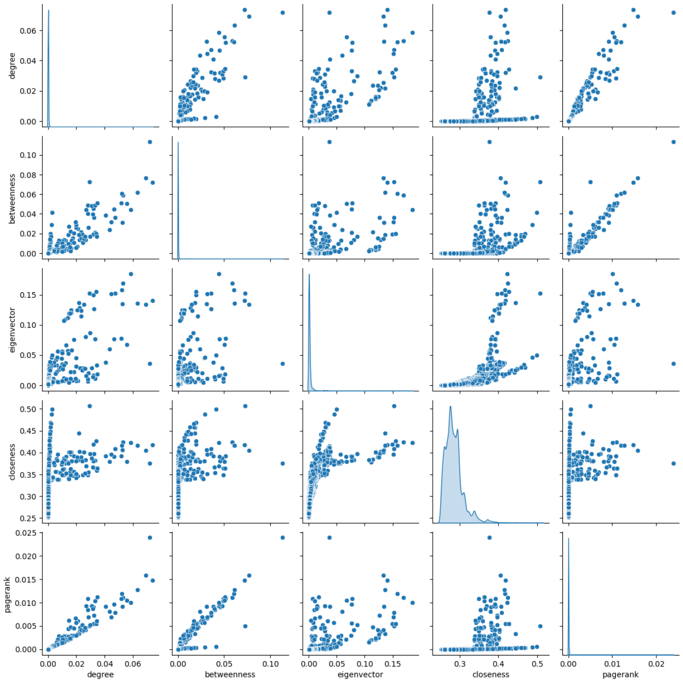

Fig. 3. Correlations and distributions of centrality measures for the English Wikipedia sample

B. Most influential articles

The most influential articles in both the English and Estonian samples are completely different. Also, within the English and Estonian samples themselves the most influential articles depend greatly based on the centrality measures used.

The results in Estonian seem much more plausible, perhaps because due to the smaller overall size of the Estonian Wikipedia a larger percentage of the whole articles were represented in the sample.

It seems that the results are influences by some patterns in Wikipedia which we could not foresee. For example, almost all articles in the top 10 in English start with “Ni” in case on degree centrality, with “Os” in case of eigenvector, and with “Ty” in case of pagerank.

It can be assumed that such anomalies can be attributed to large alphabetically ordered tables where all values are hyperlinks and which thus greatly contribute to the overall number of links that the centrality metrics are calculated by.

The top ranking articles in Estonian do not have such alphabetical anomalies. Here it seems that each centrality metric is picking up some kind of specific topic: articles ranking high based on eigenvector centrality are related on linguistics and phonetics (such as “Kanji”, “Fonology”, “A (letter)”, “Nasal”) whereas articles ranking high based on closeness are related with criminalistics (including “Crime”, “Criminal Investigation”, “Criminal Law”) and articles raking high based on pagerank seem to be associated with philosophy (including “Classical Philosophy”, “Roman Philosophy”, “Ancient-Greek”).

Also, same none-articles pages such as template pages have made their way to the top ranking pages. This would definitely be the case for category pages as well, but we removed them from the dataset since it was obvious that they are the pages with most outbound links.

C. Link Prediction

First, we present the results for the baseline approach. As can be seen on Figure 4, the 6 952 sampled text embedding pairs’ cosine similarity distributions for both 1 (linked) and 0 (non-linked) classes have a huge overlap. By iterating through all thresholds between 0 and 1.0 with a step of 0.01, we were able to calculate the best threshold for separating the distributions, which proved to be 0.8. The accuracy received by this binary separation was 65.4%.

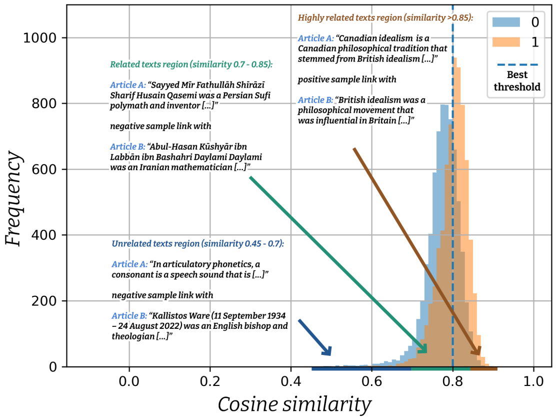

Fig. 4. Cosine similarity distribution between embeddings with examples from each region of similarities. Dashed blue line shows the best threshold for separating the negative and positive sample distributions.

The task clearly becomes difficult for the baseline in the overlap of the distributions, where texts are semantically quite related, but don't clearly exhibit any discernible features of linkedness or non-linkedness. An example of this is shown on Figure 4 where two text embeddings of Persian and Iranian mathematicians are similar, but notably their similarity of 0.85 is wrongly classified as linked articles.

Considering this huge overlap, the GraphSAGE approach which also gathers understanding of text embeddings in the neighborhoods of nodes through node embeddings has a fair chance of improving upon the baseline.

