Matching and Linking Entries in Historical Swedish Encyclopedias

Simon Börjesson*, **Erik Ersmark***, **Pierre Nugues**
Lund University
Lund, Sweden
{si7405bo-s, er5612er-s}@student.lu.se, pierre.nugues@cs.lth.se

Abstract

The *Nordisk familjebok* is a Swedish encyclopedia from the 19th and 20th centuries. It was written by a team of experts and aimed to be an intellectual reference, stressing precision and accuracy. This encyclopedia had four main editions remarkable by their size, ranging from 20 to 38 volumes. As a consequence, the *Nordisk familjebok* had a considerable influence in universities, schools, the media, and society overall. As new editions were released, the selection of entries and their content evolved, reflecting intellectual changes in Sweden.

In this paper, we used digitized versions from *Project Runeberg*. We first resegmented the raw text into entries and matched pairs of entries between the first and second editions using semantic sentence embeddings. We then extracted the geographical entries from both editions using a transformer-based classifier and linked them to Wikidata. This enabled us to identify geographic trends and possible shifts between the first and second editions, written between 1876–1899 and 1904–1926, respectively.

Interpreting the results, we observe a small but significant shift in geographic focus away from Europe and towards North America, Africa, Asia, Australia, and northern Scandinavia from the first to the second edition, confirming the influence of the First World War and the rise of new powers. The code and data are available on GitHub at https://github.com/sibbo/nordisk-familjebok.

1 Introduction

Encyclopedias are semi-structured, information-rich bodies of knowledge. In the field of knowledge extraction, their organization into articles with a headword makes them easier to process.

Before the advent of the internet, major encyclopedias like the *Encyclopædia Britannica*, *Brockhaus Enzyklopädie*, and *Nordisk familjebok* regularly released new printed editions for decades or even centuries. Largely written by academics and experts, each edition reflects the knowledge base of the educated class in their respective region at that time. Through digitization efforts, many of these editions are available online.

The *Nordisk familjebok* is widely recognized as the most comprehensive and influential Swedish encyclopedia (Aronsson, 2003; Simonsen, 2016). The encyclopedia was published in four main editions between 1876 and 1993, with over 100 volumes and several hundred thousand articles. Starting in 2003, volunteers at *Project Runeberg*[1] scanned the paper volumes, applied an optical character recognition (OCR) to the images, and proofread a part of the entries.

Linking entries between editions to identify shared, added, and removed articles can indicate changes in the perception of information value or importance due to, e.g., world events or new technologies. One way of looking at this is the geographical spread of entries, i.e., if locations in some countries receive more or less attention over time. Linking entries to a graph database like Wikidata, which has coordinates listed for most entities tied to a location, can help highlight these trends.

The main contributions of our paper are:

1. We scraped and segmented the first and second editions of the *Nordisk familjebok* OCRed by *Project Runeberg*;
2. We classified the segmented entries to identify the locations and cross-references;
3. We matched pairs of entries between the two editions (first and second);
4. We linked entries from both editions to unique Wikidata identifiers;

*Equal contribution

[1] https://runeberg.org/nf/

Proceedings of LaTeCH-CLfL 2025, pages 1–10

5. We provide a brief interpretation of the changes in geographic focus from the first to the second edition.

Our code is available on GitHub: https://github.com/sibbo/nordisk-familjebok.

2 Previous Work

This work addresses three main problems: classifying entries, matching them across editions, and linking each entry to its counterpart in a knowledge graph like Wikidata. We outline relevant techniques and review previous work. Many of them use models trained on English. We also describe models specific to Swedish.

2.1 Categorizing Entries

In this work, we only considered entries describing a location. We extracted these entries using a supervised text categorization technique. Lewis et al. (2004) is an early example of such a technique with a large corpus, where the authors describe the annotation of one million newswires and baseline techniques to classify them.

CLD3[2] is a compact model created for language classification. It uses character n-grams as input to train a two-layer neural network model. Beyond language detection, CLD3 can be applied to other text classification tasks.

The transformer architecture (Vaswani et al., 2017) with the BERT encoder component (Devlin et al., 2019) reported state-of-the-art performances in the GLUE benchmark (Wang et al., 2018) for classification tasks. Through language model pretraining, BERT achieves an impressive understanding of language, enabling it to grasp complex semantic and contextual nuances. It thus decreases the necessary amount of annotated samples to reach high classification scores.

