KB (National Library of the Netherlands): (Semi-) Automatic Cataloguing of Textual Cultural Heritage Objects

Participants: Martijn Kleppe (company representative), Iris Hendrickx (academic leader), Sara Veldhoen (company representative), Alex Brandsen, Hugo de Vos, Karen Goes, Lin Huang, Hugo Huurdeman, Areumbyeol Kim, Sepideh Mesbah, Myrthe Reuver, Shenghui Wang & Richard Zijdeman

Summary

As a National Library, the KB collects and stores all publications of and about the Netherlands. These publications are described by manually assigning keywords and other metadata by KB's annotators. While for centuries, the KB stored physical objects, recently digital publications are also being collected. Since these digital publications are available as full text, the KB now wants to explore the possibilities to automate the labelling of publications. During the ICT with Industry Workshop 2019, the main research question was: "*Can we automatically label scientific texts with relevant keywords?*". For this case study we used a set of dissertations manually tagged with keywords from the Brinkman thesaurus ('Brinkeys'), and a set of metadata scraped from university websites, combined with abstracts and full text where available. The data is sparse and heterogeneous, and partly English and partly Dutch, which complicates any classification attempt as the Brinkman keywords list is completely in Dutch.

We investigated a range of approaches, for example testing existing tools, matching (translated) keywords in titles, and applying multilingual semantic embeddings. These initial experiments show promising results and possibilities for future support tools of the KB annotators. A list of suggested keywords can then be used by KB employees to quickly select the relevant keywords, which could provide consistency and a significant time saving in the day to day work of the annotators.

Context and problem

The KB | National Library of the Netherlands has been digitising its collections at a rapid pace for a number of years now. Large amounts of scans and machine-readable text created from e.g. historical newspapers, periodicals and books are made available to end users through portals such as Delpher. At the same time, the amount of content deposited by publishers or harvested from the web in digital form, such as e-books, e-journals, and web pages, is growing quickly as well.

Rich and accurate descriptive metadata, ranging from title and author on the one hand to specialist scientific subject headings on the other, form an essential prerequisite for enabling users to effectively navigate these collections. The current practice of creating such metadata manually, however, has become prohibitively time-consuming and, in some cases, prone to error. We therefore explored possibilities for automatically extracting relevant metadata from the objects in our digitized and born digital collections, using methods and techniques from the field of Artificial Intelligence – the subfields of Machine Learning and Natural Language Processing in particular.

More specifically, the research question is: “*Can we automatically label scientific texts with relevant keywords?*”. The goal of the KB use case is to automatically suggest keywords from the Brinkman thesaurus (‘Brinkeys’) to PhD dissertations from six Dutch universities. This list of suggested keywords can then be used by KB employees to quickly select the relevant keywords, which could provide a significant time saving in the day to day work of the annotators as well as increasing the consistency of assigned keywords between annotators.

Data

The data available to answer the research question is a set of dissertations manually tagged with Brinkeys (the gold standard), and a set of metadata scraped from university websites, combined with abstracts and full text where available. This metadata also contains keywords, but there is no mapping available between these keywords and the Brinkeys. The university metadata is sparse and heterogeneous. The data in both the gold standard and university metadata is partly English and partly Dutch, while the Brinkman topic list consists of Dutch keywords and this posed an interesting cross-lingual challenge.

Data preparation

To facilitate experiments, it was required to first transform the available data into a useable format. The first step was to link the information in the manually annotated gold standard, provided by the KB, to the information scraped from university depositories. The gold standard contains (among other data) a PPN identifier, ISBN code, the dissertation title, author and assigned Brinkeys. The metadata from the universities contains does not contain the PPN identifier, but does contain the ISBN, title, author and other information. To match one to the other, we implemented a method that first tries to match on ISBN. This resulted in a match of only 10%. To increase this number, for each row in the gold standard we found all entries in the university metadata with the same author surname. Then out of all the possible matches, we calculated the Levenshtein edit distance between the title of the entry and the title in the gold standard. If any candidates are above the threshold of 60%, we pick the one with the highest score and add it as a match. This resulted in a total match of 58%, or roughly 7,000. We also removed some entries; those without Brinkeys, without a title, or without the ‘gd’ flag, which indicates that they aren’t in the KB repository. This cleaned and linked dataset is the basis for all experiments mentioned below and will be referred as ‘gold standard set’.

Research approach

We decided to take a bottom-up approach and experiment with a range of different approaches to investigate what methods would work for the problem at hand. We first discuss several enhancements of the Brinkman thesaurus that were made to provide insights in the structure and depth of the thesaurus and to open it up to a broader multi-lingual framework provided by Wikidata. Next we discuss the most straightforward approaches that we evaluated which we refer to as ‘baselines’, followed by the experiments with more advantaged approaches.

