Dolgoročni načrt za razvoj strojnega prevajalnika predavanj v slovenskem jeziku v angleščino za nesprotno prevajanje in načrta za posodabljanje

Razvoj slovenščine v digitalnem okolju (RSDO)
DS4: Strojno prevajanje

Uvod

V okviru projekta RSDO smo razvili nevronski strojni prevajalnik za jezikovni par slovenščina-angleščina. Prevajalnik na očiščeni testni množici referenčnega prevajalnika dosega bistveno višjo kvaliteto prevodovpo oceni BLEU (46,69 EN⇢SL; 51,48 SL⇢EN). Referenčni prevajalnik je po čiščenju testne množice in ponovni evalvaciji rezultatov dosegal oceno BLEU 38,34 EN⇢SL oz. 42,9 SL⇢EN. Evalvacija s testno množico Slobench pokaže še večje razlike. Prevajalnik, ki je bil razvit v okviru RSDO, je na tej testni množici povsem primerljiv s prevajalniki eTranslation, Google, Microsoft in DeepL. Na domeni prava v smeri SL->EN jih celo prekaša. Slednje potrjujejo tudi ročne evalvacije.

V tem dokumentu podajamo predloge za nadaljnji razvoj prevajalnikov za jezikovni par slovenščina-angleščina.

Pridobivanje novih podatkov za razvoj prevajalnika

Sodobni algoritmi za umetno inteligenco na osnovi globokih nevronskih mrež, med katere spadajo tudi algoritmi za strojno prevajanje, za svoje delovanje potrebujejo velike količine podatkov. Slovenščina je v tem pogledu kot jezik z relativno majhnim številom govorcev v podrejenem položaju glede na nekatere druge jezike z večjim številom govorcev. Po koncu projekta RSDO je stanje na področju podatkov, ki so na voljo za učenje strojnega prevajalnika, boljše kot na začetku: če je bilo ob začetku projekta na voljo okoli 37 milijonov poravnanih stavkov (tj. stavkov v enem jeziku s pripadajočim prevodom v drugem jeziku), je ob koncu projekta RSDO takih parov več kot 100 milijonov[1]. Vendar je to še vedno na spodnji količinski meji podatkov, ki morajo biti na voljo za dobro delujoč strojni prevajalnik.

Večjo količino besedil za učenje strojnih prevajalnikov je mogoče zagotoviti s pridobivanjem dodatnih obstoječih prevodov (tj. originalnih besedil in njihovih prevodov) ali z namenskim prevajanjem še neprevedenih besedil. Druga opcija je cenovno zelo potratna in jo je smiselno uporabiti zgolj za zelo specifične namene, kot so posebne testne ali validacijske množice manjšega obsega. Tako za širjenje nabora podatkov ostane samo zbiranje že obstoječih prevodov, kar pa je tudi povezano z določenimi stroški. Lastniki avtorskih pravic pogosto niso naklonjeni uporabi njihovih besedil za namene učenja strojnih prevajalnikov in drugih algoritmov umetne inteligence oziroma pričakujejo za to ustrezno nadomestilo. Kadar gre za osebe zasebnega prava je tako pričakovanje upravičeno, izkušnje iz projekta RSDO

[1] Kar je posledica aktivnosti projekta RSDO in pojava novih odprtih naborov podatkov.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

pa kažejo tudi, da take osebe, zlasti tiste, ki imajo v svoji lasti velike količine podatkov, niso zainteresirane za prodajo podatkov.

