Delovni sklop 4: Strojno prevajanje

Iztok Lebar Bajec, Andraž Repar

www.eu-skladi.si

www.slovenščina.eu

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI STRUKTURNI IN INVESTICIJSKI SKLADI
NALOŽBA V VAŠO PRIHODNOST

Aktivnosti sklopa 4

- Zbiranje besedil za korpus prevodov sl-en, en-sl https://zbiranje.slovenscina.eu/prevodi
- Razvoj evalvacijske metodologije
- Preučevanje različnih ogrodij za NMT
- Učenje novih NMT modelov glede na povečevanje korpusa
- Razvoj portala za preizkusno uporabo prevajalnika
- Priprava dolgoročnega načrta za nadaljnji razvoj strojnega prevajalnika

Ključni partnerji

- *UL FRI, Laboratorij za podatkovne tehnologije*
- *UM FERI, Laboratorij za digitalno procesiranje signalov*
- *VITASIS d.o.o.*
- AIKWIT d.o.o.
- UL FF, Oddelek za prevajalstvo

*poševno: tehnični del

Obstoječi/referenčni model

- Razvit na inštitutu Jožef Stefan
- Temleji na ogrodjih Nematus + Marian
- Učni korpus velikosti cca 40M
- Testna množica velikosti cca 2K
- Objavljen BLEU AN-SL 40.49, SL-AN 44.42

Odkrite pomanjkljivosti:

- Učna množica z izjavami, ki namesto šumnikov uporabljajo sičnike
- Testna množica vsebuje napake in izjave v drugem jeziku
- Testna množica delno zajeta v učni množici

NeMo

- Odprtokodno ogrodje proizvajalca NVIDIA
- Zasnovano na ogrodjih Pytorch, Lightning in Hydra
- Pokriva segment *Conversational AI*, ki zajema ASR, NLP in TTS
- V aktivnem razvoju od 2018
- Trenutna različica v1.13.0
- Nevronsko strojno prevajanje (del NLP) temelji na transformer sequence-to-sequence arhitekturi po vzoru AAYN

Učna množica

Predobdelava

- Filtriranje po dolžini in razmerju
- Izločanje nepravilnih, nepopolnih in netekočih izjav (orodje bicleaner)
- Deduplikacija (orodje bifixer)
- Normalizacija ločil (orodje moses)
- Izločeni pari, ki so v referenčni testni množici
- Preostali korpus po čiščenju ~33M, 8192 parov izločenih za validacijo

Učenje BPE tokenizerja

- Predtokenizacija ločil (orodje moses)
- Učenje skupnega BPE tokenizerja velikosti 64000 tokenov (orodje YTTM)

Zbiranje dodatnih učnih podatkov

Novi viri ves čas nastajajo – portal Opus

- CCMatrix

Potencialni viri dodatnih učnih podatkov

- Javne ustanove
- Podjetja

Končni rezultat: množica Paralelni korpus RSDO4 2.0

- 3143624 prevodnih parov
- Randomiziran vrstni red
- Delna anonimizacija

Testna množica

Testna množica Asistent

- Avtomatsko izbrani segmenti iz javno dostopnih virov
- Slaba segmentacija
- Tuji jeziki
- Duplikati

Testna množica SloBench

- Izbrana besedila iz petih področij
- Ustrezne avtorske pravice
- Novi prevodi, ki ne obstajajo nikjer drugje

Ročna evalvacija

Model Adequacy/Fluency

- Evalvacija vseh razvitih modelov in javno dostopnih prevajalnikov

Model Error Annotation

- V sodelovanju z oddelkom za prevajalstvo na FF (članek in predstavitev na JTDH2022)

Rezultati

en->sl	
orodje	**BLEU**
nemo-v1.2.6	**46.69**
marian-v8	**45.27**
eTranslation	44.9
javni-1	42.21
javni-2	42.03
javni-3	41.08
referenčni	*38.34*
javni-4	36.77

sl->en	
tool	**BLEU**
nemo-v1.2.6	**51.48**
marian-v8	**49.27**
eTranslation	47.89
javni-2	46.91
javni-4	45.73
javni-3	44.94
referenčni	*42.9*
javni-1	42.8

