Načrt razvoja in vzdrževanja semantičnih tehnologij

Razvoj slovenščine v digitalnem okolju (RSDO)
DS3: Semantične tehnologije

V okviru projekta Razvoj slovenščine v digitalnem okolju (RSDO) (2020-2023) je v Delovnem sklopu 3 (DS3) potekala izdelava različnih jezikovno-tehnoloških orodij. Orodja in naučeni modeli, ki so na voljo zainteresirani javnosti, so dvignila nivo opremljenosti slovenskega jezika z jezikovnimi tehnologijami. Poleg tega orodja dosegajo primerljive rezultate z najboljšimi metodami v angleškem jeziku. Tehnologije so bile deloma prilagojene slovenskemu jeziku, prvenstveno pa so za izgradnjo zahtevala različne učne množice, ki so bile razvite v okviru DS1. V prihodnosti se pričakuje, da je poleg avtomatske gradnje korpusov (kot smo poskusili problem manjkajočih korpusov reševati v nekaterih primerih) še vedno pomembna izgradnja in vzdrževanje specializiranih učnih korpusov. Ločeno je bilo tekom projekta vzpostavljeno evalvacijsko orodje SloBench, ki omogoča neodvisno primerjavo tehnologij za posamezne jezikovno-tehnološke naloge. Slednje še dodatno omogoča jasen vpogled v stanje uspešnosti in razvoj novih tehnologij.

Ta dokument predstavlja trajnostni načrt za vzdrževanje in nadaljnji razvoj jezikovno-tehnoloških orodij za slovenski jezik. Dokument je razdeljen na tri dele, ki ustrezajo posameznim gradnikom trenutno aktualnih tehnologij. Trenutne tehnologije slonijo na uporabi učinkovitih jezikovnih modelov (Razdelek 1), ki vsebujejo širše leksikalno in semantično znanje. Jezikovne modele lahko uporabimo v specializiranih nalogah obdelave naravnega jezika, kjer pa je potrebno biti inovativen, uporabiti dodatne jezikovne vire in izdelati tehnologije, ki rešujejo izbrano nalogo, kot je na primer izdelava pametnih asistentov (Razdelek 2). Širše se lahko orodja uporabljajo v okviru večjih projektov, kjer je potrebno zagotoviti, da so na voljo ogrodja in podporna orodja ter avtomatsko zgrajene baze znanja (Razdelek 3). Vsi trije vidiki predstavljajo celoten ekosistem, ki povezuje raziskovalce na področju jezikovnih tehnologij ter zainteresirano javnost, ki lahko ta orodja enostavno uporablja.

1. Napredni jezikovni viri, jezikovni modeli in analize jezika

Jezikovni viri, modeli in analize jezika na nižjih nivojih služijo razumevanju naravnega jezika kot podlaga za napredne semantične analize. Poleg ročne gradnje korpusov in njihovega zbiranja je zelo pomembno, da se s pomočjo naprednih metod in korpusov v drugih jezikih avtomatsko izdelujejo tudi korpusi v slovenskem jeziku. Večje količine besedil morajo biti ustrezno organizirane, da so primerne za gradnjo uspešnih jezikovnih modelov. Pri tem je potrebno slediti najnovejšim trendom (ki se posodabljajo na mesečni osnovi) in izdelovati različne domenske modele. Pomembno je preverjati vsebovano znanje naučenih modelov ter zagotavljati, da modeli spoštujejo etične vidike ob morebitni širši uporabi. Analize naučenih modelov služijo tudi naprednim raziskavam jezikoslovcev, ki lahko nato predlagajo specifične izboljšave za izgradnjo najnovejših semantičnih tehnologij.

- Izgradnja večjih jezikovnih modelov (LaMDA, BLOOM, GPT, Galactica, LLaMA...) za slovenski jezik.

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.