2.2 Matching Entries

Text matching refers to the quantification of the semantic similarity of a pair of documents, here encyclopedia entries. Applications of text matching include information retrieval and question answering. The TF-IDF document vectorization (Spärck Jones, 1972) is a baseline technique for representing documents, and the cosine similarity of two document vectors is a standard measure for evaluating their relatedness.

Dense vector representations of sentences or documents (Cordier, 1965) have proven to be better than sparse ones such as TF-IDF to encapsulate their semantics. Reimers and Gurevych (2019) showed they could train transformer models from pairs of similar sentences and embed them in the form of dense vectors reflecting their semantic proximity.

In our setup, we want to match pairs of corresponding articles between editions, which requires comparing similarity scores of embeddings. In the context of the *Nordisk familjebok*, the brute force method of comparing each article in one edition to all articles in the other quickly becomes unmanageable. With more than 100,000 articles per edition, this results in over 10^{10} comparisons.

Vector databases allow for much faster comparisons through efficient storage and indexing of vectors, employing algorithms like the hierarchical navigable small world algorithm and R-trees (Kukreja et al., 2023). Vector databases can use SBERT models to vectorize the documents or more elaborate algorithms such as those of Xiao et al. (2024), Meng et al. (2024), or Lee et al. (2024).

2.3 Adapting Models to Swedish

KB-BERT (Malmsten et al., 2020) is one of the Swedish BERT models developed at *Kungliga biblioteket* (KB), the National Library of Sweden. It is trained on a corpus of Swedish texts created between 1940-2019, including newspapers, government publications, e-books, social media posts, Swedish Wikipedia, and more. Using a teacher-student model with KB-BERT as the student model, they also created a Swedish sentence transformer, KB-SBERT v2.0 (Rekathati, 2023).

2.4 Linking Entries

Wikidata is a free online knowledge graph containing over 115 million items at the time of this study[3]. Each item has a unique QID and a number of property-value pairs that describe it. For example, Sweden's capital, Stockholm, has the QID `Q1754`, and its properties include `P625`, describing its coordinate location.

A few works have explored the task of linking named entities to Wikidata. Shanaz and Ragel (2021) linked persons mentioned in newspapers, and Nugues (2022) linked location entries from the French dictionary *Petit Larousse illustré* to their

[2] https://github.com/google/cld3

[3] https://www.wikidata.org/wiki/Special:Statistics

corresponding coordinates in Wikidata. Ahlin et al. (2024) undertook a similar task to this study, linking location entries from the second edition of the *Nordisk familjebok* to Wikidata.

3 Preprocessing

Project Runeberg is an online archive of old Scandinavian literature (Aronsson, 2023). This archive provides complete digital facsimiles and OCR texts of the first, second, and fourth editions of the *Nordisk familjebok*, and parts of the third. Volunteers have carried out a manual proofreading on the vast majority of the OCR texts of the first edition, and parts of the second edition, as well as creating a currently incomplete index over the entry headwords on each page.

3.1 Scraping

We scraped the web pages of the first and second editions of the *Nordisk familjebok* on the *Project Runeberg* website, with the exception of the supplements. We parsed the HTML pages so that we could extract the index of entries on each page, extracted the raw OCR text, and finally removed or replaced most HTML tags and uncommon Unicode characters.

3.2 Segmenting

The segmentation of the raw scraped text revealed a complex problem. While the entry headwords in the physical copies of the *Nordisk familjebok* are always in bold characters, there is often no corresponding markup in the digitized text from *Project Runeberg*, probably due to a rudimentary OCR conversion. This is especially true for the second edition, which at the time of this study had undergone less proofreading than the first edition. To deal with this, we devised a three-step approach:

1. **Bold matching**: If the paragraph begins with a bold tag, it is an entry.
2. **Index matching**: Else, if the paragraph does not begin with a bold tag but starts with a headword present in the index, it is an entry.
3. **Entry classification**: Otherwise, utilize a binary classifier model for entry classification.

Following Ahlin et al. (2024), who observed that excessively long texts negatively impacted the performance of their location classifier, we truncated entry texts to a maximum of 200 characters.

