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 Þorsteinsson & 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?

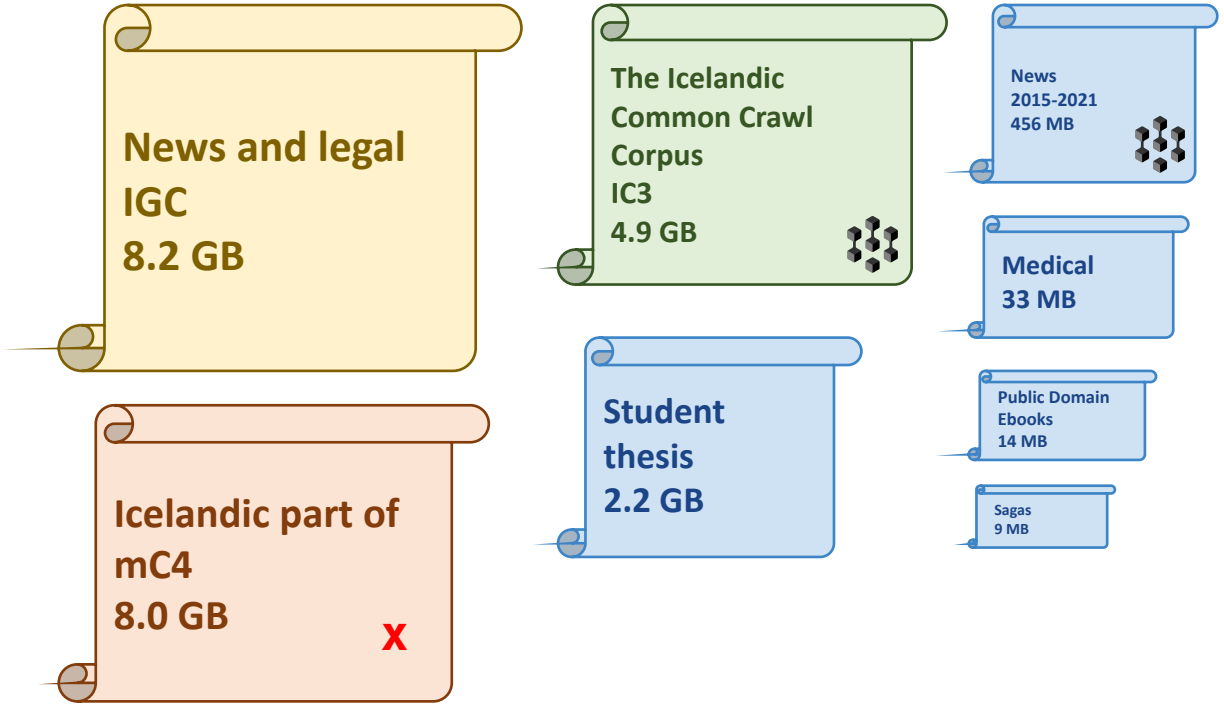
The Common Crawl hosts a searchable index

a) Search for ***.is** – response contains byte ranges in dumps (huge files)

