MIÐEIND

Máltækni og gervigreind fyrir íslensku

Haukur Páll Jónsson

ský / apríl 2023

Um Miðeind

Máltækni og gervigreindar fyrirtæki stofnað 2015

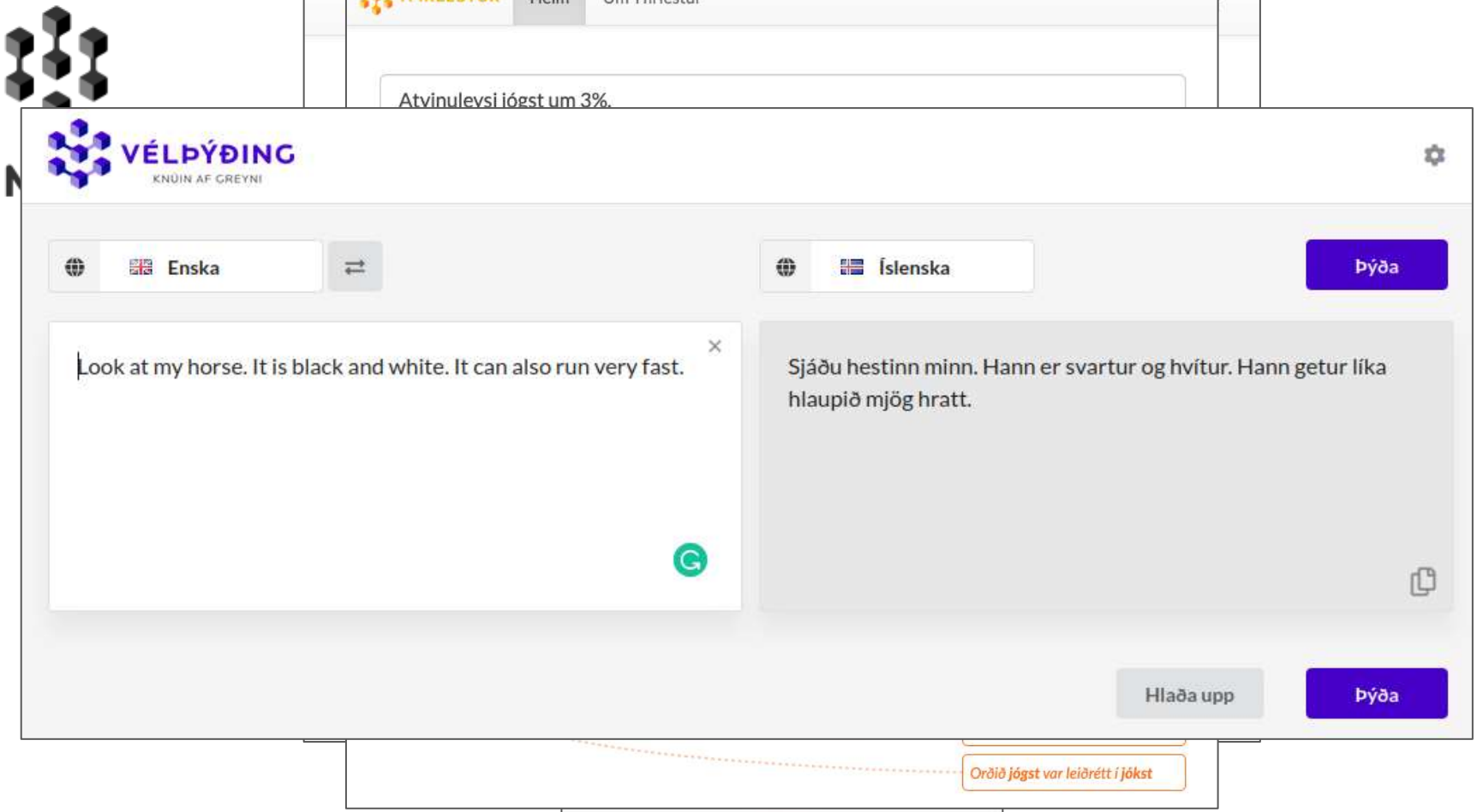
12 starfsmenn

Tókum þátt í Máltækniáætlun stjórnvalda

Ýmis máltæknitól, málrýni, þýðingar

OpenSource

Viljum tryggja að íslenska sé nothæf í stafrænum heimi


YFIRLESTUR
Heim
Um Yfirlestur
Atvinnulevsi jógst um 3%.
VÉLÞÝÐING
KNÚIN AF GREYNI
Enska
Íslenska
Þýða
Look at my horse. It is black and white. It can also run very fast.
Sjáðu hestinn minn. Hann er svartur og hvítur. Hann getur líka hlaupið mjög hratt.
Hlaða upp
Þýða
Orðið jógst var leiðrétt í jókst

MIÐEIND

Ada

Þar sem galdrarnir gerast

8 x nVidia A100 GPUs @ 40GB

1 TB RAM

Við höfum einnig fengið aðstoð frá þýskum vinum okkar við hjá Forschungszentrum Jülich sem rekar ýmsar ofurtölvur

MIÐEIND

Mállíkön

Grunnmállíkön þjálfuð með því að fylla inn í eyður:
Bjarni Benediktsson lagði <blank> fyrir Alþingi í gær.