Brinkman thesaurus enhancement

From the KB, we obtained an XML representation of the Brinkman thesaurus. This representation contained PPN identifiers, the accompanying labels, as well as the associated broader terms. Two derived sets were created based on the XML file:

- A "human-readable" representation of the thesaurus structure, where parent and child terms were mapped to an indented structure, including their textual labels and PPNs. This facilitated better understanding of the thesaurus contents and structure to team members.
- A CSV representation of the data, which includes textual labels, the hierarchy of underlying terms, and the depth of each term within the hierarchy (indicated by a number). This was useful for analyzing the properties of the Brinkeys, as well as the items in the gold standard set.

In total, 14,426 Brinkeys were assigned to items in the gold standard dataset, of which 2,215 unique terms. Leading to an average of 2.1 assigned Brinkeys per item. Each Brinkey occurred on average 6.51 times (median: 3, s.dev: 12.06). The hierarchy depth of terms in our set was relatively shallow on average: 2.14 (median: 2, s.dev: 0.93), and varied between a minimum depth of 1 and a maximum depth of 6.

Figure 1 illustrates the thesaurus structure for the top 25 terms in the gold standard set. As can be observed, its depth is quite varied: some of the top terms are on the highest root level, thus have no parent terms (e.g. *immunologie*, or *bacteriën*), while other terms are deeply nested (e.g. *chemie > biochemie > dna*). In addition, figure A1 (Appendix A) shows a simplified visual representation of the top 100 terms.

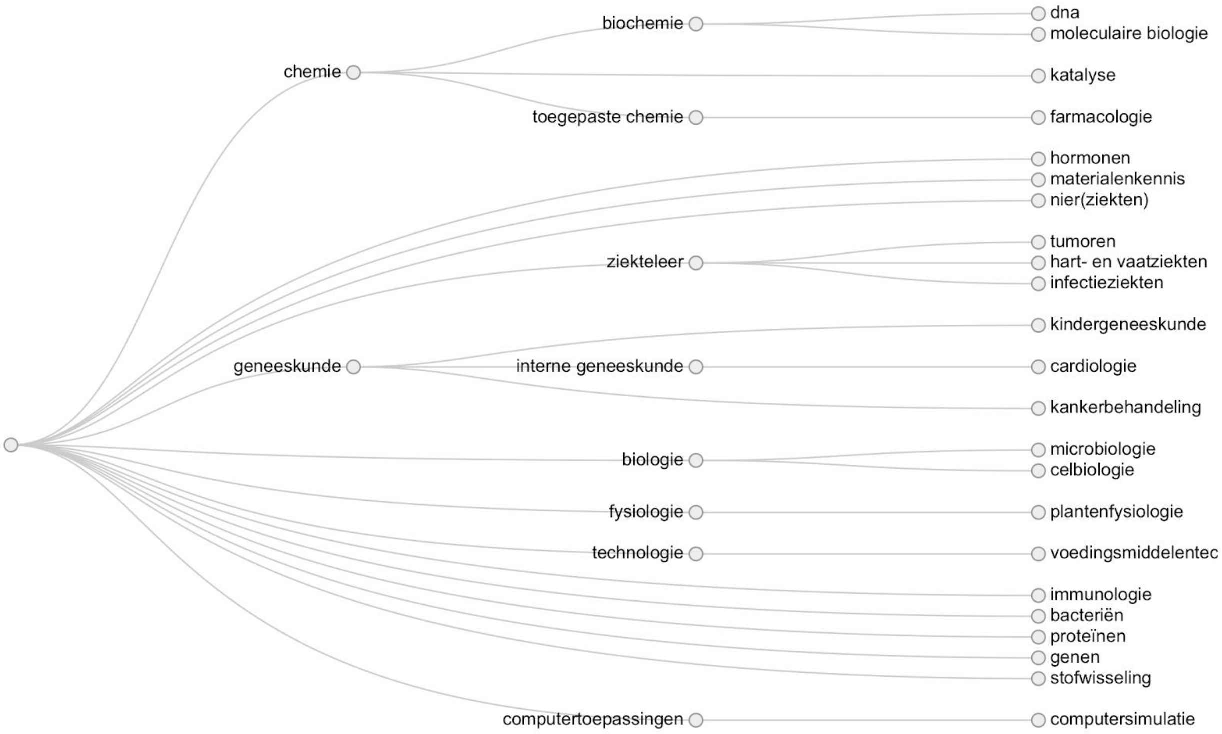


Figure 1: Representation of the full structure of the top 50 of Brinkeys. Terms on the right are the analysed (child) terms, with their parent terms represented towards the left.

Another analysis was done by connecting Brinkeys to university keywords, using the documents occurring in both gold standard set and the university dataset. Figure 2 is a visualisation representing the found connections. The alluvial diagram shows the connection between universities (e.g. *RUG*, Rijksuniversiteit Groningen), the keywords assigned to a thesis by a university (e.g. *urologie, hart- en vaatziekten*), and the corresponding Brinkman thesaurus keywords for the same thesis (*proteinen, hart- en vaatziekten*).