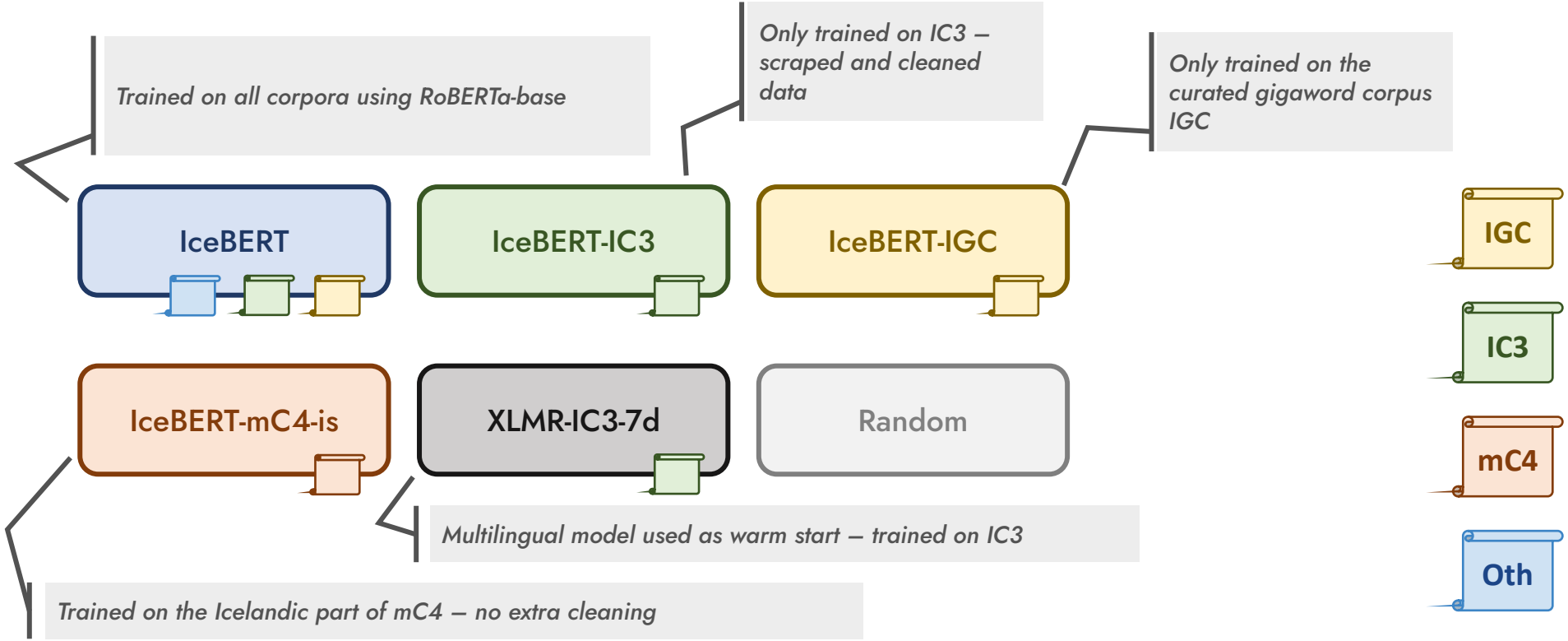
b) Fetch only data in byte ranges

c) Clean up and deduplicate data

Icelandic Corpora



Pretraining LMs for Icelandic



Part of Speech

n: noun
v: feminine
e: singular
n: nominative
g: definite article
s: proper

Part-of-speech (POS) tagging

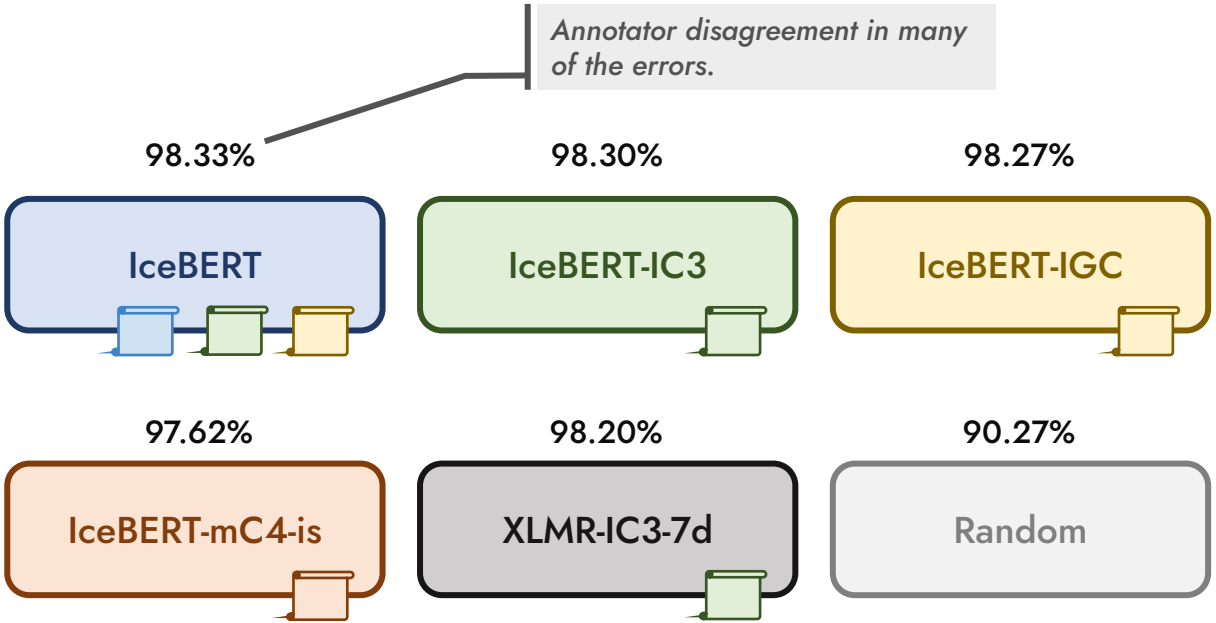
- Word type, declension, tense etc.
- Combined tags taken apart – multi label task
    - (word type, gender, case, etc)
- **Example:** *Gamla konan ók grænni rútu*

