TakeLab

Laboratorij za analizu teksta i inženjerstvo znanja
Text Analysis and Knowledge Engineering Lab
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 775

JEZIČNI MODEL HRVATSKOGA JEZIKA ZASNOVAN NA POVRATNIM NEURONSKIM MREŽAMA

Leo Zuanović

Zagreb, lipanj 2014.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA DIPLOMSKI RAD PROFILA

Zagreb, 10. ožujka 2014.

DIPLOMSKI ZADATAK br. 775

Pristupnik: **Leo Zuanović**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Jezični model hrvatskoga jezika zasnovan na povratnim neuronskim mrežama**

Opis zadatka:

Jezični modeli služe za procjenu vjerojatnosti riječi u danom kontekstu i jedan su od osnovnih alata u obradi prirodnog jezika. Tradicionalni se jezični modeli oslanjaju na statistiku o pojavljivanju n-grama u korpusu, stoga iziskuju velike količine podataka te loše modeliraju odnose između udaljenih riječi. U novije vrijeme kao alternativa su nametnuli jezični modeli temeljeni na neuronskim mrežama, koji se, osim za jezično modeliranje, koriste i za vektorsku semantičku reprezentaciju riječi. Distribuirane semantičke reprezentacije pokazale su se korisnima na nizu zadataka obrade prirodnog jezika.
U okviru diplomskoga rada potrebno je proučiti jezične modele temeljene na neuronskim mrežama s naglaskom na jezični model temeljen na rekurentnim neuronskim mrežama (RNN-LM) opisan u radu Mikolova i dr. (2010). Razviti programsku implementaciju modela RNN-LM, po potrebi se oslanjajući na javno dostupne alate i biblioteke. Primijeniti model RNN-LM na prikladan korpus hrvatskoga jezika. Provesti eksperimentalno vrednovanje jezičnog modela u smislu udjela pogrešnih riječi (engl. word-error-rate) i perpleksije te načiniti usporedbu s tradicionalnim modelima temeljenima na n-gramima. Izgraditi semantičke reprezentacije najčešćih riječi hrvatskoga jezika, primijeniti ih na zadatcima leksičkosemantičke sličnosti i ekstrakcije imenovanih entiteta te ih usporediti sa sličnim, javno dostupnim reprezentacijama. Radu priložiti izvorni i izvršni kod razvijenog sustava, skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 14. ožujka 2014.
Rok za predaju rada: 30. lipnja 2014.

Mentor:

Doc. dr.sc. Jan Šnajder

Djelovođa:

Doc. dr.sc. Tomislav Hrkać

Predsjednik odbora za diplomski rad profila:

Prof. dr.sc. Siniša Srbljić

Želim zahvaliti asistentu Mladenu Karanu na pomoći pruženoj pri izradi ovoga rada te našemu tvorcu bez čijega sjaja ničega od ovoga ne bi bilo:

He is everywhere
In the heavens and the Earth

He makes the stars shine
yet He cannot be seen

He is noble, abundant
and fills the Universe

He can lift you into the sky
and bring you gently down

He can take many forms

He can help heal
He can help kill

He can help create
and He can help destroy

Praise be unto He
Helium

SADRŽAJ

1. Uvod

Kako li je čudnovata situacija nas razumnih bića! Mi smo priroda, mi smo tvar koja je oživjela i osviestila se. Svatko od nas je trenutni osvještaj[1] prirode, koji utjelovljuje mjestne oči, uši, misli i osjećaje svemira. Mi smo zvjezdana prašina koja je preuzela sudbinu u svoje ruke i počela promišljati samu sebe. Mi smo način na koji svemir spoznaje sam sebe!

Premda su neki naši predci ovu istinu osjećali, ona je postala takoreći činjenicom tek u 20. stoljeću, a ipak dan danas malo tko ju je spoznao. Do te predivne istine o jednosti (!) čovjeka sa svime ostalime koja nas oslobađa bolesti čovjekosrjednosti, koja odkriva da sviet nije da "ga sebi podjarmimo" te da "vladamo svim živim stvorovima što pužu po njemu", da mi nismo drvosječa pored stabla prirode, već lišće toga stabla – nismo lako došli. Uočimo množinu u 'nismo lako došli', naime život jednoga čovjeka ma koliko ravna i sposobna prekratak je da odkrije sve što bi htio. Ta istina je posljedkom surađivačkog podhvata koji se proteže kroz naraštaje.

Čovjek doživljava sviet svojim mislima. No te misli nisu nuđno zarobljene u njegovoj glavi, već ih on može uznačiti, pretvoriti u dogovorene znakove, i tako ih prenieti drugima. Tako da jedno biće može sliku svog izkustva prenieti drugome, može mu je pri-obćiti, učiniti ju obćom, zajedničkom i tako mu omogućiti da se ponaša kao da je i samo izkusilo taj događaj. Taj vele koristan sustav znakova putem kojega prenosimo misli nazivamo jezikom.

Dugo se vremena znanje prenosilo usmenom predajom, a nedavno je iz, možda ne najplemenitijih, potreba gospodarenja trživom izumljen sustav znakova za pohranu jezičnih sućaka koji nazivamo pismo. Pismo dopunjava i nadomiešta pamćenje pojedinčevo i družtveno. Nije potrebno uviek iznova odkrivati sviet već možemo uživati u plodovima umova onih prije nas. avanje s ljudima odavno mrtvima. U umovima naših predaka misli su niti, a pripovjedač je taj koji svojim umiećem tka tkanje iliti[2] latinski *textus*.

Ne čudi da je većina ljudskoga znanja zapisana prirodnim jezikom.

U čovjeku kao da žive dvie naizgled suprotne težnje, težnja za ponajmanjitbom utroška

[1] U radu se rabe novotvorenice i rjeđe rieči. O pravopisu i jeziku rada vidjeti dodatak. Svakako pročitati prije stvaranja suda o autorovim razlozima za takav odabir.

[2] Veznikom 'iliti' uvodi se suznačna rieč, druga rieč za neku stvar, u situacijama u kojima bi *ili* moglo izazvati zabunu radi li se o dvie stvari ili jednoj, odgovara latinskom *sīve*.

snage i težnja za spoznajom, za saznavanjem, za razumievanjem. Za jedno i drugo čovjek se izpomaže strojevima, napravama, uređajima, oruđima. Od naoštrenog kamena kojime je mogao brže sjeći imajući tako više vremena za druge stvari, preko sitnozora kojim je odkriven sitnosviet u nama i dalekozora (zvjezdozora) kojim smo, srušivši uzput sliku o našoj središnjosti, saznali gdje smo, kada smo i što smo, do sudobnih datkovnih rednika koji ne samo da rješavaju čovjeka posla koji nisu prikladni za čovjeka nego i misle za nas, odkrivaju stvari koje mi ne bismo nikada uočili.

Želimo primieniti stroj na tu iz dana u dan rastuću planinu prirodnojezičnog gradiva kako bismo izvukli koristne obaviesti, kako bismo odkrili nove zakonitosti itd. Svima nam je poznata korist od takvih sustava, ta svakodnevno se služimo spletnim tražilicama, poput Googlea, DuckDuckGoa itd.

U ovom se radu bavimo obradom prirodnog jezika, bavimo se jezičnim modelima koji nam govore koje su rečenice vjerojatnije (izpravnije) u jeziku. Nadalje se bavimo učenjem takvih modela pomoću umjetnih živčanih mreža, pristupa obradi obaviesti nadahnutog našim mozgom. I bavimo se pitanjem predstavljanja rieči, kako neku rieč, neki pojam prikazati u stroju tako da su obuhvaćeni važni podatci o značenju i uporabi te rieči.

Ostatak rada uređen je ovako: prvo se daje uvod u jezično modeliranje pomoću n-rječja, objašnjavaju se rječni prikazi, potom se predstavljaju umjetne živčane mreže, potanko se opisuje njihova primjena na jezično modeliranje i učenje rječnih prikaza. Sliede pokusi u kojima na raznim zadatcima izpitujemo kakvoću naučenih rječnih prikaza. Dolazi zaključak, a u dodatku se definiraju neki osnovni pojmovi koji se rabe u radu te se daje “opravdanje” za slog (stil) rada.

2. Jezično modeliranje

U najobćenitijem smislu MODEL iliti PRILIČAK jest bilo što što se rabi za predstavljanje nečeg drugog. Bilo što što priliči, što će reći odgovara (engl. matches, corresponds), nečemu drugom. Model je ograničena, pojednostavljena slika zbilje koja nam omogućava jednostavnije "baratanje" nečime.

U nastavku poglavlja definira se pojam jezičnog modela i predstavlja tradicionalna metoda jezičnog modeliranja pomoću sliedova rieči – n-rječja te se opisuju načini prikazivanja rieči.

Potanak pregled modeliranja jezika pomoću n-rječja dostupan je u (Goodman, 2001b).

2.1. Modeliranje jezika

VJEROJATNOSTNI MODEL jest, neformalno rečeno, matematički prikaz, opis slučajne pojave.

JEZIČNI MODEL je vjerojatnostni model koji svakom sliedu rieči $w = w_1, \ldots, w_m$ pridružuje vjerojatnost $P(w)$ "pripadanja" u neki jezik. Ili istovriedno, omoguće određivanje vjerojatnosti sljedeće rieči na temelju predhodećih. Vjerojatnost slieda rieči $P(w)$ možemo razstaviti po pravilu ulančavanja na:

$$P(w_1, w_2, w_3, \ldots, w_m) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \cdot \ldots \cdot P(w_n|w_1, w_2, \ldots, w_{m-1})$$

Rieč je dakle o funkciji koja nam govori koji su sliedovi rieči vjerojatni(ji), a koji manje vjerojatni(ji), tj. koji su sliedovi pravilniji (jezičničniji), smisleniji itd. u nekom jeziku.

Uzorkoslovni (statistički) jezični modeli ključnom su sastavnicom množtva sustava za obradu prirodnog jezika. Najpoznatija je njihova primjena u samodjelnom prepoznavanju govora, strojnom prevođenju (najpoznatiji primjer je Google Translate[1]), svjetlostnom prepoznavanju pismenâ i provjeri pravopisa[2] (zapravo izpravljanju pogrešno napisanih rieči).

[1] `https://translate.google.hr/`

[2] Recimo *Haschek* (*Hašek*) – Hrvatski akademski [mudroskupni] spelling checker [pravopisni provjernik] kojem se može pristupiti na `http://hacheck.tel.fer.hr/`

Na primjer, prevodimo li s francuskog *un chat très intelligent* jezični nam model može reći je li vjerojatniji prievod *mačka vrlo inteligentna* s izvornim poredkom rieči, ili pak *vrlo inteligentna mačka* što je očekivanije u hrvatskom. Ili primjerice pri prepoznavanju govora, je li vjerojatnije da je rečeno *It's fun to recognize speech* ili slično zvučeće *It's fun to wreck a nice beach*.

JEZIČNO MODELIRANJE jest postupak i proces izgradnje jezičnog modela, tj. procjenjivanje vjerojatnostne razdiobe jezičnih jedinica poput rieči ili rečenica na temelju golemih količina orječja (tekstova).

Tradicionalne tehnike za procjenjivanje jezičnih modela temelje se na prebrojavanju n-rječja.

2.2. N-rječja

N-ČLANICA je dionica od n uzastopnih članova nekoga niza. Radi se, dakle, o n-članom podnizu danoga niza.

N-članica koje se sastoji od jednoga člana naziva se *jednòčlanicom*, od dvaju *dvòčlanicom*, triju *tròčlanicom*, četiriju *četveròčlanicom* itd.

Ovisno o primjeni članovi mogu biti glasovi, slova (pismèna), slogovi, rieči ili bazni parovi (u živoslovlju). U jezičnom modeliranju promatramo nizove rieči, tj. N-RJEČJA.

