Identifying Articles About People in the "Nordisk Familjebok" Encyclopedia

Victor Truong
Lund University
Lund, Sweden
vi2840tr-s@student.lu.se

Abstract

This paper presents an attempt to classify entries from the Swedish "Nordisk Familjebok" encyclopedia into those about a "Person" or "Non-Person". The encyclopedia, first published in 1876, is available online through Project Runeberg. We implemented methods for scraping the encyclopedia and extracting and classifying the entries, utilizing the Swedish KB-BERT model, based on BERT, to create contextual word embeddings. We find that the biggest challenge lies in extracting entries from inaccurate and inconsistent text. Our classifier (logistic regression) however, achieves promising results with relatively minimal training.

1 Introduction

Throughout history there have been a number of attempts to document and preserve human knowledge. Such repositories, which attempted to encapsulate "everything" we "know so far", can be found in various encyclopedias around the world. One such encyclopedia is the Swedish *Nordisk Familjebok*, first published in 1876, which contains a vast number of entries authored and reviewed by various experts in different fields of, amongst others, science and arts, and is generally considered the most extensive encyclopedic work within Swedish literature. The historical significance of encyclopedias cannot be understated, and now there exists numerous efforts of digitization in the interest of preserving these works.

The result of digitization has additional benefits however; in the current information paradigm of the Internet, *search engines* are an essential tool for quickly finding information on any topic. Making historical works digitally available not only contributes to vastly more accessible knowledge, but also allows for their content to be *searchable* and *processable*, making it easier to sort through and organize their information. Relevantly for this paper, a digital format also allows for these immense bodies of text to be ingested for machine learning tasks.

We attempt to build a simple pipeline (Section 3) for extracting entries from two editions of Nordisk Familjebok (the first and fourth editions) and train a classifier for identifying whether an entry in the encyclopedia is about a *Person* or a *Non-Person*. We achieve seemingly promising results, and turn to the validation set in hopes of interpreting our classifier (Section 5.3).

2 Related Work

Our work relies on the digitization efforts of *Project Runeberg*, as well as the Swedish model of *KB-BERT*, which in turn is based on the popular *BERT* that has proven to be state-of-the-art proficient in several natural language processing tasks. In this section, we further, and briefly, underline these previous works that are the foundation for this project.

2.1 Project Runeberg

Project Runeberg was started in 1992 in an effort to specifically digitize Nordic literature (Aronsson and the Lysator Computer Association at Linköping University, 1992). The work to digitize Nordisk Familjebok was started in 2001. As a non-profit organization it relies heavily on external funding and crowd-sourced efforts, where the brunt of the volunteer work lies in the manual proofreading of pages.

The reason for this lies in the digitization process, where a picture of the original page of the encyclopedia is scanned and its image processed by an OCR (optical character recognition) program in order to transcribe the raw text – the OCR itself however is prone to various errors which can, for example, be a result of smudges or unclear text in the original picture. This had major implications

for our process, and details on how we decided to handle anomalous text is described under Section 3.2.

2.2 Nordisk Familjebok

To make clear the terminology of this paper, an example of an *entry* in Nordisk Familjebok is:

> **Dramatik**, den art av diktkonst, som har form av skådespel; ...

Which is translated as:

> **Dramatics**, the kind of poetry that has a form of acting; ...

The word in bold ("**Dramatics**") is the *headword* for the entry itself, and an edition consists of entries for each Swedish letter A-Ö, collected across a number of volumes. Important to note is that some entries lends themselves to longer descriptions that can span several paragraphs and pages. An *entry* in the encyclopedia is therefore a description for a certain *headword*.

2.3 BERT

BERT (*Bidirectional Encoder Representations from Transformers*) is a transformer-based language model developed by Devlin et al. (2019) at Google in 2018. In recent years it has become the go-to model for various natural language processing tasks. Notably, BERT omits the decoder layers present in traditional transformer architecture (Vaswani et al., 2023) which allows for it's defining feature: bi-directionality. In contrast to unidirectional left-to-right or right-to-left language pre-training, BERT can "fuse the left and right context" of a text, allowing it to, in theory, more efficiently learn contextual semantics.