Þýðingarlíkön þjálfuð á pörun:
<is>Sólin mun skína á morgun. <en>The sun will shine tomorrow.

Spunalíkön þjálfuð með því að spá fyrir næsta orði:
Bjarni Benediktsson lagði fjárlagafrumvarp fyrir <?>

Nógu stór mállíkön sýna ***nýja hæfileika***
(zero-shot, few-shot)

MIÐEIND

Íslensk ~~risa~~ mállíkön

IceBERT: BERT líkan fyrir íslensku

verkefni: Þáttun, orðaflokkun, textaflokkun...

Grein í LREC: A Warm Start and a Clean Crawled Corpus

mBART-enis: subword sequence-to-sequence

verkefni: Þýðingar

velthyding.is

ByT5: byte-level sequence-to-sequence

Downstream task: Grammar and spelling correction

ai.yfirlestur.is

https://huggingface.co/mideind

https://api.greynir.is

MIÐEIND

OpenAI samstarf

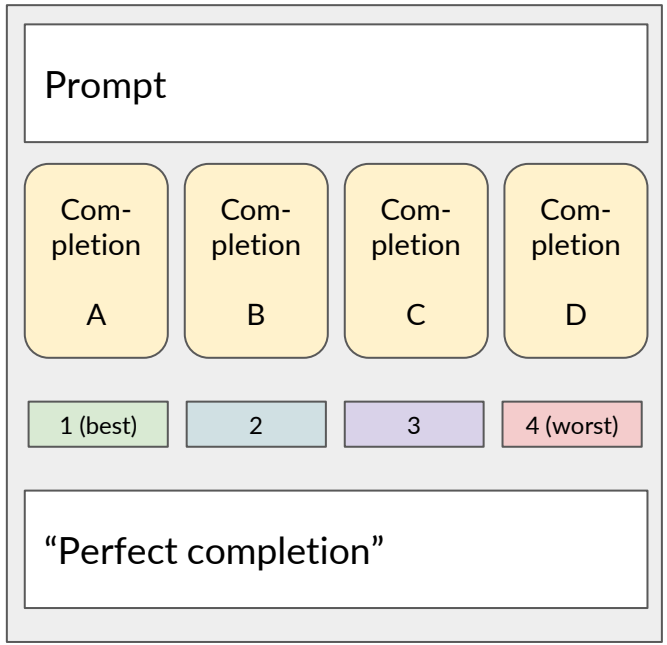
Í kjölfar forsetaferðar til Kísilsdals maí 2022

Samstarf um að styðja íslensku sem tilraunaverkefni fyrir önnur smærri tungumál.

Fyrsta verkefnið: Fínþjálfa GPT-3

MIÐEIND

GPT-4 þjálfun



- *Reinforcement Learning with Human Feedback (RLHF)* fyrir íslensku
- ~40 sjálfboðaliðar útbjuggu “prompts” og röðuðu niðurstöðum frá mismunandi líkönum
  - Útbjuggu einnig “fullkomin svör”
- Niðurstöður:
  - Skilur íslensku vel en getur ekki skilað frá sér réttri íslensku
  - Úttakið er á íslensku og vísar frekar til íslenskrar menningar í stað ameríkskrar (?)

MIÐEIND

GPT-4

Að vinna með líkaninu er ótrúlegt

Geta líkansins er ekki í jafnvægi:

Þýðingar *is* → *en* eru frábærar

Þýðingar *en* → *is* eru slæmar

Svör á ensku eru oftar betri en svör á íslensku, jafnvel þegar efni spurningar er íslenskt

https://mideind.is/gpt.html

MIÐEIND

Forþjálfun

Meiri, og betri, íslenskur texti í næstu forþjálfun

Þetta er í vinnslu

Þar með talið er Risamálheildin (IGC), IC3 og fleira

Af hverju er forþjálfunin svona mikilvæg?

- Íslenskur texti var of sjaldgæfur í fyrri forþjálfun
- Kenningar að vélþýdd íslenska hafi verið í síðustu forþjálfun

MIÐEIND

Tækifæri fyrir aðra en OpenAI

BLOOM, LLaMA, OPT, GLM, Dolly-v2

AI Sweden á stórt skandinavískt GPT líkan (spunalíkan)

Evrópusambandsverkefni

Máltækni- gervigreindaráætlun 2.0

MIÐEIND

Framtíðin

Augmented retrieval - t.d. Embla, spurningasvörun

Multimodality - hljóð, texti, mynd inn og út

Bjagi, öryggi og fleira - líka fyrir íslensku

Eitt líkan fyrir öll verkefni