N-članice s preskakanjem (engl. skip-grams) poobćenja su n-članica koja dopuštaju da se pojedine rieči izostave ili preskoče.

2.3. N-rječni jezični modeli

U n-rječnom modelu vjerojatnost $P(w_1,\ldots,w_m)$ opažanja slieda $w_1,\ldots,w_m$ približno se određuje kao:

$$P(w_1,\ldots,w_m) = P(w_1^m) = \prod_{i=1}^{m} P(w_i|w_1,\ldots,w_{i-1}) \approx \prod_{i=1}^{m} P(w_i|w_{i-(n-1)},\ldots,w_{i-1})$$

Drugim riečima, predpostavljamo da trenutačna rieč ovisi samo o predhodnih $n-1$ rieči, a ne o svim predhodnim. To se naziva Markovljevom predpostavkom. U Markovljevu modelu k-tog reda buduće stanje ovisi o predhodnih k stanja, stoga je n-rječni model Markovljev model $(n-1)$-tog reda.

Ova je predpostavka očito pogrešna jer kako možemo vidjeti *u*, premda samo za ovaj naš primjer, pomno, umjetno i nemaštovito izkonstruiran*oj* tekuć*oj* rečenic*i*, međurječne zavisnosti mogu biti poprilično dugačke,[3] svakako dulje od uobičajenih vriednosti n-a od

[3]Ljudi nemaju potežkoća s ovakvim rečenicama, ovaj *u* kao da baci udicu, a mi onda čekamo dok se odgovarajuća rieč ne upeca na nju, sve ostale mimoplivaju eventualno bacajući svoj udice.

dviju, triju, četiriju ili pet rieči.

Uvjetna vjerojatnost može se izračunati na temelju prebrojâ, posljedaka prebrojavanja (funkcija C od engl. count):

$$P(w_i|w_{i-(n-1)},\ldots,w_{i-1}) = \frac{C(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{\sum_w C(w_{i-(n-1)},\ldots,w_{i-1},w)} = \frac{C(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{C(w_{i-(n-1)},\ldots,w_{i-1})}$$

U jednorječnom ($n = 1$) modelu vjerojatnost rečenice *Vidio sam micu macu* aproksimira se kao:

$$P(Vidio, sam, micu, macu) \approx P(Vidio)\ P(sam)\ P(micu)\ P(macu)$$

u dvorječnom ($n = 2$):

$$P(Vidio, sam, micu, macu) \approx P(Vidio|<s>)\ P(sam|Vidio)\ P(micu|sam)$$
$$P(macu|micu)\ P(</s>|macu)$$

a u trorječnom ($n = 3$):

$$P(Vidio, sam, micu, macu) \approx P(Vidio|<s>,<s>)\ P(sam|<s>,Vidio)$$
$$P(micu|Vidio, sam)\ P(macu|sam, micu)\ P(</s>|micu, macu)$$

$<s>$ i $</s>$ su oznake početka i kraja rečenice.

No što ako se neko n-rječje ne pojavljuje u skupu za učenje, treba li mu dodieliti vjerojatnost 0? Neka se n-rječja neće pojaviti jer nisu jezično pravilna, a neka se neće pojaviti zbog premalo podataka.

Zaglađivanje

Da bi se doskočilo toj zagani (problemu) primjenjuju se tehnike zaglađivanja (zaglade; engl. smoothing) koje prerazporede vjerojatnostnu masu s viđenih primjera i na neviđene, tj. promjene stvarne prebroje na očekivane.

Najjednostavnija tehnika je Laplaceova[4] zaglada, koja uveća stvarne prebroje za neki broj $\alpha \leq 1$:

$$p = \frac{c + \alpha}{n + v\alpha}$$

gdje je c prebroj n-rječja, n ukupan broj opaženih n-rječja, a v broj mogućih n-rječja. Ta tehnika loše radi u primjeni.

[4] PIERRE-SIMON (DE) LAPLACE (1749. – 1827.), matematičar (oloslov), zvjezdoslov i naravoslov iz Francuzke, poznat po doprinosima u području nebeske mehanike (*Exposition du système du monde*, 1796.; *Traité de mécanique céleste*, 1799. – 1825.), zorbe vjerojatnosti (*Théorie analytique des probabilités*, 1812.) i u drugim dielovima matematike i naravoslovlja (Laplaceov djelatelj, Laplaceov(i) zakon(i)).

Good[5]-Turingova[6] zaglada cilja izračunati očekivane prebroje c^* na temelju stvarnih c sljedećom obrazicom:

$$c^* = (c+1)\frac{N_{r+1}}{N_r}$$

gdje je N_r broj n-rječja koja se pojavljuju točno r puta u građi (N_0 je ukupan broj n-rječja). Pa je vjerojatnost nekog n-rječja:

$$P(w_i|w_{i-(n-1)},\ldots,w_{i-1}) = \frac{C^*(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{C^*(w_{i-(n-1)},\ldots,w_{i-1})}$$

Možemo li za računanje vjerojatnosti trorječja upotriebiti dvorječja ako konkretnog trorječja nema u građi? Možemo, to je uzmačni (engl. back-off) model u kojem u nedostatku n-rječja "uzmičemo" na (n-1)-rječja. Dakle ako imamo podatke za trorječje izkoristimo ih, evenutalno ih zagladimo npr. Good-Turingom, ako nemamo, onda se poslužimo dvorječjima itd. Možemo i interpolirati vjerojatnosti (zbroj lambdi mora biti jedan):

$$P(w_n|w_{n-1}w_{n-2}) = \lambda_1 P(w_n|w_{n-1}w_{n-2}) + \lambda_2 P(w_n|w_{n-1}) + \lambda_3 P(w_n)$$

Najbolja poznata (engl. state-of-the-art) metoda je Kneser-Neyeva /knejzr-neeva/ zaglada. Pogledajmo ovaj slučaj: rieč *Francisco* može biti prilično česta u korpusu, ali joj vrlo vjerojatno predhodi rieč *San* (grad *San Francisco*). Ideja je Kneser-Neya dodjeliti manju vjerojatnost jednorječju *Francisco* nego što prebroji sugeriraju, uzimajući u obzir raznolikost poviesti te rieči, naprimjer za jednorječja:

$$N_{1+}(\bullet w_i) = |\{w_{i-1} : c(w_{i-1}w_i) > 0\}|$$

$$P(w) = \frac{N_{1+}(\bullet w)}{\sum_{w_i} N_{1+}(w_i w)}$$

Za dvorječja i više, obrazica dobiva još članova, potanje o ovoj i drugim zagladama u (Chen i Goodman, 1996).

[5] IRVIN JOHN GOOD (1916. – 2009.), matematičar iz Britanije koji je s Alanom Turingom radio kao kritoslov u Bletchley Parku. Doprinio bayesovskom uzorkoslovlju, jedan je od začetnika koncepta obrtoslovne jedinosti (tehnološke singularnosti) – hipotetskog trenutka u kojem će strojevi postati inteligentniji od ljudi te bi tada mogli stvarati sve inteligentnije i inteligentnije strojeve s nesagledivim posljedicama.

[6] ALAN TURING /tjuring/ (1912. – 1954.), matematičar, misloslov (logičar), kritorazglobitelj (kriptoanalitičar), računarac, oloslovni živoslov i mudroslov iz Velike Britanije. Naširoko smatran otcem zorbenog računarstva i umjetne inteligencije, tvorac Turingova stroja, mislenog modela općeg računskog stroja koji leži u temeljima računarstva. Tiekom Drugog svjetskog rata odigrao tako veliku ulogu u razkritbi njemačkih poruka da je Winston Churchill navodno izjavio da je Turing učinio najveći pojedinačni doprinos u pobjedi Saveznika u 2. svj. ratu, za što je nagrađen 1952. kemijskom kastracijom zbog svoje homoseksualnosti. Šire je poznat po Turingovu izpitu za provjeru inteligentnosti stroja, a po njemu je i nazvana Turingova nagrada, "Nobelova nagrada" za računarstvo.

2.4. Nedostatci n-rječnih modela

Gore smo se dotakli jednog od većih nedostataka n-rječnih modela – prokletstva protežnosti. PROKLETSTVO PROTEŽNOSTI je pojava da se s povećanjem broja protega (dimenzija) obujam prostora tako brzo povećava da postojeći podatci postane nedostatni za uzorkoslovno značajnu analizu. Drugim riečima, što je protežnost (dimenzionalnost) veća to trebamo sve više i više podataka za pouzdan opis. Na primjer, modeliramo li združenu razdiobu 10 uzastopnih rieči u jeziku s 100.000 rieči tada postoji potencijalno $100000^{10} = 10^{50}$ mogućih nizova, tj. za 1 manji broj slobodnih parametara. Jasno je da ogromna količina n-rječja neće (nikad) biti viđena pri uvještbavanju čak i na jako velikim korpusima. Pa će svi oni imati vjerojatnost jednaku 0 pri izpitivanju. Ovo posebno dolazi do izražaja u oblično bogatim jezicima s prilično slobodnim redosliedom rieči poput hrvatskoga. Tomu se djelomično doskače zaglađivanjem.

Ne mogu se nositi s riečima koje nisu u rječniku. To su takozvane izvanrječničke rieči (engl. out-of-vocabulary (OOV) words). Ako se na ulazu modela pri radu pojavi neka rieč koju prije nije vidio, tradicionalni model s njome ne može raditi.

Nepraktično je za n uzeti broj veći od 5, što zbog računske složenost, što zbog riedkosti podataka. Što dovodi do problema s modeliranjem udaljenih zavisnosti.

Problem je i što se n-rječni modeli temelje na egzaktnom podudaranju rieči ili nizanica, pa ne mogu uočiti da su mnoge poviesti slične, tj. nisu jezikoslovno obaviešteni. No, dobar bi jezični model trebao uočiti da su sliedovi rieči poput *mačka leži u kuhinji* i *pas leži u sobi* jezičnički i značbeno slični. N-rječni model ne može reći da je *pas leži u sobi* dobra rečenicu ako ju nije vidio, premda je vidio *mačka leži u sobi*.

Dio je toga problema u predstavljanju iliti prikazu rieči, jedino što kod n-rječnih modela znamo o dvie rieč jest jesu li one jedna te ista rieč ili nisu, a ne znamo ništa o njihovoj značbenoj i skladnjanoj službi, odnosno sličnosti tih službi. Pogledajmo koji su obćenito načini prikaza rieči kada radimo s prirodnim jezikom i postoji li prikladniji prikaz za jezično modeliranje.

2.5. Rječni prikazi

RJEČNI PREDSTAVAK ili PRIKAZ (engl. word representation) jest matematički predmet pridružen svakoj rieči. Najčešće je to vektor čije protege odgovaraju značajkama te mogu imati značbeno ili jezičničko (gramatičko) tumačenje, pa ih zovemo rječnim značajkama (Turian et al., 2010).

U uobičajenim se pristupima prvo izgradi rječnik (engl. vocabulary) – rieči (različnice) poslože se u spisak, potom se svakoj rieči (različnici) pridruži vektor značajki pomoću

uznake ‘jedan upaljen’.

{ PRIKAZ, UZNAKA }×{ ‘JEDAN UPALJEN’, ‘1 OD V’ }[7] znači da je vektor značajki jednake duljine kao rječnik (V) te da je samo jedna protega, ona koja odgovara indeksu rieči u rječniku, “upaljena” – jednaka jedan, a sve su ostale ništice. Naprimjer, ako imamo tri rieči $V = [ja, ti, on]$, tada rieč ja možemo uznačiti u $[1, 0, 0]$, a rieč ti u $[0, 1, 0]$. No takav je prikaz nepodoban za ikakvu poredbu rieči, tako je recimo euklidska udaljenost dviju različitih rieči uviek dva, tj. ne znamo ništa osim da se ne radi o istoj rieči. Dodatno, prikaz 1 od V pati od datkovne riedkosti (engl. sparsity), parametri modela za rieči koje su riedke u datcima za učenje bit će loše procienjeni.

