Building Question-Answer Data Using Web Register Identification

Anni Eskelinen, Amanda Myntti, Erik Henriksson, Sampo Pyysalo and Veronika Laippala
TurkuNLP
University of Turku
{aeeske, amanda.a.myntti, erik.henriksson, sampo.pyysalo, mavela}@utu.fi

Abstract

This article introduces a resource-efficient method for developing question-answer (QA) datasets by extracting QA pairs from web-scale data using machine learning (ML). Our method benefits from recent advances in web register (genre) identification and consists of two ML steps with an additional post-processing step. First, using XLM-R and the multilingual CORE web register corpus series with categories such as QA Forum, we train a multilingual classifier to retrieve documents that are likely to contain QA pairs from web-scale data. Second, we develop a NER-style token classifier to identify the QA text spans within these documents. To this end, we experiment with training on a semi-synthetic dataset built on top of the English LFQA, a small set of manually cleaned web QA pairs in English and Finnish, and a Finnish web QA pair dataset cleaned using ChatGPT. The evaluation of our pipeline demonstrates its capability to efficiently retrieve a substantial volume of QA pairs. While the approach is adaptable to any language given the availability of language models and extensive web data, we showcase its efficiency in English and Finnish, developing the first open, non-synthetic and non-machine translated QA dataset for Finnish – Turku WebQA – comprising over 200,000 QA pairs.

Keywords: question-answer, web genre identification, web register, XLM-R, Web-as-Corpus, language resource

1. Introduction

Recent progress in large language models (LLMs) has attracted widespread interest from the NLP community and beyond. These models are utilized for question answering and interactive discussions with humans. To do so effectively, they require training with high-quality question-answer (QA) datasets.

While the number of QA datasets for English has grown significantly, thanks to initiatives such as the open source OpenAssistant[1], many languages still face a shortage of such resources. This is mainly due to the considerable time and effort required to compile these datasets, a challenge that is particularly acute for smaller language communities. To address this, previous studies have explored strategies for producing QA datasets with reduced manual annotation. For instance, Alberti et al. (2019) developed a method for creating synthetic QA data by combining models of question generation and answer extraction, Kalpakchi and Boye (2023) used GPT-3 to create and evaluate a synthetic QA dataset of Swedish, Kylliäinen and Yangarber (2023) introduced the first QA dataset for Finnish by machine translating the English SQuAD, and Fan et al. (2019) used Reddit QA pages to extract QA pairs. While crowd-sourcing can be impractical for low-resource languages, an initiative to get QA data for Finnish has also been publicized by the name of Avoin Avustaja[2], based on OpenAssistant.

In this paper, we present another method to reduce the human effort required to create QA data. We take advantage of recent advancements in web register (genre) identification and utilize the newly released CORE corpus series (Laippala et al., 2023; Skantsi and Laippala, 2023; Repo et al., 2021), which encompasses web registers in four languages and includes several register categories featuring QA pairs. Using these datasets as our starting point, we outline a machine learning-based pipeline to extract clean QA pairs from a substantial volume of web-crawled data. We demonstrate the feasibility of the method using English and Finnish, with a focus on Finnish which currently has very limited QA resources.

As the first step of the pipeline, we train a multi-class classifier with two labels to isolate documents associated with the targeted register categories, which are likely to contain questions and answers. Then, we train an NER-style token classification model to identify the questions and answers within these documents. Lastly, we post-process and pair up the extracted questions and answers to clean QA pairs. For developing this model, we explore various data sources: a semi-synthetic English QA dataset that we constructed on top of the LFQA dataset (Blagojevic, 2022) by adding synthetic noise, a small Finnish QA dataset where we extracted QA pairs from web

[1] https://open-assistant.io/

[2] https://avoin-avustaja.fi/

LREC-COLING 2024, pages 2595–2611

documents using ChatGPT (OpenAI, 2023), and small batches of manually cleaned QA pairs from web documents in both Finnish and English. The entire pipeline is illustrated in Figure 1.

The evaluation demonstrates the effectiveness of our process in retrieving a substantial volume of QA pairs. The best results for extracting the pairs from web documents were achieved using a model trained on data augmented with annotations generated by ChatGPT.

The method can be applied to any language, provided there is a web-scale dataset available and the language is supported by a multilingual masked language model like XLM-R (Conneau et al., 2020). Ideally, the language would also be supported by a generative model such as ChatGPT, as it reduces the need for manual annotation. As a result of our work, we present the first freely available, non-synthetic and non-machine translated Finnish QA dataset, consisting of 237,000 QA pairs. Both the final QA dataset and the evaluation sets developed during our research are available at `https://github.com/TurkuNLP/register-qa`.

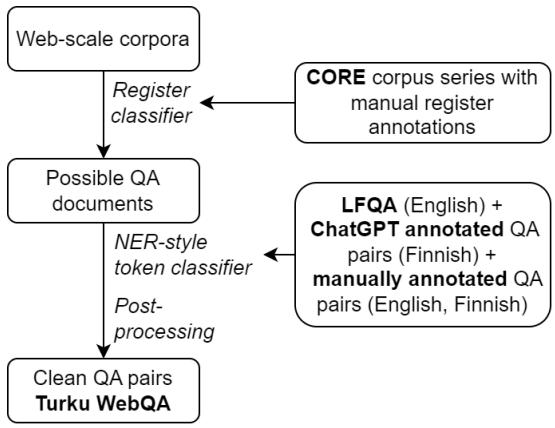

Figure 1: Overview of our method for extracting clean QA pairs from web-scale corpora.

2. Related Work

2.1. Creating QA datasets

Cambazoglu et al. (2021) distinguish between three kinds of QA tasks and datasets. In abstraction tasks, the answer is generated without relying on the vocabulary of the question or the given context. In extraction tasks, the answer is identified from the context, while in retrieval tasks, the goal is to rank text segments based on their likelihood of containing an answer. Our task falls in the extraction category.

Crowd-sourcing is commonly used to create extractive QA datasets from the web. For the SQuAD dataset (Rajpurkar et al., 2016), paid workers formulated questions based on a Wikipedia document and extracted answers as text passages. For TyDiQA (Clark et al., 2020), a similar methodology was employed, wherein paid workers devised questions about a shortened Wikipedia article.

Crowd-sourcing has been effectively applied also to languages such as French (d'Hoffschmidt et al., 2020) and Korean (Lim et al., 2019). However, for lower-resourced languages, this is often unfeasible. Alternate strategies to minimize manual effort have included machine translation (Kylliäinen and Yangarber, 2023; Ivanova et al., 2023), using GPT-3 (Kalpakchi and Boye, 2023), and sourcing QA pairs from websites like Reddit, which are specifically designed for QA interactions (Fan et al., 2019; Blagojevic, 2022).

In the domain of QA modelling, Finnish qualifies as a low-resource language. The only Finnish QA datasets currently available consist of machine translations of the SQuAD (Kylliäinen and Yangarber, 2023) and SQuAD 2.0 (Nuutinen et al., 2023) datasets as well as the now finished crowd-sourced OpenAssistant dataset[3] (Köpf et al., 2023) which consists of only 138 Finnish messages. The OpenAssistant dataset includes human-generated, human-annotated conversations which are divided into "prompter" and "assistant", where participants were instructed to answer like an AI. To the best of our knowledge, there is currently no Finnish QA dataset sourced from authentic person-to-person interactions on the web that are originally written in Finnish.