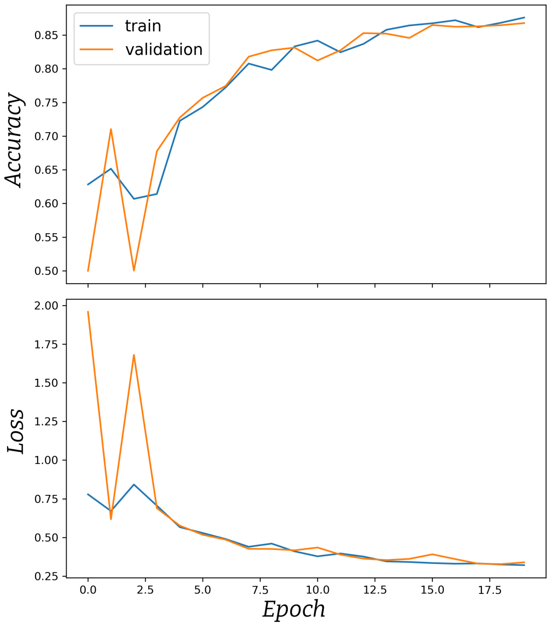

Fig. 5. Model accuracy and loss over 20 epochs.

With the predescribed approach, we train the GraphSAGE and node embedding pair classifier model with the Adam optimizer and learning rate of 0.003 for 20 epochs. The accuracy and loss metrics throughout the epochs can be seen on Figure 5. Using a small split of the training set as a separate validation set, we can see that no overfitting seemed to take place. Of note is the staggering change in validation accuracy after the 3rd epoch, but as this buffed out by itself, it can likely be written off as some initial computational instability or a very unfortunate training batch.

When evaluating the resulting model on the independent testing dataset, we retrieved an accuracy of 86.7%, providing a 21% increase from the baseline. Overall, this could be a good start for a solution that recommends similar Wikipedia articles, because even the remaining 14% of wrongly predicted links between articles could actually point to articles that are similar to the initial article - thus potentially providing users more insights into a topic or person of interest.

VI. Conclusion

This article used various methods from graph theory and network science to analyse the relations between Wikipedia articles in English and Estonian.

The main findings of the article are that in both English and Estonian betweenness, degree, and pagerank centrality measures are highly correlated. Given that the last two are multiple magnitudes faster to compute they should be preferred when analysing a larger Wikipedia sample.

It was also found that various internal structures within Wikipedia such as big alphabetically sorted tables, category, and template pages can influence the accuracy of the centrality measures and data pre-processing is needed to counteract those influences.

In terms of link prediction, we showcased a significant increase in performance when utilizing the neighborhood information aggregation of a given node. We used the GraphSAGE model which improved our simple baseline approach to link prediction by 21% in accuracy, from 65.4% to 86.7%. This model could be used to find other relevant articles to a certain topic or theme past what Wikipedia provides in its "See also" sections.

VII. Github

The dataset as well as the code for scraping, pre-processing, analysis, and figure drawing can be found in this repository.

References

[1] Gustavo A. Schwartz. Complex Networks Reveal Emergent Interdisciplinary Knowledge in Wikipedia. *Humanities and Social Sciences Communications*, 8(1):127, May 2021.

[2] Armin Sajadi. Semantic Analysis using Wikipedia Graph Structure. 2018.

[3] Dwaipayan Roy, Sumit Bhatia, and Prateek Jain. A Topic-Aligned Multilingual Corpus of Wikipedia Articles for Studying Information Asymmetry in Low Resource Languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2373–2380, Marseille, France, may 2020. European Language Resources Association.

[4] Maxime Gabella. Cultural structures of knowledge from wikipedia networks of first links. *IEEE Transactions on Network Science and Engineering*, 6(3):249–252, 2019.

[5] Xing Wang and Alexander Vinel. Benchmarking graph neural networks on link prediction, 2021.

[6] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs, 2018.

[7] Wikimedia. Pywikibot. https://github.com/wikimedia/pywikibot, 2003.

[8] Mark Ibrahim, Christopher M. Danforth, and Peter Sheridan Dodds. Connecting every bit of knowledge: The structure of Wikipedia's First Link Network. *Journal of Computational Science*, 19:21–30, 2017.

[9] Reputation X. The big fat guide to wikipedia statistics. https://blog.reputationx.com/wikipedia-statistics, 2023.

[10] Estonian Wikipedia. Estonian wikipedia — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Estonian_Wikipedia#cite_note-1, 2023.

[11] CSIRO's Data61. StellarGraph Machine Learning Library. https://github.com/stellargraph/stellargraph, 2018.

[12] Facebook Research. Language-Agnostic SEntence Representations (LASER). https://github.com/facebookresearch/LASER, 2019.

[13] Yinfei Yang and Fangxiaoyu Feng. Language-Agnostic BERT Sentence Embedding. https://ai.googleblog.com/2020/08/language-agnostic-bert-sentence.html, 2020.

[14] yannvgn. LASER Multilingual Sentence Embeddings as a pip package. https://github.com/yannvgn/laserembeddings, 2019.