Text Filtering Classifiers for Medium-Resource Languages

Jón Friðrik Daðason, Hrafn Loftsson

Department of Computer Science, Reykjavik University, Iceland

LREC-COLING 2024

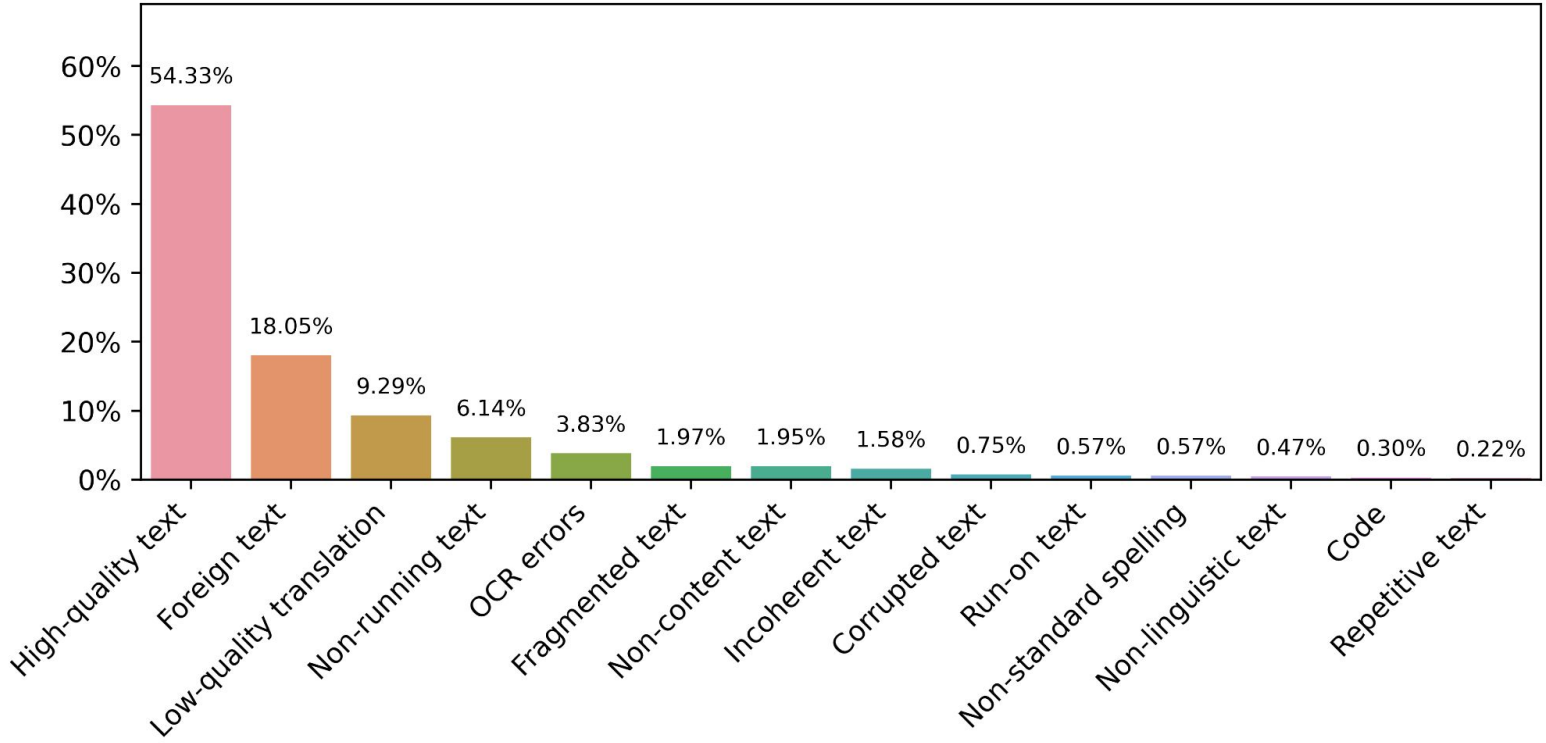
Text Quality Filtering

- Researchers increasingly rely on web-crawled corpora for pre-training
- These corpora often contain a great deal of low-quality text, such as:
  - HTML or JavaScript code fragments
  - Boilerplate text
  - Low-quality machine translations
  - Character encoding errors
- Filtering noisy corpora can improve downstream performance
  - Usually performed using heuristic rules or classifiers

TQ-IS: An Icelandic Text Quality Dataset

- TQ-IS is a new text quality dataset for Icelandic
  - Consists of 2,000 documents sampled from web-crawled corpora
  - The dataset has been manually annotated with regard to text quality
- Span-level annotations
  - Within each document, low-quality text spans have been annotated
  - Each low-quality span is labelled as one of 13 different categories
  - Includes foreign text, low-quality translations, non-linguistic text, etc.
- Document-level labels
  - Each document has been annotated as low or high-quality
  - TQ-IS contains 1,000 examples of each category

TQ-IS: Distribution of Text Span Categories



Text Quality Classifiers

- We evaluate three previously proposed text quality classifiers
  - Perplexity-based classifier
  - Supervised classifier
  - Self-supervised classifier
- First, we compare how well the classifiers perform on TQ-IS
- Next, we use the classifiers to filter web-crawled corpora for Icelandic, Estonian and Basque and compare the impact that each classifier has on three downstream tasks
  - Part-of-speech (POS) tagging
  - Named entity recognition (NER)
  - Dependency parsing (DP)

Perplexity-Based Classifier

- Based on the implementation by Wenzek et al. (2020)
- Train an n-gram language model on a high-quality corpus that has been processed by a subword tokenizer
- Web-crawled documents with a perplexity value that exceeds a predetermined threshold are classified as low-quality

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Supervised Classifier

- A document classifier trained on a manually labeled text quality dataset
- We pre-train an ELECTRA-Small model on a high-quality corpus and fine-tune it on TQ-IS

Self-Supervised Classifier

- Based on the implementation by Brown et al. (2020)
- Sample a large number of documents from two corpora
  - A high-quality, curated corpus
  - A noisy, web-crawled corpus
- We use an ELECTRA-Small model, pre-trained on a high-quality corpus, and fine-tune it to predict which corpus each document originated from
- The more confident the classifier is that a document originates from the web-crawled corpus, the more likely it is to be of low quality

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Results: Filtering TQ-IS

- We use 10-fold cross-validation to evaluate the perplexity-based classifier
- For the supervised classifier, we use 10-fold cross-validation to find optimal parameters, and evaluate it on a hold-out test set
- For the self-supervised classifier, we use 80% of TQ-IS to find optimal thresholds and evaluate it on the remaining 20%

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

Results: Downstream Performance

	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 present TQ-IS, a new text quality dataset for Icelandic
- The supervised classifier outperforms the others when evaluated on TQ-IS
- Filtering web-crawled corpora in Icelandic, Estonian and Basque did not significantly improve downstream results in our experiments
  - The impact of text quality filtering may be limited due to the size of the web-crawled corpora compared to the high-quality corpus
  - The small size of the models may be a larger bottleneck than the quality of the pre-training corpus
  - But even so, reducing the size of the pre-training corpus without degrading downstream results can result in a more computationally efficient pre-training process

Thank you

- We release the TQ-IS dataset with an open license
  - https://github.com/jonfd/tq-is
- Contact me at jond19@ru.is
- Acknowledgements
  - This research was supported with Cloud TPUs from Google's TPU Research Cloud (TRC).