2.2. Extracting QA pairs from Data

Extracting QA pairs from web pages or other sources is not novel. For web data, many have applied simple rule-based solutions for QA pair identification. Jijkoun and de Rijke (2005) identified QA pairs from Frequently Asked Question websites by utilizing the HTML formatting of the web pages and heuristics, such as identifying a question mark, to select the relevant text spans. Both Kwong and Yorke-Smith (2009) and Cong et al. (2008) utilized similar heuristics, combined with regular expression patterns, part-of-speech tagging, as well as supervised rule induction learning, to extract QA pairs from emails and online forums.

Amir Pouran Ben Veyseh and Nguyen (2022) annotated video transcriptions for questions and answers using BIO taggings and trained a NER-style token classifier to pinpoint the QA elements. During evaluation, their classifier significantly out-

[3] `https://huggingface.co/datasets/OpenAssistant/oasst2`

Name	Documents	Availability
Parsebank	6,581,550	Upon request from authors
mC4-Fi	16,089,579	https://huggingface.co/datasets/mc4
CC-Fi	40,074,961	Luukkonen et al. (2023)
Falcon RefinedWeb	(8M) 968,000,015	https://huggingface.co/datasets/tiiuae/falcon-refinedweb

Table 1: Web-scale dataset sizes and sources. From the Falcon dataset, we only took 8M documents due to the size of the dataset.

performed a rule-based method. Fan et al. (2019) extracted QA pairs from the popular question-answering forum *Explain It Like I'm Five* on Reddit, using user engagement as a measure of question and answer quality. A similar approach was also used in the *LFQA* dataset (Blagojevic, 2022).

2.3. QAs in Web Register Identification

Recent progress in web register identification has greatly enhanced the utility of web-scale data in crafting extensive QA corpora. The pioneering work in this field was the Corpus of Online REgisters of English (CORE) (Biber and Egbert, 2016; **?**), which was the first to include the unrestricted web and featured specific register categories for QA pairs, thus enabling the identification of documents with QA content across the web. Since the release of CORE, datasets with similar register annotations have been published for other languages (Laippala et al., 2022; Rönnqvist et al., 2021; Kuzman et al., 2022). In particular, the Finnish Corpus of Online Registers covers 10,000 documents (Skantsi and Laippala, 2023), and also the French and Swedish CORE collections include approximately 2,000 documents respectively (Repo et al., 2021).

The CORE corpora employ a hierarchical register annotation scheme with eight main registers, tens of subclasses and a technical category for machine-translated texts. The subregister categories are always annotated together with the main register label. For instance, documents labelled under the subregister Question-Answer Forum also fall under the primary register category Interactive Discussion. Additionally, documents combining characteristics of several registers or subregisters can be assigned multiple labels. These are referred to as *hybrid documents*.

Three of the CORE subclasses are likely to include QA pairs: Question-Answer Forum, FAQ about How-To, and FAQ about Information. In the Finnish, French and Swedish corpora, the two FAQ categories are combined into simply FA. Previous studies have shown that although the task of register identification presents challenges due to the noisiness of the data and fuzziness of the categories, the classifiers can reach a performance of nearly 80% F1-score (Skantsi and Laippala, 2023; Rönnqvist et al., 2021; Kuzman et al., 2023). This opens up the possibility of developing a classifier for detecting questions and answers within web datasets.

3. Data

3.1. Web-scale Datasets

The web-scale datasets we use for identifying QA pairs cover all the cleaned Finnish web-crawled corpora we have access to. Additionally, to demonstrate the generalisability of the method, we use the first 8 million documents of the English Falcon Redefined Web dataset (Penedo et al., 2023).

CC-Fi is a Finnish Common Crawl taken from crawls between 2013 and 2022. The dataset was introduced in Luukkonen et al. (2023).

mC4-Fi is the Finnish-language subset of the mC4 corpus (Xue et al., 2021), which in turn is derived from Common Crawl.

Finnish Internet Parsebank, originally introduced in Luotolahti et al. (2015), is a corpus of Finnish collected between 2015 and 2016 from Common Crawl and by crawling `.fi` domains.

As all of these resources are Common Crawl based, in the post-processing stage we make sure that duplicate QA pairs are removed.

Falcon RefinedWeb (Penedo et al., 2023) is based on all the English material found in all the Common Crawl releases and has undergone a strict cleaning process. The public extract on Huggingface is 600GT from the full 5000GT. Due to the volume of the dataset, we only used the first 8 million documents to demonstrate the efficiency of the harvesting method for further languages.

The sizes of the datasets and their sources are displayed in Table 1.

3.2. Register Corpora for QA Document Detection

To develop a classifier for identifying web documents with QA pairs, we merged the four primary CORE series corpora that had been manually annotated for register: English, Finnish, French and Swedish[4]. Additionally, for French and Swedish,

[4] https://github.com/TurkuNLP/FinCORE_full
https://github.com/TurkuNLP/CORE-corpus

	Not QA	QA	Total
Train	43,052	1,261	44,313
En	32,783	1,122	33,905
Fi	6,469	82	6,551
Fre	1,900	24	1,924
Swe	1,900	33	1,933
Dev	7,154	195	7,349
En	4,683	161	4,844
Fi	926	10	936
Fre	777	12	789
Swe	768	12	780
Test	13,539	376	13,915
En	9,360	326	9,686
Fi	1,853	22	1,875
Fre	1,168	9	1,177
Swe	1,158	19	1,177
Total	63,745	1,832	65,577
Total%	97.2%	2.8%	100%

Table 2: Class distribution in the register identification data after filtering and mapping the labels. Displayed are the joined dataset numbers as well as language specific numbers for the English, Finnish, Swedish and French datasets.

we had access to unpublished, larger versions of the corpora.

During the data preprocessing, all QA-related register labels—Question-Answer Forum, FAQ about How-To, FAQ about Information and FA—were first mapped to *QA*. Then, we deleted documents that were labelled as machine translations and disregarded any additional register labels in hybrid documents. Finally, all documents without a QA label were categorized as Not QA.

The size of the resulting dataset, along with its train/dev/test splits, is shown in Table 2.

3.3. Semi-Synthetic English QA Pairs

To develop the NER-style token classifier identifying QA text spans from the web documents we built a semi-synthetic dataset on top of the English long-form question answering (LFQA) dataset (Blagojevic, 2022), available on Huggingface[5]. LFQA comprises a total of 239,167 QA pairs from several QA subreddits: AskHistorian, AskScience, and Explainlikeimfive. Each question in the original dataset includes several answers, each accompanied by a quality score determined by Reddit users' votes. From this dataset, we selected specific columns: title, selftext and answers. Title and selftext were joined to create the question elements of the final dataset. From the answers column, we chose the answer with the highest score.

In order to apply the LFQA dataset for identifying the QA pairs and their text spans in the web-crawled documents, we introduced synthetic noise to the data. First, we manually identified noise from the entire QA documents found by the QA document classifier. These included phrases such as *3 mo. ago* or *Answer 06/10/2023*. We then prefixed some of the QA-pair documents with these phrases. Finally, we adopted the NER-style token classification formatting for the documents, labelling each token as Q, A, or O. The BIO taggings that Amir Pouran Ben Veyseh and Nguyen (2022) used in their experiment (see Section 2.2) were excluded, as they clearly decreased the model performance in initial experiments.

3.4. Curated Web QA Datasets

To further develop the NER-style token classifier, we compiled three domain datasets by sampling and annotating documents from the web-scale datasets, resulting in the curated datasets shown in Table 3. In addition to traditional manual annotation, we also explored using ChatGPT (OpenAI, 2023) as a QA annotator. This proved successful and reduced the workload of annotating data. In this test, we focused on Finnish due to the previously discussed limited availability of QA resources (see Section 2.1).

