Vigaste ja korrektsete lausete paralleelkorpuse loomine edasi-tagasi masintõlke abil

MARTIN MÕTUS & KAIS ALLKIVI-METSOJA

16.06.2022

TALLINNA ÜLIKOOL
Digitehnoloogiate instituut

Eestikeelse teksti automaatkorrektuur

Õigekirjakontroll

- Sõnatasandi vead
- Eesti keelele olemasolevad tööriistad vajavad edasiarendust

Grammatikakontroll

- Lausetasandi vead
- Eksperimentaalsed tööriistad arenduses
- Masintõlkepõhine

TALLINNA ÜLIKOOL
Digitehnoloogiate instituut

Ühekeelne masintõlge

- Vigane lause teisendatakse sama keele normipäraseks lauseks
- Tugineb transformer-keelemudelile

- *Zero-shot* tõlge
  - Ei vaja samakeelseid paralleelandmeid

- Ühekeelne masintõlge
  - Kasutab sünteesitud samakeelset paralleelkorpust

(Luhtaru, 2022)

Sünteetilised vead

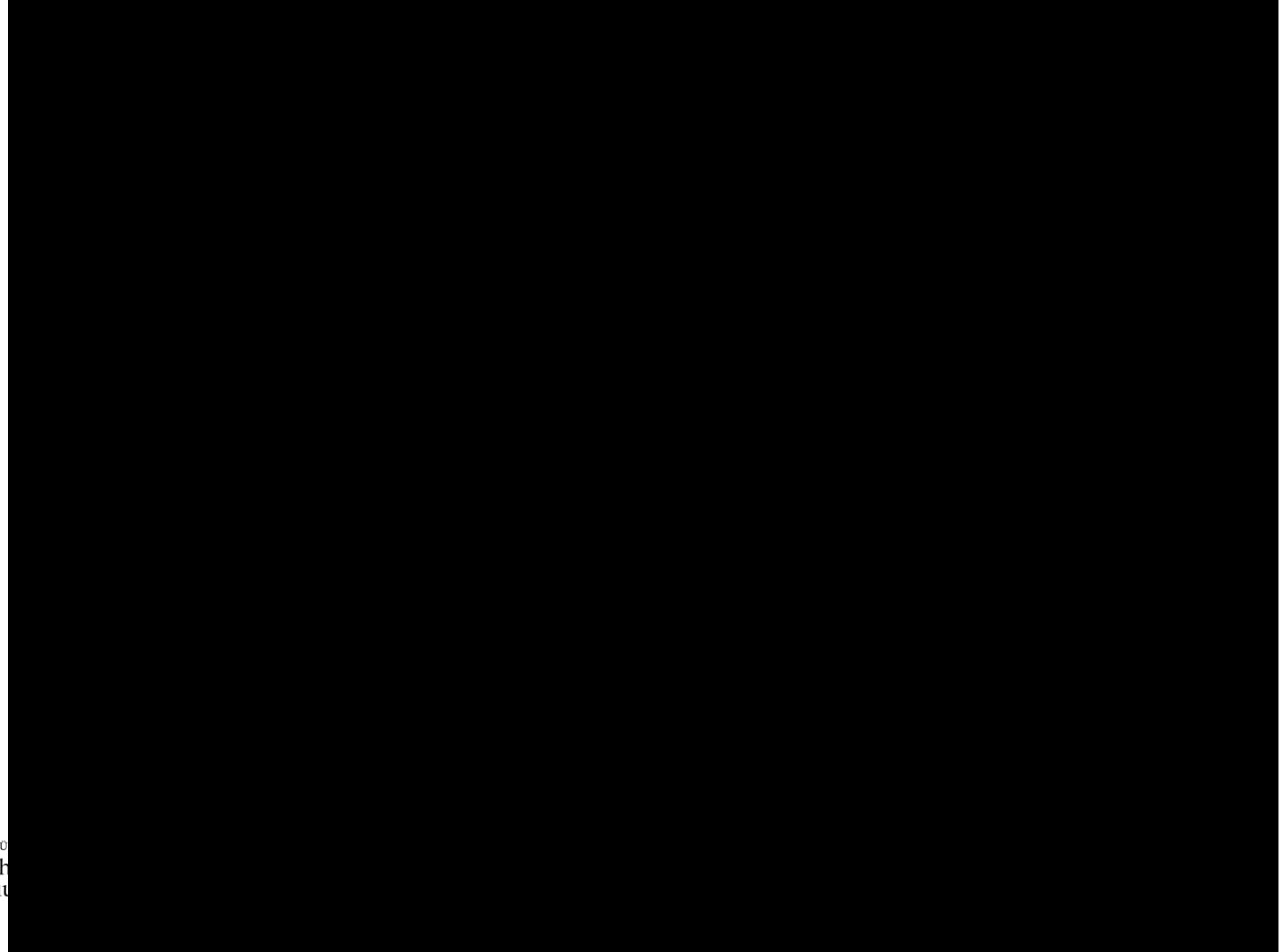
- Automaatne vigade lisamine korrektsetesse tekstidesse
- Mahuka ühekeelse rööpandmestiku loomine ja kasutamine ühekeelsete korrektuurimudelite treenimiseks
- Üks võimalikest meetoditest: edasi-tagasi masintõlge (ingl *round-trip translation*)