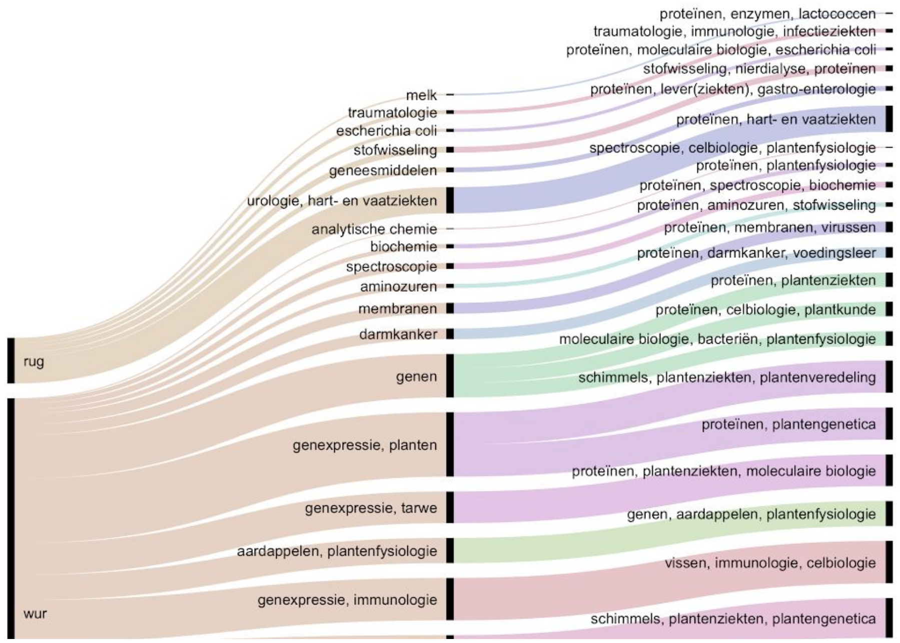


Figure 2: Cropped visual representation of the connection between university, university keywords and Brinkman terms (the size shows the relative frequency of university terms in the whole set). (Full figure A2 available in Appendix A).

Mapping Brinkeys to Wikidata

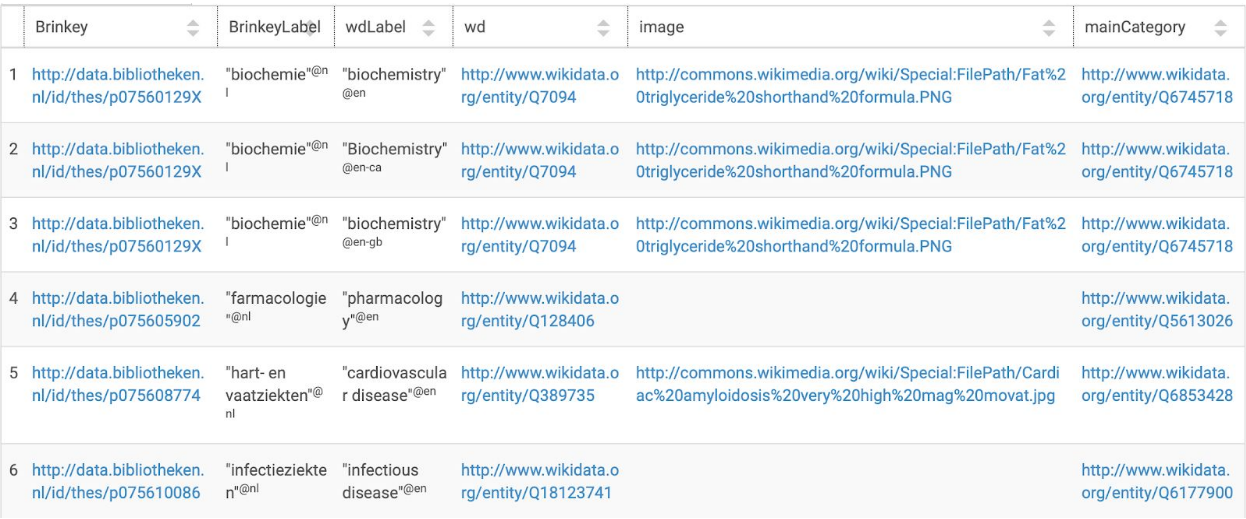
One of the hurdles of mapping Brinkeys to books, is the cross-lingual aspect. Brinkeys are provided in Dutch, while books can be provided in a multitude of languages. One possible solution (described below as one of the baseline approaches) is to transpose the language of the sources into Dutch in order to match with the Brinkeys. Obviously, we could try another route, by transposing the Brinkeys into the language of the source. This method provides two practical advantages. One, unlike book titles, the set of Brinkeys is finite. Transposing the set to English is, safe future additions, a limited task, unlike book titles that appear every day a new. Moreover, Brinkeys labels are relatively short compared to book titles, requiring less data preparation and decision on relevant terms within the dissertation title.