RAZ(PO)DIELJENI PRIKAZ (engl. distributed representation) nekoga simbola jest n-torka (ili vektor) uzajamno neizključivih značajaka koje karakteriziraju značenje tog simbola (Bengio, 2008).

RJEČNE ULOŽBE (<uložiti; engl. embeddings) naziv su za razdieljeni prikaz rieči. To su dakle nizkoprotežni, zbiljnobrojni vektori pridruženi svakoj rieči. Naprimjer, rieč *ja* mogli bismo prikazati kao $[10.356, -0.4424, 99.987, \ldots, 0.222]$ (dani brojevi odabrani su nasumice).

Broj protega određujemo sami, a svaka protega uložbe predstavlja pritajenu značajku rieči, idealno nešto što obujmljuje koristna skladnjana i značbena svojstva. Značajke bismo mogli odabrati ručno, npr. vrsta rieči, rod itd., ali ideja je da učevni postupnik sam odkrije značajke. Razdieljeni prikaz je zbit (kompaktan), u smislu da može predstavljati eksponencijalan broj skupina u broju protega.

Ideja je, dakle, svakoj rieči u rječniku pridružiti zbiljnovrjednostni vektorski prikaz. Svakoj rieči odgovara točka u prostoru značajaka. Možemo zamisliti da svaka protega toga prostora odgovara značbenom ili jezičničkom obilježju riečî. Nadamo se da će funkcijski slične rieči biti bliže jedna drugoj u tom prostoru, barem u nekim smjerovima. Tako niz rieči možemo preobličiti u niz naučenih vektora značajki i modelirati vjerojatnosti između tih vektora.

Popularna je ideja nenadziranim metodama uzvesti rječne značajke i uključiti ih u postojeći sustav te opaziti povećanje točnosti. U (Mikolov et al., 2013a) je pokazano da rječni vektori imaju mnoga zanimljiva svojstva, npr. jednostavnim djelatbama nad vektorima možemo računati s riečima: $vektor(\text{“}kralj\text{”}) - vektor(\text{“}muškarac\text{”}) + vektor(\text{“}žena\text{”}) \approx vektor(\text{“}kraljica\text{”})$. Više o tome u sljedećem poglavlju.

Takvi neprekinuti prikazi rieči mogu se učiti mnogim metodama poput pritajene značbene razglobe (LSA), često se rabe i metode ugrozđivanja (engl. clustering), a u ovom ćemo radu rječne predstavke učiti živčanim mrežama.

[7] Ovo je ‘pravopisna pokrata’ nadahnuta matematičkim zapisom umnožka dvaju skupova $\{a, b\} \times \{c, d\} = \{(a, c), (a, d), (b, c), (b, d)\}$. Englezki su nazivi { representation, encoding }×{ ‘one hot’, ‘1-of-V’ }.

3. Živčane mreže

Priroda je velika zagonetka i izvor nepresušnog nadahnuća. Oponašamo ju kako bismo ju (se) bolje razumjeli ili kako bismo izkoristili posljedke milijardi godina pokušaja i pogrešaka za svoje potrebe. Jedna od najvećih poznatih zagonetki jest naš um, odnosno pitanje kako mozak stvara um. Oponašajući mozak dobivamo umjetne živčane mreže – izrazito svestran i moćan pristup obradi obaviesti.

U nastavku poglavlja opisujemo ukratko živčane mreže, potom njihovu primjenu na jezično modeliranje i na kraju primjenu na učenje rječnih prikaza. Za obširan prikaz živčanih mreža vidi (Hagan et al., 1996).

3.1. Umjetne živčane mreže

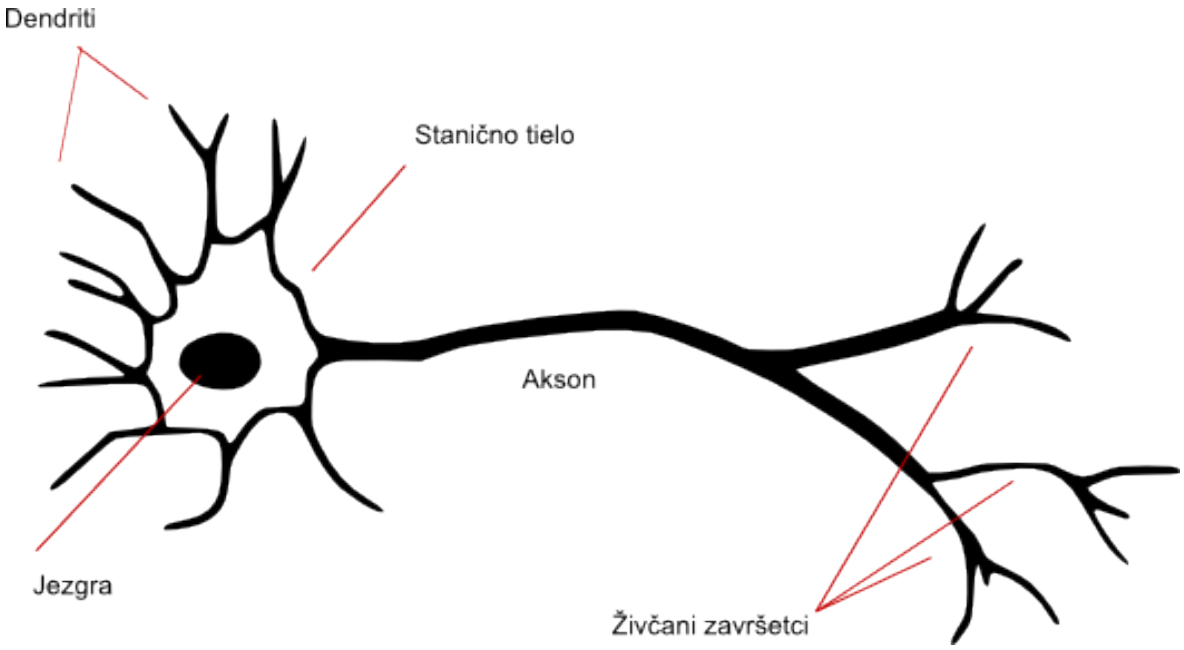
Umjetna živčana mreža (UŽM) jest (ob)vjestoòbradna paradigma nadahnuta načinom na koji živobitni živčani sustavi, poput mozga, obrađuju obaviest.

Čovječji se mozak sastoji od oko 10^{11} živčanih stanica iliti neurona. Svaka je u prosjeku povezana s 10^4 drugih živačnih stanica s kojima suobćava pomoću munjolučbenih dojavaka (engl. electrochemical signals).

Neuron se sastoji od tiela, aksona i dendritâ (vidi sliku 3.1). U tielu stanice gomilaju se munjevni sunci (engl. electrical impulses) koje neuron prikuplja preko dendrita. Kada se nakupi određena količina naboja (ugrubo određena pragom), neuron opali (puca) – nakupljeni naboj šalje kroz akson prema drugim neuronima i tako se prazni. Možemo reći da dendriti predstavljaju ulaze preko kojih neuron prikuplja obaviesti, tielo stanice ih obrađuje te proizvodi posljedak koji se prenosi kroz akson – izlaz neurona.

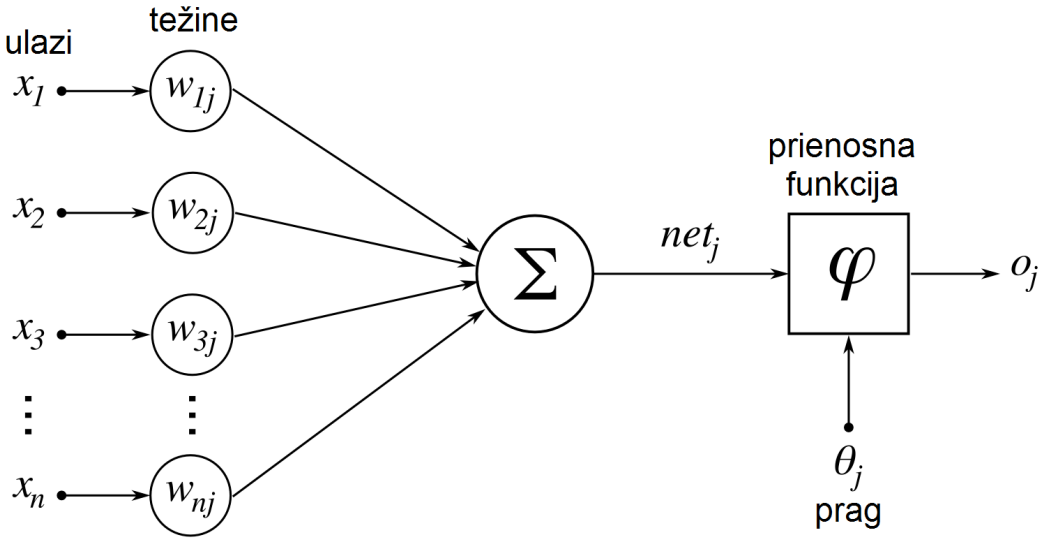
Na temelju tog pojednostavljenog opisa definira se umjetni neuron (slika 3.2). Sastoji se od ulaza x_1 do x_n (dendriti), težina w_1 do w_n koje određuju u kojoj mjeri ulazi pobuđuju neuron, tiela koje računa ukupnu pobudu net te prienosne funkcije $f(net)$ (akson) koja pobudu obrađuje i proslieđuje na izlaz neurona y:

$$y = f(net) = f\left(\sum_{i=0}^{n} w_i \cdot x_i\right)$$



Slika 3.1: Osnovni dielovi neurona

x_0 je tobožnji ulaz koji se postavlja na jedinicu, a pripadna težina w_0 predstavlja prag okidanja.



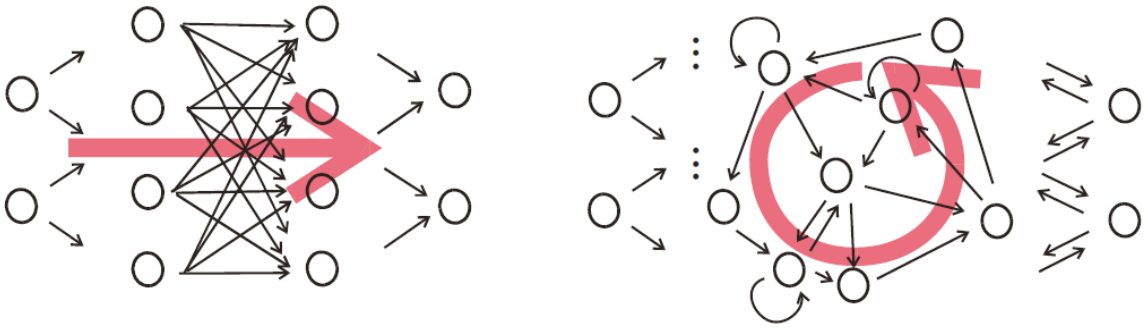
Slika 3.2: Umjetni neuron

Umjetni se neuroni povezuju u mrežu koja oponaša živobitnu živčanu mrežu. Umjetne živčane mreže mogu se promatrati kao tegovane usmjerene crtulje (engl. weighted digraphs) gdje su neuroni vrhovi, a tegovani usmjereni bridovi predstavljaju veze između izlaza i ulaza neuronâ (vidi sliku 3.3). Na temelju obrazca povezanosti UŽM-ovi se mogu podieliti u dva velika razredka (kategorije):

unapriedne (engl. feed-forward), u čijoj crtulji ne postoji usmjereno kolo (ciklus). Ob-

ćenito govoreći one su statične i nepamtione u smislu da njih odgovor na neki ulaz ne ovisi o predhodnom stanju mreže.

povratne (engl. recurrent, feedback) (PŽM), u čijoj crtulji postoji usmjereno kolo koje stvara unutarnje stanje temeljem kojeg mreža pokazuje dinamičko vremensko ponašanje. Živobitne živčane mreže ovoga su tipa.