    - Prilagoditev večjih večjezikovnih podatkovnih modelov slovenščini z dodatnim učenjem, pripravo vhodnega slovarja žetonov.
    - Zagotovitev dovolj računskih virov (10-100x kar je zdaj na voljo pri DGX100 in dosegljivo na Vegi).
    - Zagotovitev/kontinuirano pridobivanje podatkov podatkov (10x - 100x kar je zdaj na voljo) tako preko več slovenskih originalnih virov kot preko strojnega prevajanja angleških virov.
    - Razvoj modelov z upoštevanjem dodatnih jezikovnih virov/baz znanja/semantičnih grafov.
    - Prilagoditev večjih/kompleksnih modelov z namenom izdelave kompaktnih/manj zahtevnih modelov, ki ohranjajo nivo uspešnosti.
    - Zagotovitev ekipe, ki se bo ukvarjala samo z izgradnjo jezikovnih modelov (npr. 2 FTE).
- Sodelovanje in vključevanje v evropske iniciative za izgradnjo velikih jezikovnih modelov (npr. OpenGPT-X), ki bodo omogočili preboje in večjo interoperabilnost med evropskimi jeziki.
- Izdelava učnih množic za semantične tehnologije, ki še niso na voljo za slovenščino.
    - Primeri nalog so na primer poenostavljanje besedil, analiza čustev, zbirke po zgledu BIG-bench (204 naloge).
    - Bistveno povečanje množice za razdvoumljanje, sklepanje o pomenskem sosledju (angl. natural language inference), parafraziranje, avtomatsko generiranje razlag, zdravorazumsko sklepanje, logično sklepanje, ipd.
    - Upoštevanje etični vidikov pri pripravi množic.
    - Izgradnja ročnih/pol-avtomatskih leksikonov.
- Sistematična nadgradnja slovenskega Wordneta (sloWNet) z ročnim pregledom in povečanjem na celotni obseg Princeton WordNet in/ali English WordNet. Izdelava orodij za gradnjo in čiščenje korpusa.
- Zbiranje nestandardnih besedil (npr. narečnih) in izdelava orodij za semantično procesiranje nestandardnih besedil ali avtomatsko standardizacijo besedil. Zbiranje in analiza besedil zamejskih Slovencev in Slovencev po svetu (npr. časniki).
- Analiza specifičnosti (npr. leksikalne, morfološke lastnosti ali zakonitosti) slovenskega jezika v pomembnejših nalogah razumevanja jezika (našteto zgoraj). Priprava podatkovnih množic, ki bi bile primernejše za gradnjo kompleksnih modelov za slovenski jezik (npr. specifike pri odkrivanju koreferenčnosti, Winogradove sheme).

2. Izdelava jezikovno-tehnoloških orodij

Jezikovno-tehnološka orodja predstavljajo jedro obdelave naravnega jezika. Te tehnologije so lahko uporabljene samostojno ali integrirane v različne produkte, kot na primer avtomobili. Ta orodja so ključen pokazatelj opremljenosti jezika v času digitalizacije. Kombinacija in večja uspešnost delovanja teh orodij omogoča, da bo slovenski jezik primerljiv z ostalimi jeziki in ne bo zaostal v času digitalizacije. Spremljanje svetovnega razvoja in razvijanje/prilagajanje trenutno najuspešnejših postopkov bo ključno, da se bo slovenščina lahko ohranila tudi v digitalni dobi,

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.

sicer lahko (po-)ostane omejena le na lokalno komunikacijo znotraj skupnosti. Vsa spodaj predlagana orodja jo lahko približajo interoperabilnosti z ostalimi večjimi jeziki.