To link Brinkeys to English topics we used Wikidata a Linked Open Data thesaurus that provides in multiple languages. Through the R package "WikidataR" we approached the Wikidata API and requested to provide an English keyword for a Brinkey label. For example, asking for 'biochemie' provides 'biochemistry' as one possible answer. Because of the time window, we accepted lower quality matches, and just accepted the first alternative provided by Wikidata. Despite this anomaly, the 'translation' seems to have worked rather well for the dissertation set, providing related English terms as seen in the screenshot below. (For a limited time, the query can be executed at: https://api.druid.datalegend.net/s/wxlyoflVc).

Knowing the link between the Brinkeys and the Wikidata keywords, we are able to retrieve additional data from Wikidata. In the screenshot below we see links to images that go with the topics at Wikidata. More importantly, because the Brinkman thesaurus proved to be rather flat, we are able to retrieve information from the Wikidata thesaurus hierarchy. In the screenshot the top concept ('mainCategory') of the Wikidata key words is provided.

The information retrieved, especially the English keywords and hierarchy was next used to inform the machine learning models to link thesis' titles to Brinkman keywords. In future work, this method can be expanded to the keywords, abstracts and even full text. Not just for English, but for any language in which the keywords are provided at Wikidata. This allows researchers from across the world to find resources at the KB in their own language.

In addition to its results, this approach is worthwhile to mention from a Data Science point of view, for it combines methods from structured data and textual data to solve the problem at hand.



	Brinkey	BrinkeyLabel	wdLabel	wd	image	mainCategory
1	http://data.bibliotheken.nl/id/thes/p07560129X	"biochemie"@nl	"biochemistry"@en	http://www.wikidata.org/entity/Q7094	http://commons.wikimedia.org/wiki/Special:FilePath/Fat%20triglyceride%20shorthand%20formula.PNG	http://www.wikidata.org/entity/Q6745718
2	http://data.bibliotheken.nl/id/thes/p07560129X	"biochemie"@nl	"Biochemistry"@en-ca	http://www.wikidata.org/entity/Q7094	http://commons.wikimedia.org/wiki/Special:FilePath/Fat%20triglyceride%20shorthand%20formula.PNG	http://www.wikidata.org/entity/Q6745718
3	http://data.bibliotheken.nl/id/thes/p07560129X	"biochemie"@nl	"biochemistry"@en-gb	http://www.wikidata.org/entity/Q7094	http://commons.wikimedia.org/wiki/Special:FilePath/Fat%20triglyceride%20shorthand%20formula.PNG	http://www.wikidata.org/entity/Q6745718
4	http://data.bibliotheken.nl/id/thes/p075605902	"farmacologie"@nl	"pharmacology"@en	http://www.wikidata.org/entity/Q128406		http://www.wikidata.org/entity/Q5613026
5	http://data.bibliotheken.nl/id/thes/p075608774	"hart- en vaatziekten"@nl	"cardiovascular disease"@en	http://www.wikidata.org/entity/Q389735	http://commons.wikimedia.org/wiki/Special:FilePath/Cardiac%20amyloidosis%20very%20high%20mag%20movat.jpg	http://www.wikidata.org/entity/Q6853428
6	http://data.bibliotheken.nl/id/thes/p075610086	"infectieziekten"@nl	"infectious disease"@en	http://www.wikidata.org/entity/Q18123741		http://www.wikidata.org/entity/Q6177900

Figure 3. Screenshot of Wikidata retrieved Brinkeys labels in English.

Baselines

It was decided on the first day to develop and test a number of possible straight-foward approaches to investigate the complexity of the task and to be able to compare different approaches to each other.One of these approaches is to use an existing tool, *Annif*, an off the shelf keyword generation tool developed by the National Library of Finland. We denote this as Baseline 1. We also investigated two direct mapping approaches, namely performing a lexical match between (translated) tokens in the dissertation title and the Brinkeys (denoted as baseline 2) and secondly a lexical match between the university metadata keywords and the Brinkeys (baseline 3).

Baseline 1 - Existing Keyword Extraction System; Annif

For the first baseline we decided to try an existing system as this was thought to be an easy way to provide a baseline. Annif[1] is an open-source software tool for automated subject (keyword) indexing

[1] http://annif.org/

that is developed by the National Library of Finland. The tool is flexible and can be used with different keyword vocabularies. However, installing and configuring the tool took longer than expected, partly due to the challenge of extensive but uninformative documentation. When it was up and running, we first processed our data with the built-in corpora as a first test. We did not perform formal evaluation, but the results were inspected manually and seemed relevant. We then integrated our own Brinkman thesaurus and trained the system on the dissertation titles and abstracts (where available). This produced a recall at 10 of 16.7 and a precision at 3 of 16.7. Again, when checking the output, it looks like the suggested labels make sense generally, and at least 1 correct label is included. However, some overfitting took place; the method caught on to low frequent features in the training data, which caused e.g. the label 'Twente' to be applied to almost all Dutch titles/abstracts. We think that using more data, and perhaps using full text, should increase the accuracy of this method.