2.4 KB-BERT

We make use of the National Library of Sweden's language model, *KB-BERT*, which is based on the aforementioned BERT architecture. Specifically tailored for the Swedish language, at it's release KB-BERT outperformed all contemporary Swedish language models as well as Google's multilingual M-BERT model on Swedish tasks (Malmsten et al., 2020).

Given the National Library's vast literary resources, the biggest challenges described in its paper was the collection and formatting of pretraining data from both old and more recent Swedish texts.

3 Method

The digitized editions of Nordisk Familjebok are available online at https://runeberg.org/nf/ (2024). Each edition consists of a number of volumes that in turn contains a number of pages, each with its own unique URL: https://runeberg.org/nf{*volume*}/{*page*}.html.

Project Runeberg include three important components for each page in the encyclopedia: 1) An *Index* - a list of the headwords for which the page contains entries for. 2) An image scan of the real page from the original encyclopedia copy. 3) The text transcribed from the image by the OCR (which may or may not have been proofread).

The process from which we go from the raw text of the entire encyclopedia to classifying entries can be summarized into four main parts: *1) scraping*, *2) entry extraction*, *3) embedding* and *4) classification*.

3.1 Scraping

For this paper, we focused on the first and fourth editions, published 1876-1899 and 1951 respectively. For a sense of scale: the first edition contains 20 volumes with a total of ~15 000 pages, while the fourth edition contains 20 volumes and ~10 000 pages in total, that needed to be scraped.

For each page, we downloaded the HTML from which we then extracted the *index* and *OCR text*. The HTML was pruned of almost all HTML elements/tags, *except* for stylization tags denoting **bold** (‘<b>‘) and *italicized* (‘<i>‘) text. The reason for this being that these elements often highlight important information: headwords are often given in bold text for example, and are therefore often surrounded by the ‘<b>‘ HTML tag. All other HTML tags were removed, with only their textual content (if any) preserved.

Separate text files were created for each page where we stored the index of headwords provided by Project Runeberg along with the OCR text.[1]

3.2 Entry Extraction

After scraping, we needed to compile the information of the encyclopedia into a more convenient format, instead of as raw text. Each entry should be represented by an object containing its headword and corresponding description. For our specific purposes, a label denoting whether the entry is about a

[1] Available on GitHub. See Appendix A for further details.

Person or a Non-Person is needed as well. For this we used JSON.

However, this was by far the most challenging part of the process. Recall that there is no guarantee that a page's OCR text has been proofread, nor is the existence of a complete index. At the time of writing[2], about 97.7% of the entire first edition has been marked as proofread, but only about 50% of the index has been completed. The fourth edition has a complete index across all it's volumes, but less than 1% of all it's pages has been proofread.

Complications we encountered during the entry extraction were:

- Most pages' texts are split into paragraphs, but not every paragraph represents a unique entry. Some headwords have entries that can span several paragraphs or even several pages. We could therefore not rely on just differentiating between paragraphs.
- The formatting of headwords is inconsistent. In the ideal case, headwords are highlighted by appearing between bold (‘<b>‘) tags, but this is not guaranteed, which is especially true for the fourth edition. Furthermore, the format of headwords in the index may differ from their format in the actual text.
- OCR errors. A lot of headwords end up being misspelled, or having unnecessary apostrophes nestled between it's characters as a result of smudges in the original image, for example.

Thus, in order to extract entries we settled for some heuristical methods that checks the *first line* of each paragraph for the (potential) existence of a headword, that can signal whether the paragraph is a new entry.

3.2.1 Bold Tags

If a word appears in bold, i.e. ‘<b>*word*<\b>‘, then we assume it's a headword and extract it along with the paragraph as a new entry.

3.2.2 Index Matching

If no bold tags could be found, we start checking words against the index. If one of the first words of the paragraph matches exactly with a headword found in the index; we assume the paragraph to be the headword's corresponding entry.