The manually annotated English dataset contains 100 documents sampled from the Falcon RefinedWeb (see Section 3.1). We identified 345 questions and 192 answers in this sample. Specifically, 41 documents had full QA pairs, 24 had only questions, and 35 were labelled as empty. The "empty" classification includes potential standalone answers, which were excluded due to the challenges in identifying them without accompanying questions.

The manually annotated Finnish dataset was sampled from each of the web-scale datasets (3.1). It comprises 218 documents: 107 with QA pairs, 28 with only questions, and 83 categorized as empty. In total, this dataset has 376 questions and 333 answers.

The ChatGPT-annotated Finnish dataset, sourced from the web-scale datasets, includes 3,424 randomly selected documents. These were annotated by ChatGPT, using the Finnish manual annotations as training data, covering a total of 2,919 questions and 2,491 answers.

Each dataset uses the NER-style format, where every token is labelled as either Q(uestion), A(nswer), or O(ther). The manual annotations were made by two annotators with experience in linguistics. For the Finnish documents, 121 were double-annotated to compute the inter-annotator

https://github.com/TurkuNLP/multilingual-register-labeling

[5]https://huggingface.co/datasets/-vblagoje/lfqa

Language	Sources	Annotator	Documents	Questions	Answers
English (total)	Falcon RefinedWeb	Human	100	345	192
Dev			40	200	70
Test			60	145	122
Finnish (total)	mC4-Fi, CC-Fi, Parsebank	Human	218	376	333
Train			100	206	164
Dev			50	66	63
Test			68	104	106
Finnish (total)	mC4-Fi, CC-Fi, Parsebank	ChatGPT	3,424	2,919	2,491
Train			3,424	2,919	2,491

Table 3: Sources and sizes of the curated QA datasets.

agreement, using the overlap F1-score detailed in Section 4.4. This agreement scored 0.85 for questions and 0.88 for answers, averaged across annotations. Excluding empty documents, the scores were 0.74 for questions and 0.79 for answers. The seqeval accuracy (see Section 4.4) was also measured between the annotators and resulted in 0.83 for all documents and 0.78 for non-empty documents.

4. Methods

4.1. QA Document Identification

We approached the QA document identification from the web-scale datasets as a multi-class classification task with two labels, using the register datasets detailed in Section 3.2 as data. Our choice for a pre-trained model was XLM-R, available from HuggingFace[6], due to its consistent superiority over other multilingual models (Repo et al., 2021; Rönnqvist et al., 2021). To manage inference costs on large data, we opted for the base version over the larger variant. The implementation was done using the Huggingface Transformers library, Pytorch version.

The XLM-R was fine-tuned for text classification with the training sets detailed in Table 2. Given the transformer's limitation of processing only 512 tokens simultaneously, we truncated longer texts during tokenization. Furthermore, to address the issue of class imbalance in our data, we calculated class weights using the *compute_class_weight* method from the sklearn[7] library, integrating them into our loss function.

We optimized the learning rate hyperparameter using a grid search with values of {1e-5, 4e-6,5e-6, 7e-5, 8e-6}. The best learning rate was 4e-6 with a batch size of eight and ten epochs. While the initial setting specified 10 epochs, the models often trained for fewer than 2, given the early stopping criterion set at 5 and evaluations being conducted every 500 steps. The optimization was done on the development set, and the final evaluations on the test set.

4.2. QA Pair Annotation with ChatGPT

We fine-tuned OpenAI's ChatGPT[8], specifically the GPT3.5 Turbo model, using our manually annotated examples to acquire additional training data for the token classifier. For this, we used the manually annotated Finnish QA dataset mentioned in Section 3.4. We had to exclude some texts from the fine-tuning process due to ChatGPT's token limit of 4,096, leaving us with a training set of 84 texts for the task. Similarly, our development and test sets were reduced to 43 and 56 texts, respectively (on evaluation, see 5.2 below).

For fine-tuning, we incorporated each training document into a sequence that included a system prompt[9], the text to be annotated, and the existing annotations labelled as Q, A, or O. We adjusted hyperparameters during this process, specifically the number of epochs (2–4) and the temperature affecting the model's output variability. Based on our tests using the development set, the optimal settings were 3 epochs with a temperature of 0.0.

4.3. Question and Answer Extraction with XLM-R

The task was modelled as a NER-style token classification, where each token is labelled as Q, A or O. Again, we used the base version of XLM-R and the Huggingface Transformers library. We experimented with different combinations of the datasets described in Sections 3.4 and 3.3. Long texts were truncated due to the 512 token limit.

We optimized the learning rate using a grid search with values of {1e-5, 4e-6,5e-6, 7e-5, 8e-6}. The other hyperparameters were a batch size of eight and ten epochs. Early stopping was set at five and evaluations were conducted either every

[6] https://huggingface.co/xlm-roberta-base

[7] https://scikit-learn.org/stable/

[8] https://platform.openai.com/docs/guides/fine-tuning

[9] See Appendix A

25, 250 or 2500, depending on the size of the training set. With English training data, we used 2500, with Finnish data cleaned with ChatGPT, 250, and with only manually annotated Finnish, 25.

The hyperparameter selection was evaluated against the development sets, and the final models against the test sets, both consisting of the manually annotated English and Finnish datasets. After the evaluations, we chose the two best models to run inference on – one for English and one for Finnish.

4.4. Evaluation metrics

For evaluating the document-level QA identification (see Section 4.1), we used F1-score, accuracy, precision and recall calculated using the sklearn library[10].

For evaluating the identification of the possible questions and answers and their text spans in the documents previously identified as QA (see Section 4.3), we followed Rajpurkar et al. (2016) and the SQuAD dataset and used the macro averaged overlap F1-score. In addition, we also report the seqeval accuracy measure, available in the seqeval library [11]. The F1-score was calculated as the F1-score of the overlaps of the tokens between the predictions and the ground truth, with overlapping sections marked as true positives. This evaluation was done on the raw predictions, as opposed to the post-processed QA pairs (see below), to specifically evaluate the token classifier's performance, not the post-processing step.

To evaluate our final step, post-processing and pairing up the questions and answers, we manually evaluate a sample of extracted QA pairs. The pairs are evaluated using three metrics: noisiness (remaining dates, other errors), sufficiency of the answer (do the question and answer form a coherent pair), and the presence of context required to understand the pair. Each metric was scored as either 0 (no error) or 1 (error), except noisiness, where pairs with minor errors (a few characters) were scored 0.5. Evaluations were done by the annotators of the curated datasets, to preserve consistency.

5. Evaluation

5.1. QA Document Identification

The best results for the QA document-level identification as well as the class-specific scores are found in Table 4.

The overall micro-averaged F1-score taking into account the class imbalance is high, 0.98, and also macro-averaged F1 is 0.79. However, the class-specific scores reflect the difficulty of the task, the F1-score being 0.60 for the QA class. In particular, the recall is lower than the precision—this can, however, be an advantage for ensuring the higher quality of the documents for further steps.

	F1	Precision	Recall
QA	0.60	0.82	0.47
Not QA	0.99	0.99	1.00
	F1-micro	F1-macro	F1-weighted
Avg.	0.98	0.79	0.98

Table 4: QA document classifier performance.

	Accuracy	Overlap F1
Evaluation 1	0.67	
Questions		0.67
Answers		0.73
Evaluation 2	0.76	
Questions		0.74
Answers		0.82
Evaluation 3	0.69	
Questions		0.55
Answers		0.69

Table 5: ChatGPT performance in cleaning QA pairs from web documents. Evaluation 1: 68 docs, zero F1 for texts beyond ChatGPT token cap; Evaluation 2: 56 docs, excluding overlong texts; Evaluation 3: Texts with at least one QA annotation in both manual and ChatGPT sets.