Some entries have numbered subentries under the same headword. This is notably the case with entries for noble lineages and royal houses, containing a list of people under the same family name, as for instance the *Leijonhufvud*[4] and *Natt och Dag*[5] families. For sake of simplicity, we did not consider subentries in this paper.

3.2.1 Bold Matching

We applied the rule that a paragraph is an entry if it begins with an HTML bold tag, `<b>`. The headword is chosen as the text between the opening bold tag `<b>` and the closing bold tag `</b>`, removing any trailing punctuation.

3.2.2 Index Matching

The index contains the headwords of all entries on a page. They are manually added by proof-readers, which invariably gives rise to human errors. This, together with OCR errors, makes strict character comparisons of index words and entry texts impractical.

We utilized the Levenshtein distance (Levenshtein, 1966) to match the index words to the raw text. We found that many of these index words were too long for absolute edit distance to fairly represent the similarity of these words. Therefore, we extended the Levenshtein distance metric to be relative to word length and, through manual testing, set a match threshold of 0.15.

With these prerequisites, the method greedily attempts to match the longest index word to a substring of the same character length, starting at the beginning of the paragraph. In the event of a match, the index word is chosen as the entry headword.

3.2.3 Entry Classification

We created an entry classifier from a reimplementation of Google's CLD3 architecture. This provided us a foundation for a general classification model that is well-suited for exploiting small semantic details in the texts.

Paragraphs in the scraped text that were indeed articles often contained distinctive features, such as punctuation and different types of parentheses. Therefore, we determined that a logistic head, instead of a two-layer network, would suffice for entry classification.

To create a training set, we leveraged the structure of the encyclopedias. Given that a paragraph

[4] https://runeberg.org/nfai/0520.html
[5] https://runeberg.org/nfbs/0318.html

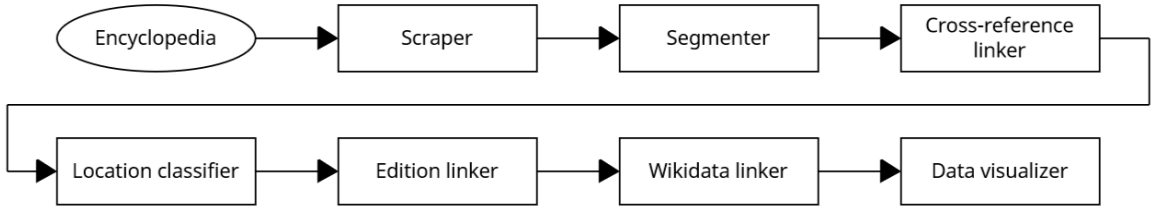


Figure 1: Overview of the pipeline.

beginning with a bold tag is almost certainly a valid entry, we used these paragraphs as ground truth for entries, removing bold tags in the process. Additionally, we used the fact that an encyclopedia is alphabetically ordered to find ground truth for non-entries. For example, in a volume, where all entries begin with the letter *K*, a paragraph starting with any other capital letter is a non-entry.

3.3 Cross-references

Many entries in the *Nordisk familjebok* are cross-references, entries that refer to another entry and provide little to no information on their own, e.g.:

> ***Nervtumör.*** *Se Nervsjukdomar.*
> "***Nerve tumor.*** *See Neurological disorder.*"

For the goals of the study, cross-references provided no value. Therefore, we developed a rule to annotate an entry as a cross-reference if the text was shorter than 60 characters and contained the substring ␣Se "See". We then extracted the word after ␣Se and matched this word to an entry with that exact headword. Some cross-references are longer than 60 characters, but these entries usually provide some information on their own, so we left them as is.

4 Method

Figure 1 shows the processing pipeline consisting of scraping, segmenting, linking cross-references, location classification, edition linking, Wikidata linking, and data visualization.

We described the preprocessing modules, scraping, segmenting and linking cross-references in Section 3. In this section, we describe the rest of the architecture.

4.1 Location Classifier

To determine the location entries, we trained a binary classifier. We manually annotated 200 entries to create a training set of locations and non-locations. We used KB-BERT to tokenize the entry texts and encode them as in Ahlin et al. (2024). We then fitted a logistic regression to the hidden states of the `[CLS]` token.

