Raziskovalna infrastruktura projekta RI-SI CLARIN

Darinka Verdonik,* Matej Rojc,* Izidor Mlakar,* Danilo Zimšek,* Zdravko Kačič,* Milan Ojsteršek* in Tomaž Erjavec†

* Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška 46, 2000 Maribor
darinka.verdonik@um.si; matej.rojc@um.si; izidor.mlakar@um.si; danilo.zimsek@um.si;
zdravko.kacic@um.si; milan.ojstersek@um.si
† Odsek za tehnologije znanja, Institut Jožef Stefan« Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

1 Uvod

Cilj evropske infrastrukture CLARIN (CLARIN ERIC, 2020) je spodbujanje raziskovalne dejavnosti na področju humanističnih in družbenih ved. Ta vizija se uresničuje z gradnjo in delovanjem raziskovalne infrastrukture v skupni uporabi, ki raziskovalnim skupnostim zagotavlja jezikovne vire, tehnologije in strokovno znanje. Slovenska raziskovalna infrastruktura CLARIN.SI (Erjavec et al. 2014) je vzpostavljena kot dolgoročni infrastrukturni projekt v okviru evropskega konzorcija CLARIN ERIC (European Research Infrastructures Consortium) ter naj bi olajšala in spodbujala v mednarodno okolje vpeto raziskovanje slovenščine, podporo razvoju digitalnih jezikovnih virov za slovenščino, digitalnih pripomočkov za komunikacijo med računalnikom in človekom ali namenjenih rabi v izobraževalnih okoljih ter lajšanju komunikacijskih zadreg oseb s posebnimi potrebami.

Konzorcij CLARIN.SI v mednarodnem projektu CLARIN zagotavlja prisotnost slovenskega jezika v virih in storitvah infrastrukture CLARIN. Institut »Jožef Stefan« (IJS) kot tehnični center konzorcija vzdržuje repozitorij jezikovnih virov, ki ustreza zahtevnim merilom, kakršna postavlja evropski CLARIN ERIC, in vsebuje že več kot 180 odprto dostopnih jezikovnih virov. CLARIN.SI ponuja tudi storitve za korpusno jezikoslovje in jezikovne tehnologije, na prvem mestu spletne konkordančnike z bazo več kot 50 jezikoslovno označenih korpusov. Center za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT UL) omogoča skozi portal viri.cjvt.si dostop do referenčnih korpusov slovenskega jezika (Gigafida 2.0 in Kres) tako za raziskave kot za poučevanje slovenščine, poleg tega pa njihovi strežniki ponujajo tudi dostop do slovarskih baz, npr. do Slovarja sopomenk sodobne slovenščine ter do Kolokacijskega slovarja sodobne slovenščine. S CLARIN povezano delo na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru (UM) se osredotoča na razvoj govornih tehnologij in virov.

Institut »Jožef Stefan« je skupaj s partnericama Univerzo v Ljubljani in Univerzo v Mariboru pridobil projekt »Razvoj raziskovalne infrastrukture za mednarodno konkurenčnost slovenskega RRI prostora – RI-SI – CLARIN« (s kratkim imenom »RI-SI CLARIN«) v skupni vrednosti 467.000 EUR za nakup nove raziskovalne opreme. Projekt traja od decembra 2018 do septembra 2021, financirata pa ga Ministrstvo za izobraževanje, znanost in šport in Evropski sklad za regionalni razvoj. Operacija je namenjena predvsem izboljšanju strojne opreme CLARIN.SI. Ta je pred začetkom projekta zajemala gručo treh računalnikov za namene portala www.slovenscina.eu in strežnik, ki je pokrival vse druge storitve CLARIN.SI (domača stran, repozitorij, konkordančniki, spletne storitve). Navedena strojna oprema ni omogočala posodabljanja infrastrukturnih storitev za znanstveno delo na področju računalniškega jezikoslovja in digitalne humanistike in je bila popolnoma neprimerna za eksperimentalno delo. Nova oprema, nabavljena v okviru RI-SI CLARIN, omogoča oz. bo omogočala bistveno hitrejše, intenzivnejše in visoko kakovostno vključevanje v velike mednarodne projekte. S tem prispeva h krepitvi mreže raziskovalne infrastrukture človeških virov v znanosti ter prostemu pretoku ljudi, zamisli in znanja v evropskem raziskovalnem prostoru.

Cilji že izvedenih in še načrtovanih nabav so med drugim:

- zagotoviti nadaljnje delovanje tehničnih storitev infrastrukture CLARIN.SI,
- omogočiti hranjenje velikih multimodalnih jezikovnih podatkov,
- omogočiti, da CLARIN.SI sledi paradigmi »velepodatkov« (angl. big data),
- omogočiti, da CLARIN.SI ponuja javno dostopne spletne storitve za obdelavo velikih količin slovenskih besedil,
- vzpostaviti namensko gručo računalnikov s pospeševalniki GPGPU za potrebe globokega učenja.

