A Warm Start and a Clean Crawled Corpus — A Recipe for Good Language Models

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Vilhjálmur Þorsteinnsson & Hafsteinn Einarsson

MIÐEIND EHF and UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
Summary and Main Contributions

Summary

- Developed a set of language models for Icelandic – both from scratch and based on a multilingual model
- Compared performance on downstream tasks across different base corpora and models

Contributions

a) Language models for Icelandic
b) Adaptations of LM for Icelandic with SOTA results
c) Icelandic WinoGrande dataset
d) Icelandic Common Crawl Corpus (IC3)
TLD (.is) Targeted Scraping of the Common Crawl

Icelandic is a medium resource language - how can we get more data?

The Common Crawl hosts a searchable index

a) Search for *.is – response contains byte ranges in dumps (huge files)
b) Fetch only data in byte ranges
c) Clean up and deduplicate data
Icelandic Corpora

News and legal IGC 8.2 GB

The Icelandic Common Crawl Corpus IC3 4.9 GB

Student thesis 2.2 GB

Medical 33 MB

Public Domain Ebooks 14 MB

Sagas 9 MB

Icelandic part of mC4 8.0 GB X

News 2015-2021 456 MB
Pretraining LMs for Icelandic

- **IceBERT**
  - Trained on all corpora using RoBERTa-base

- **IceBERT-IC3**
  - Only trained on IC3 – scraped and cleaned data

- **IceBERT-IGC**
  - Only trained on the curated gigaword corpus IGC

- **IceBERT-mC4-is**
  - Trained on the Icelandic part of mC4 – no extra cleaning

- **XLMR-IC3-7d**
  - Multilingual model used as warm start – trained on IC3

- **Random**
  - Trained on all corpora using RoBERTa-base
Part of Speech

Part-of-speech (POS) tagging

• Word type, declension, tense etc.
• Combined tags taken apart — multi label task
  • (word type, gender, case, etc)
• Example: Gamla konan ók grænni rútu

Data

• MIM-Gold (Loftsson et al. 2010, Developing a PoS-tagged corpus using existing tools)
Annotator disagreement in many of the errors.

- IceBERT: 98.33%
- IceBERT-IC3: 98.30%
- IceBERT-IGC: 98.27%
- IceBERT-mC4-is: 97.62%
- XLMR-IC3-7d: 98.20%
- Random: 90.27%
Named Entity Recognition

Named entity recognition (NER)

- The task of labelling entities in text
- **Example:** Katrín Jakobsdóttir (PER) vinnur í Stjórnarráðinu (ORG)

Data

- **MIM-GOLD-NER**
- Svanhvít Lilja Ingólfsdóttir, Ásmundur Alma Guðjónsson, Hrafn Loftsson
NER – F1-score

- **IceBERT**: 91.43%
- **IceBERT-IC3**: 91.29%
- **IceBERT-IGC**: 91.10%
- **IceBERT-mC4-is**: 89.57%
- **XLMR-IC3-7d**: 92.52%

This model has seen the most text, and probably the most entities.
Constituency parsing
  • Labeling sentence constituents

Data
  • GreynirCorpus (Miðeind)
Parsing – F1-score

- IceBERT: 90.02
- IceBERT-IC3: 89.37
- IceBERT-IGC: 89.66
- IceBERT-mC4-is: 88.60
- XLMR-IC3-7d: 89.01

*Has seen the least Icelandic data so performance is not surprising.*
Grammatical Error Detection

Grammatical error detection
  • Spelling
  • Grammar
  • Style

Data
  • Icelandic Error Corpus (IceEC)
  • Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, Xindan Xu
Grammatical Error Detection - F1-score

Largest model is by far the best for this task.
Coreference resolution

- **Example:** The cup could not fit into the suitcase because it was too big.

**Data**

- We translated a part of the WinoGrande corpus and adapted it for Icelandic

```json
{"qID": "3VP4ZJHPQH09K508N09089NH08-2", "sentence": "Porlákína sagði Agnetu að hún kemist ekki út að borða vegna ristilbólgu. _ sýndi því skilning.", "option1": "Porlákína", "option2": "Agnetu", "answer": "2"}
{"qID": "3VP4ZJHPQH09K508N09089NH08-2", "sentence": "Porlákína sagði Agnetu að hún kemist ekki út að borða vegna ristilbólgu. _ var afsókuð.", "option1": "Agnetu", "option2": "Porlákína", "answer": "2"}
{"qID": "3M47JKRCKZJDSUGN4KNOMEAQ86S-2", "sentence": "Jón elskaði að borða sushi en Bjarti fannst það ógeðslegt. _: pantaði salat í matinn.", "option1": "Jón", "option2": "Bjartur", "answer": "2"}
{"qID": "3TVMRGPCKS7PCXN3FNSGBLZ842-1", "sentence": "Hurvín var háværari en glugginn því hjarir _ voru betur smurða.", "option1": "hurðarinnar", "option2": "gluggans", "answer": "2"}
{"qID": "3TVMRGPCKS7PCXN3FNSGBLZ842-2", "sentence": "Hurvín var háværari en glugginn því hjarir _ vorur verr smurða.", "option1": "hurðarinnar", "option2": "gluggans", "answer": "1"}
{"qID": "306996CF6YYM26Q0XROB8GLZ842-1", "sentence": "Hann setti skýringu við myndina sína og reyndi að nota flókið mál en þurfti að endurskrifa hana því _ var of stutt.", "option1": "skýringin", "option2": "málið", "answer": "1"}
{"qID": "3PCPFX4U40O22NQ3AMO03MKZ1K3FQ8-2", "sentence": "Vaka þurfti aðstoð hjá Emíliönu til að hefja brjóstaugjöf, en _ gat ekki aðstoða hana.", "option1": "Vaka", "option2": "Emíliana", "answer": "1"}
```
Even the large model does not do much better than chance.
Key Results

Feasible to extract good training data for language models from the Common Crawl

Cleaning of the data is important

Similar performance in many tasks using models trained on curated vs. scraped corpora

Using a multilingual model as a warm start leads to comparable performance to training from scratch, at a fraction of the cost

The Icelandic WinoGrande dataset is a tough task and should serve well to measure future progress