Some headwords can consist of multiple words. Take for example: "Page, Walter", in which case we'll match against two words in succession. In the same page[3] there is a headword for "Page", which would also give a match for the paragraph containing "Page, Walter":

> Page [péid$], Walter Hines, amerikansk författare och diplomat (1855—1918). P. ägnade sig efter univ.-studier fr. o. m. 1880 åt journalistisk verksamhet.

Note that we ignore any phonetic text, such as "[péid$]". In order to differentiate which headword a potential entry is actually about, we implemented a "candidate system", where each matched headword is added to a list of "candidates". If multiple candidates are found, we *assume* the appropriate headword to be the longest candidate, i.e. the headword with most words in it that matched exactly.

3.2.3 String Similarity

Where neither bold tags or an index match is found, we apply string similarity scoring using *Levenshtein distance*[4] between a paragraph and each headword in the index. Before each scoring between a paragraph and a headword, we normalize the strings of both by removing spacing, apostrophes and punctuation etc... The reason for this being to make the comparison as resistant to OCR errors as possible. As an example, the headword "Dagens nyheter" (*"Daily news"*) becomes "dagensnyheter" during scoring which would match well with a paragraph that, for example, begins with "D'agens n'yhete'r ..." and becomes "dagensnyheter...". We score using the first characters of the paragraph equal the length of the normalized headword.

Consider then the headword for letter 'B', which would match perfectly against any other headword in the same page that starts with 'B' (which is every other headword in that page...). Any headword that is technically a sub-word of another headword would encounter this issue (since the length of the headword decides how many characters we compare from a paragraph). So in the same vein as the index matching, we implemented a candidate

[2](May 2024) Information on the progress of proofreading and index completion available at https://runeberg.org/nf/

[3](Vol. 16, pp. 711-712, Fourth edition) https://runeberg.org/nffp/0454.html

[4]https://en.wikipedia.org/wiki/Levenshtein_distance

system where the longest matching candidate is picked.

We use an initial similarity scoring threshold for candidates that allows for at most one single character error between headword and beginning of paragraph. If no candidate is found that matches this criteria, we simply go for the headword that scored the highest above a lesser but still "reasonable" (i.e. "similar enough") threshold.

3.2.4 Families

Some headwords can have an entirely different format in the index than in the text, in which case neither index matching or similarity scoring will work. One such case that we discovered, which is particularly important to us, is that of people of a certain family.

In Nordisk Familjebok, there exists a fair share of entries about families (usually nobility) followed by entries of notable members of that family. In the fourth edition the headwords for family members are usually formatted like "Hammarskjöld, 2. Carl Gustaf" in the index, while the corresponding paragraph would begin with[5]:

> 2) Carl Gustaf H., den föreg:s brorsons son, ämbetsman (1838—98), 1877 prof, i nationalekonomi och finansrätt i Uppsala.

Luckily, this pattern seemed to at least be consistent for a large number of family entries. The first and fourth editions have differing formatting, but seem at least consistent within their own edition.

We used regex to recognize said patterns in both index and text. Perhaps surprisingly, here we also implemented a candidate system with string similarity scoring, mostly as a precaution. Seemingly, matching the first names and numbering, i.e. "Carl Gustaf" and "2." with "Carl Gustaf" and "2)" from the previous example *should* be enough, but given the unpredictable nature of the errors of our corpus this felt like good practice.

3.3 Embedding

In order to start classifying our entries, we first need to turn them into embeddings via KB-BERT. Before feeding the entries into KB-BERT, we limited the text to $\leq$200 characters. Too much text would end up muddling the semantics of the entry text (Ahlin et al., 2023), and risk important context being diluted. Another benefit of this, is that most Person entries contain their most useful information at the beginning of the entry. For example consider the following entry[6]:

> 2. Galilei, Galileo, italiensk fysiker och astronom, den föregåendes son, föddes i Pisa d. 15 . . .

Which translates to:

> 2. Galilei, Galileo, italian physicist and astronomer, son of the preceding, born in Pisa d. 15 . . .

