Nevronski strojni prevajalnik AN-SL, SL-AN

www.eu-skladi.si

Iztok Lebar Bajec

www.slovenščina.eu

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI STRUKTURNI IN INVESTICIJSKI SKLADI
NALOŽBA V VAŠO PRIHODNOST

Projekt RSDO, DS4

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

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:

- OpenSubtitles2016 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 ogrodijh Pytorch Lightning in Hydra
- Pokriva segment *Conversational AI*, ki zajema ASR, NLP in TTS
- V aktivnem razvoju od 2018
- Trenutna različica v1.8.2
- Nevronsko strojno prevajanje (del NLP) temelji na transformer sequence-to-sequence arhitekturi po vzoru AAYN

Korpus

množica	velikost (M)
corpus	*40,722*
TC3	24,420
novo	0,420
javno	35,261
interno	3,770
poravnano-navodila	0,284
poravnano-openscience	0,531
prevodi	1,000
ccmatrix	27,407

*skupaj, vključujoč duplikate, 133M parov
**_corpus_ je bil uporabljen za gradnjo referenčnega prevajalnika
***TC3 je _corpus_ očiščen OpenSubtitles 2016 in testne množice

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)

Učenje

Uporaba Vega HPC in distribuiranega učenja na 64x A100 40GB GPU (16 vozlišč) za eno smer. Pri trenutni velikosti učne množice in uporabi numerike s polovično preciznostjo en cikel traja 4 dni (~24 epoh oz. 600k iteracij).

Hiperparametri

- Ahritektura *aayn_base, batch_size=1024*
- Encoder *hidden_size=1024, num_layers=24, inner_size=4096, attention_heads=16, pre_ln=True*
- Decoder *hidden_size=1024, num_layers=6, inner_size=4096, attention_heads=16, pre_ln=True*
- Otpimizer *name=AdamW, lr=2.0*
- Scheduler *name=NoamAnnealing, min_lr=1e-6*

Rezultati

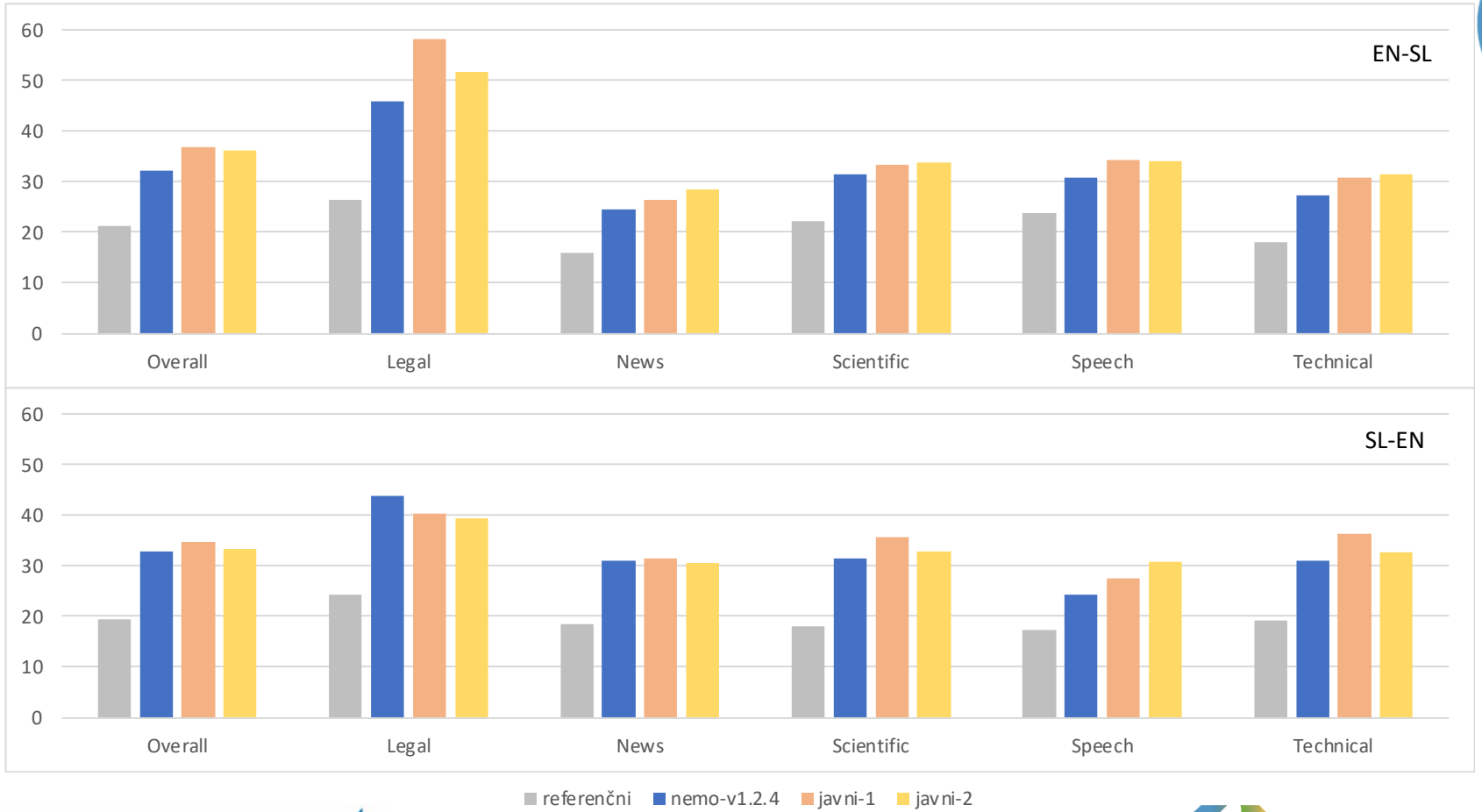
en->sl	
orodje	BLEU
nemo-v1.2.4	**46.53**
eTranslation	44.9
javni-3	42.21
javni-2	42.03
javni-1	41.08
referenčni	*38.34*
javni-4	36.77

sl->en	
tool	BLEU
nemo-v1.2.4	**51.24**
eTranslation	47.89
javni-2	46.91
javni-4	45.73
javni-1	44.94
referenčni	*42.9*
javni-3	42.8

*rezultati nad očiščeno testno množico, odstranjeni in zamenjani napačni prevodi

www.slovenščina.eu
REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO
EVROPSKA UNIJA
EVROPSKI STRUKTURNI IN INVESTICIJSKI SKLADI
NALOŽBA V VAŠO PRIHODNOST

Rezultati



www.slovenščina.eu

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI STRUKTURNI IN INVESTICIJSKI SKLADI
NALOŽBA V VAŠO PRIHODNOST