Text Filtering Classifiers for Medium-Resource Languages

Jón Friðrik Daðason, Hrafn Loftsson
Department of Computer Science, Reykjavik University, Iceland
{jond19, hrafn}@ru.is

Introduction

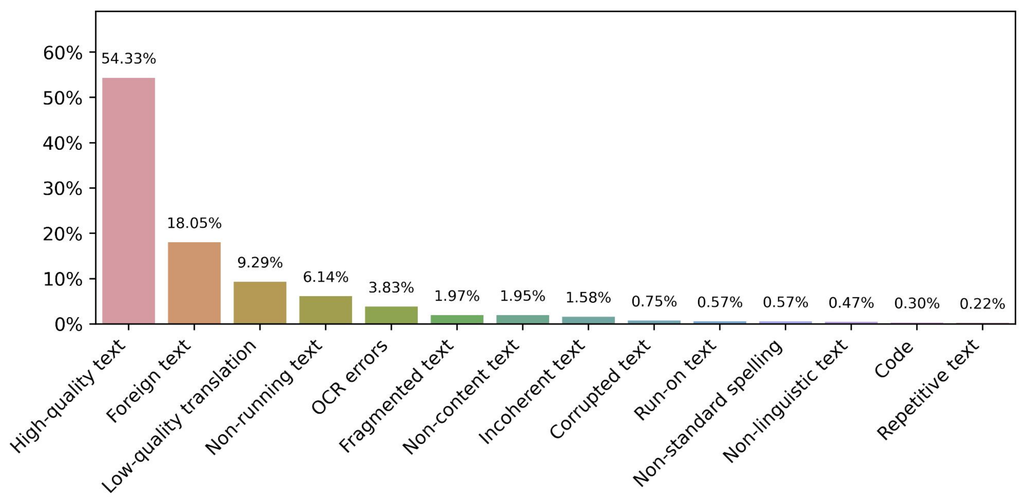
- We present TQ-IS, a new text quality dataset for Icelandic
- Three previously proposed text quality classifiers are evaluated on TQ-IS
- We use the same classifiers to filter web-crawled corpora for Icelandic, Estonian and Basque and measure their impact on downstream tasks

Text Quality Filtering

It has become standard practice to pre-train language models on large, web-crawled corpora. Such corpora are often noisy, containing a large amount of low-quality documents. It has been shown that filtering noisy pre-training corpora can improve the downstream performance of language models.

TQ-IS: An Icelandic Text Quality Dataset

- TQ-IS consists of 2,000 documents, sampled from Icelandic web-crawled corpora, which have been manually annotated with regard to text quality
- Within each document, low-quality text spans have been manually identified and labeled as one of 13 categories
- Documents have been annotated as low or high-quality
  - TQ-IS contains 1,000 examples of each category
- The ratio of non-space characters belonging to each low-quality category is shown in the graph below



Text Quality Classifiers

We evaluate three previously proposed classifiers:

- Perplexity-based classifier
  - If the perplexity of a document exceeds a predetermined threshold, then it is classified as low-quality and discarded
- Supervised classifier
  - Trained on a manually labeled text quality dataset
- Self-supervised classifier
  - Trained to discern between documents that originate from curated and web-crawled corpora

Filtering TQ-IS

We find that the supervised classifier significantly outperforms the other two when evaluated on TQ-IS.

Classifier	F1 score
Supervised classifier	99.01%
Perplexity-based classifier	94.48%
Self-supervised classifier	93.40%

Filtering Web-Crawled Corpora

- We use the three classifiers to filter the Icelandic, Estonian and Basque subsets of the mC4 corpus
- For each language, we supplement a high-quality corpus with filtered and unfiltered versions of the web-crawled corpus
- We pre-train an ELECTRA-Small language model on each resulting corpus and evaluate it on three downstream tasks
  - Part-of-speech (PoS) tagging, named entity recognition (NER) and dependency parsing (DP)
- Our results show that filtering does not significantly impact downstream results

	PoS			NER			DP		
Corpora	**IS**	**ET**	**EU**	**IS**	**ET**	**EU**	**IS**	**ET**	**EU**
HQ	**96.95**	97.93	**96.88**	**91.30**	**91.36**	**83.13**	84.79	88.38	84.31
+mC4	96.80	**97.93**	96.82	**91.08**	91.14	81.32	**84.89**	**88.66**	85.13
+mC4-PPL	**96.90**	**97.95**	96.84	**91.39**	**91.7**	**82.66**	84.75	**88.75**	**85.27**
+mC4-SC	96.86			**91.42**			**84.79**		
+mC4-SSC	96.85	**97.96**	**96.92**	**91.27**	91.01	**83.01**	**84.82**	88.44	85.03

Conclusions

- We release TQ-IS, an Icelandic text quality dataset, with an open license
  - https://github.com/jonfd/tq-is
- The supervised classifier obtains the highest F1 score on TQ-IS
- Text quality filtering doesn't appear to have a significant impact on downstream results for the languages in our evaluation
  - It is possible that the web-crawled corpora aren't large enough compared to the high-quality corpora they were supplemented with to have a great deal of impact
  - The models in this experiment may be more limited by their size than the quality of the pre-training data
  - However, we still reduced the size of the web-crawled corpora by half, which could make the pre-training process faster and more computationally efficient

Acknowledgements

This research was supported with Cloud TPUs from Google's TPU Research Cloud (TRC).