Situacija je drugačna pri javnih ustanovah oz. organizacijah, ki se financirajo iz državnega proračuna, občinskih proračunov in drugih javnih sredstev. Javni sektor v Sloveniji na dnevni ravni producira velike količine prevodov, ki bi lahko služili za učenje strojnih prevajalnikov. Kot sledi iz poročila projekta European Language Resource Coordination[2], je organizacija prevajalskih procesov v slovenski javni upravi ugodna v primerjavi z marsikatero drugo evropsko državo, saj gre velika večina prevodov skozi Sektor za prevajanje pri Generalnem sekretariatu Vlade Republike Slovenije. To pomeni, da so prevodi že zbrani znotraj ene organizacije in jih ni potrebno iskati po posameznih ministrstvih, službah, oddelkih in sektorjih. Kljub temu do konca projekta RSDO ti prevodi niso bili na voljo za učenje strojnih prevajalnikov. Identificirali smo dve glavni oviri, ki preprečujeta širšo razpoložljivost prevodov. Prva ovira je povezana z lastniškimi pravicami nad prevodi, saj javne ustanove, ki naročajo prevode pri zunanjih izvajalcih, pogosto niso (ali ne ne vedo, ali so) lastnice avtorskih pravic prevodov. Ta vidik je potrebno upoštevati pri pripravi razpisov, naročil in pogodb za izvajanje prevodov, v katere je treba vključiti ustrezno določilo o prenosu avtorskih pravic na naročnika oz. o odpovedi uveljavljanja avtorskih pravic s strani izvajalca. Druga ovira pa je povezana z vsebino prevedenih besedil, ki občasno lahko vsebujejo osebne podatke posameznikov, podatke, povezane z nacionalno varnostjo, in druge občutljive vsebine. Organizacijo prevajalskih procesov je treba zastaviti tako, da je take prevode mogoče enostavno izločiti iz nabora podatkov za učenje strojnega prevajalnika oz. da se take občutljive podatke anonimizira.

Na področju pridobivanja novih podatkov za razvoj prevajalnika smo identificirali tri ključne naloge:

1. Nadaljevati je treba s širjenjem vzporednega korpusa splošnega jezika, saj novi modeli zahtevajo vedno več podatkov. Poleg tega je treba začeti izdelovati domenske korpuse za podporo postopkom za adaptacijo modelov na izbrano strokovno domeno. Nekatere domene bi bilo mogoče izločiti iz splošnega korpusa, pri drugih pa prevodov ni dovolj in jih bo treba dopolniti z novimi, pri čemer lahko pomembno vlogo igrajo prevodi, ki nastajajo v okviru rednih delovnih procesov v javnem aparatu.

   Rezultat: povečan vzporedni korpus splošnega jezika

2. Testna množica se uporablja za evalvacijo razvitih modelov. Za čim bolj verodostojne rezultate je potrebno razširiti nabor domen (trenutno 5 domen v okviru množice Slobench) in dodati alternativne referenčne prevode (trenutno samo en referenčni prevod v množici Slobench), ki so lahko ročno prevedeni ali prevedeni strojno in naknadno popravljeni oz. post-editirani.

   Rezultat: Na voljo so dodatne evalvacijske množice za različne domene in naloge

[2] https://lr-coordination.eu/sites/default/files/Reports%202021/ELRCWhitePaper.pdf?lang=el

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

3. Vzpostaviti je potrebno cevovod za reden dotok besedil, ki se prevajajo v okviru delovanja javne uprave, kar vključuje identifikacijo in izločitev ali anonimizacijo besedil z občutljivimi/osebnimi podatki ter nato prenos prevodov na ustrezno lokacijo, kjer so na voljo za učenje strojnih prevajalnikov. V zvezi s tem je potrebno izvesti tudi aktivnosti za povečanje zavedanja o pomenu prevodov in drugih besedil za strojno prevajanje in storitve umetne inteligence.

   Rezultat: Vzpostavljen je cevovod za reden dotok besedil, ki se prevajajo v okviru delovanja javne uprave.

Tehnične nadgradnje algoritma strojnega prevajalnika

Postopek adaptacije modela prevajanja na izbrano domeno

Splošni model je naučen na korpusih splošnega jezika. Izkušnje kažejo, da imajo splošni modeli težave pri prevajanju besedil izven domene, zato je za prevajanje domensko specifičnih besedil ključnega pomena adaptacija na domeno. Pogoj za adaptacijo je razpoložljivost dovolj velike množice domensko specifičnih besedil, ki so lahko monolingualna ali bilingualna. Smiselno bi bilo proučiti različne metode, ki bi jih lahko uporabili za dodatno učenje modelov splošnega prevajalnika na teh besedilih in izboljšali kvaliteto prevajanja. Dodaten bilingualni korpus je lahko neposredno uporaben za učenje še nekaj dodatnih epoh. Monolingualni korpus ni neposredno uporaben. Lahko pa iz njega zgradimo sintetični vzporedni korpus tako, da uporabimo prevajanje nazaj/naprej (ang. back translation) in potem še nekaj dodatnih epoh učimo na sintetičnem korpusu.