2 Nova oprema na Institutu »Jožef Stefan«

IJS v sklopu operacije nabavlja sledečo raziskovalno opremo:

1. Gruča za spletne storitve: dva strežnika z dvema 32-jedrnima procesorjema in dva strežnika z dvema 24-jedrnima procesorjema, vsi štirje z 2,3 GHz delovne frekvence, 1 TB pomnilnika, RAID krmilno kartico, 4 diski, diskovnim poljem s 24 TB in kartico za priklop na zunanja diskovna polja prek povezave FC z redundantnimi napajalniki.
2. Strežnik repozitorija: dva redundantna strežnika (eden kot produkcijski, drugi kot razvojni in pomožni) z dvema 16-jedrnima procesorjema, 2,3 GHz delovne frekvence in 768 GB pomnilnika, RAID krmilno kartico, diskovnim poljem s 24 TB in kartico za priklop na zunanja diskovna polja prek FC povezave ter redundantnimi napajalniki.
3. Diskovno polje za varnostne kopije: razširitveno diskovno polje z redundantnim priklopom in redundantnimi napajalniki s 16 diski po 10 TB kapacitete, certificirano za vgradnjo s strani proizvajalca diskovnega polja.
4. Stikalo za optični kanal: 2x nadgradnja stikala FC z 12 vrati HP SN 3000B na 24 vrat za povezavo obstoječih in novih diskovnih polj na nove strežnike.

3 Nova oprema na Centru za jezikovne vire in tehnologije UL

Center za jezikovne vire in tehnologije Univerze v Ljubljani je za potrebe izdelave in testiranja novih storitev nabavil strežniško rezino z dvema procesorjema po 12 jeder s 512 GB pomnilnika in predpomnilniški disk SSD velikosti vsaj 960 GB. Strežniška rezina je povezana z obstoječim diskovnim poljem s 10 TB efektivne kapacitete.

4 Nova oprema na Univerzi v Mariboru

Nova oprema na Univerzi v Mariboru, nabavljena do septembra 2020, vključuje gručo GPU-strežnikov, ki je optimirana za izvajanje aplikacij, temelječih na uporabi globokega učenja, strežnike za obdelavo velikih jezikovnih podatkov in diskovno polje za hranjenje velikih količin jezikovnih podatkov. V prihodnje je predvidena samo še ena nabava, tj. nadgradnja gruče GPU-strežnikov z novejšimi enotami GPU v letu 2021.

4.1 NVIDIA DGX-1

Sistem NVIDIA DGX-1-V100-EDU/32GB sestoji iz programske in strojne opreme, ki je optimirana za globoko učenje. Vključuje 8 grafičnih kartic Tesla V100 s tehnologijo NV-Link, vsaka grafična kartica ima 32 GB grafičnega pomnilnika, 2 centralni procesni enoti Intel Xeon E-2698 v4 20-core 2.20 GHz s skupno 40 procesorskimi jedri. Skupno zajema platforma 40.960 jeder NVIDIA CUDA® (FP32) in 20.480 jeder NVIDIA CUDA® (FP64). Poleg tega zagotovi 5120 Tensor jeder in 512 GB sistemskega pomnilnika DDR4 LRDIMM frekvence 2133 MHz. DGX-1 vključuje za shranjevanje podatkov še 1x 480 GB SSD (Intel S3610), na katerem je nameščen sistem, in 4x 1,92 TB SSD v polju RAID 0, ki zagotavlja hitro zapisovanje in branje uporabnikovih podatkov. Mrežna povezljivost je zagotovljena z dvema priključkoma 10 GbE. Infrastruktura teče na operacijskem sistemu Ubuntu Server Linux OS DGX-1 s priporočljivim grafičnim gonilnikom. Maksimalna poraba moči sistema pri polni obremenitvi je 3200 W.

4.2 Uporaba DGX-1

Dostop do DGX-1 imajo osebe, ki se uvrstijo na seznam uporabnikov. Na ta seznam se lahko uvrstijo zaposleni pri projektnih partnerjih (Univerza v Mariboru, Insitut !Jožef Stefan« in Univerza v Ljubljani), njihovi magistrski in doktorski študenti ter gostujoči raziskovalci, ki bodo opremo uporabljali v namene, skladne s projektom RI-SI CLARIN. Uporabniki so dolžni pri uporabi slediti navodilom uporabe in izvajati naloge skladno s tem, kar napovejo. Prav tako uporabniki na sistemu ne smejo hraniti podatkov, temveč jih morajo prenašati na lastne diskovne nosilce.

Dostop do sistema je omogočen preko upravljalca programskih bremen SLURM, ki omogoča upravljanje z viri in razvrščanje nalog v vrsto glede na parametre, ki jih določi uporabnik. Gre za sistem, pri katerem uporabnik kvantitativno določi vire, ki jih za izvajanje neke naloge potrebuje. Naloga se uvrsti v vrsto. Ko so na voljo specificirani viri, se naloga izvede.

