FIN-CLARIAH 2024/25

D:3.3.3 a. Interaction in web content

A case study

Authors:
Anni Eskelinen
Erik Henriksson
Amanda Myntti
Veronika Laippala
TurkuNLP

WP 3.3. Data Enrichment
D 3.3.3 Machine-learning-based enrichment of social media
a. Interaction in web content

Introduction

- Goal to get a form of interaction, question-answer data, for Finnish from noisy web data (web-scale corpora) using register identification and token classification
- Possible use case large language model (LLM) training/fine-tuning

Pipeline

1. **QA document identification from web-scale corpora with Register classifier**
   a. Trained using the CORE Corpora + XLM-R
2. **QA span extraction from predicted QA documents with NER-style token classifier**
   a. Trained XLM-R using synthetic and annotated QA pair training data
   b. ChatGPT and manual work for data annotation
3. **Post-processing**
   a. Combining short Q/A spans with surrounding ones
   b. Discarding short spans and duplicates, and combining subsequent questions/answers to form a pair

Final Data

CLEAN QA PAIRS

- Over 200,000 Finnish QA pairs from 180,000 QA documents
- Wide variety of topics
- Natural Finnish language use

Code available in GitHub, models and data in Huggingface (annotated data in GitHub).

What next?

LLM training: Co-operation with LLM people at TurkuNLP

Build new machine learning models to get meaningful data from social media: e.g., clustering and topic modeling to give further structure to the corpora, still figuring out the specifics!

WEB-SCALE CORPORA

Names	Documents
Finnish Parsebank	6,581,550
mC4-Fi	16,089,579
CC-Fi	40,074,961
Falcon RefinedWeb	8,000,000

Hugging Face

TurkuNLP **Turku-WebQA**

Q	Can someone help me plan the care for a child aged 3 for a full day at nursery?
A	Let him play educational toys like number blocks, Lego, memory card games.

Q	Pitääkö maksaa laskuja kirjasta, joka on jäänyt yliajaksi käyttöön, vaikka olisi alle 15-vuotias?
A	Alle 15-vuotiailta ei peritä myöhästymismaksuja. Kun käyt seuraavan kerran kirjastossa, voit pyytää virkailijaa poistamaan kortillesi kertyneet myöhästymismaksut.

Manual evaluation of pairs

	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