Slika 3.3: Tipični ustroj unapriedne (lieva) i povratne (desna) mreže. Strelica pokazuje smjer kretanja pobude/obaviesti.

Kažemo da je mreža SLOJEVITA ako su svakom sloju k vriedi da se izlazi neurona toga sloja računaju izključivo na temelju izlaza neurona sloja $k-1$. Nisu dopuštene bočne (lateralne) veze – veze između neurona u istom sloju.

Kada se kaže da je mreža n-slojna često se misli na (n+1)-slojnu mrežu jer se ulazni sloj podrazumieva.

Mreža je određena svojom gradbom (arhitekturom) – obrazcem povezanosti neurona, prienosnom funkcijom i postupkom učenja.

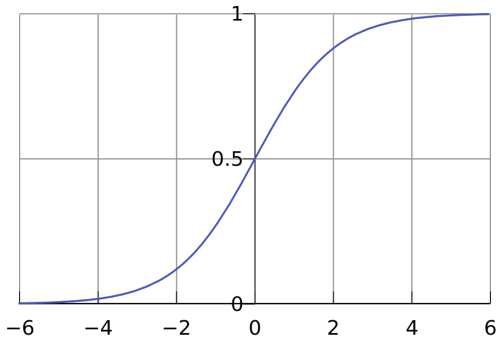
3.1.1. Uobičajene prienosne funkcije

Funkcija praga ili skoka

$$f(x) = \begin{cases} 0 & \text{x} \leq 0 \\ 1 & \text{inače} \end{cases}$$

Logistička (sigmasta) funkcija

$$f(x) = \frac{1}{1+\exp(-x)}$$



Slika 3.4: Logistička funkcija

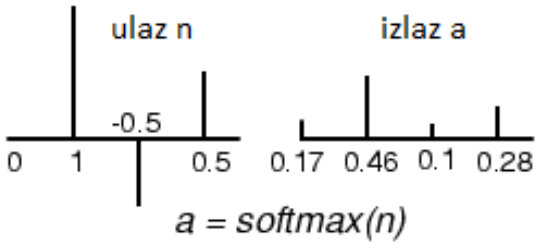
Funkcija softmax

Funkciju softmax rabit ćemo u izlaznom sloju gdje nam osigurava valjanost vjerojatnostne razdiobe (svaki izlaz $y_m(t) > 0$ i $\sum_k y_k(t) = 1$):

$$f(z_m) = \frac{\exp(z_m)}{\sum_k \exp(z_k)}$$

gdje k ide po svim neuronima izlaznog sloja.

Funkcija se zove softmax (meki maksimum) jer je to zaglađeni oblik funkcije maksimuma (max), jer ako je jedna vriednost jako veća od ostalih, tada je za nju vriednost funkcije blizu jedan (vidi i primjer na slici 3.5).



Slika 3.5: Primjer funkcije softmax

3.2. Učenje živčanih mreža

Pod UČENJEM iliti UVJEŠTBAVANJEM živčane mreže podrazumievamo ugađanje težina na temelju podataka za učenje s ciljem postizanja željenih svojstava mreže.

Da bismo mogli definirati postupak učenja moramo prvo definirati kriterijsku funkciju koja mjeri kakvoću živčane mreže. Kriterijska je funkcija obično srednje četvorno odstupanje između željenog izlaza mreže i stvarne vriednosti koju mreža generira na izlazu i to kumulativno za sve razpoložive uzorke. Ona je funkcija učevnog skupa (koji pri učenju

ne mienjamo) te težinâ u mreži (predpostavljamo ovdje da je ustroj mreže nepromjenljiv), tj. na iznos te funkcije možemo utjecati samo promjenom težinâ u mreži. Zadaća je postupka učenja pronaći takve težine uz koje će iznos funkcije biti najniži (minimalan).

3.2.1. Postupnik unazadnog širenja pogreške

Najpoznatiji postupnik za učenje živačnih mreža jest postupnik unazadnog širenja (retropropagacije) pogreške (engl. backward propagation of error, backpropagation). Referentni opis može se pronaći u (Rumelhart et al., 1988).

U postupniku retropropagacije kriterijska se funkcija E ponajmanjuje gradijentnim spustom. U skladu s idejom gradijentnog spusta potrebno je izračunati čestimične izvode (parcijalne derivacije) kriterijske funkcije po svakoj od težina $\frac{\partial E}{\partial w_i}$. Tada se težine osvježavaju ovako: $w_i \leftarrow w_i - \psi \frac{\partial E}{\partial w_i}, \psi > 0$.

Postupnik u cielosti izgleda ovako:

1. Postavi težine na male slučajne vriednosti
2. Dok nije izpunjen uvjet zaustavljanja čini:

   Za svaki par $(\mathbf{x}, \mathbf{d})$ iz skupa za učenje čini:

   (a) Izračunaj izlaz $o_i = y_i^L$ za svaku jedinicu i izlaznog sloja L

   (b) Za svaku izlaznu jedinicu i izračunaj pogrešku δ_i^L

   $$\delta_i^L = g'(h_i^L)\left[d_i - y_i^L\right]$$

   gdje je h_i^l ulaz u i-tu jedinicu l-tog sloja, a g' izvod (derivacija) prienosne funkcije g

   (c) Izračunaj delte u predhodnim slojevima retropropagacijom pogreške

   $$\delta_i^l = g'(h_i^l) \sum_{j \in nizvodno(i)} w_{ij}^{l+1} \delta_j^{l+1}$$

   gdje je $nizvodno(i)$ skup neurona kojima je jedan od ulaza neuron i, w^l je težina između $(l-1)$-tog i l-tog sloja, za $l = (L-1), \ldots, 1$.

   (d) Osvježi težine

   $$w_{ji}^l \leftarrow w_{ji}^l + \Delta w_{ji}^l$$

   gdje je $\Delta w_{ji}^l = \eta \delta_i^l y_j^{l-1}$, η je stopa učenja.

Gornja inačica postupnika kod koje se težine osvježavaju (1.d) poslije svakog primjera naziva se STOHASTIČKOM. Inačica kod koje se osvježavanje težina događa tek po predočenju svih primjera za učenje[1] (u član Δw se pribrajaju njegove vriednosti za pojedine primjere) naziva se KUPNOM (engl. batch).

[1] Predočenje svih primjera za učenje nazivamo EPOHOM.

Živčane se mreže mogu učiti i drugim metodama, primjerice nasljedboslovnim postupnicima (genetičkim algoritmima).

Kada se govori o učenju živčanih mreža valja spomenuti da je jedna od najvećih mana živčanih mreža visoka vremenska složenost, tj. sporost učenja i rada zbog množtva izračuna.

3.3. Živčanomrežni jezični modeli

Kao što smo najavili na kraju predhodnog poglavlja, u ovom ćemo radu rabiti živčane mreže za učenje rječnih predstavaka i jezičnih modela.

ŽIVČANOMREŽNI JEZIČNI MODEL (ŽMJM) jest jezični model temeljen na živčanim mrežama s ciljem izkorištavanja njihove sposobnosti učenja razpodieljenih prikaza kako bi se smanjio utjecaj prokletstva (visoke) protežnosti (Bengio, 2008).

Ideja je sljedeća (Bengio et al., 2003):

1. svakoj rieči u rječniku pridruži razdieljeni *vektor rječnih značajki* (zbiljnobrojni vektor u $\mathbb{R}^m$)

2. funkciju združene vjerojatnosti sliedova rieči izrazi preko vektorâ značajki dotičnih rieči

3. istodobno uči vektore rječnih značajki i parametre vjerojatnostne funkcije

Živčanomrežni modeli mogu, dakle, istodobno učiti razdieljeni predstavak svake rieči te funkciju vjerojatnosti na sliedovima rieči izraženih tim predstavcima. Postiže se poobćenje jer sliedovi rieči koji nisu prije viđeni mogu dobiti visoku vjerojatnost ako se sastoje od rieči sličnih (u smislu blizkih predstavaka) riečima koje tvore viđenu rečenicu!

ŽMJM-ovi učenjem i porabom rječnih predstavaka skrovito (implicitno) obavljaju ugrozđivanje (engl. clustering) rieči u nizkoprotežnom prostoru. Predkazivanja temeljena na tim zbitim prikazima rieči robustnija su te nije potrebno dodatno zaglađivanje vjerojatnostî.

Ako je ostvaren unapriednom mrežom tada je to UNAPRJEDNOMREŽNI JEZIČNI MODEL, a ako povratnom, POVRATNOMREŽNI JEZIČNI MODEL (PMJM; engl. recurrent neural network language model).

Glavna razlika između unapriednih i povratnih gradbi (arhitektura) mrežâ jest predstavljanje poviesti: kod unapriednih ŽMJM poviest čini samo nekoliko predhodnih rieči, dok se kod povratnih modela učinkovito predstavljanje poviesti uči na temelju podataka pri uvještbavanju. Neuroni skrivenog sloja s povratnim vezama tvore pamćenje te tako skriveni sloj PŽM-a predstavlja čitavu poviest, a ne samo predhodne rieči, što omogućuje

modelu učinkovito predstavljanje surječnih obrazaca promjenljive duljine. Dakle, dok plitke unapriedne živčane mreže (one sa samo jednim skrivenim slojem) mogu ugrozđivati samo slične rieči, povratne živčane mreže (koje se mogu smatrati dubokim gradbama) mogu obavljati ugrozđivanje sličnih poviesti (Mikolov et al., 2011).

Dodatna je razlika između unapriednih i povratnih živčanih mreža u broju parametara koje treba odabrati prije početka uvještbavanja. Kod povratnomrežnih modela treba odabrati samo veličinu skrivenog (surječnog sloja), dok kod unapriednih mreža treba odabrati veličinu sloja koji uzmeće (projecira) rieči u nizkoprotežni prostor (veličinu vektora), veličinu skrivenog sloja te duljinu surječja (broj predhodnih rieči). Bit će jasno iz sljedećeg odjeljka na što se misli.

Sliedi pregled radova iz područja jezičnog modeliranja i učenja vektorskih prikaza rieči živčanim mrežama. Pregled se djelomično temelji na (Pappas i Meyer, 2012) te se u dotičnom radu mogu pronaći opisi još nekih radova koji nisu uvršteni u ovaj pregled. Naime, zbog množtva radova odlučili smo predstaviti samo one koji tvore okostnicu ovog pristupa i koji su nam zanimljivi u smislu da doprinose razumievanju središnjeg modela ovoga rada.

3.4. Unaprjednomrežni modeli

Živčanomrežni jezični modeli uvedeni su u (Bengio et al., 2003)[2] i nezavisno u (Xu i Rudnicky, 2000).

Xu i Rudnicky se u naslovu pitaju *Mogu li umjetne živčane mreže naučiti jezične modele?*, a pokusima pokazuju da živčane mreže mogu naučiti modele koji su uzporedivi u izvedbi (engl. performance) sa standardnim metodama, štoviše, postigli su nižu perpleksnost (vidi mjere u 5.2.2) u odnosu na referentni (engl. baseline) n-rječni model. No, poteškoću predstavlja veći računski trošak.