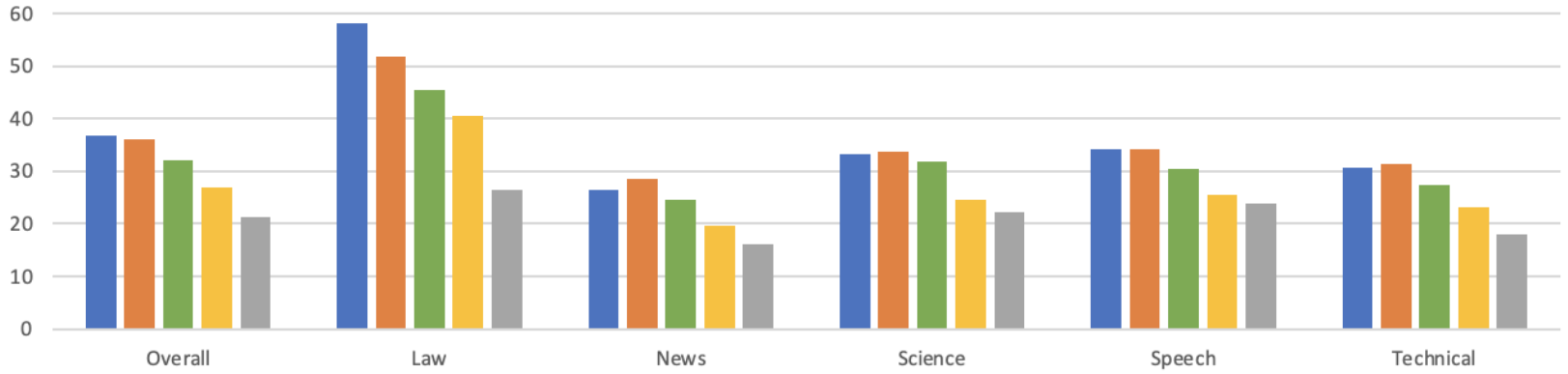
en->sl	
orodje	**BLEU**
javni-3	36.83
javni-2	36.02
eTranslation	32.15
nemo-v1.2.6	**32.07**
marian-v8	**26.88**
referenčni	*21.32*
javni-1	--
javni-4	--

sl->en	
tool	**BLEU**
javni-3	34.73
javni-2	33.36
nemo-v1.2.6	**32.97**
eTranslation	31.19
marian-v8	**27.92**
referenčni	*19.42*
javni-1	--
javni-4	--

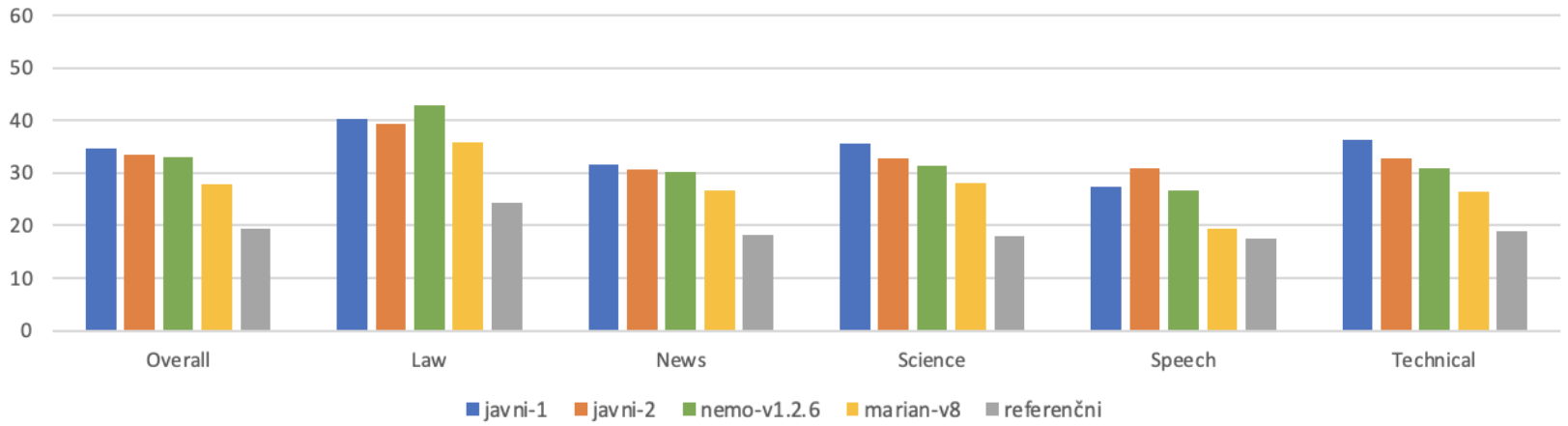
*modra: očiščena in izboljšana referenčna testna množica
**oranžna: testna množica slobench, ki bolj sovpada z ročno evalvacijo

Rezultati

EN-SL



SL-EN



DEMO

Dosegljiv na

https://slovenscina.eu/prevajalnik oz. https://storitve.cjvt.si/prevajalnik

Portal za oddajo prevodov

Ali želimo, da se razvoj jezikovnih tehnologij za slovenščino ne dogaja samo v ameriških tehnoloških gigantih?

Ali želimo, da se slovenščina uporablja v vseh novih oblikah komunikacije, ki jih ponuja (in jih še bo ponudila) tehnologija?

Če je odgovor DA → https://zbiranje.cjvt.si/prevodi/