Edasi-tagasi masintõlge

- Lähtekeelne lause tõlgitakse esmalt sihtkeelde ja seejärel tagasi lähtekeelde
  - Inglise keele automaatkorrektuuris rakendust leidnud meetod (Lichtarge jt, 2019)
- Eri sihtkeelte kasutamine
- Tõlge mitme sihtkeele kaudu

Edasi-tagasi masintõlge

- Eesti keele ühendkorpus
- Sihtkeeled
  - soome
  - läti
  - leedu
  - inglise
  - vene
  - saksa
  - ...
- Tõlketeenused
  - TartuNLP Neurotõlge
  - OpenNMT-l põhinevad keelemudelid



Masintõlke väljund

- Sisaldab nii vigu kui ka ümbersõnastusi:
  - sõnade ja fraaside asendused, kustutamine ja lisamine
  - sõnavormide asendused
  - sõnajärje muutused
  - kokku- ja lahkukirjutuse ja kirjavahemärkide kasutuse vead

Tõlkeväljundi analüüs eesti-soome-eesti tõlke näitel

- 100 juhuvalikuga leitud lause tõlgete analüüs
- Muutustest 50% vead, 46% ümbersõnastused ja 4% parandused
  - sõna- ja fraasiasendused: 52% (sh sõnaasendused 42%)
  - vormiasendused: 13% (sh käändevormiasendused 10%)
  - sõnajärje muudatused: 7%
  - sõnade ja fraaside kustutused: 15%
  - sõnade lisamised: 6%
  - kirjavahemärkide lisamised, kustutused ja asendused: 4%
  - õigekiri ja kokku-lahkukirjutus: 3%

TALLINNA ÜLIKOOL
Digitehnoloogiate instituut

Tõlkenäited (1)

alglause

Tallinna linnapea Edgar Savisaare nõunik Arvo Teder ütles BNS-le, et konsortsiumi ettepanek ei ole veel linnavalitsusse jõudnud.

eesti-inglise-eesti

Tallinna linnapea **nõunik Edgar Savisaare, Arvo Teder,** ütles BNSile, et konsortsiumi ettepanek ei ole veel **linnavolikoguni** jõudnud.

eesti-soome-eesti

Tallinna linnapea **nõunik Edgar Savisari nõunik** Arvo Teder ütles BNSile, et konsortsiumi ettepanek ei ole veel **linnavalitsusele** jõudnud.