5.2. QA Pair Annotation with ChatGPT

In Table 5, we report three evaluations of ChatGPT's performance in annotating QA pairs from Finnish documents. In Evaluation 1, we evaluate against the entire test set (68 texts), assigning a zero F1-score for the 12 texts that were too long for ChatGPT to annotate (see Section 4.2). In Evaluation 2, a test set of 56 texts was used, omitting the 12 overlong texts. Finally, in Evaluation 3, only the 33 texts with QA annotations in both manual and ChatGPT sets were included.

ChatGPT's performance in annotating clean QA pairs varies by the evaluation setting. Evaluations 2 and 3 seem most informative to us. Evaluation 2, with F1-scores ranging from 0.74 to 0.82, demonstrates the quality of the generated annotations because it excludes texts that ChatGPT couldn't process. This represents more accurately the task of providing additional training data than Evaluation 1, which is calculated over all documents, had F1-scores between 0.67 and 0.73. Evaluation 3 only includes texts with some human-annotated QA content and leaves out documents marked as empty. This highlights the ChatGPT's ability to filter out noise. The F1-scores show decent per-

[10] https://scikit-learn.org/stable/
[11] https://pypi.org/project/seqeval/

formance in cleaning questions (0.55) and slightly better results for answers (0.69).

5.3. Question and Answer Extraction with XLM-R

The performance of the token classifier on the manually annotated English and Finnish test sets can be found in Table 6. For English, the best model was trained using the Finnish human annotations and the ChatGPT-cleaned data, achieving an 88% accuracy. For Finnish, the best-performing model was trained with English semi-synthetic data, Finnish human annotations and the ChatGPT-cleaned data, reaching an accuracy of 85%. This model only slightly numerically outperformed the model trained solely with manually annotated Finnish data and Finnish data cleaned by ChatGPT in the overlap F1 measure.

As expected, increasing the amount of training data generally increased model performance. Even the small set of Finnish manual annotations brought improvements. However, in the case of the semi-synthetic English data built on top of LFQA, the advantage was marginal for Finnish, and for English, the performance decreased by a large margin. For Finnish, we believe the improvement was modest because the semi-synthetic dataset differs too much from the web QA pairs in the test set. For English, we suspect the dramatic drop is caused by the model adopting an English-specific QA structuring from the LFQA data, which adversely affected its ability to extract QA pairs from English web documents.

Finally, using data annotated by ChatGPT proved beneficial. Despite some errors in the annotations and the dataset's limited size, it brought clear improvements for both languages.

6. Resulting QA Dataset

6.1. Compiling the Dataset

First, we used the best-performing QA document classifier (see Section 5.1) to identify documents likely containing questions and answers within the web-scale datasets (see Section 3.1). Only documents labelled as QA with a probability of over 0.5 were included. This step made it possible for us to avoid processing the entire corpus with the token classifier, and focus solely on the documents that were likely to contain questions and answers. Table 7 lists the amount of the retrieved documents. Though a relatively small fraction of documents was retrieved, the actual number of documents obtained is still substantial due to the large sizes of the corpora.

Next, we extracted questions and answers from the documents predicted as QA utilizing the best-performing token classifiers (see Section 5.2). Each token was classified with label Q, A or O. Spans of label Q were identified as questions and similarly spans of label A as answers. In some cases, the predictions of the token classifier were fluctuating: labels were predicted for very short text spans, and adjacent tokens were predicted differently inside words or sentences. An example of this behaviour can be seen in Table 9. One potential cause for these errors, especially for Finnish, could be the XLM-R tokenizer splitting Finnish words into a high number of tokens.

Then, we post-processed the results of the token classifier. To combat the issue of fluctuating labels mentioned above, we experimented with two simple heuristics to clean questions and answers in the documents: averaging the predictions over sentences weighted with the prediction scores and combining short sections labelled as O with the surrounding question or answer spans. In our manual evaluation (see Section 4.4), the sentence-averaging method yielded better results. Aggregations were specifically done at the sentence level to address situations where one clause contains the actual question while the others contain important context. Particularly, we observed a considerable improvement in the Finnish pairs with the averaging method compared to the second approach. For English, the improvement was more subtle, and the performance of the second method was also satisfactory. We recognize that different post-processing strategies may be necessary for different languages.

Lastly, to pair the identified questions and answers, we assumed that a subsequent question and answer constitute a QA pair. If two questions or two answers appeared consecutively, we followed the approach of Kwong and Yorke-Smith (2009) and considered them as a continuation of the same item. However, the pairs are presented in a format from which the original divisions can be reconstructed. Finally, we discarded pairs where the length of either question or answer is fewer than 15 characters (1–3 Finnish words).

The evaluation of final QA pairs (using the sentence-averaging method) is presented in Table 10. Notably, there is variation between the different Finnish corpora: the more quality-controlled Parsebank yields better quality QA pairs on average. There are also noticeable differences between Finnish and English.

The numbers of the extracted clean QA pairs can be found in Table 8, which also shows the final size of the Turku WebQA dataset. From Tables 7 and 8 it can be seen that most of the documents that were extracted with the register model did contain QA pairs.

Train	Dev & Test	Accuracy	Question F1	Answer F1
Fi	Fi	0.68	0.57	0.49
En/LFQA + Fi	Fi	0.68	0.59	0.55
En/LFQA + Fi + ChatGPT	**Fi**	**0.85**	**0.82**	**0.75**
Fi + ChatGPT	Fi	0.85	0.78	0.76
En/LFQA + ChatGPT	Fi	0.82	0.74	0.73
En/LFQA	Fi	0.50	0.44	0.34
En/LFQA	En	0.29	0.21	0.21
En/LFQA + Fi	En	0.28	0.29	0.21
En/LFQA + Fi + ChatGPT	En	0.32	0.24	0.20
Fi	En	0.68	0.62	0.41
Fi + ChatGPT	**En**	**0.88**	**0.77**	**0.81**
En/LFQA + ChatGPT	En	0.31	0.22	0.24

Table 6: Results for the NER-style token classifier experiments, evaluated against the manually annotated test sets in Finnish and English. The emphasized models performed the best.

Dataset	QA labelled docs	Proportion
Parsebank	31,654	0.48%
mC4-Fi	66,134	0.41%
CC-Fi	212,604	0.53%
Falcon	82,261	1.03%

Table 7: Documents predicted as QA by the register model. Proportion refers to the proportion of the QA-labelled documents in the full dataset.

Dataset	Documents	QA-pairs
Parsebank	25,101	30,106
mC4-Fi	45,498	71,406
CC-Fi	117,801	135,339
Finnish Total	188400	236,851
Falcon RefinedWeb	49,028	87,049

Table 8: Extracted QA documents and clean pairs after filtering low-quality pairs in the final datasets.

6.2. Analyzing the QA Pairs

To gain insight into the contents of the retrieved QA pairs, we first ran a simple topic modelling solution on the dataset using Gensim's LdaModel[12] (for the parameters, see Appendix B). Selected topics, their keywords, as well as example QA pairs are given in Appendix C. These topics cover a range of themes commonly found in QA forums and other web pages with questions and answers, demonstrating that our pipeline maintains the diversity found in the source corpora.

Second, we manually analyzed examples of the QA pairs in the final dataset, illustrated in Table 11. These samples show that, for the most part, the QA pairs are clean and the answers align well with their corresponding questions. The pipeline can even merge two adjacent questions to create a more comprehensive question item, as seen in the mC4-Fi example, and pair it with the appropriate answer. Additionally, the model was also able to extract the related context given before the question, as seen in the CC-FI example.