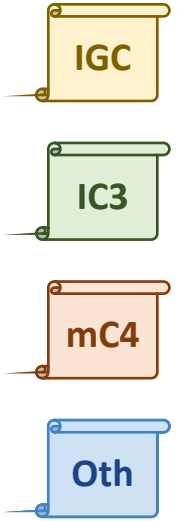
Data

- **MIM-Gold** (*Loftsson et al. 2010, Developing a PoS-tagged corpus using existing tools*)

```
<s n="2">
            <w lemma="orð" type="nvengs">Orðin</w>
            <w lemma="vera" type="sfg3fþ">voru</w>
            <w lemma="skrifa" type="sþgven">skrifuð</w>
            <w lemma="með" type="aþ">með</w>
            <w lemma="blýantur" type="nkeþ">blýanti</w>
            <w lemma="á" type="ao">á</w>
            <w lemma="blað" type="nheo">blað</w>
            <w lemma="sem" type="ct">sem</w>
            <w lemma="leggja" type="sþghen">lagt</w>
            <w lemma="vera" type="sfg3eþ">var</w>
            <w lemma="ofan" type="aa">ofan</w>
            <w lemma="á" type="ao">á</w>
            <w lemma="lík" type="nheog">líkið</w>
            <c type="punctuation">.</c>
            <w lemma="þrír" type="tfhfn">Þrjú</w>
            <w lemma="orð" type="nhfn">orð</w>
            <c type="punctuation">,</c>
            <w lemma="óskiljanlegur" type="lhfnsf">óskiljanleg</w>
            <w lemma="erlendur" type="nkeþ-s">Erlendi</w>
            <c type="punctuation">.</c>
</s>
```

Part of Speech - accuracy





Named Entity Recognition

Groups

PERSON
LOCATION
ORGANIZATION
MISCELLANEOUS
DATE
TIME
MONEY
PERCENT

```
Vel     O
gerð    O
og      O
rómantísk       O
með     O
þeim    O
Zach    B-Person
Braff   I-Person
(       O
«       O
Scrubs  B-Miscellaneous
»       O
,       O
«       O
Garden  B-Miscellaneous
State   I-Miscellaneous
»       O
)       O
,       O
Rachel  B-Person
Bilson  I-Person
(       O
«       O
O.C.    B-Miscellaneous
»       O
þættirnir       O
)       O
ofl.    O
```

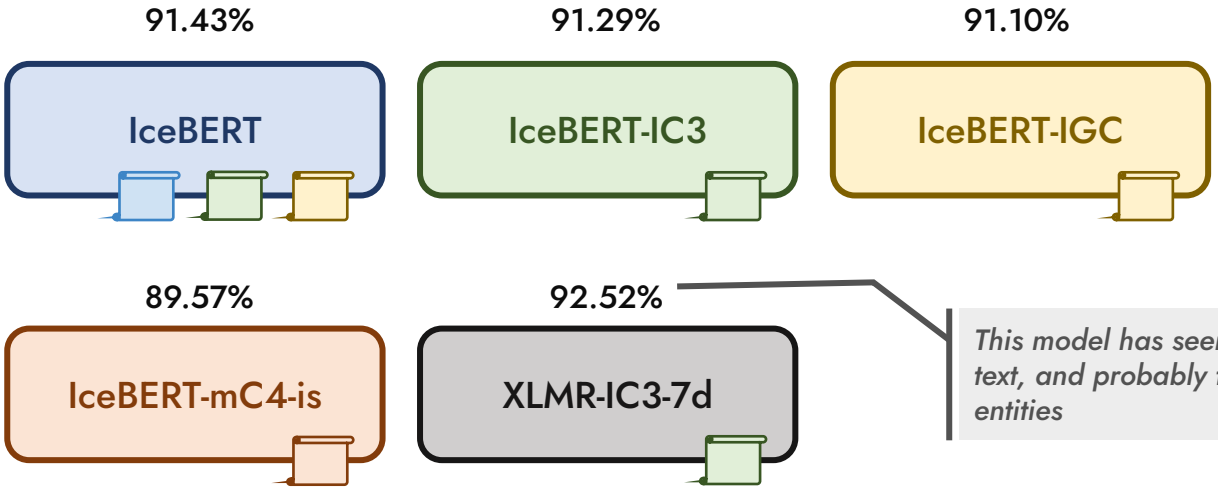
Named entity recognition (NER)