The name of a person is usually followed by their nationality, profession/title, sometimes relations and birth/death. It was therefore beneficial to limit the context for all Person entries to these properties to let KB-BERT create embeddings out of. Of course, here we make something of an assumption on the meaningfulness of BERT's word representations. The discussion around whether language models actually produce semantically groupable embeddings is a broadly researched topic, which we discuss further under Section 5.3.

3.4 Classification

For the training set: a total of 400 entries[7] were pseudo-randomly picked with an even 200/200 split between manually labelled Person and Non-Person entries. We also included a validation set of 112 entries[7], with a split of 25/87 of Person and Non-Persons. Remember that during our scraping we preserved the stylization tags (Section 3.1), the same applies here.

The training set was built in a way where we attempted to include entries from A-Ö, and also with a good variation of nationalities and professions for Person entries. The goal was for the classifier to lean towards a more name-agnostic approach of classification. Given the etymological nature of names (names can be whatever, and denote people, locations, plants etc...), the classifier should ideally focus on the parts of person entries that denote professions, nationalities, births, deaths; decisive properties of people.

We fitted a binary logistic regression model with the embeddings of the training set and tested the model against our validation set. The results of the

[5](Vol. 4, pp. 807-808, Fourth edition) https://runeberg.org/nffi/0512.html

[6](Vol. 5, pp. 797-798, First edition) https://runeberg.org/nfae/0405.html

[7]Available on GitHub. See Appendix A for further details.

	precision	recall	f1-score	suppo
Non-Person	0.99	0.95	0.97	87
Person	0.86	0.96	0.91	25
accuracy			0.96	112
macro avg	0.92	0.96	0.94	112
weighted avg	0.96	0.96	0.96	112

Table 1: Results of the validation

validation is shown in Table 1. We can see that for the recall: 95% of Non-Person entries were correctly predicted as Non-Person, and 96% of Person entries were correctly predicted as Person.

We then used our model to categorize all the entries from both the first and fourth editions of the encyclopedia.

4 Results

For the entry extraction, we managed to extract 118602 entries from the first edition, and 93777 entries from the fourth, to a total of 212379 entries (Figure 1).

After applying the classifier on the entries, we ended up with 30918 entries from the first edition being classified as Person entries, and 20307 entries of the fourth being classified as Person entries (Figure 2). Approximately 26.07% and 21.65% of the entries of the first and fourth editions respectively consisted of people. In total, 24.12% of all entries were classified as people.

The full results of the entry extraction and classification are available at GitHub, see appendix A for further details on how to reproduce the results.

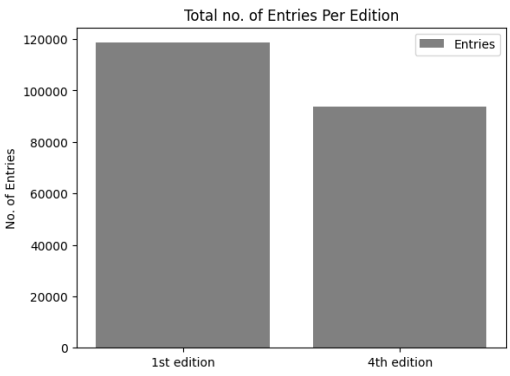

Figure 1: Results of the Entry Extraction 3.2

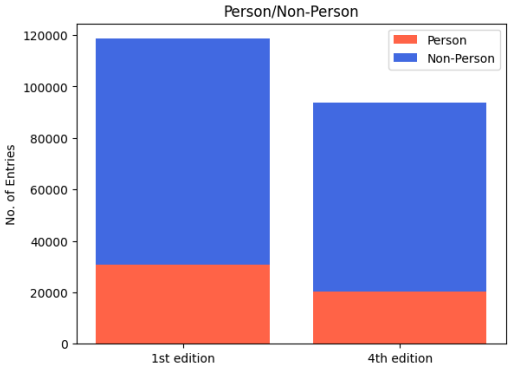

Figure 2: Results of Classification 3.4

5 Discussion and Evaluation

5.1 Entry Extraction

Questions of accuracy arises when evaluating the results for the entry extraction: how many real entries did we manage to extract? Also, how many of the entries that we extracted are even real entries?

