Neuromasintõlge

Mark Fišel
Keeletehnoloogia õppetool
TÜ arvutiteaduse instituut
tartunlp.ai

ELRC/Tilde/… KT/MT seminar
30. september 2021

TÜ keeletehnoloogia

- uurimisprojektid — Nt Horizon-2020 “Bergamot”, CEF NLTP, **MTee**, EKT projektid jt
- lahendused eesti keelele — Nt EstNLTK, Neurotõlge, Neurokõne, WordNet, Vabamorf, EstBERT jm
- praktilised rakendused — R&D, koostöö äripartneritega, **tõlkekratt**, **kõnekratt**

https://tartunlp.ai

1. “Tõlkimine on lihtne!”
2. “Masintõlge on juba lahendatud”
3. “Masintõlge/masinõpe ainult imiteerib õppimismaterjali”

Saa sisust aru + väljenda seda teises keeles

Saa sisust aru + väljenda seda teises keeles

chair → tool

Saa sisust aru + väljenda seda teises keeles

chair → tool / õppetool / juhtima / juhin / … :-(

mine metsa

Saa sisust aru + väljenda seda teises keeles

chair → tool / õppetool / juhtima / juhin / … :-(

mine metsa

ma ei täna sind ei täna ega kunagi

Saa sisust aru + väljenda seda teises keeles

chair → tool / õppetool / juhtima / juhin / … :-(

Mitmetähenduslikkus vs kontekst

TARTUNLP

Me ei tea, kuidas me tõlgime

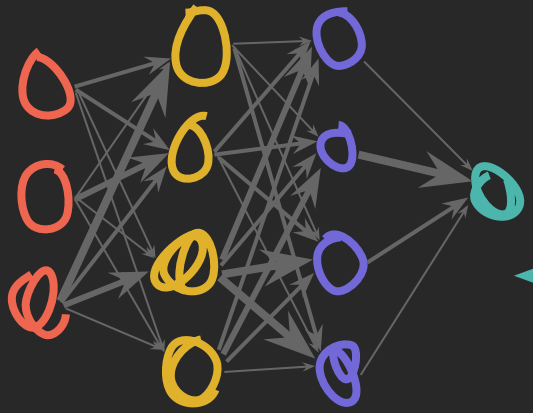
Masinõpe: ei selgita, kuidas teha, vaid lihtsalt näita

“Kurat sind võtaks, Johnny.” → “Damn it, Johnny.”

“Ta peksti oma garaažis läbi.” → “He got beaten up in his garage .”

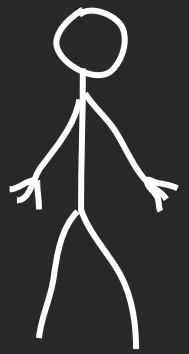
“See oli ilus.” → “That was beautiful, man.”

OpenSubtitles, opus.nlpl.eu



I like this!

See mulle meeldib

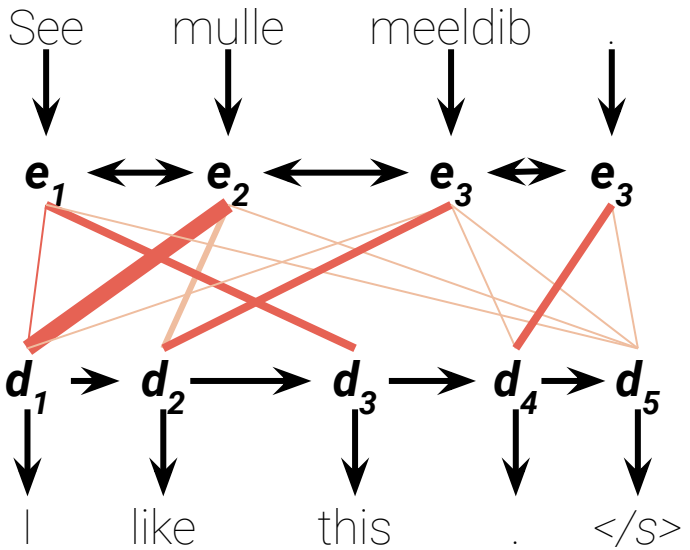


I like this

Sisend →kodeerija→ tähendusvektorid

Tähendusvektorid →dekodeerija→ väljund

Tähelepanumehhanism:



Enesetähelepanu:

Src: Google AI Blog

“Masintõlge on lahendatud”

- Google / Tilde / Neurotõlge jt. on juba päris ok
- Järeltoimetamine on levinud

“Masintõlge on lahendatud”

- Google / Tilde / Neurotõlge jt. on juba päris ok
- Järeltoimetamine on levinud

Reamees Parts → Private Parts

Private Parts → ?..

TARTUNLP

Uus tekstivaldkond = uus R&D projekt

Tilde / TartuNLP / ...

Uus tekstivaldkond = uus R&D projekt

Tilde / TartuNLP / ...

Uus tekstivaldkond = uus R&D projekt

Open-source NMT, tõlkekratt:
koodivaramu.eesti.ee/tartunlp

Mida teha kui andmeid on vähe?

nt. võru-eesti NMT, ainult 30k tõlkenäidet

- Mitmekeelsus: sarnased keeled aitavad
  - nt soome, põhja/lõunasaami
- Siirdeõpe: “eelõpe” + häälestamine
  - nt õpi kõigepealt soome-eesti, siis võru-eesti tõlget
- Sünteetilised andmed: tõlkeandmete genereerimine
  - võrukeelseid tekste: 168k lauset (×6 korda)

soome-ugri.neurotolge.ee:

TARTUNLP

TARTUNLP
SOOME-UGRI NEUROTÕLGE

Neurotõlge Info Koostöö EN ET Kõik demod

Väljundkeel eesti keel

A edesi ei näeq tükk aigo tii pääl üttegi võrokiilset silti.

60 / 700

Kuid edasi ei näe tükk aega teel ühtegi võrukeelset silti .

“Ainult imiteerib õppimisnäiteid”

- keelesegu näide
- grammatika parandamise näide

- neuromasintõlge õpib näidetel
  - tähelepanu
  - tõenäosused
  - närvivõrgud
- uus valdkond = uus NMT süsteem
- võimalusi on palju! koostöö / valmis lahendused / jne

TARTUNLP

Aitäh!

soome-ugri.neurotolge.ee
translate.ut.ee
tartunlp.ai