Baseline 2 - Lexical Match Brinkeys - Titles

The second baseline was to try and match words in the titles of the dissertations directly to the Brinkeys based on string matching. The main problem was that the titles are in English and Dutch, while the Brinkeys are all in Dutch. So to make this method work, the English words first needed to be translated. Due to time constraints and practical issues (with the Google translate API for example), the title words were translated individually and out of context. This string-to-string translation leaves room for semantic error and ambiguity that might be solved with semantic mapping, or a more semantic method of translation.

Even though there is mismatch in language, it was first attempted to match the Brinkeys and titles as is (after lowercasing and punctuation stripping). With this method 1162 out of 24019 unique title words match a Brinkey, roughly 5%. Due to the low result, we then translated the title words to Dutch using Google Translate which increased the result to 11.5%. The most glaring problem with this method is the translation of title words to Dutch: the translation is cumbersome, and is not always a 1 on 1 match with a BrinKey keyword because the translation results to a phrase instead of a keyword, or a different word.

Pre-translation, the recall@1 is 16.9% and the precision@1 is 30.5%. After translation, these figures go down to 18.9% and 20.2% respectively, indicating that the translation process actually hinders matching Brinkeys to the title.

A quick qualitative check shows that, without translation, a surprising number of English keywords are correctly recognized, especially medical terms. And although some relevant new matches exist after translation, the process introduces too much noise and false matches to be of real use. Problems arise from ambiguities as well as practical issues that tend to lead to weird, sometimes even inexplicable solutions.

Baseline 3 - Lexical Match Brinkeys - University Keywords

The third baseline is very similar to the 2nd, except it is using the keywords assigned by the universities, instead of the dissertation titles. This method ran into the same problems with translation. In total 2597 out of 6856 dissertations had a (translated) university keyword that directly matched a Brinkey.

The recall at 20 is 11.6% and the precision at 3 is 14%, which is surprisingly significantly lower than the previous method. This could possibly be explained by the fact that the KB annotators are not experts in all scientific fields, and are more likely to pick words out of the title as Brinkeys, while the scientists themselves will use more nuanced keywords. This would explain the higher scores for the title matching.

The results from these baselines give a low limit which our methods should be able to improve upon.

Experiments

After the brainstorming session on the first day of the workshop, a couple of possible methods were discussed. Method 1 and 2 are the result of this session, while method 3 was added later. All of these methods were evaluated using a 20/80 split train and test set.

Method 1 - Naive Bayes Classifier

As a first attempt, we trained a Naive Bayes classifier, using TF-IDF word weights as features. This was trained on 18,537 instances, using the title text and the abstract where available, with 11,302 possible labels (the Brinkeys). Due to this being a multi class classifier, it only predicted one Brinkey per dissertation, so we only have recall and precision at 1; 3.5% and 6.5% respectively. This is still significantly worse than the baselines, so we decided not to work on this method further.

Method 2 - BrinKeysEmb - Multilingual Word Embeddings

The second method is to using multilingual word embeddings to measure the semantic relatedness between concepts, specifically between the concepts in the dissertation and the Brinkeys. In short, the BrinKeysEmb approach is organised in the following steps:

- Use the pre-trained model trained on English and Dutch Wikipedia which were trained using Fasttext.
- Learn a mapping from the English to the Dutch embedding space using the adversarial training and Procrustes refinement.
- Use the university keywords existing in the metadata to find the most similar Brinkeys to them, otherwise extract all concepts from the textual representation of a given PhD dissertation using a state-of the art multilingual concept extractor and use them instead of the university keywords.
- Measure the Semantic Relatedness of a given concept in the given PhD dissertation, compared to all Brinkman keywords.
- Rank the list of Brinkeys according to their embedding similarity to the set of concepts of a given PhD dissertation.

This method performed reasonably well, with a recall at 20 of 25% and a precision at 3 of 7%. For future work, additional qualitative evaluation is of importance to better understand the range of applicability of this approach. Also, we can benefit from the hierarchy of the Brinkman topics to understand whether we are able to predict more general or specific topics. Furthermore the ranking of the final Brinkeys is designed in a very simple way and deserve further discussion and refinement. Finally in order to boost the overall performance we might consider employing an ensemble approach which consists of exact matching, partial matching and the word embeddings.

Method 3 - Fasttext Classifier

Finally, we used the fasttext tool to label dissertations. We created the input for the tool by concatenating the dissertation title, abstract, university keywords and institution. This, together with the correct labels (Brinkeys) was used to train and evaluate the method. The precision at 3 is 16.2% and the recall at 20 is 40.3%, much higher than previous methods.