To demonstrate the drawbacks of the pipeline, Table 11 also features a QA pair we consider as noisy. This can be observed in the initial part of the Parsebank example where the date was mistakenly incorporated into the question by the QA extractor. Despite this error, the overall quality of the question remains acceptable.

7. Conclusion

In this paper, we have presented a resource-efficient method for developing QA datasets by utilizing web register identification to harvest documents with questions and answers from web-scale data, followed by machine learning techniques to extract the actual QA pairs from these documents.

Our evaluations emphasize the importance of domain-specific data in effectively training models to extract clean QA pairs: using only the English LFQA dataset resulted in low performance. Furthermore, generative models like ChatGPT can help with annotation, reducing the human effort needed to create these domain-specific datasets.

The pipeline consists of three steps: Firstly, document classification is used to refine our web-scale corpora to those likely containing questions and answers. Secondly, the token classifier extracts the questions and answers as spans within the documents. Lastly, our post-processing step aggregates the token classifier's results and pairs the questions and answers together.

Overall, our methodology effectively retrieves a substantial number of QA pairs with minimal noise and a variety of topics. As an outcome, we release the first Finnish, non-synthetic non-machine translated QA dataset, Turku WebQA. This openly available dataset consists of 237,000 QA pairs and is freely accessible at `https://github.com/TurkuNLP/register-qa/tree/main/Turku-WebQA`.

[12] `https://radimrehurek.com/gensim/`

O: Kysymykset ja vastaukset 14 kysymystä Hei, Ottaisi Q: tko O: yhteyttä Q: laitteiden O: tark Q: kojen O: mallimerkin Q: töjen O: kanssa sähköpostitse osoitteeseen email@example.com
O: Questions and answers 14 Questions Hi, Would you mi Q: nd O: contacting me Q: about the devices' O: ex Q: act O: model na Q: mes O: by email at email@example.com

Table 9: A problem case with our token classifier model on a Finnish document. The prediction fluctuates between question (Q) and other (O). English translation by us.

Language	Source	Noisy artefacts	Insufficient Answer	Missing context
Fi	Total (N=73)	0,29	0,22	0,08
	CC-Fi (N=25)	0,36	0,22	0,03
	mC4-Fi (N=25)	0,28	0,28	0,14
	Parsebank (N=22)	0,23	0,14	0,07
En	Falcon (N=22)	0,17	0,07	0,10

Table 10: Results of our manual evaluation on the extracted QA pairs. Results averaged over two evaluators. Finnish total is micro averaged over Finnish corpora.

Source	Question	Answer
CC-FI	Eli, tossa kun selailin noita prosesseita, pisti silmään sellanen kun iexplore.exe - muistia 58115Kt, isoimmalla. Niitä siis on 6. Apuja ?	explorer nähtävästi käsittelee jokaisen välilehden omana prosessinaan, jos yksi välilehti kaatuu niin se ei kaada kaikkia muitakin..
	So, as I was browsing the processes, I noticed something called iexplore.exe - the largest takes up 58115Kt of memory. There are 6 of these. Help ?	explorer apparently handles every tab as a separate process, so that if one of them crashes it won't crash every one of them..
mC4-FI	Miksi järvien ja lampien vesi on vähäsuolaista?Eikö niistä haihdu vettä niinkuin meristäkin?	Järvien ja merien suolaisuuden eroista meiltä on kysytty ennenkin. Laitan Lähteitä ja lisätietoja -kenttään linkin vanhaan vastaukseen.
	Why do lakes and ponds have low salinity?Does the water not evaporate the same way as in the ocean?	Salinity of lakes and ponds has been covered on this page before. I will link you the old answer and in the "Sources and More Information" box.
Parsebank	3.2.2013 Koska se pelaajaesittelyvideo tulee tänne?	Videosalissa on nähtävillä ko. video. Päivityksiä pelaajistoon on tullut sitten edellisen version, mutta korjailemme taas jossain vaiheessa asian kuntoon.
	Feb. 3rd 2013 When will you upload the player introduction video?	The video in question is available in the videohall. We've had player changes since the lastest version, but we'll be fixing this at some point.
Falcon RefinedWeb	Can someone help me plan the care for a child aged 3 for a full day at nursery?	Let him play educational toys like number blocks, Lego, memory card games.

Table 11: Examples of cleaned QA pairs from the web-scale datasets. These represent a shorter variety of cleaned pairs, chosen for ease of visualisation. Original spelling mistakes reflected in the translations.

Looking ahead, as our process should work on any language that is supported by a multilingual masked language model and a web-scale dataset, extending this approach to further low-resource languages with even fewer resources than Finnish could bring great benefits to the research community. Furthermore, using the new QA dataset for fine-tuning is a natural next step, as in the era of generative language models QA datasets play a significant role in LLM training.

8. Acknowledgements

We wish to acknowledge FIN-CLARIAH (Common Language Resources and Technology Infrastructure), and CSC – IT Center for Science for compu-

tational resources.

This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Academy of Finland under grant numbers 358720 and 331297.

9. Limitations

The study has at least the following limitations:

- We focus on two languages only. It is possible that some parts of the pipeline do not work on a multilingual scale. However, excluding the final post-processing (which we found to have language-dependent aspects) we do not think that this is very likely.
- The pipeline's overall performance could be better. Especially the QA document retrieval recall leaves room for improvement, leading to the exclusion of many potential QA pairs.
- Due to the transformer's limitation of processing only 512 tokens simultaneously we might be missing some questions and answers from the documents, or they might be incomplete.
- Despite the relatively high QA pair extraction performance, the resulting QA pairs do still have some noise, as illustrated in Table 11.

10. Bibliographical References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Franck Dernoncourt Amir Pouran Ben Veyseh, Viet Dac Lai and Thien Huu Nguyen. 2022. BehanceQA: A New Dataset for Identifying Question-Answer Pairs in Video Transcripts. In *Proceedings of LREC*.

Douglas Biber and Jesse Egbert. 2016. Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 2(1):3–36.

Vladimir Blagojevic. 2022. Long-form qa beyond eli5: an updated dataset and approach. *towardsdatascience.com/long-form-qa-beyond-eli5-an-updated-dataset-and-approach-319cb841aabb*.

B. Barla Cambazoglu, Mark Sanderson, Falk Scholer, and Bruce Croft. 2021. A review of public datasets in question answering research. *SIGIR Forum*, 54(2).

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. 2008. Finding question-answer pairs from online forums. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 467–474, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Sardana Ivanova, Fredrik Andreassen, Matias Jentoft, Sondre Wold, and Lilja Øvrelid. 2023. NorQuAD: Norwegian question answering dataset. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 159–168, Tórshavn, Faroe Islands. University of Tartu Library.

Valentin Jijkoun and Maarten de Rijke. 2005. Retrieving answers from frequently asked questions pages on the web. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05,

page 76–83, New York, NY, USA. Association for Computing Machinery.

Dmytro Kalpakchi and Johan Boye. 2023. Quasi: a synthetic question-answering dataset in Swedish using GPT-3 and zero-shot learning. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 477–491, Tórshavn, Faroe Islands. University of Tartu Library.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The GINCO training dataset for web genre identification of documents out in the wild. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1584–1594, Marseille, France. European Language Resources Association.

Helen Kwong and Neil Yorke-Smith. 2009. Detection of imperative and declarative question-answer pairs in email conversations. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI'09, page 1519–1524, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ilmari Kylliäinen and Roman Yangarber. 2023. Question answering and question generation for Finnish. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 529–540, Tórshavn, Faroe Islands. University of Tartu Library.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 57(3):1045–1079. Funding Information: Open Access funding provided by University of Turku (UTU) including Turku University Central Hospital. Funding was provided by Academy of Finland (Grant No. 331297), Emil Aaltosen säätiö, National Science Foundation (Grant No. 1147581). Publisher Copyright: © 2022, The Author(s).

Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika, and Sampo Pyysalo. 2022. Towards better structured and less noisy web data: Oscar with register annotations. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 215–221, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1.0: Korean qa dataset for machine reading comprehension. *ArXiv*, abs/1909.07005.

Risto Luukkonen, Ville Matti Johannes Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kristiina Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Le Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, and Samuel Antao. 2023. FinGPT: Large generative models for a small language. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Emil Nuutinen, Iiro Rastas, and Filip Ginter. 2023. Preprint https://www.researchsquare.com/article/rs-3251566/v1.

OpenAI. Chatgpt (september 25 version) [online]. 2023. Accessed on 2023-10-16.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the english web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*.

Samuel Rönnqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. 2021. Multilingual and zero-shot is closing in on monolingual web register classification. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 157–165, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

Lokesh Shrestha and Kathleen McKeown. 2004. Detection of question-answer pairs in email conversations. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 889–895, Geneva, Switzerland. COLING.

Valtteri Skantsi and Veronika Laippala. 2023. Analyzing the unrestricted web: The finnish corpus of online registers. *Nordic Journal of Linguistics*, page 1–31.

11. Language Resource References

Juhani Luotolahti, Jenna Kanerva, Veronika Laippala, Sampo Pyysalo, and Filip Ginter. 2015. Towards universal web parsebanks. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 211–220, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Guilherme Penedo and Quentin Malartic and Daniel Hesslow and Ruxandra Cojocaru and Alessandro Cappelli and Hamza Alobeidli and Baptiste Pannier and Ebtesam Almazrouei and Julien Launay. 2023. *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A. Appendix: ChatGPT system prompt

Given the following raw, web-scraped text, your task is to identify and label each segment as either a question, an answer, or other text (boilerplate, such as navigation text, links, headers, footers, titles, usernames, and any other non-content text). The output should be a JSON array. Each segment should be represented as “q”: “[question content]” for questions, “a”: “[answer content]” for answers, and “t”: “[other text]” for any other text. Do not remove, omit, translate, or alter any part of the input text, including single characters or words. Every symbol, character, and piece of content from the input must be precisely and completely represented in the output. Crucially, also escape sequences (such as newline characters \n) should be retained in the output.

Consecutive questions (“q” key), answers (“a” key) and boilerplate text (“t” key) should be grouped under a single “q”, “a” or “t” key, respectively. Newline characters should be grouped at the ends of questions or answers. Do NOT try to split the text into sentences.

A single raw text may contain either a single question, multiple questions, a question and an answer, or multiple questions and answers. If the text seems to have just an answer (without a question), leave it unannotated. It is also possible that the text does not contain any questions or answers. Some texts contain multiple forum comments, where there might be many persons involved in the discussion. Try to recognize all the questions and answers in the text, and label them as instructed above. Advertisements, such as sales ads, are not to be treated as questions; label them with “t”. If there are no questions or answers in the text, or just a single answer, wrap the entire text in [“t”: ...]. Note that an answer should never be present in the output without an associated question.

Remember that everything in the outer output array should be contained either in a “q”: ... array, “a”: ... array, or “t”: ... array. No plain text strings in the outer array.

The next message includes the input text (delimited with “‘). Please proceed with the labeling, ensuring that every single character from the input is included in the structured output without any omission or alteration.

B. Appendix: Gensim parameters

For topic modeling with Gensim’s LDAmodel (see Section 6.2), we used the following parameters:

- num_topics (the number of extracted topics

from the corpus): 20 (CC-FI, mc4-FI, Parsebank), 15 (Falcon RefinedWeb)

- num_words (the number of words output per topic): 5
- passes (the number of training iterations): 30

These settings were selected based on preliminary experiments within our text corpora.

C. Appendix: Sample QA pairs from selected topics

We present selected topics with illustrative QA pairs from our dataset, grouped by their source corpus (CC-FI, mc4-FI, Parsebank, and Falcon RefinedWeb). For each corpus, we list five topics along with two QA pairs per topic. Topic names are based on our interpretation of the keywords generated by Gensim. Finnish QA pairs and keywords have been translated into English. To save space, longer answers have been truncated, and newlines (\n) have been removed; no other modifications have been made to the original data.

C.1. Corpus: CC-FI

Accommodation Keywords: “hotelli” [*hotel*], “hotellin” [*hotel’s*], “km” [*kilometers*], “lähellä” [*near*],

- **Q**: Onko majoitusliikkeessä 082 - Island Splash tai sen lähiseudulla ravintoloita? [*Are there any restaurants at or near the accommodation 082 - Island Splash?*] **A**: Kyllä. Läheisiin ravintoloihin kuuluu Mango’s Café & Grill (5 minuutin kävelymatkan päässä). [*Yes. Nearby restaurants include Mango’s Café & Grill (a 5-minute walk away).*]
- **Q**: Salliiko Domaine du Clos Fleuri - Spa lemmikit? [*Does Domaine du Clos Fleuri - Spa allow pets?*] **A**: Lemmikkejä ei valitettavasti sallita. [*Unfortunately, pets are not allowed.*]

Applications and exams Keywords: “kirjallisesti” [*in writing*], “vaatimukset” [*requirements*], “viimeinen” [*last*],

- **Q**: Milloin on viimeinen hakupäivä talvikurssille 2019? [*When is the last application deadline for the winter course 2019?*] **A**: Viimeinen hakupäivä on 30.2019. Lähetä hakemuksesi täältä » [*The last application deadline is January 30, 2019. Submit your application here »*]
- **Q**: Mitä teoriakoe sisältää? [*What does the theoretical exam include?*] **A**: Liikennetilannetehtäviä. Monivalintatehtäviä. Ikäpoikkeusluvalla B-luokan ajo-oikeutta suorittava saa mennä kokeeseen jo 17 vuotiaana [*Traffic situation tasks. Multiple choice questions. Individuals with an exemption for age can take the B-class driving test at the age of 17.*]

Businesses Keywords: “myynti” [*sales*], “oy” [*Ltd.*], “tuotteilla” [*with products*], “yrityksen” [*company’s*]

- **Q**: Mikä on yrityksen JH-Suunnittelu Oy Uusimaa verkko-osoite (URL)? [*What is the website address (URL) for the company JH-Suunnittelu Oy Uusimaa?*] **A**: Verkkosivusto yritykselle JH-Suunnittelu Oy Uusimaa on www.jhsuunnittelu.fi [*The website for the company JH-Suunnittelu Oy Uusimaa is www.jhsuunnittelu.fi.*]
- **Q**: Mikä on yrityksen JT Thermo Steel Oy vuosittainen myynti? [*What is the annual sales of the company JT Thermo Steel Oy?*] **A**: Yrityksen JT Thermo Steel Oy vuosimyynti on noin EUR 5 000,00. [*The annual sales of the company JT Thermo Steel Oy is approximately EUR 5,000.00.*]

Car maintenance Keywords: “ajaa” [*to drive*], “auto” [*car*], “moottorin” [*engine’s*], “vika” [*fault*]

- **Q**: Lisääkö auton vakionopeuden säätimen käyttö polttoaineen kulutusta? Autoni on VW Golf Plus 1,6 FSI. [*Does using the cruise control increase fuel consumption? My car is a VW Golf Plus 1.6 FSI.*] **A**: Periaatteessa kyllä. Mäkisellä tiellä se painaa kaasua taloudellisen ajon kannalta väärissä paikoissa. Tasaisella säädin pitää nopeuden tasaisena, mikä vähentää kulutusta, koska turhat nopeusmuutokset jäävät pois. [*In principle, yes. On hilly roads, it may press the accelerator in the wrong places for economical driving. On flat terrain, the controller keeps the speed constant, which reduces consumption because unnecessary speed changes are avoided.*]
- **Q**: Seat v01 starttaa mutta sammuu samantien. Kaasu pohjassa tartatessa kovat kierrokset sekunnin ajan jonka jälkeenn sammuu. [*Seat v01 starts but immediately shuts off. Holding the gas pedal down, it revs up for a second before stalling.*] **A**: Viittaisi siihen, että bensapumppu ei käy moottorin käydessä, mutta hyräyttää kuitenkin järjestelmään paineen käynnistettäessä.. [*This suggests that the fuel pump is not running when the engine is running, but it*

does generate pressure in the system when started.]