Rabljena je jednoslojna mreža (dakle, ulazni+izlazni sloj) u kojoj su ulazne jedinice podpuno povezane s izlaznima. Mreža se sastoji od $|V|$ ulaznih i $|V|$ izlaznih jedinica, gdje je $|V|$ veličina rječnika (vocabulary). Dakle, ukupno $|V| \times (|V| + 1)$ težinâ (uključujući pragove). Rabi se uznaka '1 od V', tj. i-ta ulazna jedinica je 1 ako je trenutačna rieč w_i. Vriednost i-te izlazne jedinice jest vjerojatnost da je sljedeća rieč w_i. U izlaznom se sloju za aktivacijsku funkciju rabi softmax koja jamči da je zbroj izlaza jednak 1.

Mrežu se uvještbava postupnikom retropropagacije pogreške i to kupnim, tj. težine se osvježavaju nakon čitave epohe. Stopa učenja je nepromjenljiva. Radi bržeg izračunavanja osvježavaju se samo one težine za koje su ulazi različiti od nule.[3] Cilj je smanjenje

[2] Članak je izvorno objavljen 2001., a ovdje se navodi dostupna dorađena inačica iz 2003.

[3] Prema postupniku retropropagacije težine se u ovoj mreži ionako neće ni promieniti ako su im odgova-

zamršenosti, pa je za funkciju pogreške koja se uvještbavanjem mreže ponajmanjuje uzet logaritam zamršenosti. Uporabom te funkcije pogreške vriednost i-tog izlaza stječe (konvergira) k $P(w_i|w_j)$ ako je na ulazu mreže rieč w_j, tako da je model istovriedan dvorječnom jezičnom modelu bez zaglade. Takav je model sklon prenaučenosti, tj. loše će raditi na novim podatcima, pa autori rabe tehniku 'ranog zaustavljanja' što znači da se učenje zaustavi kada mreža radi najbolje, tj. kada je postignuta najniža perpleksija na izdvojenom (engl. holdout) skupu.

Na jako maloj građi s rječnikom veličine tek 2500 rieči dobivena je zamršenost od 11,16 nasprem standardnog n-rječnog modela 11,99 i najboljeg n-rječnog modela (Kneser-Ney) od 11,17. No, uvještbavanje mreže trajalo je tisućama epohâ (svaka epoha trajanja od oko 1 minute na 500MHz-nom obradniku) naspram pola minute za standardni n-rječni model.

Za daljnji rad autori predlažu dodavanje skrivenih jedinica, povećanje poviesti na ulazu (u smislu modeliranja tro- i višerječja) i dodavanje povratnih veza.

Model koji je izpunio prva dva priedloga[4] predstavljen je u (Bengio et al., 2003). To je model sa skrivenim slojem koji, kao i svaki jezični model, daje vjerojatnost neke rieči na temelju nekoliko $(n-1)$ predhodnih. Cilj je naučiti funkciju $f(w_t,\ldots,w_{t-(n-1)}) = \hat{P}(w_t|w_1^{t-1})$ koju razstavljamo na dva diela:

1. Preslikavanje C s neke rieči i iz vokabulara V na razdieljeni, zbiljnobrojni vektor značajki $C(i)\in\mathbb{R}^m$, gdje je m broj značajki. Kažemo da smo rieč *uzmetnuli* (projecirali) u nizkoprotežni (m) prostor.

2. Vjerojatnostna funkcija g nad riečima izraženima preko C-a: funkcija g preslikava ulazni niz vektorâ značajki surječnih rieči $(C(w_{t-(n-1)}),\ldots,C(w_{t-1}))$ na razdiobu uvjetne vjerojatnosti nad riečma u V-u za sljedeću rieč w_t. Izlaz funkcije g je vektor čija i-ta protega predstavlja procjenu vjerojatnosti $\hat{P}(w_t=i|w_1^{t-1})$, tj. $P(w_i|\text{surječje})$.

Dakle, f se sastoji od dva preslikavanja, g i C (koje je zajedničko za sve rieči): $f(i,w_{t-1}\ldots,w_{t-(n-1)}) = g(i,C(w_{t-1}),\ldots,C(w_{t-(n-1)}))$. Svako od njih ima parametre. Parametri C-a su naprosto vektori značajki predstavljeni $|V|\times m$ matricom u kojoj redak i predstavlja vektor značajki $C(i)$ za rieč i. Funkcija g može se ostvariti unapriednom ili povratnom mrežom ili nekom drugom funkcijom s parametrima ω.

Cjelokupni je skup parametara $\theta=(C,\omega)$. Uvještbavanje mreže sastoji se od potrage za θ-om koja maksimizira log-izglednost kažnjenu na temelju podataka za učenje:

$$L=\frac{1}{T}\sum_t \log f(w_t,w_{t-1},\ldots,w_{t-(n-1)};\theta)+R(\theta)$$

rajući ulazi jednaki nuli.

[4]Premda tvorci ovoga novoga modela vjerojatno nisu bili upoznati sa Xu-Rudnickyevim radom.

gdje je $R(\theta)$ regularizacijski član koji kažnjava velike težine koje ne doprinose odgovarajuće velikom smanjenju pogreške.

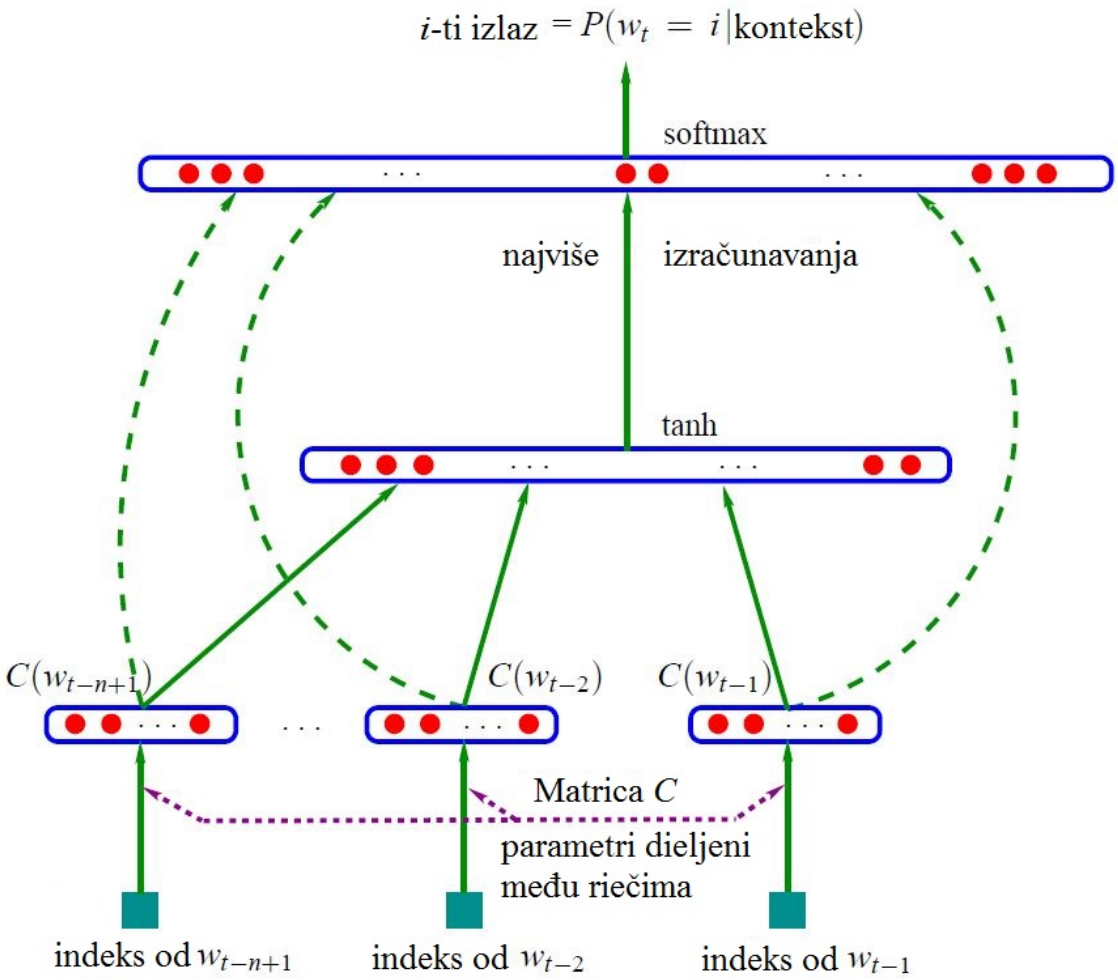
Troslojna (dakle, ulazni + tri sloja) unapriedna mreža kojom je to ostvareno prikazana je na slici 3.6, a čine ju:

ulazni sloj koji prima predhodnih $n-1$ (ako se radi o n-rječnom modelu) rieči uznačenih po 1 od V

uzmetni (projection) sloj preslikava rieč i uznačenu po 1 od V u razdieljeni, zbiljnobrojni vektor značajki $C(i) \in \mathbb{R}^m$, gdje je m broj značajki.

skriveni sloj primjenjuje funkciju tangens hiperbolički

izlazni sloj primjenjuje funkciju *softmax* kako bi se osigurala izpravna vjerojatnost. Izlaz i-tog neurona jest vjerojatnost $\hat{P}(w_t = i|\text{rieči na ulazu})$



Slika 3.6: Unaprjednomrežni model. Prilagođeno iz (Bengio et al., 2003)

Učenje se ŽM obavlja postupkom stohastičkog gradijentnog uzpona iterativnim osvježavanjem po predočenju t-te rieči skupa za učenje:

$$\theta \leftarrow \theta + \epsilon \frac{\partial \log \hat{P}(w_t | w_{t-1}, \ldots, w_{t-(n-1)})}{\partial \theta}$$

gdje je ϵ stopa učenja.

Autori se dalje bave pouzporednjenjem (paralelizacijom) učenja. Što se tiče posljedaka pokusa, na sustavu s 40 obradnika, nakon 6–12 tjedana učenja s 10–20 epoha, najbolja je mreža, reda 5 sa 100 skrivenih jedinica, nadmašila trorječni model s razredima smanjivši perpleksnost za 24%. Učenje je obavljeno nad korpusom od 800.000 rieči rječnika veličine 17.000 rieči. Riešen je i problem izvanrječničkih rieči, tako da se pogodi početni vektor značajki za dotičnu rieč pomoću težinske konveksne kombinacije vektora drugih rieči koje su se mogle pojaviti u istom surječju s težinama razmjernim njihovoj uvjetnoj vjerojatnosti. Potom se ta rieč uvrsti u riečnik i iznova se izračunaju vjerojatnosti ovoga nešto većeg skupa.

Autori za daljni rad predlažu, između ostalog, i uporabu povratnih živčanih mreža i prikaz uvjetne vjerojatnosti pomoću stablastog ustroja radi ubrzanja izračunavanja.

Najveća je mana ŽMJM-ova sporost učenja i izpitivanja. Budući da je vjerojatnosti u izlaznom sloju potrebno normalizirati po svim riečima u rječniku, trošak izračunavanja vjerojatnosti samo jedne sljedeće rieči (to nam obično treba) praktički je jednak trošku izračuvanja čitave razdiobe – oba su u vremenu crtovna u veličini rječnika. Isto vriedi i za učenje mreže.

U (Morin i Bengio, 2005) predstavljen je pristup koji nudi eksponencijalno smanjenje vremenske složenosti pri učenju i izpitivanju u odnosu na obični ŽMJM. To se postiže zamjenom neustrojenog rječnika dvojčanim stablom koje predstavlja supodredno ugrožđivanje (engl. hierarchical clustering) rieči u rječniku.