Rezultat: postopek domenske adaptacije modela

Razvoj modelov za avtomatsko post-editiranje

Strojni prevodi običajno niso neposredno uporabni in jih je potrebno ročno popraviti do željene kakovosti. V strojnih prevodih se lahko pojavljajo določeni vzorci napak. Iz originalnih strojnih prevodov in popravljenih strojnih prevodov lahko ustvarimo neke vrste vzporedni korpus, na katerem učimo, namesto strojnega prevajalnika, strojni popravljalnik napak. Za učenje popravljalnika je potrebna velika množica popravljenih strojnih prevodov. Obstaja tudi nevarnost prekomernega popravljanja, ki v strojne prevode vnaša dodatne napake. Kot učinkovite so se izkazale metode, ki temeljijo na uporabi trojčkov: izvorna poved, strojni prevod, popravljen strojni prevod)[3]. Avtomatsko popravljanje prevodov je smiselno, če je kakovost strojnih prevodov dobra in ni potrebno kompleksno popravljanje. Preveriti bi veljalo različne metode učenja popravljalnika: na osnovi pravil, SMT in NMT.

Rezultat: model za avtomatsko post-editiranje

[3] http://dx.doi.org/10.18653/v1/P19-1292

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Razvoj faktoriziranih modelov prevajanja

Razvoj faktoriziranih modelov prevajanja in preučevanje njihovega delovanja na testni množici splošnega jezika in domensko specifičnih testnih množicah. V povedi lahko besednim oblikam dodajamo različne dodatne informacije (npr. leme, oznake MSD), s katerimi bi lahko izboljšali uspešnost prevajanja. Raziskave kažejo, da te informacije k uspešnosti prevajanja pri velikih učnih množicah ne doprinesejo dosti, so pa koristne, če so učne množice majhne. Torej bi lahko bilo faktorizirano prevajanje uporabno pri domensko specifičnih prevajalnikih.

Rezultat: faktoriziran model prevajanja

Razvoj podpornih orodij za predprocesiranje in postprocesiranje

V preteklosti je uspešnost strojnega prevajanja bila odvisna od ustrezno izvedenega predprocesiranja, ki zmanjša pomanjkanje podatkov (data scarcity) pri učenju prevajanja. Predprocesiranje zajema normalizacijo ločil (zamenjavo nekaterih ločil iz nabora znakov UTF-8 v ustrezna ločila manjšega nabora, popravilo pogostih napak), tokenizacijo (ločevanje stičnih ločil od besed) in truecasanje (pretvorba velikih začetnic glede na besedo in ne glede na to ali je beseda prva v povedi). Postprocesiranje obsega obratno pretvorbo po prevajanju. Z razpoložljivostjo velikih učnih množic za prevajalnik v splošni domeni je pomanjkanje podatkov postalo manj pereč problem, lahko pa ima še vedno vpliv v domensko specifičnih prevajalnikih. Z razvojem podpornih orodij za predproceisranje in postprocesiranje, ki bodo prirejena za uporabo v slovenskem jeziku, lahko podpremo uspešnost domensko specifičnih prevajalnikov in izvedemo raziskave glede uporabnosti predprocesiranja in postprocesiranja v različnih domenah.

Rezultat: podporna orodja za pred/post-procesiranje

Razvoj in/ali uporaba orodij za avtomatsko čiščenje učnih množic

V razpoložljivih vzporednih korpusih, ki jih običajno uporabljamo za učenje modelov prevajalnika se večkrat lahko pojavljajo segmenti z napakami. Napake je lahko neprimerna poravnava segmentov, kjer segment v enem jeziku ni prevod soležnega segmenta v drugem jeziku. To je lahko posledica napake pri poravnavi ali pa precejšnjega prirejanja besedila pri prevajanju. Pojavljajo se lahko tudi segmenti besedila v tretjem jeziku. V nekaterih vzporednih korpusih najdemo tudi poravnane segmente, kjer je razvidno, da je bila vsebina strojno prevedena. Z razvojem podpornih orodij za čiščenje vzporednih korpusov lahko takšne segmente izločimo iz učne množice prevajalnika in naknadno naučimo novi model za prevajanje ter primerjamo njegovo delovanje z modelom naučenim na neprečiščeni učni množici.