Social security Keywords: “hakea” [*to apply*], “kela” [*Social Insurance Institution of Finland*], “maksetaan” [*is paid*], “työnantaja” [*employer*]

- **Q**: Onko minulla oikeus työttömyyskorvaukseen lomauttamisen ajalta? [*Do I have the right to unemployment benefits during a layoff?*] **A**: Lomautuksen ajalta maksetaan työttömyyskorvausta omavastuuajan jälkeen. Koronatilanteen vuoksi on kuitenkin säädetty määräaikainen laki, jonka perusteella omavastuuajalta kuitenkin maksetaan työttömyyspäivärahaa, mikäli ensimmäinen omavastuupäivä on 16.– 6.2020. [*During a layoff, unemployment benefits are paid after the waiting period. However, due to the coronavirus situation, a temporary law has been enacted, under which unemployment benefits are paid for the waiting period if the first waiting day falls between June 16 and 6, 2020*.]
- **Q**: Saanko opiskeluun opintotukea? [*Can I receive student financial aid for studying?*] **A**: Avoimen AMKin opiskeluun ei voi saada opintotukea eikä muitakaan opintososiaalisia etuja, kuten ateriatukea tai matka-alennuksia. [*Open University of Applied Sciences studies are not eligible for student financial aid or any other student social benefits, such as meal subsidies or travel discounts*.]

C.2. Corpus: mc4-Fl

Childcare Keywords: “ihmiset” [*people*], “lapsi” [*child*], “lapsen” [*child’s*], “kanssa” [*with*]

- **Q**: Mikä on Tanssila? [*What is Tanssila?*] **A**: Tanssila on lasten- ja nuorten tanssiin keskittynyt tanssiateljé. Koulumaisuuden sijaan on etusijalla luovuus, esiintyminen sekä itse tekemällä tai paremmin sanottuna, itse tanssimalla oppiminen. [*Tanssila is a dance studio focused on children and youth dance. Instead of formality, emphasis is placed on creativity, performance, and learning through doing or, more precisely, learning through dancing itself*.]
- **Q**: Eikö työnantajalla ole tarjota sellaista osaaikatyötä, jota pystyt tekemään? [*Doesn’t the employer have any part-time work available that you could do?*] **A**: Jos olet liittynyt ammattiliittoon, niin mene puhumaan pääluottamusmiehen kanssa. Jos et ole liittynyt lue Kunnallisen esimiehen työsuhdeopasta. [*If you are a member of a trade union, talk to the chief shop steward. If you are not a member, read the Municipal Supervisor’s Employment Guide*.]

Christian faith Keywords: “jeesus” [*Jesus*], “jumala” [*God*], “raamatun” [*Bible’s*]

- **Q**: Kuka on Pyhä Henki? [*Who is the Holy Spirit?*] **A**: Pyhän Hengen identiteetistä on monia väärinkäsityksiä. Joidenkin mielestä Pyhä Henki on mystinen voima. Toisten mielestä Pyhä Henki on persoonaton voima, jonka Jumala antaa Kristuksen seuraajien käyttöön. [*There are many misconceptions about the identity of the Holy Spirit. Some think the Holy Spirit is a mystical force. Others believe the Holy Spirit is an impersonal force that God gives for the use of Christ’s followers*.]
- **Q**: Keitä ovat kerubit? Ovatko kerubit enkeleitä? [*Who are the cherubim? Are cherubim angels?*] **A**: Kerubit ovat enkeliolentoja, jotka palvovat ja ylistävät Jumalaa. He laulavat ylistystä Jumalalle ja muistuttavat meitä Jumalan majesteettisuudesta, kirkkaudesta ja läsnäolosta. [*Cherubim are angelic beings who worship and praise God. They sing praises to God and remind us of the majesty, glory, and presence of God*.]

Courses Keywords: “kielen” [*language’s*], “koulutuksen” [*education’s*], “kurssin” [*course’s*], “peruuttaa” [*to cancel*]

- **Q**: Miten ilmoittaudun? [*How do I register?*] **A**: www-sivuillamme on lomake, jonka täyttämällä kurssille voi ilmoittautua. [*On our website, there is a form that you can fill out to register for the course*.]
- **Q**: Kuka on järjestelmänvalvojani? Miten poistan tai arkistoin kurssin tai kumoan arkistoinnin? [*Who is my system administrator? How do I delete or archive a course or undo archiving?*] **A**: Lisätietoja on kohdassa Kurssin arkistoiminen ja poistaminen. [*More information is available in the section Archiving and Deleting a Course*.]

Devices Keywords: “laite” [*device*], “laitteen” [*device’s*], “toimii” [*works*]

- **Q**: Mikä on puhelimen tukiasema? [*What is a phone base station?*] **A**: Puhelimen tukiasema on laite, johon puhelin muodostaa yhteyden. [*A phone base station is a device to which the phone establishes a connection*.]
- **Q**: Toimiiko Shield näissä kaikissa? [*Does Shield work in all of these?*] **A**: Kyllä. Shield toimii sekä PC että Mac-tietokoneissa. [*Yes. Shield works on both PC and Mac computers*.]

Water and plumbing Keywords: “paine” [*pressure*], “piipun” [*pipe’s*], “vesi” [*water*]

- *Q*: Mikä on veden pH ja veden dH? [*What is the pH and dH of water?*] **A**: Veden pH tarkoittaa veden happamuutta ja dH veden kovuutta. Kulutukseen jaettavan veden pH on alle tai hieman yli 8, dH on 0,46 mmol/litrassa (pehmeä vesi). Asiakkaille jaettava veden happamuus ja kovuus korjataan alkaloinnin avulla suositusten mukaiseksi. [*The pH of water refers to its acidity, and dH refers to the water hardness. The pH of the water distributed for consumption is slightly below or above 8, and the dH is 0.46 mmol/liter (soft water). The acidity and hardness of the water distributed to customers are adjusted to comply with recommendations using alkalization.*]

- **Q**: Mikä on lauhdutinkuivain? [*What is a condenser dryer?*] **A**: Lauhdutinkuivain on kuivain, jossa kondensoitunut vesi kerääntyy säiliöön. [*A condenser dryer is a dryer in which condensed water accumulates in a reservoir.*]

C.3. Corpus: Parsebank

Products Keywords: “koira” [*dog*], “tuote” [*product*], “tuotteen” [*product’s*], “tuotteita” [*products*]

- **Q**: Missä maissa Marimekon tuotteet valmistetaan? [*In which countries are Marimekko products manufactured?*] **A**: Valtaosa Marimekon tuotteista valmistetaan Euroopassa ja noin kolmannes sen ulkopuolella. Korkealuokkaista käsityö- ja valmistusosaamista on eri puolilla maailmaa, ja Marimekon hankinnan lähtökohtana on löytää aina kullekin tuotteelle sopivin ja osaavin valmistaja. [*The majority of Marimekko products are manufactured in Europe, with about one-third produced outside of Europe. High-quality craftsmanship and manufacturing expertise are found in various parts of the world, and Marimekko’s procurement approach is to always find the most suitable and skilled manufacturer for each product.*]