Worth noting is that most of our methods for entry extraction relies on the existence of an index for a page. For the fourth edition this seems to work relatively well, given it's completed index. In contrast, recall that merely ~50% of the first edition had been indexed.

However, most of the entries in the first edition seems to have proper bold-tagging; simplifying the extraction of headwords significantly. Out of the 118602 first edition entries, we extracted 114955 entries through bold-tags alone, which was ~97% of the entries. Interestingly, this coincides well with the fact that ~97.7% of the first edition has been proofread. When glancing over a few of the non-proofread pages, we find that many headwords are indeed not in bold. For the fourth edition, only 785 entries were found through bold-tags alone (out of 93777 entries).

An important reminder: the proofreading for Project Runeberg is a public effort, open to any volunteer who wishes to contribute. As such, a page being proofread is still prone to various human errors. A page may have been *marked* as proofread and its spelling errors corrected, but this does not guarantee that the formatting of the page and paragraphs are consistent.

To the best of our knowledge, there does not exist an official count of the number of headwords for the first and fourth editions of Nordisk Familjebok (that we could find). At least for the fourth edition, Project Runeberg has a complete index from which

we might derive a count. Summing all of the unique headwords found in all of its indexes amounts to 94459 entries, assuming this is an accurate count: this means that the entry extraction for the fourth edition had a discrepancy of $94459 - 93777 = 682$ entries that we might have missed. Here we place heavy emphasis on *might* have missed.

5.1.1 Fake Entries

Some paragraphs that did not denote an entry still ended up being extracted as entries. An example of such a fake entry is the following paragraph[8]:

> Falu grufva var länge, såsom statsegendom, ..." (*"Falu mine was for a long time, as state property, ...*)

Which ended up being matched with the headword: "Falun" (a city in central Sweden to which Falu mine belongs to). "Falu grufva" does have it's own entry, but it is first introduced in the preceding page and spans several paragraphs (including the paragraph in question). In the index, "Falu grufva" actually appears as "Falu grufva 1. Stora Kopparbergs grufva", leading to the headword "Falun" being *very* similar to the beginning of the paragraph, and thus the paragraph is extracted as a fake entry.

The extent of these fake entries is largely unknown to us and are difficult to verify. It is not hard to see how cases like these might bloat our results and lead to an overblown amount of entries. However, we reason that cases like this has no significant impact on the classification either (other than possibly slowing the process). They are not included in the training of the classifier, and any additional paragraph about a Non-Person that gets extracted as a fake entry would most likely not include phrases/context that could be confused with the properties of a Person entry. Any additional paragraph in Person entry that gets extracted as a fake entry does not negatively impact the classification of the real Person entry itself (since they are seen as separate entries).

In some cases, real entries will have been extracted correctly albeit with the wrong headword attached. This has no bearing on the results but is nonetheless a slight annoyance.

5.1.2 Missed Entries

Due to unforeseen formatting and errors there exists entries that were not extracted. For example the entry for Napoleon Bonaparte in the first edition is[9]:

> 2. Napoleon B., f. 1769, d. 1821. Se Napoleon I.

While the corresponding headword in the index is "Bonaparte, Napoleon". Since we rely on numbering for members of families in both index and text (Section 3.2.4), the entry ends up being missed.

In the fourth edition, a similar entry for Napoleon exists[10]:

> 4) Napoléon B., den föreg:s bror (1769—1821), se Napoleon I.

For which the headword in the index is "Bonaparte, 4. Napoléon (Napoleon I)". While a numbering exists for our method to detect, the entry *still* ends up being missed due to "... (Napoleon I)"; the string similarity between "Napoléon B., den före" and "Napoléon (Napoleon I)" ends up being too dissimilar. Here we discover that the candidate system + string similarity scoring we added alongside extracting family members (Section 3.2.4) backfires in cases like this.

Other entries are completely undetectable for our algorithms. Recall that we rely on there being clear separate paragraphs for each body of text. If entries (most often "sub-entries") are mentioned inline and not on their own paragraphs, it will not be found. An example of such a case is "1. Anckarsvärd (Cosswa), Mikael"[11]. Whether this is intentional or not is unclear to us.