Vsak uporabnik DGX-1 za vsako svojo nalogo napove potrebe po kapacitetah: število GPU-jev, število CPU-jeder, količino delovnega pomnilnika in čas trajanja naloge. Če viri niso na voljo, se uporabniki razvrščajo glede na to, koliko virov zahtevajo. Tisti, ki zahtevajo veliko virov, dobijo nižjo prioriteto.

4.3 Strežniška infrastruktura za obdelavo in hranjenje velepodatkovnih jezikovnih virov

Drugi sklop opreme na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru vključuje strežniško infrastrukturo za obdelavo velikih jezikovnih podatkov in diskovno polje za hranjenje velikih podatkov. Strežniška infrastruktura je sestavljena iz treh fizičnih strežnikov (vozlišč). Vsak izmed teh je opremljen z dvema procesorjema Intel Xeon Gold 5218 2.30 GHz (skupaj 32 jeder) in 512 GB (Dual Rank x4 DDR4-2933) sistemskega pomnilnika. Krmilnik diskov je Smart Array P408i-a/2 GB. Za sistemski disk sta uporabljena dva SSD-diska v polju RAID 1 in s skupno kapaciteto 800 GB. Za obdelavo podatkov je uporabljenih pet SSD-diskov v polju RAID 5 in skupno kapaciteto 7,6 TB ter en dodaten trdi disk kapacitete 2,4 TB. Povezljivost z omrežjem je zagotovljena s štirimi mrežnimi vmesniki. Na dveh strežnikih sta nameščena operacijska sistema Linux, medtem ko ima tretji strežnik nameščen operacijski sistem Windows Server. Infrastruktura vključuje še dve diskovni polji NAS Synology DS1819+ DiskStation za hrambo in obdelavo velikih jezikovnih podatkovnih zbirk. Skupna kapaciteta obeh diskovnih polj je 224 TB (16x NAS trdi disk 16 TB IronWolf v polju RAID 5).

S predlagano strežniško infrastrukturo se bodo izboljšale računske zmogljivosti, kar bo zelo povečalo število in obseg izvajanja raziskav na področju ugotavljanja pomena in ekstrakcije terminologije ter znanja iz besedil v slovenskem in drugih jezikih. Lotili pa se bomo lahko tudi najtršega oreha na področju detekcije plagiatov, ki je vezan na ugotavljanje direktnega prevajanja besedil iz drugih jezikov. Hranili bomo lahko precej več vhodnih jezikovnih virov, iz katerih bomo lahko pridobivali kvalitetne podatke in na njih izvajali analize ter iz njih ekstrahirali znanje.

5 Zaključek

Infrastrukturo CLARIN.SI smo načrtovali kot storitveno in razvojno infrastrukturo, ki je komplementarna SLING (SLING 2020) in HPC RIVR (Univerza v Mariboru 2019). CLARIN.SI bo razvil svojo obstoječo platformo stabilnih storitev in repozitorijev. Sistemi za razvoj in interaktivne obdelave pa so načrtovani kot komponente v omrežju SLING, ki bodo omogočile soopravilnost s sistemi v omrežju in bodo že v začetku lahko uporabljali kapacitete HPC RIVR za večje obdelave. Ker sta razvoj in vzdrževanje CLARIN.SI in HPC RIVR do določene mere sorodna, so potrebna sorodna znanja in veščine, zato lahko namestitev gruče GPU na Univerzi v Mariboru izkoristimo tudi za prenos znanja in tvorbo jedra znanja v okviru Univerze v Mariboru. Tako pričakujemo pozitivne sinergijske učinke sodelovanja, izmenjave znanja, skupne projekte ter visoko stopnjo soopravilnosti. CLARIN.SI bo prek HPC RIVR tudi uresničeval zahteve za dolgoročno varovanje svojih raziskovalnih podatkov na oddaljeni lokaciji, s člani SLING pa bo sodeloval prek skupnega vlaganja v infrastrukturni razvoj Slovenije prek izdelave tehničnih publikacij.

Literatura

CLARIN ERIC (2020) CLARIN - European Research Infrastructure for Language Resources and Technology. Dosegljivo na https://www.clarin.eu/. Obiskano 1. 6. 2020.

Tomaž Erjavec, Jan Jona Javoršek, Simon Krek (2014) Raziskovalna infrastruktura CLARIN.SI. *Zbornik Devete konference Jezikovne tehnologije. Informacijska družba – IS 2014*, 9. – 10. 10. 2014, Institut "Jožef Stefan", Ljubljana.

SLING (2020) Slovensko nacionalno superračunalniško omrežje. Dosegljivo na http://www.sling.si/sling/. Obiskano 1. 6. 2020.

Univerza v Mariboru (2019) Projekt HPC RIVR. Dosegljivo na https://www.hpc-rivr.si/ Obiskano 1. 6. 2020.