4.2 Matching Pairs of Entries

We matched the location entries of the first and second editions. We created sentence embeddings of the entries with the KB-SBERT model and used a Qdrant vector database[6] to store them. We then calculated the closest match using cosine similarity. For an entry from the first edition, we finally obtained a list of ranked candidates from the second. We used a greedy strategy and kept the first candidate.

Since always using the closest match leads to many false positives, especially for entries that only exist in one of the editions, we used a cosine similarity threshold value of 0.9 that maximized the F1 score on a manually annotated dataset of 200 entries.

This resulted in a list of matching pairs in the first and second editions of the *Nordisk familjebok* as well as lists of removed and added entries.

4.3 Wikidata Linking

We linked entries marked as locations to Wikidata items and retrieved their geographical coordinates. This consisted of two steps: querying Wikidata and linking texts.

4.3.1 Querying Wikidata

We queried the Wikidata API[7] with the entry headwords and chose the first five results. For each Wikidata item, we retrieved the first 200 characters of the corresponding Swedish Wikipedia article if available[8]. Otherwise, we used the Swedish Wikidata description. We designed our program to prefer Wikipedia texts, assuming that the more

[6] https://qdrant.tech/
[7] https://www.wikidata.org/w/api.php
[8] Using the Wikipedia API: https://sv.wikipedia.org/w/api.php

encyclopedic Wikipedia text would better match the entry texts.

4.3.2 Linking Texts

We encoded the segmented entry text and the retrieved texts of each Wikidata item with the KB-SBERT model, and we compared the encyclopedia entry to each item to find the highest cosine similarity score. Due to the limited search space of five items, we extended the method with a matching threshold, chosen through evaluation on two test sets consisting of 25 random locations from each edition, respectively. We achieved the best F1 scores with a threshold of 0.6.

Lastly, we retrieved the QID and the geographical coordinates using the coordinate location property (`P625`) of the best match that passed the threshold.

5 Results and Evaluation

Table 3 shows the precision and recall scores of all parts of the pipeline where applicable. Most precision and recall scores were acquired by evaluating validation sets of 25, 50 or 100 random entries either in the encyclopedias or in the JSON files. These validation sets should give a general idea of the performance of each part. Nonetheless, their size is relatively small and larger sets would certainly improve their reliability and statistical significance.

5.1 Segmenter

In Table 1, we can see that the second edition has roughly double the number of entries compared to the first one. The number of matches we obtained with the index and classifier strategies is very low in the first edition, since it has been proofread almost completely.

Christensson (2005) estimates the number of entries in the first edition to 103,000. The disparity between this and our 84,534 entries is likely due to not segmenting supplemental volumes.

Ahlin et al. (2024) report the extraction of 130,383 entries when segmenting the second edition, while Simonsen (2016) estimates over 182,000 headwords. Both included supplemental volumes, which we chose to exclude, but like us, they also omitted subentries. We believe the difference from the former is due to using index matching and a binary classifier for entries without bold tags, and the discrepancy from the latter again is mainly due to not segmenting the supplemental volumes.

In combination with the recall and precision scores for segmenting in Table 3, we can be relatively certain that these numbers are good estimates for the total number of entries in the encyclopedias, excluding subentries and supplemental volumes.

5.2 Cross-references

Table 3 shows the performance of linking cross-references to their referenced entry. The method was quite simple, and gave rise to some errors, most notably linking the cross-reference to an incorrect entry with the same headword. For example, in the second edition, *Bajesid* is listed as an alternate spelling of a lineage of sultans in the Ottoman Empire:

> ***Bajesid,*** *turkiska sultaner. Se Bajasid.*
> "***Bajesid,*** *Turkish sultans. See Bajasid.*"

However, when trying to find the referenced entry *Bajasid*, another cross-reference for the city of the same name is matched:

> ***Bajasid,*** *stad. Se Bajaset.*
> "***Bajasid,*** *city. See Bajaset.*"

This is because the first entry with an exact headword match is chosen. For the purpose of removing redundant entries, we believe the performance of our method is satisfactory, but it could probably be improved by using a named entity recognizer.