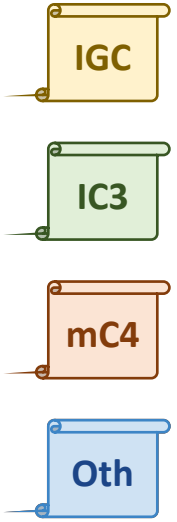
- The task of labelling entities in text
- **Example:** Katrín Jakobsdóttir (PER) vinnur í Stjórnarráðinu (ORG)

Data

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

NER – F1-score


91.43%
IceBERT
91.29%
IceBERT-IC3
91.10%
IceBERT-IGC
89.57%
IceBERT-mC4-is
92.52%
XLMR-IC3-7d
This model has seen the most text, and probably the most entities


IGC
IC3
mC4
Oth

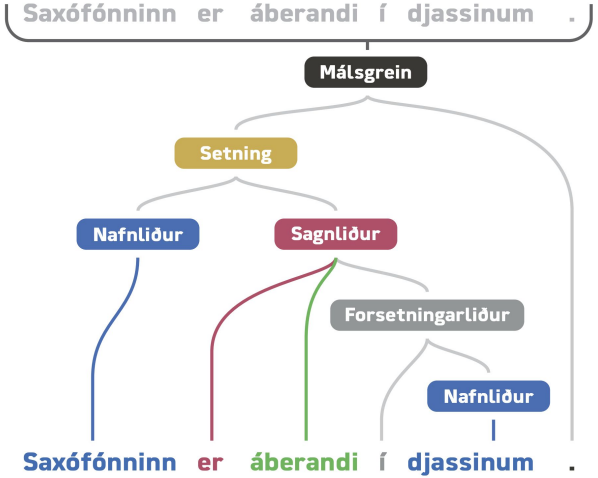
Parsing

Constituency parsing

- Labeling sentence constituents

Data

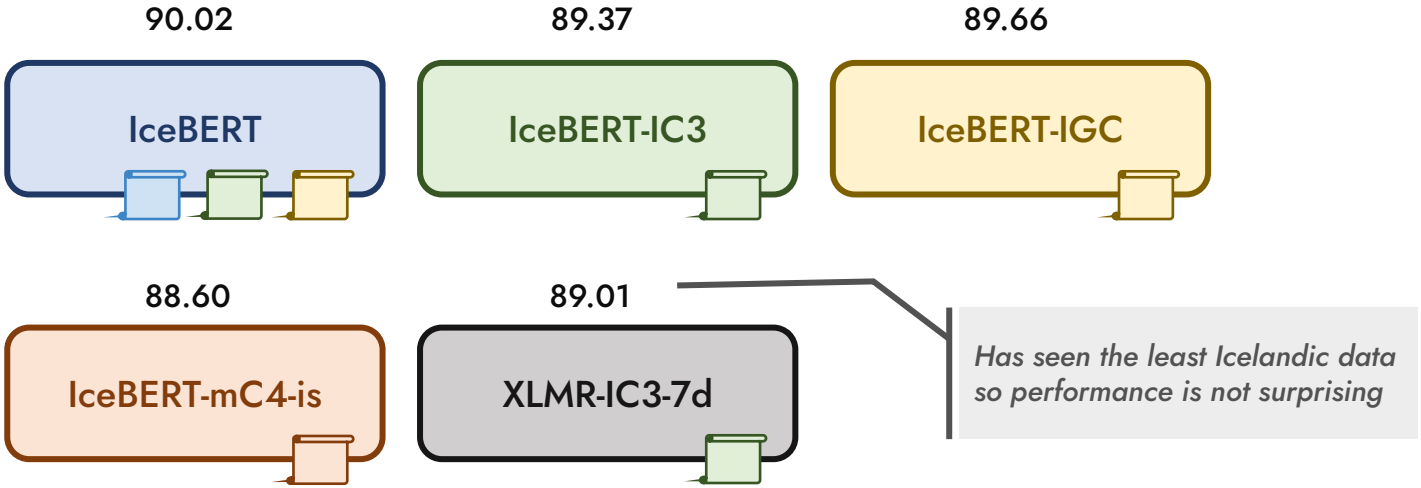
- **GreynirCorpus (Miðeind)**

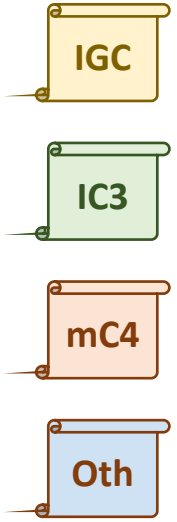



```
(S0 (S-PREFIX (C (st En (lemma en))))
    (S-QUE (NP-OBJ (fn_et_þf_hk hvað (lemma hver)))
           (IP (VP (VP (so_1_þf_fh_p3_et_nt_gm gerir (lemma gera)))
                   (NP-SUBJ (no_et_nf_kvk_gr ríkisstjórnin (lemma ríkisstjórn)))