Results

To evaluate the results, it was decided to use the following metrics across all baselines and further experiments:

- Recall at 20 predicted labels
- Precision at 3 predicted labels

This should make it possible to directly compare the different methods, giving an overview of accuracy. However, we found that some of the methods used do not offer 20 predicted Brinkeys per dissertation, in which case the most relevant metric is given. Something to keep in mind when viewing these results is that the goal of the work is to provide a list of *possible* Brinkeys to the KB employee which they then use to pick actually *relevant* keywords. This means that recall is more important than precision.

An important fact to consider is that in some cases, the Brinkeys suggested by the method might actually be relevant, but not added by the KB employee in the gold standard. The arbitrariness of keyword assignment can lead to false false positive, resulting in a lower recall. Besides the quantitative evaluation, it would therefore be useful to manually inspect the results of some of the methods, obtaining a more qualitative view of the results. Ideally, manual evaluation of results by KB employees would give us an insight into the severity of this arbitrariness and give more accurate results (also see Future Work).

Evaluation

Table 1 gives an overview of the quantitative results of the different methods.

Method	**Recall**			**Precision**	
	At 1	**At 10**	**At 20**	**At 1**	**At 3**
Baseline 1 (Annif)		16.7			16.7
Baseline 2 (Title - Brinkey matching)	16.9			**30.5**	
Baseline 3 (University keyword - Brinkey matching)	11.6			14	
Method 1 (Naive Bayes classifier)	3.5			6.5	

Method 2 (Multi-lingual word embeddings)			24.8		6.6
Method 3 (FastText classifier)			**40.3**		16.2

Table 1: Results of 6 experiments, in percentages

These results might seem quite low, especially in those cases where the advanced methods cannot outperform the baselines (most strikingly when it comes to precision). We should remember, however, that the task is quite challenging: as many as 2200 labels (Brinkeys) occur in the data, some of which are very rare. Also, the dataset is relatively small with roughly 7k items that have full data. In addition, most of the methods were applied to title and abstract only, in some cases adding university keywords. Applying them to full text might increase performance significantly, as most machine learning based techniques are 'data hungry'.

Conclusions

During the ICT with Industry Workshop 2019 we investigated a range of methods to automatically label dissertations with relevant keywords from the Dutch Brinkman thesaurus (brinkeys). We looked at the Brinkman thesaurus itself and worked on visualising the structure and depth of the thesaurus an opening it up to a broader multi-lingual framework provided by Wikidata.
We experimented with an existing tool for keyword generation developed by the Finnish national library that gave promising results. As baseline methods we investigated straightforward matching of words in the thesis titles or university keywords against the brinkeys and in case of English dissertations, the automatically generated Dutch translations. We investigated multi-lingual word embeddings and fasttext as more advanced methods for keyword prediction. It is hard to make a comparison and pick the most promising method based on the results, because of the above-mentioned issues: neither approach has been pushed to its maximum capability, mostly due to time constraints. Furthermore, qualitative evaluation is needed in order to determine the strengths and weaknesses of the systems. In the end, an ensemble of methods might actually perform best.

Future work

While the workshop was a very fruitful week, there is still more work that could be done to further improve the methods, which will hopefully result in a useable system for the KB. Below we have listed some possible avenues of research that might benefit this project.

The main drawback for the experiments was the small data size. With only 7000 dissertations and a potential keyword list of 2200 keywords, the sparsity was just too high. A larger data set will certainly lead to better results. Due to the time constraints we worked only with the text from titles and abstracts. Exploring methods to learn keywords on the basis of full texts is another possible direction for future steps.

OCLC has also experimented with an in-house tool called `Ariadne[2]' that yielded very promising results (precision@3 as 29.2% and recall@20 as 56.9) and is certainly worth to be investigated further.

Another logical next step to pursue is to further analyse the results of the methods, to better understand and evaluate their outcomes. For instance, inspecting the suggested Brinkeys for a set of dissertations with one or more KB employees, to see if the result are relevant in the eyes of the employee. This should tentatively provide a more accurate evaluation of the methods, as in these multi-label classification tasks multiple labels can be viewed as 'correct'.

It would also be very useful to have multiple KB employees annotate a set of the same dissertations. This set can be used to calculate the inter-rater agreement, a measure that shows the average overlap between different annotators, and is an indication of how hard the task is for humans, which can be interpreted as an upper limit that automatic methods can reach.