5.2 KB-BERT and the Swedish Language

It is also important to acknowledge certain language reforms. The first edition of Nordisk Familjebok is more than a century old, it's volumes published during the years 1876-1899. Some linguistic differences are easily found between it and the fourth edition, published in 1951.

Essential context also are the texts used to train the KB-BERT model. KB-BERT is trained only on "modern Swedish", which Malmsten et al. (2020) limits to texts from the 1940's until late 2019.

The Swedish Institute for Language and Folklore divides the Swedish language from 1732-1900

[8](Vol. 4, pp. 1003-1004, First edition) https://runeberg.org/nfad/0508.html

[9](Vol. 2, pp. 847-848, First Edition) https://runeberg.org/nfab/0432.html

[10](Vol. 3, pp. 455-456, Fourth Edition) https://runeberg.org/nffc/0286.html

[11](Vol. 1, pp. 691-692, First edition (https://runeberg.org/nfaa/0691.html)

from the more modern Swedish of 20th century to today (for Language and Folklore", 2020). Since the first edition was published closer to the turn of the century (1876-1899), the changes in language between the editions does not appear to be too drastic. The most notable differences are a result of the spelling reform in 1906, where the spelling of the 'v'- and 't'-sounds were changed (for Language and Folklore", 2020):

- "haf">"hav" and "grefve">"greve"
- "godt">"gott" and "speladt">"spelat"
- For some words, 'å' was exchanged for 'o': "båll">"boll".

Fortunately, this does not seem to have had any noticeable impact on our classification. Most likely because the words for, and spelling of, titles, professions, nationalities etc... has largely remained the same through the various language reforms.

5.3 Classification

Assessing the accuracy of our final classification is difficult. The large volume of data makes manual verification infeasible, but skimming over the entries classified as people makes it *seem* like an overwhelming majority are labelled correctly[12], but this is borderline guesswork. Nevertheless, the validation results are easily available to us and analyzing the entries the model predicted wrongly provides some interesting insights.

From the validation results (Table 1), we see that 5% (4 entries) of the Non-Person entries were predicted as Person while 4% (1 entry) from the Person entries was predicted as Non-Person.

5.3.1 Non-Persons as Person

The four Non-Person entries that were wrongly labelled as people were (roughly translated):

- "Novatianists, schismatic party within the Roman church, in the 2nd century, named after the presbyter Novatianus. During Decius' persecution many had apostatized, ..."
- "Xeros. See previous entry."
- "Yxkull (Üxkull), Swedish noble families. See Meyendorff. "
- "York, the house of Y., the branch of the Plantagenet family, which descended from Edward IH of England's son Edmund (d. 1402), Duke of York"

Our focus on professions/titles and nationalities being closely tied to Person entries constitutes a large majority of the training data. In other words, entries of that type are overrepresented. We speculate that this plays a significant role in the above prediction errors.

For "Novatinaists", "Yxkull" and "York", we have mentions of nationalities and titles: "Roman", "presbyter", "Swedish", "noble", "son" and "Duke". It should be noted that the Swedish word for "noble" is "adel" (noun) or "adlig(a)" (adjective) and holds no other meaning than being a human title for an upper socioeconomic class, unlike in English where "noble" can also be used as an adjective to denote a virtuous person. In addition, certain names ("Novatianus", "Decius", "Edward", "Edmund") might have been implicitly learned and more closely tied to the notion of "Person".

The thought that BERT, and by extension KB-BERT, should closely group together the resulting contextual embeddings of Person entries thanks to their semantic similarity is a natural one to have. But how credible is this intuition? Extensive research already exists regarding the question of interpreting neural networks (Zhang et al., 2021), with the ubiquitous transformer architecture being an especially hot topic (Bibal et al., 2022, Chefer et al., 2021). Here, we are specifically interested in the ontological capabilities of BERT.