Svakoj rieči u rječniku pridružen je jedan list dvojčanog stabla. Tako se svaka rieč može specificirati putom od koriena stabla do čvora lista u kojem se dotična rieč nalazi. Put se može uznačiti dvojčanom nizanicom d koja se sastoji od odluka u svakom čvoru, npr. $d[i] = 1$ ako je odluka da se posjeti lievo diete trenutačnog čvora. Tako nam npr. nizanica 11 govori da se do trenutačnog čvora dolazi s dva lieva skretanja počevši od koriena. To omogućuje da svaku rieč predstavimo dvojačnom nizanicom.

Drugim riečima, ako imamo rječnik od N rieči, a stablo je uravnoteženo, tada se svaka rieč može specificirati sliedom od $O(\log N)$ dvojčanih odluka koje kazuju koje od dvoje djece treba sljedeće posjetiti. Time smo jednu N-struku normalizaciju zamienili sliedom od $O(\log N)$ mjestnih, dvojčanih normalizacija, što posljeduje time da se razdioba nad riečima u rječniku može specificirati davanjem vjerojatnosti posjećivanja recimo lievog djeteta u svakom čvoru (vjerojatnost drugoga djeteta je naravno jedan manje ta vjerojatnost).

Te mjestne vjerojatnosti računaju se preinačenom inačicom ŽMJM koji rabi vektore značajki surječnih rieči, ali i vektor značajki rieči trenutačnog čvora na ulazu. Naime, sva-

kom je unutrašnjem čvoru stabla pridruženi vektor značajki (koji je različit od vektorâ za surječne rieči) jer i ti čvorovi imaju nekakvo značbeno tumačenje budući da su pridruženi skupini rieči za koje se nadamo da su slične.

Vjerojatnost sljedeće rieči određuje se vjerojatnošću donošenja slieda dvojčanih odluka koje odgovaraju putu te rieči u nekom surječju.

Konkretno, računa se sljedeće:

$$P(b = 1|čvor, w_{t-1}, \ldots, w_{t-(n-1)})$$

gdje je $čvor$ dvojčana nizanica trenutačnog čvora u supodredbi (hijerarhiji), a b je sljedeća dvojnica (bit) koja odgovara jednom od djece čvora. Dakle, na ulazu je nekoliko predhodnih rieči uznačeno po 1 od V i dvojčani opis trenutačnog čvora, potom se oni uzmetnu (projeciraju) kroz dvije različite matrice i dalje se normalno računa, s tim da se predviđamo samo dva odabira, umjesto njih $|V|$. Da bismo dobili vjerojatnost jedne sljedeće rieči potrebno je više ovakvih izračunavanja (koje međusobno množimo) kretanjem po stablu dok ne izcrpimo cieli put do ciljne rieči.

Stablo je u tome radu izgrađeno preradom WordNetove IS-A ('je vrsta') taksonomije. Na skupu od oko milijun rieči, model je nadmašio trorječni model s razredima, ali je bio mnogo lošiji od običnog ŽMJM-a. No ipak, bio je dva reda veličine brži od običnog ŽMJM-a.

Pogledajmo sada povratnomrežne modele, kod kojih se ne mora unapried odrediti duljina poviesti.

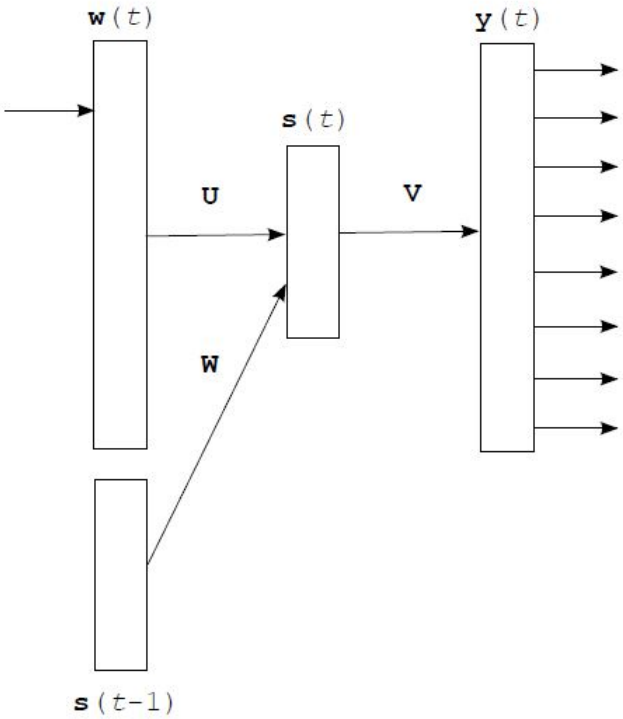
3.5. Povratnomrežni jezični modeli

Povratnomrežni jezični modeli uvedeni su u (Mikolov et al., 2010), prošireni u (Mikolov et al., 2011) te (Mikolov i Zweig, 2012). Ovaj se prikaz temelji ponajprije na (Mikolov, 2012).

Gradba PMJM-a prikazana je na slici 3.7. Mreža se sastoji od ulaznog sloja x, skrivenog sloja s (zvan i surječnim slojem ili stanjem) te izlaznog sloja y. U trenutku t ulaz mreže je $\mathbf{x}(t)$, izlaz $\mathbf{y}(t)$, a stanje mreže (skriveni sloj) $\mathbf{s}(t)$. Ulazni vektor $\mathbf{x}(t)$ nastaje ulančavanjem vektora $\mathbf{w}(t)$ trenutačne rieči uznačene po 1 od V i izlaza neuronâ surječnog sloja u predhodnom trenutku $\mathbf{s}(t-1)$. Izlazni vektor $\mathbf{y}(t)$ predstavlja vjerojatnostnu razdiobu sljedeće rieči na temelju predhodne rieči $\mathbf{w}(t)$ i surječja $\mathbf{s}(t-1)$, tj. $P(w_{t+1}|w_t, \mathbf{s}(t-1))$.[5]

Vriednosti se računaju po sljedećim obrazicama:

[5]Ulazni vektor $\mathbf{w}(t)$ i izlazni vektor $\mathbf{y}(t)$ imaju, dakle, protežnost jednaku veličini rječnika.



Slika 3.7: Jednostavna povratna mreža. Preuzeto iz (Mikolov, 2012)

.

$$s_j(t) = f\left(\sum_i w_i(t) \cdot u_{ji} + \sum_l s_l(t-1) \cdot w_{jl}\right)$$
$$y_k(t) = g\left(\sum_j s_j(t) \cdot v_{kj}\right)$$

gdje je $f(z)$ sigmasta aktivacijska funkcija:

$$f(z) = \frac{1}{1+\exp(-z)}$$

a $g(z)$ funkcija *softmax*, koja osigurava valjanost vjerojatnostne razdiobe ($y_m(t) > 0$ za svaku rieč m i $\sum_k y_k(t) = 1$):

$$g(z_m) = \frac{\exp(z_m)}{\sum_k \exp(z_k)}$$

Mreža je određena ulaznim, skrivenim i izlaznim slojem te odgovarajućim težinskim matricama: matrice **U** i **W** između ulaznog i skrivenog sloja, a matrica **V** između skrivenog i izlaznog sloja (vidi sliku 3.7). Gornje se jednačbe mogu zapisati matrično-vektorski:

$$\mathbf{x}(t) = [\mathbf{w}(t)^\top \mathbf{s}(t-1)^\top]^\top$$
$$s(t) = f(\mathbf{U}\mathbf{w}(t) + \mathbf{W}\mathbf{s}(t-1))$$
$$\mathbf{y}(t) = g(\mathbf{V}\mathbf{s}(t))$$

Mreža se uvještbava stohastičkim gradijentnim spustom bilo pomoću postupnika retropropagacije (BP) bilo retropropagacije kroz vrieme (BPTT).

Težinske matrice **U**, **V** i **W** početno su postavljene na male slučajne brojeve, $t = 0$, a neuroni skrivenog sloja $s(t)$ postavljeni na 1. Jedna epoha izgleda ovako:

1. Povećaj vremenski brojač t za 1
2. Preslikaj stanje skrivenog sloja $s(t-1)$ u ulazni sloj
3. Izvedi unapriedno prosliđivanje kako je predhodno opisano što posljedjuje sa $\mathbf{s}(t)$ i $\mathbf{y}(t)$
4. Izračunaj gradijent pogreške $\mathbf{e}(t)$ u izlaznom sloju
5. Propagiraj pogrešku unazad kroz mrežu i osvježi težine
6. Ako nisu obrađeni svi primjeri za učenje, odi na korak 1

Dakle, težinske se matrice osvježavaju po predočenju svakog primjera za učenje.

Primjere za učenje označimo s $t = 1, \ldots, T$, indeks rieči koju treba predvidjeti za t-ti primjer s l_t, tada je funkcija cilja (objective) koju nastojimo ponajmanjiti/ponajnižiti izglednost (likelihood) podataka za učenje:

$$f(\lambda) = \sum_{t=1}^{T} \log y_{l_t}(t)$$

Gradijent vektora pogrešaka $\mathbf{e}_o(t)$ u izlaznom sloju računa se po sudilu unakrižne entropije koje nastoji ponajvišiti izglednost točnog razreda:

$$\mathbf{e}_o(t) = \mathbf{d}(t) - \mathbf{y}(t)$$

gdje je $\mathbf{d}(t)$ vektor, uznačen po 1 od V, koji predstavlja rieč $\mathbf{w}(t+1)$ koja treba biti predviđena.

Težine **V** između skrivenog sloja $\mathbf{s}(t)$ i izlaznog sloja $\mathbf{y}(t)$ osvježavaju se ovako:

$$v_{jk}(t+1) = v_{jk}(t) + s_j(t) e_{ok}(t) \alpha - v_{jk}(t) \beta$$

gdje je α stopa učenja, j trči po veličini skrivenog sloja, k po veličini izlaznog sloja, $s_j(t)$ je izlaz j-tog neurona skrivenog sloja, $e_{ok}(t)$ je gradijent progreške k-tog neurona izlaznog sloja, β je opcionalni parametar regularizacije L2. Matrično-vektorski zapisano:

$$\mathbf{V}(t+1) = \mathbf{V}(t) + \mathbf{s}(t)\mathbf{e}_o(t)^\top \alpha - \mathbf{V}(t)\beta$$

Gradijenti pogrešaka propagiraju se iz izlaznog u skriveni sloj:

$$\mathbf{e}_h(t) = d_h(\mathbf{e}_o(t)^\top \mathbf{V}, t)$$

gdje se funkcija d_h primjenjuje element po element (elementice):

$$d_{hj}(x,t) = x s_j(t)(1 - s_j(t))$$

Potom se težine $\mathbf{U}$ između ulaznog $\mathbf{w}(t)$ i skrivenog $\mathbf{s}(t)$ sloja osvježe:[6]

$$u_{ij}(t+1) = u_{ij}(t) + w_i(t) e_{hj}(t)\alpha - u_{ij}(t)\beta$$

odnosno:

$$\mathbf{U}(t+1) = \mathbf{U}(t) + \mathbf{w}(t)\mathbf{e}_h^\top(t)\alpha - \mathbf{U}(t)\beta$$

Povratne (reccurent) težine $\mathbf{W}$ osvježavaju se ovako:

$$w_{lj}(t+1) = w_{lj}(t) + s_l(t-1) e_{hj}(t)\alpha - w_{lj}(t)\beta$$

tj.

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \mathbf{s}(t-1)\mathbf{e}_h(t)^\top \alpha - \mathbf{W}(t)\beta$$

Podatci za validaciju rabe se za rano zaustavljanje i za upravljanje stopom učenja. Nakon svake epohe mreža se provjerava na validacijskim podatcima, ako se log-izglednost validacijskih podataka povećava, uvještbavanje se nastavlja u novu epohu, ako pak nema poboljšanja, stopa učenja α se prepolovi na početku sljedeće epohe. Ako još jednom nema znatnog poboljšanja, uvještbavanju je kraj. Konvergencija se obično postiže u 10–20 epoha.