[2] Wang, S. & Koopman, R. Scientometrics (2017) 111: 1017. https://doi.org/10.1007/s11192-017-2298-x

Appendix A: Additional figures

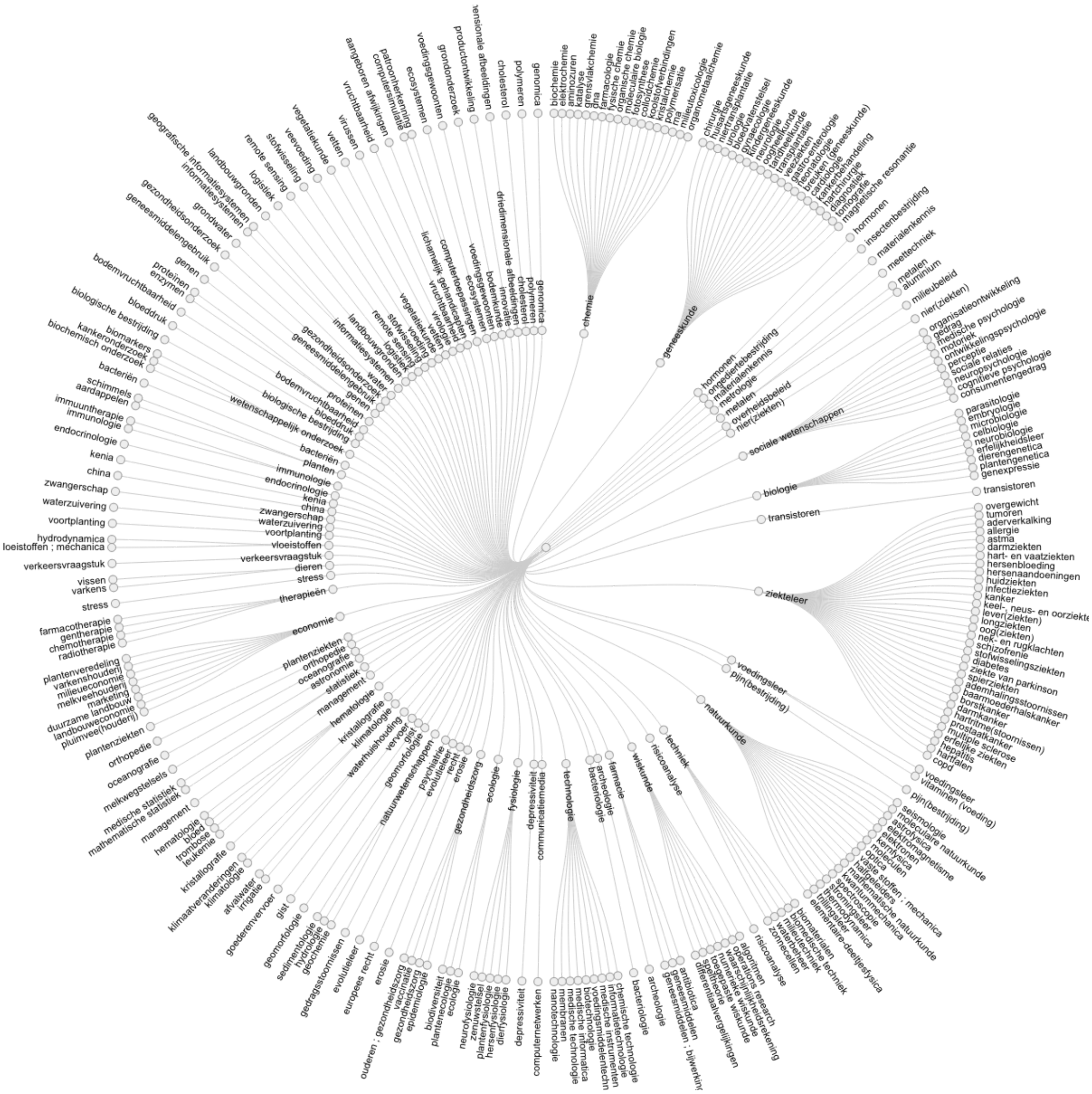


Figure A1: Simplified visual representation of the structure of the top 100 most popular Brinkman terms occurring in the theses dataset. The root and leaf node are presented in the figure, intermediate nodes omitted.