5.3 Location Classifier

In Table 2, the ratio of locations in both editions is very similar, and the ratio in the second edition is almost identical to that of Ahlin et al. (2024)

Ed.	Entries	Bold	Index	Classifier
1^{st}	84,534	97.7%	2.14%	0.17%
2^{nd}	150,340	76.0%	11.5%	12.5%

Table 1: The total number of entries segmented for both editions, and the proportion of entries segmented using each of the three strategies.

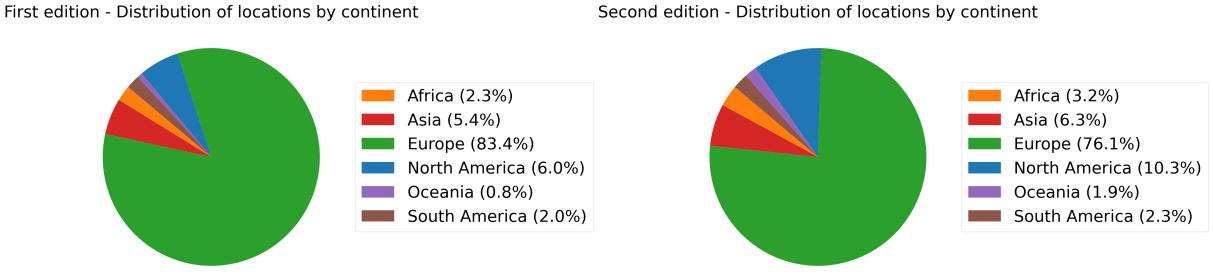
Ed.	Entries	Locations	Proportion
1^{st}	84,534	18,932	22.4%
2^{nd}	150,340	32,378	21.6%

Table 2: The total number of entries segmented for both editions, the number of entries classified as locations, and the corresponding proportions.

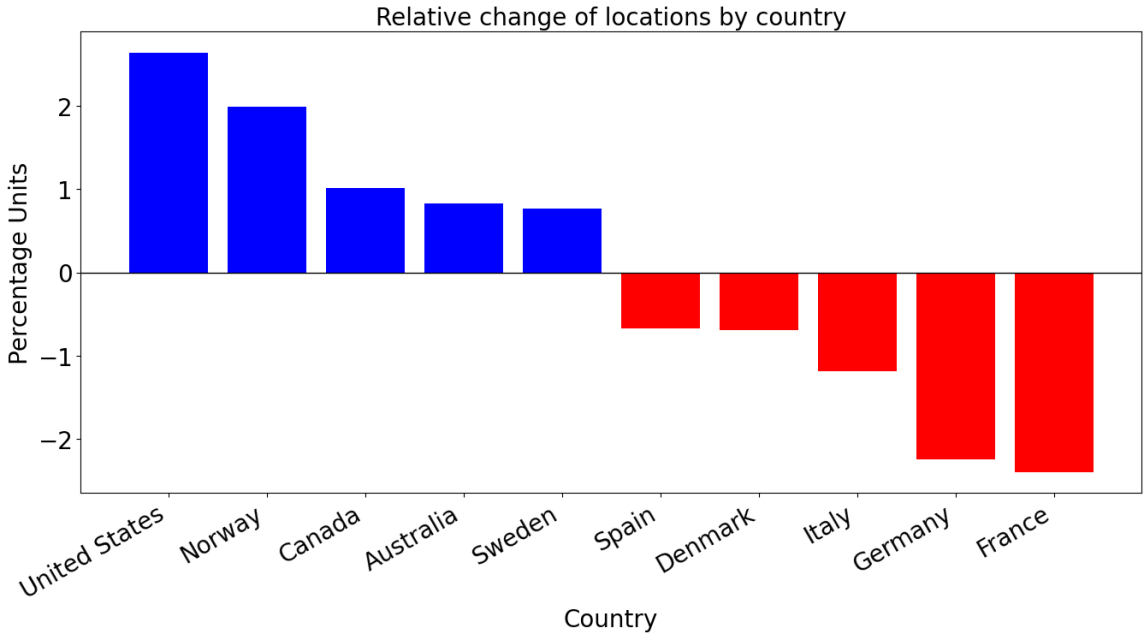
Method	First edition			Second edition		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Segmenter, weighted mean[2]	≈1.0	1.0	1.0	0.99	0.94	0.96
Bold matching[2]	1.0	1.0	1.0	1.0	1.0	1.0
Index matching[2]	0.96	-	-	0.94	-	-
Entry classifier[4]	0.95	0.95	0.95	*	*	*
Cross-references[3]	1.0	0.85	0.92	1.0	0.75	0.86
Location classifier[1]	0.84	0.96	0.90	0.92	0.92	0.92
Entry matching[4]	0.85	0.83	0.83	*	*	*
Baseline: headword match[3]	0.74	0.81	0.76	*	*	*
Wikidata linking						
QID match[1]	0.40	0.52	0.45	0.48	0.16	0.24
Within 25 km[1]	0.76	0.64	0.69	0.84	0.40	0.54