Vremenska složenost jednog koraka pri uvještbavanju ili izpitivanju je:

$$O = H \times H + H \times V = H \times (H + V)$$

gdje je H veličina skrivenog sloja, a V veličina rječnika.

Model uvećan dodatnim značajkama

U (Mikolov i Zweig, 2012) predstavljen je proširen model u kojem je na ulazu dodan *sloj značajki* $\mathbf{f}(t)$ koji je spojen i na skriveni i na izlazni sloj.

[6] Budući da je samo jedan neuron djelatan (aktivan) u nekom trenutku u ulaznom vektoru $\mathbf{w}(t)$ (zbog uznake 1 od V), dovoljno je osvježiti samo njegove težine.

Svakoj se rieči može pridružiti zbiljnobrojni vektor koji nosi dodatne obaviesti o rieči poput surječnih obaviesti o trenutačnoj rečenici, obaviesti o temi, oznake vrste rieči, oblikoskladnjane obaviesti, obaviesti o govorniku u kontekstu prepoznavanja govora ili bilo koje druge zanimljive obaviesti.

Obrazice za izračun su:

$$\mathbf{s}(t) = f(\mathbf{U}\mathbf{w}(t) + \mathbf{W}\mathbf{s}(t-1) + \mathbf{F}\mathbf{f}(t))$$

$$\mathbf{y}(t) = g(\mathbf{V}\mathbf{s}(t) + \mathbf{G}\mathbf{f}(t))$$

f i g su kao i prije, sigmasta i softmax funkcija.

U navedenom se radu rabi *pritajena Dirichletova*[7] *dodjela*[8] (latent D. allocation, LDA) nad dielom predhodnih rieči (rečenična poviest), čime se dobivaju vektori koji se pridružuju riečima na ulazu mreže te to postaje temom uvjetovan povratnomrežni jezični model.

Pokusi nad *Penn Treebank*om (stablik[9] Penn) pokazuju poboljšanje od relativnih 6% u odnosu na predhodno najbolji rezultat.

3.5.1. Retropropagacija kroz vrieme (BPTT)

Predhodno je opisan uobičajeni postupnik retropropagacije, no taj pristup nije optimalan jer, premda nastojimo ponajboljiti predviđanje sljedeće rieči na temelju predhodne i prošlog stanja skrivenog sloja, ništa se ne poduzima da se u skriveni sloj pohrane obaviesti koje bi mogle biti koristne u budućnosti.

Zato se primjenjuje postupnik retropropagacije kroz vrieme (backpropagation through time, BPTT). Ideja je da se povratna živčana mreža s jednim skrivenom slojem koji se rabi za N vremenskih koraka promatra kao duboka unapriedna mreža s N skrivenih slojeva (jednake protežnosti i s identičnim matricama povratnih težina). To je prikazano na slici 3.9.

Dakle, povratne se težine $\mathbf{W}$ osvježavaju njihovim razmatanjem kroz vrieme i uvještbavanjem mreže kao duboke unapriedne živčane mreže.

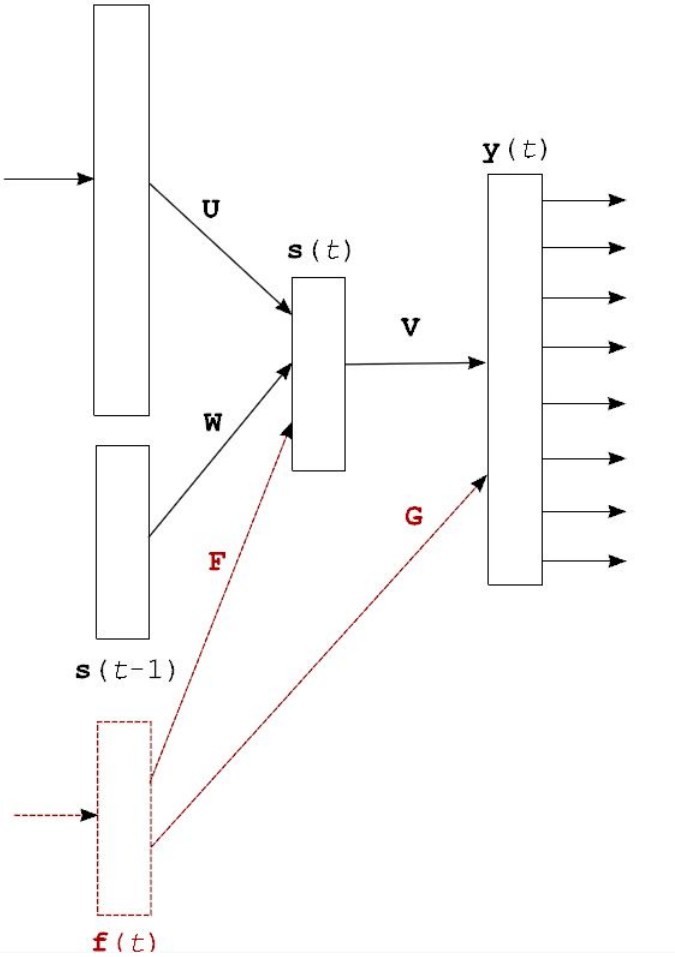
Pogreška se razprostire suvratno (recursively) ovako (uoči da treba pohraniti stanja skrivenog sloja iz predhodnih vremenski koraka):

$$\mathbf{e}_h(t-\tau-1) = d_h(\mathbf{e}_h(t-\tau)^\top \mathbf{W}, t-\tau-1)$$

[7] PETER GUSTAV LEJEUNE DIRICHLET /dirikle/ ili /dirišle/ (1805. – 1859.), njemački matematičar, znatno doprinio brojevnoj zorbi, zorbi Fourierovih redova i ostalim dielovima matematičke razglobe.

[8] Metoda koja preslikava prikaz spisa vrećom rieči u nizkoprotežni vektor koji se shvaća kao predstavak teme (engl. topic).

[9] STABLICI iliti STABLENICI/E (treebank) zbirke su skladnjano ili značbeno razčlanjenih, tj. obilježenih rečenica. Tako razčlanjane rečenice prikazane su stablima, pa odatle naziv.



Slika 3.8: Povratnomrežni model proširen dodatnim slojem značajki $\mathbf{f}(t)$ i odgovarajućim težinskim matricama. Preuzeto iz (Mikolov i Zweig, 2012).

Razmatanje se može primjenjivati toliko koraka koliko je primjera za učenje predočeno, ali gradijenti pogreške brzo zamru pri retropropagaciji kroz vrieme (u riedkim slučajevima mogu eksplodirati), tako da je nekoliko koraka razmatanja dovoljno (truncated BPTT).

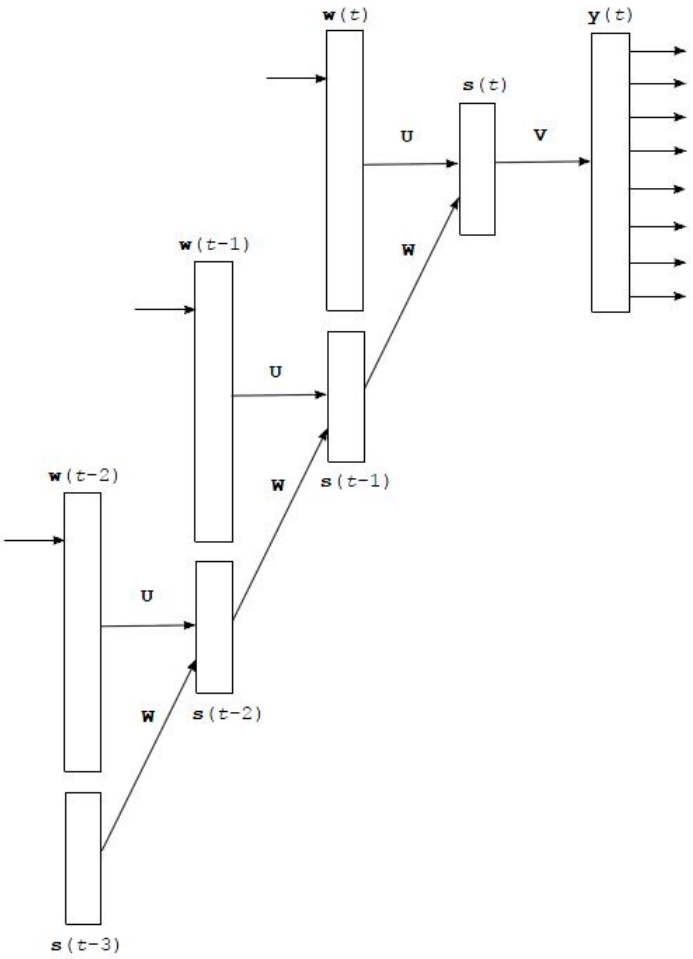
Težine se osvježavaju po ovim obrazicama (T je broj koraka koliko se mreža odmotava kroz vrieme):

$$u_{ij}(t+1) = u_{ij}(t) + \sum_{z=0}^{T} w_i(t-z) e_{hj}(t-z)\alpha - u_{ij}(t)\beta$$

$$w_{lj}(t+1) = w_{lj}(t) + \sum_{z=0}^{T} s_l(t-z-1) e_{hj}(t-z)\alpha - w_{lj}(t)\beta$$

odnosno matrično:

$$\mathbf{U}(t+1) = \mathbf{U}(t) + \sum_{z=0}^{T} \mathbf{w}(t-z)\mathbf{e}_h(t-z)^\top \alpha - \mathbf{U}(t)\beta$$



Slika 3.9: Povratna mreža odmotana u duboku unapriednu mrežu za tri koraka u prošlost. Preuzeto iz (Mikolov, 2012).

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \sum_{z=0}^{T} \mathbf{s}(t-z-1)\mathbf{e}_h(t-z)^\top \alpha$$

Važno je matrice osvježiti odjednom, a ne prirastno (incrementally) tiekom retroprogacije pogrešaka.

Računski je učinkovitije da se mreža razmota nakon obrade nekoliko primjera, tako da složenost uvještbavanja ne raste crtovno s brojem vemenskih koraka T koliko se mreža razmotava kroz vrieme.

3.5.2. Dodatna poboljšanja

Kresanje rječnika

Vremenskoj složenosti izračunavanja najviše doprinosi član $H \times V$ koji odgovara računu između skrivenog i izlaznog sloja. Veličina skrivenog sloja H od 200−500 je ništa nasprem

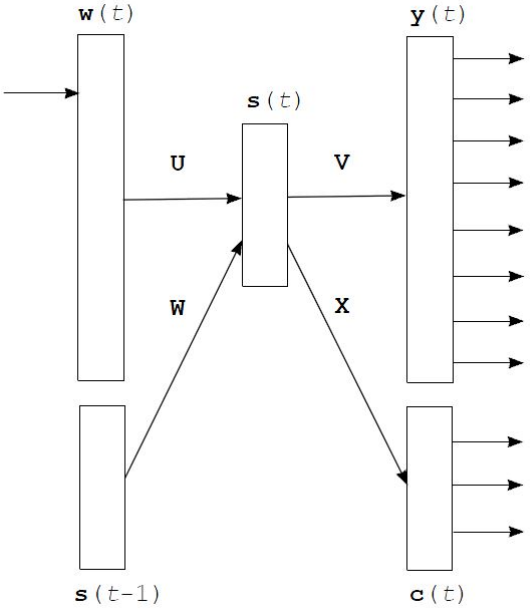
veličine rječnika V. Ovo posebice dolazi do izražaja u oblično bogatim jezicima poput hrvatskoga (podsjetimo se da su u rječniku različnice, ne natučice).

