Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU

Predavanje NUK
2023-02-22

Pregled predavanja

Uvod

Kje se srečujemo z jezikovnimi podatki

Humanistične vede

Empirično podprte jeziko(slo)vne raziskave:

- temeljijo na realnih besedilih
- za učinkovito uporabo potrebujemo analitična orodja
- jezikoslovci, slovaropisci, zgodovinarji, družboslovci...

Računalništvo

Jezikovne tehnologije:

- obdelava jezika postaja vedno bolj raziskovalno/komercialno zanimivo področje
- glavna paradigma: nadzorovano strojno učenje
- programi so večinoma jezikovno neodvisni, potrebujejo pa učne (ročno označene) podatke za učenje modela in testne podatke za evalvacijo

Kaj so jezikovni viri

Korpusi

- enovito kodirana in dokumentirana zbirka besedil
- označena (ročno ali strojno)
- referenčni/specializirani; eno/večjezični; pisni/govorni

Leksikalni viri

- besedišča jezika za uporabo v programih
- strojno berljivi slovarji

Modeli jezika

- podatki za nek program, ki mu omogoči označevanje besedil v nekem jeziku za neko raven označevanja
- npr. CLASSLA model za lematizacijo slovenščine; besedne vložitve makedonščine

Nekaj zanimivih korpusov

- Gigafida: referenčni korpus sodobne slovenščine 1990–2018
  (1 milijarda besed, samo na konkordančnikih)
- KAS / OSS: korpusa slovenskih znanstvenih besedil
  (2 milijardi besed, terminologija)
- Janes: korpus besedil družbenih omrežij
  (250 milijonov besed, normalizacija)
- IMP: korpus starejše slovenščine
  (15 milijonov besed, normalizacija)
- siParl: korpus parlamentarnih razprav 1990 - 2022
  (200 milijonov besed, bogati metapodatki)
- MetaFida: združeni korpus
  (3.5 milijarde besed, samo konkordančniki)

Ponovna uporaba

Klasični pristop

- za vsako raziskavo izdelati jezikovne vire posebej
- viri nedostopni drugim raziskovalcem

Slabosti

- izdelava jezikovnega vira je lahko zelo draga in dolgotrajna: velika izguba časa in denarja, če se to počne večkrat
- vzdržuje se monopol institucij, ki so vire izdelale
- kasnejši raziskovalci ne morejo preveriti ali poboljšati prvih rezultatov
- viri ne morejo biti uporabljeni pri razvoju produktov

Odprt dostop do rezultatov raziskovalnih projektov

- Brez ovir do publikacij in podatkov
  - prihranek denarja in časa;
  - izogibanje ponavljanju dela;
  - spodbujanje sodelovanja;
  - večja transparentnost znanstvenega procesa;
  - spodbujanje inovacij
- Odprta znanost: močan trend v EU in Sloveniji
- Problemi pri omogočanju odprtega dostopa do jezikovnih virov:
  - avtorske pravice nad besedili
  - varovanje zasebnosti (tudi pravica do pozabe): GDPR
  - pogoji uporabe spletnih portalov (npr. Twitter)

Kaj je CLARIN(.SI)

Raziskovalne infrastrukture

Kaj je RI?

Naprave, podatki in storitve, ki jih znanstvena skupnost uporablja pri raziskovanju na svojem področju.

- ESFRI: European Strategy Forum on Research Infrastructures
- Razvojni načrti: 2006 (35 RI), ..., 2018 (55), 2021 (66)
- 22 RI je organiziranih kot ERIC
  (European RI Consortium = evropska pravna oseba)
- Slovenija sodeluje v 20/22 RI, 2 s področja humanistike:
- DARIAH ERIC / DARIAH-SI = Digital Research Infrastructure for the Arts and Humanities / Digitalna raziskovalna infrastruktura za umetnost in humanistiko: INZ + ZRC SAZU
- **CLARIN ERIC / CLARIN.SI** = Common Language Resources and Technology Infrastructure / Infrastruktura za jezikovne vire in tehnologije

CLARIN: Common Language Resources and Technology Infrastructure

- Vizija: digitalni jezikovni viri in orodja za vse (evropske) jezike so dostopni prek enotne prijave za raziskovalce v humanistiki in družboslovju
- Namenjena je dolgotrajnemu in obsežnemu hranjenju ter dostopu do jezikovnih virov in tehnologij
- Prispevek k ohranjanju in podpiranju večjezične evropske kulturne dediščine
- Paradigma sodelovanja pri razvoju virov in orodij, zagotavljanje večkratne uporabnosti in prilagajanja individualnim potrebam

CLARIN ERIC

- Sedež na Nizozemskem
- Trenutno 22 držav članic + 3 opazovalke
- Podporno osebje, odbori za upravljanje, delovne skupine
- Večina dela se odvija v okviru nacionalnih konzorcijev

Kaj CLARIN ERIC ponuja slovenskim raziskovalcem?