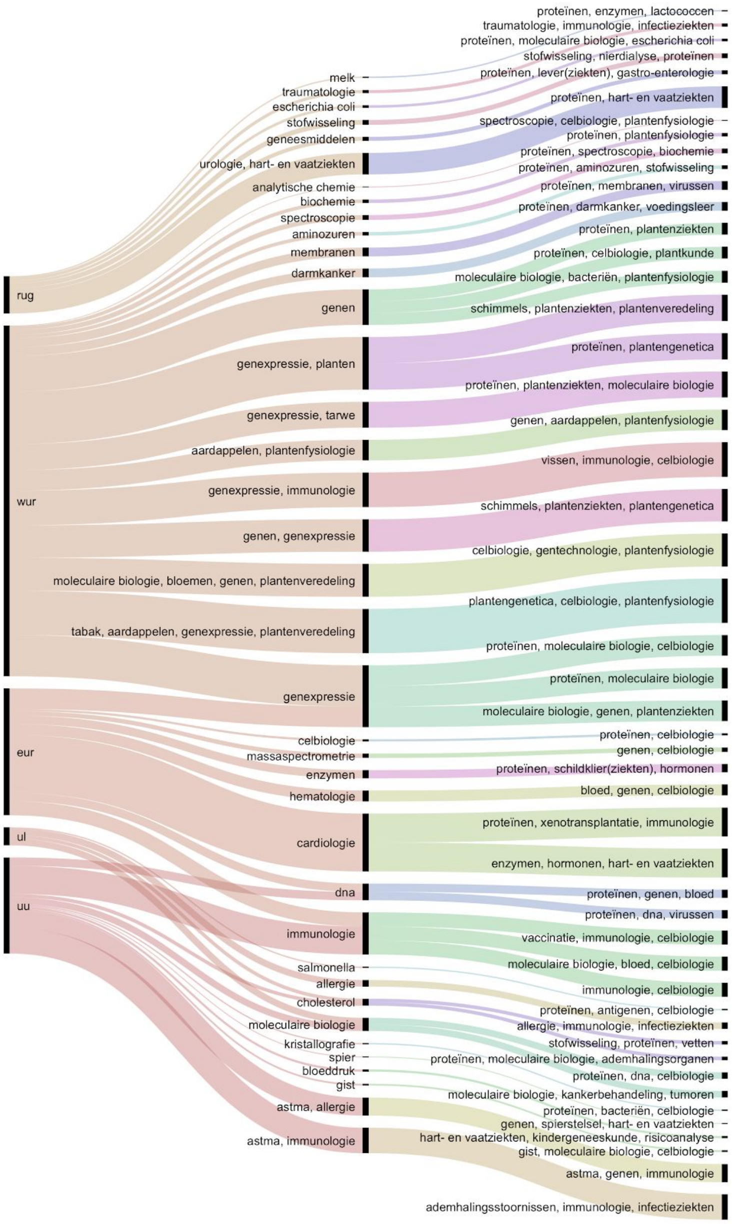

rug
wur
eur
ul
uu
melk
traumatologie
escherichia coli
stofwisseling
geneesmiddelen
urologie, hart- en vaatziekten
analytische chemie
biochemie
spectroscopie
aminozuren
membranen
darmkanker
genen
genexpressie, planten
genexpressie, tarwe
aardappelen, plantenfysiologie
genexpressie, immunologie
genen, genexpressie
moleculaire biologie, bloemen, genen, plantenveredeling
tabak, aardappelen, genexpressie, plantenveredeling
genexpressie
celbiologie
massaspectrometrie
enzymen
hematologie
cardiologie
dna
immunologie
salmonella
allergie
cholesterol
moleculaire biologie
kristallografie
spier
bloeddruk
gist
astma, allergie
astma, immunologie
proteïnen, enzymen, lactococcen
traumatologie, immunologie, infectieziekten
proteïnen, moleculaire biologie, escherichia coli
stofwisseling, nierdialyse, proteïnen
proteïnen, lever(ziekten), gastro-enterologie
proteïnen, hart- en vaatziekten
spectroscopie, celbiologie, plantenfysiologie
proteïnen, plantenfysiologie
proteïnen, spectroscopie, biochemie
proteïnen, aminozuren, stofwisseling
proteïnen, membranen, virussen
proteïnen, darmkanker, voedingsleer
proteïnen, plantenziekten
proteïnen, celbiologie, plantkunde
moleculaire biologie, bacteriën, plantenfysiologie
schimmels, plantenziekten, plantenveredeling
proteïnen, plantengenetica
proteïnen, plantenziekten, moleculaire biologie
genen, aardappelen, plantenfysiologie
vissen, immunologie, celbiologie
schimmels, plantenziekten, plantengenetica
celbiologie, gentechnologie, plantenfysiologie
plantengenetica, celbiologie, plantenfysiologie
proteïnen, moleculaire biologie, celbiologie
proteïnen, moleculaire biologie
moleculaire biologie, genen, plantenziekten
proteïnen, celbiologie
genen, celbiologie
proteïnen, schildklier(ziekten), hormonen
bloed, genen, celbiologie
proteïnen, xenotransplantatie, immunologie
enzymen, hormonen, hart- en vaatziekten
proteïnen, genen, bloed
proteïnen, dna, virussen
vaccinatie, immunologie, celbiologie
moleculaire biologie, bloed, celbiologie
immunologie, celbiologie
proteïnen, antigenen, celbiologie
allergie, immunologie, infectieziekten
stofwisseling, proteïnen, vetten
proteïnen, moleculaire biologie, ademhalingsorganen
proteïnen, dna, celbiologie
moleculaire biologie, kankerbehandeling, tumoren
proteïnen, bacteriën, celbiologie
genen, spierstelsel, hart- en vaatziekten
hart- en vaatziekten, kindergeneeskunde, risicoanalyse
gist, moleculaire biologie, celbiologie
astma, genen, immunologie
ademhalingsstoornissen, immunologie, infectieziekten

Figure A2: Visual representation of the structure of the mapping between universities, university keywords and Brinkman terms. Size represents relative frequency of university keywords in whole dataset.