Tõlkenäited (2)

alglause

Ainult väike osa hääletajaid möönis valimiskasti juurest lahkumisel, et oli välja tulnud Clintoni pärast.

eesti-inglise-eesti

Ainult vähesed valijad tunnistasid **valimiskastist lahkudes**, et nad olid **Clintonisse tulnud**.

eesti-läti-eesti

Vaid vähesed valijad, **kes lahkusid valimiskastist**, tunnistasid, et nad on **lahkunud Clintoni pärast**.

Õigekirjavigade süntees

- Kustutused, lisamised, asendused ja nihutused
- Eri liiki tähevigade sagedus põhineb eesti keele kui teise keele õppijate A2–C1-taseme loovkirjutiste veastatistikal
  - 2457 lauset, 23 090 sõna, 745 õigekirjaveaga sõna
- Tähtede lisamisel ja asendamisel on arvestatud nende sagedust eesti keeles, lähtudes ühendkorpusest 2019

Õigekirjavigade statistika

Tase	Täheveaga sõnu	Täheveaga lauseid	Tähtede kustutusi	Tähtede lisamisi	Tähtede asendusi	Tähtede nihutusi
A2	5,8%	24,0%	25,9%	20,1%	51,1%	2,9%
B1	4,3%	26,2%	27,3%	29,1%	42,4%	1,2%
B2	2,4%	20,7%	24,5%	30,3%	43,2%	1,9%
C1	2,3%	23,7%	36,6%	32,4%	29,7%	1,4%
Keskmine	**3,7%**	**23,7%**	**28,6%**	**28,0%**	**41,6%**	**1,9%**

Sünteesvigade näide

Tallinna linnapea **nuõnik** Edgar Savisari nõunik Arvo Teder ütles BNSile, et konsortsiumi **ettepanmnek** ei ole veel **linavalitsusele** jõudnud.

nõunik ⇨ nuõnik

ettepanek ⇨ ettepanmnek

linnavalitsusele ⇨ linavalitsusele

Veamärgendusega korpus

- Eesti keele tasemeeksamitel A2–C1-tasemele hinnatud loovtekstid, pärit Eesti vahekeele (õppijakeele) korpusest
- Maht ~3700 lauset, praeguseks märgendatud ~2500 lauset
  - A2: 495, B1: 673, B2: 723, C1: 566
- M2-formaadis veamärgendus (Dahlmeier & Ng, 2012)
  - tähistatud vea skoop, liik, parandus ja parandusvariant

Vealiigid

- Asendused
  - õigekirjaviga (R:SPELL), sõnavalik (R:LEX), vormivalik (R:NOM:FORM, R:VERB:FORM), kirjavahemärgi valik (R:PUNCT), algustäheviga (R:CASE), kokku-lahkukirjutuse viga (R:WS), sõnajärjeviga (R:WO)
- Lisamised
  - liigne sõna (U:LEX) või kirjavahemärk (U:PUNCT)
- Kustutused
  - puuduv sõna (M:LEX) või kirjavahemärk (M:PUNCT)

Märgenduse näide

```
  0    1  2           3     4       5 6
S|Mul|on|pesumasin|läks|katki|.|
A 1 2|||U:LEX|||-NONE-|||REQUIRED|||-NONE-|||0
A 2 4|||R:WO|||läks pesumasin|||REQUIRED|||-NONE-|||0
```

Korrektne lause:

Mul läks pesumasin katki.

Masintõlkepõhise korrektori treenimine

- Transformer-arhitektuuril põhinev keelemudel
- Sünteesvigade kasutamine keelemudeli eeltreenimiseks
- Grammatikakorrektori tulemuslikkuse hindamine
  - Saagis
  - Täpsus
  - F-skoor

Täname

- Kaisa Norak (TLÜ digitehnoloogiate instituudi analüütik)
- Karina Kert ja Silvia Maine (TLÜ üliõpilased)