                   (ADVP-DATE-REL (ao þá (lemma þá)))))
           (grm ?))))
```

Picture credit: Government LT-Programme

Parsing — F1-score





Grammatical Error Detection

Grammatical error detection

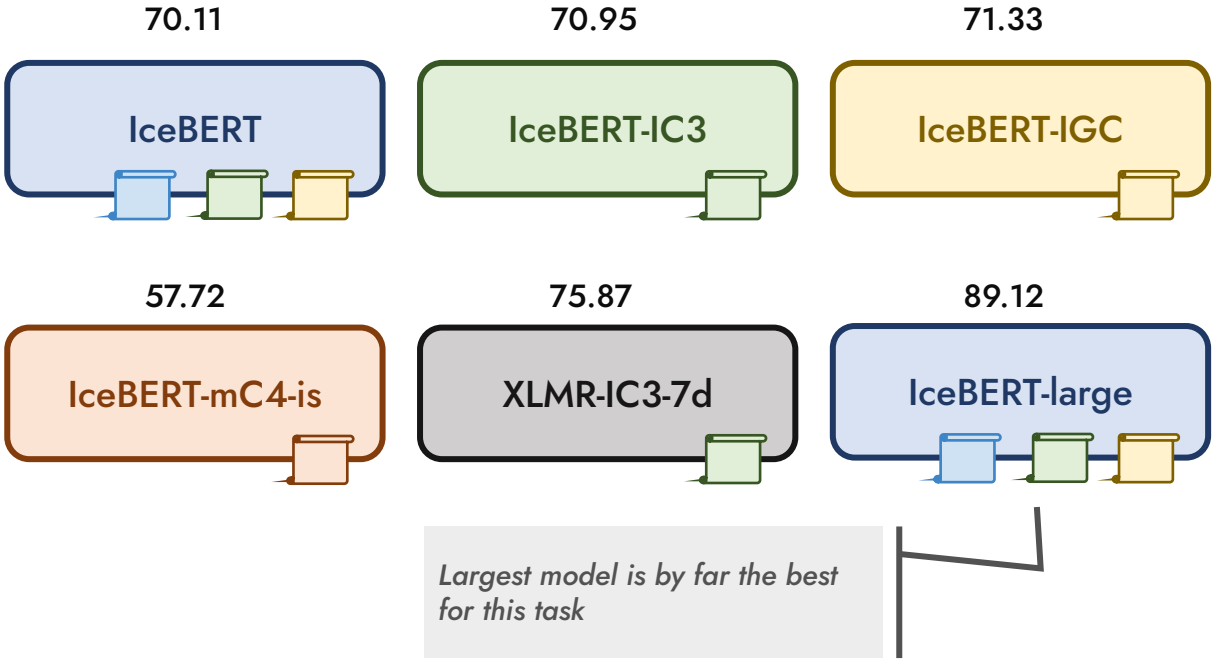
- Spelling
- Grammar
- Style

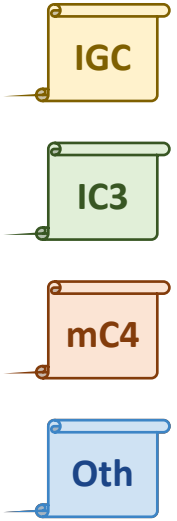
Data

- **Icelandic Error Corpus (IceEC)**
- Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, Xindan Xu

```
<s n="18">
  <w>Vissulega</w>
  <w>hafa</w>
  <w>kosningaloforð</w>
  <w>verið</w>
  <w>gefin</w>
  <w>um</w>
  <w>að</w>
  <w>auka</w>
  <revision id="10">
    <original><w>íslensku</w><w>kennslu</w></original>
    <corrected><w>íslenskukennslu</w></corrected>
      <errors>
        <error xtype="split-compound" eid="0"/>
      </errors>
  </revision>
  <w>fyrir</w>
  <w>innflytjendur</w>
<c>.</c>
</s>
```

