

Building Question-Answer Data Using Web Register Identification

Anni Eskelinen, Amanda Myntti, Erik Henriksson,
Sampo Pyysalo and Veronika Laippala
TurkuNLP
University of Turku



LREC-COLING 2024



UNIVERSITY
OF TURKU



/ Intro

Finnish is a semi low-resource language compared to English

No (big) dataset for question-answering (that is not based on machine translation)

- Objective: Question-answer (QA) data for Finnish (& English)
 - Dataset(s) that feature QA pairs from the web
 - Annotated test set(s)
- Pipeline approach:
 - One model for QA document identification using web register (genre) identification
 - One model for QA extraction using token classification from the identified texts
 - Post-processing



/ Background

- Crowdsourcing, machine translation, sourcing QA pairs from e.g., Reddit
- Finnish:
 - Finnish machine translated SQuAD and SQuAD 2.0
 - Kylliäinen and Yangarber, 2023
 - Nuutinen et al., 2023
 - OpenAssistant dataset only has 138 Finnish messages
- QA pair extraction:
 - Utilizing HTML formatting and heuristics (Jijkoun and de Rijke (2005))
 - Part-of-speech tagging, regular expressions.. (Kwong and Yorke-Smith (2009) and Cong et al. (2008))
 - Token classification (Amir Pouran Ben Veyseh and Nguyen (2022))





/ Background

- Registers
 - Corpus of Online REgisters of English (CORE)
 - Biber and Egbert, 2016, Laippala et. al, 2023
 - Now for many other languages as well
 - 8 main categories
 - More subcategories
 - Hybrid documents
- Classifiers can reach 80% F1-score performance



Methods

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

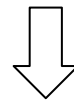
Table 9: A problem case with our token classifier model on a Finnish document. The prediction fluctuates between question (Q) and other (O). English translation by us.

Two models:

- Multi-class text classification model with two labels (QA / Not QA)
- Token classification with custom labels (Q, A, O)
- All models based on XLM-R base, transformers, pytorch

ChatGPT

- Data annotation



Post-processing

- Sentence-averaging (better for Finnish)
- Combining short sections labelled as O with the surrounding question or answer spans (worked for English)
- Discarding short pairs and combining subsequent questions or answers



/ Data for QA Document Identification

Register datasets

- English, Finnish, French, Swedish
- QA-related labels mapped under QA
 - Question-Answer Forum
 - FAQ about How-To
 - FAQ about Information
 - FAQ

	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.



/ Register data examples

NA (Narrative) SR (sports report)

QPR vs Southampton - Match preview and team news Friday 16 November 2012 The bottom two of the Premier League meet on Saturday, with Queens Park Rangers looking for their first win of the season at the 12th time of asking. Mark Hughes' side have picked up four draws so far, but have to go back 6 months for their last league win, at home to Stoke at the beginning of May. All is not lost just yet for R's fans - one win would put them two points behind 17th place Aston Villa. Southampton are sticking by the manager that led them to successive promotions, regardless of Premier League performances, yet Nigel Adkins will be desperately disappointed to have picked up just one solitary victory - now seven games back. Defeats to West Brom, Spurs and West Ham in recent weeks have kept the Saints at the foot of the table. This weekend QPR will have their squad depth tested - Stephane M'Bia is out suspended, while Fabio (hamstring), will miss the tie. The club await fitness news on Jamie Mackie, Kieron Dyer, Ryan Nelson and Park Ji-Sung. Djibril Cisse will start up-front, preferred to Bobby Zamora. Southampton have most of their first team members available, with just Frazer Richardson out with a thigh injury.



/ Register data examples

HI (how-to instructional) HT (how-to)

Do you have a hard water problem in your area? This simple test will tell you. The tester strip is totally harmless, it will measure the quantity of free calcium in your water supply. Dip the strip in a glass of water for 5 seconds. Shake off the excess water and leave for one minute, then see how many squares have changed colour. Each square represents 5 degrees of potential hardness. Whatever the reading, it will be the same after a Water-King is fitted, because none of the calcium is removed from the treated water. 1 degree of hardness = 18mg. of calcium carbonate per litre of water. If you have 10 degrees (2 squares) or over, you have a hard water and limescale problem.