- Izdelava orodja za gradnjo baz (grafov) znanja (angl. knowledge graphs). Kot izhodišče se lahko uporabi terminologijo DSB ter jo opolnomoči s semantičnimi lastnostmi (npr. povezovanje z zunanjimi viri znanja - WikiData).
- Izdelava orodij za odgovarjanje na vprašanja in pametnih asistentov (v smislu ChatGPT). Upoštevanje verzioniranja, vključevanje specifičnega znanja, zunanjih virov (podobno omenjeno zgoraj), preverjanje delovanja/kuratorstvo, ipd.
- Izdelava orodij za odkrivanje strojno/avtomatsko generiranih besedil.
- Izdelava orodij za primerjalno analizo besedil
    - Nadgradnja sistema za diahrone semantične premike (posodobitev z novejšimi viri, izboljšava metod za modeliranje, vključno z boljšo obdelavo polisemičnih besed in vizualizacije ter interpretacije).
    - Orodja za primerjavo poljubnih besedil s strani uporabnikov (npr. primerjava različnih literarnih avtorjev, različnih medijskih virov)
- Izdelava orodja za luščenje ključnih besed (npr. na podlagi vhodnega besedila).
- Izdelava orodja z krnjenje (vključitev tudi v jezikovni cevovod).
- Izdelava orodij za obdelavo literarnih besedil in lirike. Na primer identifikacija rim, asonanc, aliteracij, ipd. Izdelava orodij/pomočnikov ustvarjalcem/umetnikom.
- Izdelava orodij, ki poleg besedila uporabljajo ali generirajo večmodalne podatke (npr. izdelava slik, videov, glasbe).
- Izdelava orodij za semantično primerjanje besedil, iskanje semantične podobnosti stavkov, dokumentov. Prilagoditev obstoječih metrik slovenskemu jeziku (npr. soundex) za vključitev v orodja za združevanje entitet ali prepoznavanje duplikatov.

3. Izdelava ogrodij in vzdrževanje ter posodabljanje infrastrukture

Jezikovna infrastruktura omogoča, da jezikovne tehnologije in vire zbiramo in hranimo na način, da so dostopne vsakomur. Zaradi nabora množice različnih pristopov je pomembno zagotoviti odprtost ter javno primerjavo metod, da se lahko končni uporabniki lažje odločijo, kaj je zanje najbolj primerno. Ogrodja pa morajo omogočati povezavo več orodij med seboj ali enostaven grafični vmesnik, da lahko orodja kdorkoli uporablja za svoje namene ali jih testira nad svojimi besedili. Ker nekatera orodja potrebujejo specializirano strojno opremo, bi bilo potrebno zagotoviti, da so te tehnologije na voljo sleherniku vsaj v nekem omejenem obsegu.

- Izdelava javno dostopnih storitev jezikovnih tehnologij
    - Kontinuirano sledenje napredku na področju jezikovnih tehnologij in ponujanje orodij za omejeno prosto uporabo. Zagotovitev podpore (npr. 1 FTE), ki bi preverjala in posodabljala aktualna orodja v repozitorijih kode ter DEMO okoljih.
    - Implementacija razširljivega in enostavnega ogrodja za vključevanje in zaganjanje obstoječih orodij za jezikovne tehnologije. Uporaba načina delovnih tokov, kar bi omogočilo uporabo metod tehnološko neveščim uporabnikom.

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.

    - Zagotovitev ekipe (npr. vsaj 2-3 FTE) in ustrezne strojne opreme za razvoj, vzdrževanje in posodabljanje ogrodja (zasnova orodja in podatkovnega modela je bila pripravljena v okviru RSDO - R3.1.2).
- Sistematičen razvoj Digitalne slovarske baze pod licenco CC BY-SA (po zgledu Wikipedije), vključno s tehnološko podporo. Dopolnitev z zunanjimi viri, kot je na primer slovar sopomenk, SSKJ, ipd. - nadaljevanje aktivnosti na odkupu/pridobivanju virov.
- Aktivna in organizirana podpora nadgradnji WikiData, Wikipedia, Wiktionary itd. s slovenskimi podatki.
- Vzdrževanje in nadgradnje evalvacijskega ogrodja SloBench (https://slobench.cjvt.si/), kjer se nahajajo aktualni primerjalniki vseh orodij za pomembnejše naloge obdelave naravnega jezika.

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.