- **Q**: Mistä Lactrasen sisältämä laktaasientsyymi on peräisin [*Where does the lactase enzyme contained in Lactrase come from?*] **A**: Lactrase valmistetaan Suomessa. Maltodekstriini on peräisin maissista. [*Lactrase is manufactured in Finland. Maltodextrin is derived from corn.*]

STDs Keywords: “hiv” [*HIV*], “tarttua” [*to transmit*], “tartunnan” [*infection’s*], “välityksellä” [*via*], “seksitaudit” [*STDs*]

- **Q**: Miten klamydia tarttuu? [*How is chlamydia transmitted?*] **A**: Klamydia tarttuu suojaamattoman seksin välityksellä. Klamydiabakteeri voi joutua silmiin käsien välityksellä. Vastasyntynyt voi saada tartunnan synnytyksen yhteydessä. [*Chlamydia is transmitted through unprotected sex. The chlamydia bacteria can also enter the eyes through hand contact. A newborn can acquire the infection during childbirth.*]

- **Q**: Olin baarissa, vihainen mies sylki minua suoraan silmiin sekä suuhun, voiko tarttuva olla nyt minulla?! (Hepatiitti/hiv) [*I was at a bar, an angry man spat directly into my eyes and mouth, could I be infected now?! (Hepatitis/HIV)*] **A**: Hei. Kuvailemassasi tilanteessa ei ole hepatiitti eikä hiv riskiä. [*Hello. In the situation you described, there is no risk of hepatitis or HIV.*]

University admissions Keywords: “hakea” [*to apply*], “koulutus” [*education*], “opiskella” [*to study*], “suorittaa” [*to complete*], “yliopiston” [*university’s*]

- **Q**: Olen abi, voinko hakea yhteishaussa? [*I’m a high school senior, can I apply through joint application?*] **A**: Voit hakea yhteishaussa lukiopohjaiseen koulutukseen, mutta et peruskoulupohjaiseen koulutukseen. Sama koskee kaikkia lukion oppimäärän suorittaneita. [*You can apply through joint application for education based on high school education, but not for education based on basic education. The same applies to all those who have completed the high school curriculum.*]

- **Q**: Mitä vaaditaan arkeologian opiskelijalta ja missä sitä voi opiskella? Mitä Turun yliopiston arkeologian pääsykokeessa vaaditaan ? [*What is required from an archaeology student and where can one study it?*] **A**: Hei Tuula, arkeologiaa voit opiskella kolmessa yliopistossa Suomessa, jotka ovat: Helsingin yliopisto Oulun yliopisto Turun yliopisto [*Hi Tuula, you can study archaeology at three universities in Finland, which are: University of Helsinki, University of Oulu, and University of Turku.*]

Windows devices Keywords: “laite” [*device*], “laitteen” [*device’s*], “puhelimen” [*phone’s*], “windows” [*Windows*]

- **Q**: Mitä Windows 10 -päivityksestä on syytä tietää? [*What should I know about the Windows 10 update?*] **A**: Microsoft tarjoaa Windows 10:n ilmaisena päivityksenä sihen oikeutettuihin Windows 7-, Windows 8.1-

ja Windows Phone 8.1 -laitteisiin. Se on saatavilla 29. heinäkuuta 2015 alkaen. [*Microsoft offers Windows 10 as a free upgrade to eligible devices running Windows 7, Windows 8.1, and Windows Phone 8.1. It has been available since July 29, 2015.*]

- **Q**: Entä jos ostan tietokoneen, jossa on Windows 8 ? Voinko päivittää Windows 8.1 :ksi? [*What if I buy a computer with Windows 8? Can I upgrade to Windows 8.1?*] **A**: Kyllä. Jos käytössäsi on Windows 8 , saat ilmaisen Windows 8.1 -päivityksen Windows-kaupan kautta. [*Yes. If you have Windows 8, you can get the free Windows 8.1 update through the Windows Store*.]

Work life Keywords: "päätyy" [*ends up*], "työnantajalle" [*to the employer*], "työntekijän" [*employee's*], "rinnalla" [*alongside*]

- **Q**: Onko minulla oikeus lomarahaan? [*Do I have the right to holiday pay?*] **A**: Lomaraha ei ole työntekijän lakisääteinen oikeus, vaan sen maksaminen perustuu työehtosopimuksen määräykseen tai työpaikan käytäntöön. Lomaraha on yleensä 50 % vuosilomapalkasta. Lomarahan maksamisen ehtona on usein se, että työntekijä aloittaa loman sovittuna ajankohtana ja palaa lomalta takaisin töihin. [*Holiday pay is not a statutory right of the employee; rather, its payment is based on the provisions of the collective agreement or the practice of the workplace. Holiday pay is typically 50% of the annual holiday salary. The condition for paying holiday pay is often that the employee starts the holiday at the agreed time and returns to work after the holiday.*]

- **Q**: Miten ongelman voi korjata? [*How can the problem be fixed?*] **A**: Maski on huonosti säädetty tai istuu huonosti . Suosittelemme, että säädät maskin uudelleen sen käyttöohjeessa olevien sovitusohjeiden mukaisesti. [*The mask is poorly adjusted or poorly fitted. We recommend readjusting the mask according to the fitting instructions provided in its manual*.]

C.4. Corpus: Falcon RefinedWeb

Education Keywords: "program", "school", "students"

- **Q**: If I am planning to complete my degree through an online program, am I eligible to apply for a CKMEF? **A**: No. CKMEF does not provide support for students who pursue an education or a special program solely online. You may take a combination of both traditional classroom courses and online courses.

- **Q**: Do I need to be a member of the Defence Force to attend? **A**: No, all of UNSW Canberra's Professional Education courses are open to the public.

Healthcare Keywords: "medical", "pain", "treatment"

- **Q**: What is the difference between curative care and palliative care? **A**: Curative care involves treatment to cure or eradicate disease. Palliative care occurs when a cure is no longer possible.

- **Q**: What conditions can benefit from neural therapy? **A**: Research has shown that neural therapy can be effective in lower back pain, lateral epicondylitis (tennis elbow), and fibromyalgia in combination with an exercise program.Neural therapy has also been shown to significantly assist with chronic postoperative pain.

Measurements Keywords: "1", "2", "3", "4" "long"

- **Q**: What is half of 3/4 measure? **A**: Half of 3/4 cup would be 1/4 cup plus 2 tablespoons, or 6 tablespoons.

- **Q**: Side A of an equilateral triangle is 3 units long. What is the individual length of each of the two other sides? **A**: As stated above, an equilateral triangle has three equal sides and three equal angles.So, if one side of an equilateral triangle is 3 units long, the other sides must also be 3 units long.

Prices and valuation Keywords: "cost", "price", "rate", "value"

- **Q**: What is the price range of your apartments? **A**: The price range at Pleasant Springs is $1009-$1499.

- **Q**: What is the Cost? **A**: There is no cost. There is no paywall.

Transportation Keywords: "airport", "bus", "flight", "travel"

- **Q**: How many flights fly from Vadodara to Agartala on the daily basis? **A**: Around 30 flights are flying daily from Vadodara to Out of which 30 are connecting flights on this route. Some major airlines between this route are Air India and Indigo .

- **Q**: When does the first Vadakara to Chennai bus leaves for the day? **A**: The first bus for Vadakara to Chennai bus route leaves at 15:15. It is a volvo bus and fare for this bus is ₹1384.