Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*

in
Center za *jezikovne vire*
in *tehnologije*

Semantični jezikovni viri in tehnologije – stanje slovenščine

prof. dr. Marko Robnik-Šikonja

UL
27. 9. 2019

Kako bo mogoče? – posvet o digitalni prihodnosti slovenščine

Vloga semantičnih jezikovnih tehnologij

- semantične tehnologije so ključni del umetne inteligence na področju jezika
- omogočajo **razumevanje** naravnega jezika
- so del vsakodnevnega življenja ljudi v razvitih državah
- vloga teh tehnologij v vsakdanjem življenju se bo še povečevala
- … še posebej med mlajšimi generacijami
- … med delovno aktivnim prebivalstvom
- … in v okviru asistenčnih tehnologij tudi med starejšimi

Katere so semantične jezikovne tehnologije?

- v zadnjem času velik napredek v svetu, kot posledica tehnološkega preboja na področju globokih nevronskih mrež
  - bistveno izboljšano strojno prevajanje, predvsem za večje jezike
  - razpoznavanje govora na nivoju ljudi, solidno generiranje govora
  - solidno avtomatsko povzemanje
  - uporabni odgovori na vprašanja
  - generiranje preprostejših besedil (poročila, deli člankov)

Stanje slovenščine?

- Kako bo mogoče **uporabljati slovenščino v sklopu prihodnjih inteligentnih naprav in storitev?**
- strojno razumevanje jezika je v velikem javnem interesu
- … in je kot tako prepoznano v strateških dokumentih EU in Slovenije
- digitalni prepad med jeziki se poglablja
- EU vlaga v medjezikovne tehnologije, ki koristijo vsem jezikom
- da bi izkoristili te tehnologije, je potrebno imeti jezikovno specifične semantične vire in tehnologije
- slovenščina ima velik korpus, nima odptodostopnega temeljnega slovarja, niti omembe vrednih semantičnih virov
- način delovanja semantičnih tehnologij ni prilagojen specifičnostim slovenščine (sklanjatve, dvojina, večbesedne zveze, pravopis, pragmatika)
- slovenščina zaostaja tudi glede na jezike v svoji skupini, med jeziki manjših tehnološko razvitih držav

Kako je mogoče vzpostaviti delovanje semantičnih tehnologij v slovenščini ?

- razvoj **odprtodostopnih označenih jezikovnih virov**, s katerimi se jezikovno znanje ljudi in njihovo razumevanje sveta zapiše v strojno berljivi obliki,
- razvoj in prilagoditev najnovejših tehnologij **globokih nevronskih mrež** slovenščini, kar bo omogočilo procesiranje teh virov in izdelavo strojnih modelov za različne naloge razumevanja jezika

Kako je mogoče: viri

Razvoj virov:

- Osnovni: ročno označeni jezikovni podatki, npr. oblikoskladenjsko označene in skladenjsko razčlenjene učne zbirke
- Srednji nivo: ročno označene učnimi množicami za razločevanje različnih pomenov besed, prepoznavanje anafore in koreferenc in pomensko združevanje besed.
- Za višjenivojske naloge razumevanja jezika (strojno prevajanje, odgovori na vprašanja, povzemanje): potrebno je sistematično pridobivanje in zbiranje iz različnih človeških dejavnosti

Konkretni cilji:

- izboljšati slovenski WordNet (povečati, preveriti, dodati razlage, povezati z drugimi jeziki)
- podpirati razvoj slovenske Wikipedije in Wikislovarja (tudi z odkupom pravic)
- nadgraditi oblikoskladenjsko označene korpuse s podatki o anaforah in koreferencah, imenskih entitetah in besednih pomenih
- razviti slovenski PropBank (korpus besedil z označenimi udeleženskimi vlogami) in FrameNet (korpus konceptualno označenih besedil)
- izgradnja in vzdrževanje učnih množic za strojno prevajanje (prevodni pomnilniki), odgovore na vprašanja in povzemanje besedil.

Kako je mogoče: tehnologije

Razvoj tehnologij

- hiter razvoj jezikovnih tehnologij, viri nastajajo počasneje
- posebnostim slovenskega jezika je potrebno sprotno prilagajati najnovejše tehnologije globokih nevronskih mrež
- podpirati tehnološko sposobne ustanove, da se zagotovi dovolj usposobljenega kadra

Cilji:

- **osnovna orodja** za obdelavo jezika: nevronski lematizator, nevronski oblikoskladenjski označevalnik in nevronski skladenjski razčlenjevalnik
- **srednji sloj** semantične infrastrukture: nevronski jezikovni modeli, vektorske vložitve, označevanje udeleženskih vlog, prepoznavanje anafore in koreferenc, semantični okvirji, razdvoumljanje pomena besed v sobesedilu
- tehnologije za **razumevanja jezika**: nevronsko strojno prevajanje, avtomatsko povzemanje, odgovarjanje na vprašanja, napredni pripomočki za pisanje itn.

Kako je mogoče, da gre svet naprej?

- Sematični viri in tehnologije umetne inteligence se hitro razvijajo.
- Primer: zbirka devetih nalog GLUE (General Language Understanding Evaluation) (Wang et al, 2019)
  - Ali je stavek slovnično pravilen?
  - Kakšna je čustvena obarvanost stavka? (sentiment)
  - Ali je par stavkov semantično enakovreden? (novice)
  - Ali je par vprašanj semantično enak? (Quora)
  - Koliko sta si stavka semantično podobna? 1-5
  - Ali stavek logično sledi iz prejšnjega ali mu nasprotuje?
  - Odgovor na vprašanje iz danega odstavka?
  - Prepoznavanje vzročnosti v stavkih na podlagi prejšnjega besedila.
  - Prepoznavanje reference zaimka vira v daljšem besedilu.
- GLUE testira razumevanje mnogih kategorij (leksikalna semantika, predikate in njihove argumente, logiko, splošno znanje)
- vlaganje v semantične jezikovne tehnologije zato danes istočasno pomeni tudi vlaganje v tehnologije umetne inteligence, ki so eden strateških ciljev v EU in Sloveniji

Kako bo mogoče: predpogoji

- zagotoviti dolgoročno stabilno financiranje razvoja semantičnih jezikovnih virov in tehnologij, ki bodo odprtokodni in prosto dostopni

- s programskim financiranjem ARRS poskrbeti za **kadrovsko stabilno zasedbo jezikovnotehnoloških strokovnjakov** na institucijah, ki so sposobne razvoja teh tehnologij (Univerza v Ljubljani, Insitut Jožef Stefan, Univerza v Mariboru): vsaj **12 FTE oz. 600.000€/letno**
- z rednimi razpisi financirati **razvoj pomembnih virov in tehnologij** za razumevanja jezika; ocena: **500.000€/letno**
- to bo omogočilo tudi hitrejši in cenejši razvoj ostalih jezikovnih virov: temeljnih, terminoloških, narečnih, zgodovinskih in posebnih slovarjev, slovnic in priročnikov