[1] 25 entries used, [2] 50 entries used, [3] 100 entries used, [4] Used respective training/test data, '-' : The metric was not applicable, '*' : The values are the same for both editions.

Table 3: Performance metrics of the pipeline for both editions



(a) Distribution of locations by continent in the first edition. (b) Distribution of locations by continent in the second edition.



(c) The top five countries with the largest percentage unit increase (blue), top five countries with the largest percentage unit decrease (red), in location counts from the first edition to the second edition.

Figure 2: Location-related statistics from both editions.

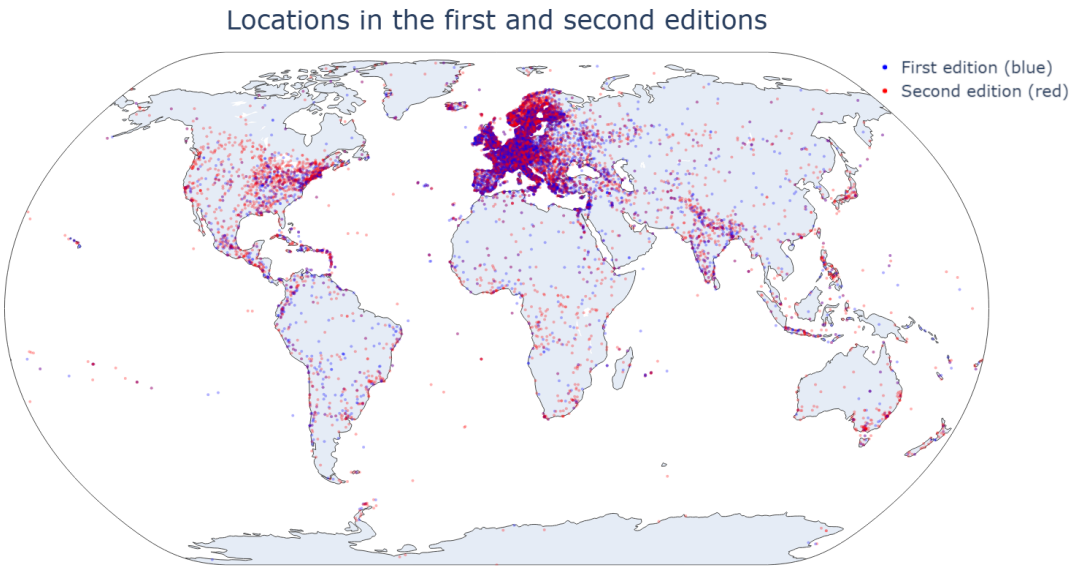


Figure 3: Geographic distribution of locations in both editions.

(21.7%), which is expected since the same method was used.

Table 3 shows the F1 scores of the location classifier for both editions. We can see that they match or surpass 0.9, which is notable considering the KB-BERT model was not fine-tuned for this task.

5.4 Matching Entries

The performance metrics presented in Table 3 demonstrate that our matching approach performs better than the baseline model (headword match) across all metrics, albeit not significantly. We had anticipated a more pronounced performance improvement from the more advanced KB-SBERT model compared to the simple baseline model.

By examining matched sentences, it becomes apparent why certain errors occur. For instance, our method erroneously matched the following two entries, *Åker* and *Åsenhöga*:

> ***Åker.*** *1. Socken i Jönköpings län, Östbo härad. Areal 15,842 har. 1,798 innev. (1892). Å. bildar med...*
> "***Åker.*** *1. Parish in Jönköping county, Östbo hundred. Acreage 15,842 ha. 1,798 res. (1892). Å. forms with...*"

and

> ***Åsenhöga,*** *socken i Jönköpings län, Mo härad. 12,960 har. 1,257 inv. (1921). Å. bildar med...*
> "***Åsenhöga,*** *estate in Jönköping county, Mo hundred. 12,960 ha. 1,257 res. (1921). Å. forms with...*"

These entries exhibit strikingly similar semantic structures, with comparable word sets, order, and article topic. Scenarios like these are understandably difficult, and frequently occur in the corpus.