Grammatical Error Detection - F1-score





Coreference resolution

Coreference resolution

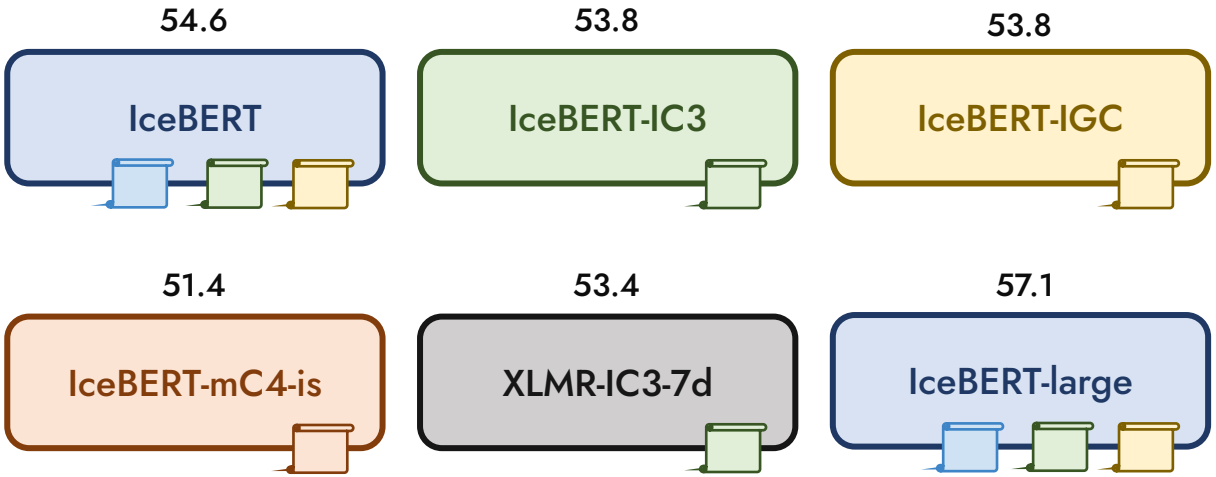
- **Example**: The cup could not fit into the suitcase because it was too big.

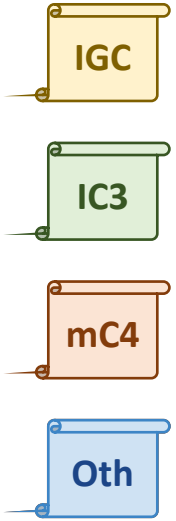
Data

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


```
{"qID": "3JVP4ZJHDPQHO0KHNK0SOB899NH0I8-2", "sentence": "Þorlákína sagði Agnetu að hún kæmist ekki út að borða vegna ristilbólgu. _ sýndi því skilning.", "option1": "Þorlákína", "option2": "Agneta", "answer": "2"}
{"qID": "3JVP4ZJHDPQHO0KHNK0SOB899NH0I8-2", "sentence": "Þorlákína sagði Agnetu að hún kæmist ekki út að borða vegna ristilbólgu. _ var afsökuð.", "option1": "Agneta", "option2": "Þorlákína", "answer": "2"}
{"qID": "3M47JKRKCXZJD5UJGN4IKNOMEAQ86S-2", "sentence": "Jón elskaði að borða sushi en Bjarti fannst það ógeðslegt. _ pantaði salat í matinn.", "option1": "Jón", "option2": "Bjartur", "answer": "2"}
{"qID": "32TMVRKDGPCKS7PCHXN3FNSGBLZ842-1", "sentence": "Hurðin var háværari en glugginn því hjarir _ voru betur smurðar.", "option1": "hurðarinnar", "option2": "gluggans", "answer": "2"}
{"qID": "32TMVRKDGPCKS7PCHXN3FNSGBLZ842-2", "sentence": "Hurðin var háværari en glugginn því hjarir _ voru verr smurðar.", "option1": "hurðarinnar", "option2": "gluggans", "answer": "1"}
{"qID": "306996CF6YYM26Q0XROB8RBF8LL1BN-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": "3PCPFX4U40OL22NQ3AOM0KMZ1K3FQB-2", "sentence": "Vaka þurfti aðstoð hjá Emilíönu til að hefja brjóstagjöf, en _ gat ekki aðstoðað hana.", "option1": "Vaka", "option2": "Emilíana", "answer": "1"}
```

Coreference resolution – accuracy





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