- S slovensko EduGain prijavo dostop do vseh virov in storitev centrov CLARIN držav članic
- Spletne storitve, npr. virtualni jezikovni observatorij
- Podpora ciljnim projektom, npr. razvoju učnih vsebin, izvedbi delavnic za uporabnike, snovanju evropskih projektnih prijav
- Infrastruktura znanja:

CLARIN.SI

- Začetek dela v 2015
- Institut "Jožef Stefan":
  - Odsek za tehnologije znanja (E8)
  - Laboratorij za umetno inteligenco (E3)
  - Center za mrežno infrastrukturo (CMI)
- Organiziran kot konzorcij 12 partnerjev:
  - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
  - 4 raziskovalni inštituti: ZRC SAZU, IJS, INZ, ZRS Koper
  - 1 društvo in 1 zavod: SDJT, Trojina
  - 2 podjetji: Amebis, Alpineon

Storitve CLARIN.SI

Delovanje CLARIN.SI

Trije stebri

1. **Repozitorij jezikovnih virov in orodij**
2. Spletne storitve (glavna: **konkordančniki**)
3. Podpora digitalni humanistiki in jezikovnim tehnologijam (prenos znanja, dogodki, projekti)

Repozitorij

- Trenutno najpomembnejša storitev CLARIN.SI
- Arhiv $>$ 500 jezikovnih virov in orodij, od tega $>$ 250 (tudi) za slovenski jezik: korpusi, slovarji, besedišča, modeli, programi
- Samoarhiviranje + uredniški pregled
- Repozitorij certificiran s strani CLARIN in Core Trust Seal
- Dolgotrajno hranjenje, avtentikacija in avtorizacija, stalni identifikatorji, eksplicitni pogoji uporabe in licence
- Pomemben doprinos k odprti znanosti

Konkordančniki CLARIN.SI

- Orodja za (spletno) analizo korpusov
- Besede v kontekstu, frekvenčni seznami, ključne besede, kolokacije, ...
- Podpirajo delo z zelo velikimi korpusi (več milijard besed)
- Korpusi so lahko bogato označeni:
  - strukture (besedilo, odstavek, termin, . . .)
  - metapodatki (leto izdaje, vrsta besedila, spol avtorja,. . .)
  - atributi pojavnic (oblikoskladenjska oznaka, lema, normalizirana oblika,. . .)
- Bogat poizvedovalni jezik
- Raznovrstni izpisi in analize
- RESTful vmesnik, tj. poizvedbe možne prek URLjev (rezultat lahko tudi v JSON ali XML)
- Ponujajo cca. 100 korpusov v 33 jezikih z 20 milijard besed

Strokovna podpora in diseminacija

- Središče znanja za južnoslovanske jezike CLASSLA: strokovna podpora pri uporabi jezikovnih virov in tehnologij za južnoslovanske jezike
- Od 2018 letna finančna podpora projektom (30k), letno izbranih na odprtem razpisu za člane konzorcija (25 uspešno zaključenih)
- Organizacija in podpora dogodkom (Jezikovne tehnologije in digitalna humanistika, snemanje predavanj JOTA, EURALEX 2108, TSD 2019)
- Obveščanje in promocija (predstavitve na konferencah, študentom, izvedba tečajev)

Vpetost v projekte in RI

- MIZŠ (2018–2021): nadgradnja strojne opreme
- EU ELEXIS (2018–2022): repozitorijska zbirka metapodatkov 143 digitalnih slovarjev
- MK RSDO (2020–2023): pregled in arhiviranje jezikovnih virov projekta
- RI CESSDA/ADP RDA Node Slovenia (2019–2020): pregled in analiza slovenskih repozitorijev raziskovalnih podatkov
- RI DARIAH-SI/INZ: sodelovanje na področju standardizacije zapisa in izdelave korpusov parlamentarnih podatkov
- CLARIN ERIC:
  - 2 manjša projekta (2016, 2019) + mednarodni delavnici
  - ključna vloga v "CLARIN Flagship" projektih: ParlaMint I (2020–2021), ParlaMint II (2022–2023).
  - sodelovanje v delovnih telesih za pravna vprašanja, za standardizacijo, za uporabniška vprašanja
  - več nagrad in priznanj slovenskim znanstvenikom za delo, povezano s CLARIN

Zaključki

Zaključki

Na kratko

CLARIN.SI nudi možnost trajnega arhiviranja jezikovnih virov, odprt in brezplačen dostop do jezikovnih virov, orodij in storitev za (slovenske) raziskovalce in (kjer le mogoče) podjetja ter podporo pri ustvarjanju, arhiviranju in uporabi jezikovnih virov in orodij.

Nadaljnje informacije

- `https://www.clarin.eu/`
- `https://www.clarin.si/`
- ERJAVEC, Tomaž, DOBROVOLJC, Kaja, FIŠER, Darja, JAVORŠEK, Jan Jona, KREK, Simon, KUZMAN, Taja, LASKOWSKI, Cyprian Adam, LJUBEŠIĆ, Nikola, MEDEN, Katja. Raziskovalna infrastruktura CLARIN.SI. V: Jezikovne tehnologije in digitalna humanistika : zbornik konference. 2022. str. 47-54. `https://nl.ijs.si/jtdh22/pdf/JTDH2022_Erjavec-et-al_Raziskovalna-infrastruktura-CLARIN.SI.pdf`

Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU

Predavanje NUK
2023-02-22