5.5 Wikidata Linking

When linking an entry to Wikidata, the best cosine similarity match was often not with the correct entity, but with a place or object not very far away, usually within only a few kilometers. A common error was matching a *socken*, an old Swedish term for a church parish, to a nearby city, municipality, or building with the same or a very similar name. For example,

> ***Öved****, socken i Malmöhus län...*
> "***Öved****, parish in Malmöhus county...*"

was linked to *Övedsklosters slott*, a castle within the borders of the parish.

It is difficult to understand why this match yielded the highest cosine similarity score, but such linking errors make little difference on a global scale. Therefore, we created a metric to check if the matched Wikidata entity was within 25 kilometers of the correct coordinates. Although this metric significantly improved performance for both editions, especially the second, the results in Table 3 remain quite poor. Even though only about half of all locations in the encyclopedias were linked within 25 km of their correct coordinates, it seems reasonable to assume that the overall distribution of locations remains roughly the same.

In Figure 2, we see a slight shift in focus away from large European countries like France, Germany, and Italy, towards primarily North America, Australia, Norway, and Sweden. We provide a brief interpretation of this in Section 6.2.

Another source of error stems from the limited search space we set to reduce computation time, which occasionally caused the program to miss the correct Wikidata item.

The search functionality in Wikidata can be unreliable, especially for uncommon entries. For instance, finding the Russian location *Migulinskaya* required using Cyrillic characters. Additionally, Sweden introduced a spelling reform around the turn of the 19th century. Among the changes was replacing the letter *q* with *k* in most words (Pettersson, 2005). For example, *Qvenneberga* in the first edition became *Kvenneberga* in the second one. Such small spelling changes can be crucial: The first term yielded no search results, while the second one resulted in a few hits. Altogether, these quirks can lead to search results missing valid entries, complicating the process of finding specific items.

6 Discussion

6.1 Applications of Entry Matching

The potential applications of matching entries across the editions of the *Nordisk familjebok* are significant, especially in the context of digitization and preserving the relevance of this cultural artifact.

One potential application is the development of a search system based entirely on the editions of the *Nordisk familjebok*. This concept is currently being explored at the Centre for Digital Humanities at Gothenburg University.[9] Such a system could greatly benefit from the inter-edition links developed in this work, enabling comprehensive search results across all editions from a single query.

Another application of our pipeline that could improve the accessibility of historical encyclopedias in the digital age is to extend Wikipedia pages with links to corresponding entries in digital facsimiles of encyclopedias.

6.2 Geographic Focus

Given the rapid globalization since the first edition, we expected a more even geographic distribution in the second edition due to its later publication date. Figures 2a and 2b confirm this hypothesis. The historical events that unfolded during the publication time frame of the editions could illuminate the reasons behind the observed changes.

The First World War involved many countries worldwide, including Canada, Australia, the United States, Japan, and various European colonies in Africa. The involvement of these regions in the war may have influenced Swedish societal discourse, consequently affecting the content of the second edition (Snape, 2018).

Figure 2c shows an increase in the number of locations situated in Norway and northern Sweden. From the late 19th century to the mid-20th century, Norway and northern Sweden underwent significant industrialization in hydroelectric (Thomson, 1938) and timber production (Sundvall, 2023), respectively. Consequently, the population of these regions increased, which may explain these additions in the second edition.

Furthermore, Figures 2c and 3 depict a relative decrease of location mentions for several European countries in the second edition. However, since the second edition contains more locations overall, it does not imply that the absolute number of location mentions has decreased for these countries.

7 Conclusion

In this paper, we compared two editions of a historical Swedish encyclopedia. We described the corpus collection, the segmentation of the raw text input into entries, the categorization of entries, and how we matched pairs of entries between the two editions. We finally reported how we linked geographical entries from both editions to Wikidata.