Rezultat: podporna orodja za čiščenje učnih množic

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Razvoj strojnega prevajalnika za področje izobraževanja

Specifičnost domene

Posebnost prevajanja za področje izobraževanja, še posebno ko govorimo o prevajanju predavanj, je že v samem jeziku. Ta je bliže spontanemu govoru kot branemu. Predavatelj med predavanjem pogosto ne govori v lepih, pravilnih knjižnih stavkih, ampak pogosto uporablja mašila, med govorom dela pavze na neobičajnih mestih (izven ločil), in verjetnost ponovitve besed ali napačnih štartov je velika. Tehnične vede predstavljajo še dodaten problem, saj pri teh pogosto nastopajo domensko specifične kratice, matematične enačbe polne eno-črkovnih spremenljivk in operatorjev, ter izrazi iz tujih jezikov. Omenjeno tipično negativno vpliva na kvaliteto strojne prepoznave govora, ki je podlaga za prevod. Prevajanje slabe podlage napake le še potencira. Javno dostopni viri poravnanih korpusov izhajajo po večini iz pisanega jezika. Za dvig primernosti prevajalnikov za področje izobraževanja bi bilo potrebno poskrbeti za povečanje količine razpložljivih poravnanih korpusov, ki so bliže spontanemu govoru ter pokrivajo različne domene področja izobraževanja.

Rezulat: domensko specifičen prevajalnik; družboslovje, tehnika

Razvoj modelov za sprotno prevajanje

Sprotno prevajanje/tolmačenje v področju izobraževanja ima velik potencial. Podpira odpiranje izobraževalnih institucij za tuje govoreče študente, Erasmus izmenjave in obenem pomaga pri doseganju ciljev dostopnosti. Tu postane pomemben vidik razpoznavalnika čim manjša latenca. V takih primerih uporabe strojni prevajalnik vedno nastopa kot korak za avtomatsko sprotno prepoznavo govora, ki z namenom doseganja čim manjše latence med izgovorjenimi besedami in njihovim izpisom napoveduje delne, a vedno daljše, hipoteze vse do trenutka, ko govorec ne naredi dovolj premora oz. se spremembe v hipotezah ne umirijo in napoved postane dokončna. V okviru te aktivnosti je potrebno razviti rešitve za sprotno prevajanje sprotno prepoznanega govora. Možne rešitve so prevajanje nastajajočih hipotez (angl. re-translation) in sprotno prevajanje (angl. streaming translation)[4]. Velik izziv pri tem je ohranjanje konteksta čez daljše besedilo, kljub prevajanju le manjših segmentov.

Rezultat: sprotni prevajalnik

Razvoj modelov za neposredno prevajanje iz govora v govor

Z namenom doseganja čim manjše latence najnovejše raziskave posegajo po metodologijah, ki skušajo zaobiti vmesne korake, npr. neposredna sprotna prepoznava in prevod govorjenega jezika v zapis v tujem jeziku[5], ter neposredno sprotno prevajanje iz govora v govor (angl. seech-to-speech)[6], ki namesto zapisa v tujem jeziku generira govor v tujem jeziku. Omenjena smer je trenutno v samem vrhu aktualnosti. V okviru te aktivnosti je

[4] https://arxiv.org/abs/2004.03643
[5] https://openai.com/blog/whisper
[6] https://arxiv.org/abs/2211.04508, https://arxiv.org/abs/2301.10606, https://arxiv.org/abs/2212.05805

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

potrebno preučiti obstoječe rešitve in jih dopolniti ter nadgraditi s specifikami, ki izhajajo iz slovenskega jezika.

Rezultat: prevajalnik iz govora v govor

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.