/ Web datasets

Name	Documents
Parsebank	6,581,550
mC4-Fi	16,089,579
CC-Fi	40,074,961
Falcon RefinedWeb	(8M) 968,000,015



/ Data for QA Extraction 1/2

Semi-Synthetic English QA Pairs

LFQA

- 239,167 examples
- several QA subreddits: AskHistorian, AskScience, and Explainlikeimfive

-> turning into a token classification dataset

- Title + selftext = Question
- Best scored answer = Answer
- Some synthetic noise = Other

-> label each “word divided by space as either Q, A or O





/ Example

Title: since oil & water don't mix, how are essential oil soaks helpful?

Selftext: Doesn't the oil just sit on the surface, like it looks, reaching the intended body parts only in the small area that intersects with the top of the water? Or does it slowly mix with the water?

Subreddit: explainlikeimfive

Answers: { "a_id": ["e01xr6g", "e01z28r"], "score": [11, 6], "text": ["As far as I know there is no scientific proof that essential oils work anyway, but yes your skin can only absorb so much.", "Essential oils are worthless for everything but smelling good anyway, so adding water certainly doesn't improve anything"] }



/ Data for QA Extraction 2/2

Annotated datasets

- **Manually annotated data**
- **ChatGPT**

- Sampled from the web-scale datasets
- Uses the token classification labels Q, A, O
- 121 Finnish documents were double-annotated
- Agreement:
 - 0.85 Questions
 - 0.88 Answers

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.



/ QA Pair Annotation ChatGPT

- GPT3.5 Turbo model
- Fine-tuned using manually annotated examples (test set introduced before)
- System prompt, text to be annotated, labels to use

	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.



/ Example

"text": [{"t": ":"}, {"q": "I have an outdoor water faucet that was installed at time of construction 16 years ago. I has several manufactured holes all the way around the opening and when I connect a hose to the faucet water sprays out in all directions. Is there some kind of adapter I can attach to cover these holes or should I replace this construction faucet with a regular faucet?"}]



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.



/ QA Document Identification Results

Label	F1	Precision	Recall
QA	0.60	0.82	0.47
Not QA	0.99	0.99	1.00

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%

Label	F1-micro	F1-macro	F1-weighted
Avg.	0.98	0.79	0.98



/ QA Extraction Results

- Macro averaged overlap F1-score
- Seqeval accuracy

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.



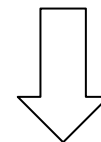
Final QA dataset

- Turku WebQA
 - Over 200,000 (Finnish) QA pairs
 - Wide variety of topics varying between corpora
 - E.g., accommodation, car maintenance, childcare, devices, products, education, transportation...

Example:

- How do I get the results of my exam?
- You will need to book a follow-up appointment with your family doctor or referring practitioner after the colposcopy exam. This is critical as there may be recommendations for future follow-up. The colposcopy doctors do not generally contact you personally with the results.

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%



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





/ Manual evaluation of final QA pairs

- Manual evaluation
 - Noisy artefacts
 - Insufficient Answer
 - Missing context
- Variation between different corpora

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.



/ Conclusions

TurkuNLP Huggingface has the models and final pairs
<https://huggingface.co/TurkuNLP>

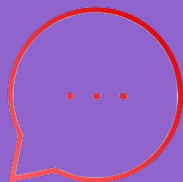
- QA register model for document identification
- QA extraction models for both Finnish and English
- Finnish QA pairs

Github includes the raw data as well as annotated data, English pairs and the codebase. <https://github.com/TurkuNLP/register-qa>

Limitations

- Poor recall for QA document identification
- Some noise remains
- Only 512 tokens
- Post-processing somewhat language specific





TURKUNLP
.ORG



UNIVERSITY
OF TURKU

Thank you for listening!