In the classification and matching tasks, we used transformer models with parameters pre-trained on modern Swedish. A possible improvement is to fine-tune the models on older Swedish texts. We could also explore alternative algorithms for matching entries, such as the Hungarian algorithm (Kuhn, 1955).

This work enabled us to identify shifts between the two editions and a few geographic trends. Most notably, the second edition reflects the evolution of the geographic awareness toward a more diverse global outlook. Beyond the historical events mentioned in Section 6.2, there may be countless societal, cultural, political, and economic factors contributing to these changes. We hope our work will invite further investigation to provide a better un-

[9] https://nordiskfamiljebok.dh.gu.se/

derstanding of the context surrounding them.

Limitations

Our evaluation of headword detection and entry matching is limited and a comprehensive study would include more data. Our validation sets should give a general idea of the performance of each part. Nonetheless, their size is relatively small and larger sets would certainly improve their reliability and statistical significance.

Large language models that we used in this research may generate classification errors or show bias. This bias may come from the corpus used for training the models, mostly contemporary Swedish, while we applied them to the *Nordisk familjebok* that uses a slightly different language.

Ethics Statement

We identified a few potential risks:

1. The *Nordisk familjebok* belongs to book history. It sometimes includes old-fashioned viewpoints and its information is dated.

2. This encyclopedia was written in a different historical context. A few entries may include content that can now be considered offensive. Potential users of our work or of applications based on it must be aware of this context.

Acknowledgments

We would like to thank the anonymous reviewers for their suggestions and comments.

This work was partially supported by *Vetenskapsrådet*, the Swedish Research Council, registration number 2021-04533.

References

Axel Ahlin, Alfred Myrne Blåder, and Pierre Nugues. 2024. Mapping the past: Geographically linking an early 20th century Swedish encyclopedia with Wikidata. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11040–11048, Torino, Italia. ELRA and ICCL.

Lars Aronsson. 2003. Preface to the digital facsimile edition. Last accessed 2024-06-07.

Lars Aronsson. 2023. About Project Runeberg. Last accessed 2024-06-03.

Jakob Christensson. 2005. I encyklopediernas trollkrets: Om Bernhard Meijer och Nordisk familjebok. *Biblis*, 2005(32):32–49.

Brigitte Cordier. 1965. Factor-analysis of correspondences. In *COLING 1965*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Sanjay Kukreja, Tarun Kumar, Vishal Bharate, Amit Purohit, Abhijit Dasgupta, and Debashis Guha. 2023. Vector databases and vector embeddings-review. In *2023 International Workshop on Artificial Intelligence and Image Processing (IWAIIP)*, pages 231–236. IEEE.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. NV-Embed: Improved techniques for training LLMs as generalist embedding models. *Preprint*, arXiv:2405.17428.

Vladimir Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the National Library of Sweden–making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. SFR-Embedding-2: Advanced text embedding with multi-stage training.

Pierre Nugues. 2022. Connecting a French dictionary from the beginning of the 20th century to Wikidata. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2548–2555, Marseille, France. European Language Resources Association.

Gertrud Pettersson. 2005. *Svenska språket under sjuhundra år: En historia om svenskan och dess utforskande*, 2nd edition. Studentlitteratur, Lund.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages

3982–3992, Hong Kong, China. Association for Computational Linguistics.

Faton Rekathati. 2023. The KBLab blog: Swedish sentence transformer 2.0. Last accessed 2024-06-05.

Abdul Lathif Fathima Shanaz and Roshan G. Ragel. 2021. Wikidata based person entity linking in news articles. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 66–70.

Maria Simonsen. 2016. *Den skandinaviske encyklopædi: Udgivelse og udformning af Nordisk familjebok & Salmonsens konversationslexikon*. Centrum för Öresundsstudier (Print), 37. Makadam i samarbete med Centrum för Öresundsstudier vid Lunds universitet, Göteborg ; Stockholm.

Michael Snape. 2018. Anglicanism and interventionism: Bishop Brent, the United States, and the British empire in the First World War. *The Journal of Ecclesiastical History*, 69(2):300–325.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Samuel Sundvall. 2023. Migration and decentralised industrialisation: The development of rural migration in northern Sweden (1850–1950). *Rural History*, pages 1–20.

Claudia Thomson. 1938. Norway's industrialization. *Economic Geography*, 14(4):372–380.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.