There is ground for the notion that certain subspaces and clusters within BERT's vector space do represent certain semantic attributes (Coenen et al., 2019). Through the semantic properties (values) of embeddings, Anelli et al. (2022) shows that meaningful ontological patterns does occur naturally within BERT's vector space, likening it to the structure of a knowledge graph.

A second experiment from the same paper involved binary classification for several ontological categories, one among them being discerning whether an entity was "human" or not. Coincidentally, this was the most successful classification which Anelli et al. (2022) claims is partly due to the larger amount of "human" data, but also because of its *unambiguous* nature. Given our identical goal, this further strengthens the validity of our own seemingly good results.

[12] All classified entries are available on GitHub. See Appendix A

The stranger case amongst the faulty entries is "Xeros", which simply directs to a previous entry. It's absent of any meaningful context, and yet is classified as a Person. A potential explanation for this is the combination of existing entries within the training set, and the previously discussed semantic workings of BERT. In our training set, an entry for "Xenokles", an architect of ancient Greece, and "Xerxes", the name of an ancient Persian monarch, are included as people, for example.

Moreover, the letter 'X' itself is already a natural outlier of the Swedish language. Swedish words beginning with 'X' in both editions of Nordisk Familjebok are mostly, if not solely, borrowed: scientific terms, ancient names of people and cities, names for species of various kinds etc... In the fourth edition, only 41 entries for 'X' exists! Its unique (and slim) position within the Swedish language of the past closely ties it to the notion of older/ancient names of *people* and *geographical locations* (especially ancient Greece, and older versions of Spanish where 'X' was used in place of 'J' for a lot of spellings).

How does this potentially impact KB-BERT? Admittedly this is more on the side of pure speculation, the *exact* contents of KB-BERT's training corpus is inaccessible to us. In it's pre-training, material from digitized newspapers, governmental reports, e-books and e-magazines, social media and Swedish Wikipedia was included, amongst others (Malmsten et al., 2020). The entirety of 2019 Swedish Wikipedia was ingested, which at the very least is a public resource. Exponentially more things exist/has been discovered since the publishing of Nordisk Familjebok. The usage of the letter 'X' has been broadly expanded; new scientific and technological terms, and products of various kinds; cars, movies, games, websites etc... Even then, things whose names explicitly start with the letter 'X' is still rare within the Swedish language.

All this is to say: the dimensionality KB-BERT gives to the word "Xeros", without any additional context, *probably* lends it closer to the "subspace of (ancient) names for humans", if such a semantic region exists. Dalvi et al. (2022)'s findings on BERT's ability to learn latent concepts certainly does not rule out such a specific possibility. We think similar reasoning can be applied to the rest of the faulty entries, whose additional contexts in this case ends up further reinforcing the classifiers wrongful belief that they are about specific people. However, furthering these claims would require a deep dive into the inner geometry of KB-BERT, which unfortunately is outside the scope of this paper.

5.3.2 Person as Non-Person

The singular Person entry that was mislabelled was (translated):

> Xanthippe, Socrates' wife, notorious for her quarrelsomeness, which is why the name Xant(h)ippa (lat. form) usually denotes a quarrelsome wife.

The entry is a clear outlier, given the lack of profession and nationality, so this might not be too surprising. Still, our previous reasoning on "Xeros" posited that it by itself makes the classifier believe it to be the name of a person. Shouldn't the same be true for the unusual name of "Xanthippe"? In this case, we're dealing with a longer entry and thus more context, which might've skewed the embedding and subsequent classification.

Naturally, "wife" in an encyclopedic context is a definitively human title, and we further chalk this error up simply due to a lack of similar entries in the training data. At the same time, the existence of similar entries at all in the encyclopedia is not guaranteed.

The classifier hasn't thus learnt that being "Socrates' wife" should speak for the human-ness of "Xanthippe", or rather: the classifier has instead learnt to specifically classify on entries merely *pertaining* to a person of a nationality, profession or title, instead of the human element itself. This would also provide an additional reasonable explanation for the faulty entries discussed in Section 5.3.1.

5.3.3 Fictional Characters

An important distinction that we've not discussed so far is what we exactly consider a *Person* to be, as Nordisk Familjebok includes several famous figures from classic novels and folktales. In the end, we decided to not count fictional characters for simplicity's sake, since their descriptions felt too unpredictable/inconsistent. Worth mentioning is that in our training set we had the following entry:

> Gambrinus, a Flemish fairy king, who is said to have invented the beer.

which is labelled as Non-Person. Notably, the entry does contain the words "Flemish" and "king", and we can't rule out the possibility that this probably lead to some confusion for the classifier during training.

6 Conclusions

There is clear room for improvement for the entry extraction. In hindsight, further thought should've gone into handling certain types of entries. Specifically, some entries are synonymous with other entries and thus resorts to only referring to another entry in their description. A clear example is the previously discussed "Xeros" (Section 5.3.1). In practice, we should be able to prune such entries without any negative consequences.

It's clear that not all entries were successfully found, and that some entries that were found were not actually real entries. In our case, the most significant impact of this is that we miss out on potential Person entries, while the impact of fake entries should not have any real bearing on performance.

For the classification, we'd like to state that with a minimal training set (for perspective: 400 entries, ~0.19% of all entries), we achieve very promising results. Due to the binary nature of our query, "Is it a Person or not?", and the relatively consistent formatting of entries of people (full name followed by nationality, profession etc...), logistic regression seems to perform very well. The innate semantic space built by BERT (and by extension KB-BERT) most likely played an essential role in this.

We argue that a lack variation and size for the training results in a classifier with a skewed notion of "Person". Indeed, our training set might instead have lead the classifier's definition of "Person" towards things merely related to people of a profession/title/nationality (Section 5.3.1, Section 5.3.2).

With comparatively little effort, we believe performance could easily be improved by increasing the size and variation of the training set. In addition, refining the training process and/or testing other classification methods are also exciting prospects.

Finally, the results of this work can easily be built upon. Acquiring a comprehensive list of people from the encyclopedia paves the way for more interesting questions regarding the distribution of professions, gender and nationalities, for example, and could provide a lot of insight into historical questions of cultural and social elements of the times in which the editions were published.

References

Axel Ahlin, Alfred Myrne, and Pierre Nugues. 2023. Mapping the past: Geographically linking an early 20th century swedish encyclopedia with wikidata.

Vito Walter Anelli, Giovanni Maria Biancofiore, Alessandro De Bellis, Tommaso Di Noia, and Eugenio Di Sciascio. 2022. Interpretability of bert latent space through knowledge graphs. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 3806–3810, New York, NY, USA. Association for Computing Machinery.

Lars Aronsson and the Lysator Computer Association at Linköping University. 1992. Project runeberg. Accessed during the months March-June 2024.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of bert. *Preprint*, arXiv:1906.02715.

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2022. Discovering latent concepts learned in bert. *Preprint*, arXiv:2205.07237.

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.

"The Institute for Language and Folklore". 2020. Svenskans historia. Accessed 5th of June 2024.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert. *Preprint*, arXiv:2007.01658.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. 2021. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742.

A Training Data and Classification Results

The exact results and structure of our data can be found on GitHub: `https://github.com/victorthedude/EDAN70-project-in-computer-science`. More info on how to reproduce our results is available in the repository's README file.

A.1 Scraping

- "`data/nf.zip`": The text we got from scraping from Project Runeberg (Section 3.1)

A.2 Entry Extraction

- "`data/json/first_ed/first_ed.json`": All extracted entries from the first edition (Section 3.2)
- "`data/json/fourth_ed/fourth_ed.json`": All extracted entries from the fourth edition (Section 3.2)

A.3 Training

- "`data/json/training/ver1/dataset_1_training.json`": Entries used for training the classifier (Section 3.4)
- "`data/json/training/ver1/dataset_1_validation.json`": Entries used for the validation set (Section 3.4). Results of the validation are shown in Table 1.

A.4 Classification

- Results of our classification (Section 4) can be found in "`data/json/classification/ver1/...`". Entries classified as Person and Non-Person, for both editions, are available.