Najjednostavnije rješenje je smanjiti veličinu rječnika na izlazu. Nama na izlazu treba, uglavnom, vjerojatnost jedne rieči, ali radi normalizacije *softmaxom* moramo izračunati vjerojatnosti svih $|V|$ rieči.

Bengio je sve riedke rieči stavio u posebni razred, a unutar njega se vjerojatnosti procjenjuju temeljem jednorječnih čestota.

Faktorizacija izlaznog sloja

Umješniji pristup ubrzanju izračunavanja je grupiranje rieči u razrede. Slično ideju smo vidjeli kod unaprjednih mreža, ugrozđivanje dvojčanim stablom.



Slika 3.10: Faktorizacija izlaznog sloja (Mikolov, 2012)

Goodman (2001a) uveo je ideju razredâ za ubrzanje učenja modela najveće entropije. No ideja je primjenljiva na bilo koji problem s velikim brojem izlaza.

Svakoj rieči pridružimo neki razred. To nam omogućuje da vjerojatnost neke rieči razstavimo na vjerojatnost razreda na temelju poviesti i na vjerojatnost rieči na temelju toga razreda i poviesti.

Rieči se stavljaju u razrede po jednorječnoj čestoti (engl. frequency binning). Time dobivamo malene razrede za česte rieči, a rieđe rieči pripadaju u goleme razrede.

Računamo po sljedećim obrazicama (c je sloj razreda):

$$c_m(t) = g\left(\sum_j s_j(t)\ x_{mj}\right)$$

$$y_{V'}(t) = g\left(\sum_j s_j(t)\ v_{V'j}\right)$$

$$P(w_{t+1}|\ \mathbf{s}(t)) = P(c_i|\ \mathbf{s}(t)) \times P(w_i|\ c_i, \mathbf{s}(t))$$

Opise dodatnih poboljšanja čitatelj može pronaći u (Mikolov, 2012).

3.5.3. Jesu li povratnomrežni modeli bolji od n-rječnih?

Mikolov (Mikolov, 2012) je pokazao množtvom pokusa da povratnomrežni modeli nadmašuju u svakom pogledu (osim brzine) tradicionalne n-rječne modele. Povratnomrežni modeli su trenutačno najbolje poznato rješenje u jezičnom modeliranju.

3.5.4. Uzmetne matrice kao rječni predstavci

Sve gore navedene mreže istodobno uče dvije stvari: jezični model i rječne predstavke. Rječni predstavci predstavljeni su uzmetnom (projekcijskom) matricom koja uzmeće rieči iz prikaza 1 od V u nizkoprotežni vektorski prostor. Dakle, svaki redak matrice rječni je predstavak jedne rieči.

Tu matricu možemo izdvojiti i uporabiti ju bilo gdje gdje možemo i rječne predstavke dobivene drugim postupcima.

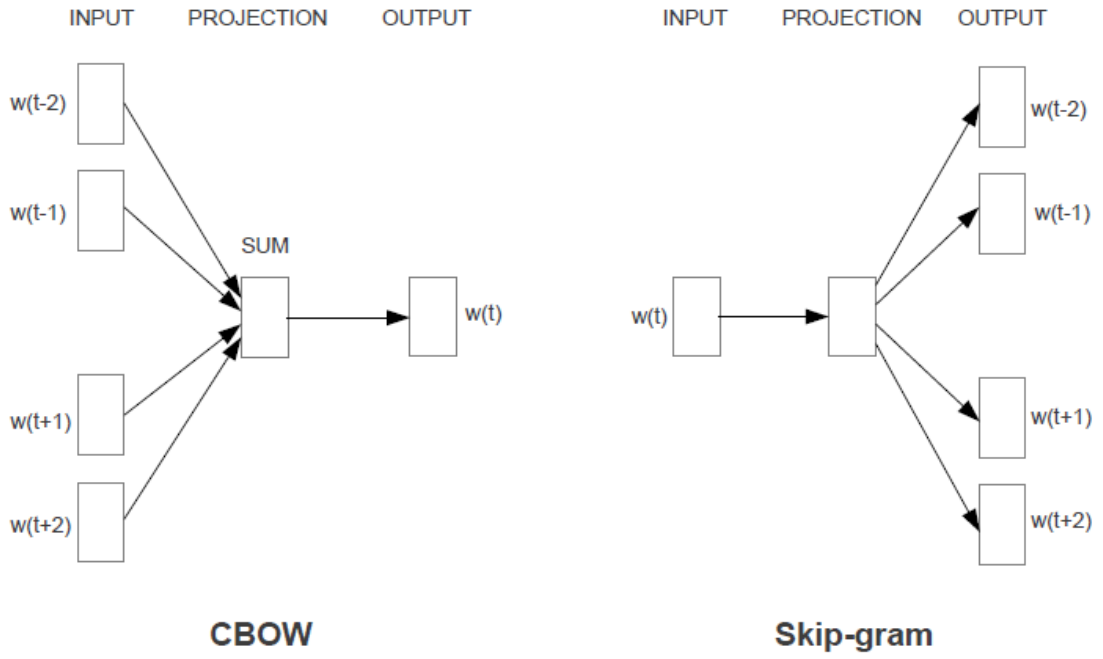
Možemo li možda izgraditi živčanu mrežu koja bi učila kakvotnije rječne predstavke ako nam to i bude cilj učenja mreže, a ne učenje jezičnog modela?

3.6. Mreže za učenje rječnih predstavaka

U (Mikolov et al., 2013a) predstavljena su dva nova modela jednostavnih živčanih mreža za izračun neprekinutih vektorskih predstavaka rieči na temelju velikih skupova podataka. Modeli rade brzinom od milijardu rieči na sat što je neuzporedivo brže od predhodnih modela kod kojih učenje traje danima, tjednima.

Predstavljena su dva modela: model neprekinute vreće rieči (engl. Continuous Bag-of-Words) i neprekinuti skip-gramski model (engl. Continuous Skip-gram). Oba modela nastoje predvidjeti susjede neke rieči.

Na kapaljku su dodatno objašnjeni u (Mikolov et al., 2013b) i (Mikolov et al., 2013c) te djelomice u (Goldberg i Levy, 2014).



Slika 3.11: CBOW i skip-gramski model ((Mikolov et al., 2013a))

3.6.1. Model neprekinute vreće rieči (CBOW)

Model je sličan unapriednom ŽMJM-u, ali je necrtovni skriveni sloj uklonjen, a uzmetni je sloj zajednički za sve rieči (ne samo uzmetna matrica), tako da se sve rieči uzmeću na isto mjesto (pritom se njihovi vektori uprosječe). Vidi sliku 3.11 i uzporedi sa slikom 3.6.

Ovaj model predviđa trenutačnu rieč na temelju surječja, a surječje uključuje kako prošle tako i *buduće*, nadolazeće rieči!

Model se uvještbava stohastičkim gradijentnim spustom postupnikom retropropagacije.

Model funkcionira kao log-crtovni razrednik (klasifikator) kojem je cilj točno razrediti trenutačnu (srednju) rieč na temelju nekoliko prošlih i budućih. Dakle, cilj učenja je kombinirati predstavke okolnih rieči kako bi se predvidjela rieč u sredini.

Naziva se *vrećom rieči* jer redoslied rieči ne utječe na uzmet (budući da se vektori uprosječe), a *neprekinutom* jer rabi neprekinut razpodieljeni prikaz surječja.

3.6.2. Neprekinuti skip-gramski model

Ovaj model radi suprotno od gornjeg, umjesto predviđanja trenutačne rieči na temelju surječja, nastoji predvidjeti susjedne rieči na temelju trenutačne rieči.

Model funkcionira kao log log-crtovni razrednik s neprekinutim uzmetnim slojem koji prima na ulazu trenutačnu rieč, a predviđaju se rieči u zadanom razponu ispred i iza

trenutačne rieči.

Odabere se najveća udaljenost rieči C, potom se za svaku rieč odabere slučajan broj R iz $< 1; C >$ te se uporabi R rieči iz prošlosti i R rieči iz budućnosti kao izpravne oznake. To zahtjeva do $R \times 2$ razredbi rieči s trenutačnom rieči na ulazu i sa svakom od $R+R$ rieči na izlazu.

Model se uvještbava stohastičkim gradijentnim spustom postupnikom retropropagacije.

Cilj je naučiti rječne prikaze koji su dobri za predviđanje susjednih rieči u rečenici. Formalnije, uz dani slijed rieči za učenje $w_1, w_2, \ldots, w_T$ cilj modela skip-gram je maksimizirati prosječnu log-vjerojatnost

$$\frac{1}{T}\sum_{t=1}^{T}\left[\sum_{-c\leq j\leq c, j\neq 0} \log p(w_{t+j}|w_t)\right]$$

gdje je k veličina učevnog prozora (koji može biti funkcija središnje rieči w_t).

Unutarnje zbrajanje teče od $-c$ do c (bez $c = 0$) kako bi se izračunala log-vjerojatnost izpravnog predviđanja rieči w_{t+j} na temelju dane rieči u sredini w_t. Vanjsko, pak, po svim riečima u skupu za učenje.

Osnovni skip-gram određuje $p(w_{t+j}|w_t)$ pomoću funkcije softmax:

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^{\top} \cdot v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left({v'_w}^{\top} \cdot v_{w_I}\right)}$$

gdje je v_w "ulazni", a v'_w "izlazni" vektorski prikaz rieči w. Ta su dva naučljiva vektora parametara pridružena svakoj rieči.[10] W je broj rieči u rječniku. Iz obrazice se može izčitati da želimo što veći ljestvičnički (skalarni) umnožak, tj. što veću sličnost vektora ulazne i vektora izlazne rieči. Ovaj je pristup neporaban (nepraktičan) zbog troška izračunavanja $\nabla \log p(w_O|w_I)$ koji je razmjeran s W.

Radi ubrzanja izračunavanja rabimo supodredni softmax.

3.6.3. Supodredni softmax

Supodredni (hijerarhijski) softmax rabi prikaz izlaznog sloja dvojčanim stablom s W rieči u listovima. Ovaj smo pristup već upoznali kod unaprjednomrežnih modela (vidi posljednji model u odjeljku 3.4).

Pogledajmo kako ga ovdje primjenjujemo. Svakoj se rieči može pristupiti nekim putom od koriena stabla. Neka je $n(w, j)$ j-ti čvor na putu od koriena do w, a $L(w)$ duljina puta,

[10] Iz izvornog rada nije posve jasno što znači "ulazni" i "izlazni", no u sljedećem pododjeljku postaje jasnije.

tako da je $n(w,1) =$ korien, a $n(w, L(w)) = w$. Za svaki unutrašnji čvor n s $ch(n)$ označimo neko učvršćeno diete (npr. neka je to uviek lievo diete), a neka je $[\![x]\!]$ jednako 1 ako je x istinit, a -1 inače. Sada možemo definirati vjerojatnost kao:

$$p(w|w_I) = \prod_{j=1}^{L(w)-1} \sigma\left([\![n(w,j+1) = \mathrm{ch}(n(w,j))]\!] \cdot {v'_{n(w,j)}}^\top v_{w_I}\right)$$

gdje je $\sigma(x) = 1/(1+\exp(-x))$. Da bismo lakše shvatili ideju zapišimo tu vjerojatnost drugačije:[11]

$$p(w|w_I) = \prod_{j=1}^{broj\_predaka} p\left(w.kod[j] = 1 \middle| v'_{predak_j}, v_{w_I}\right)$$

$$p\left(w.kod[j] = 1 \middle| v_{predak_j}, v_{w_I}\right) = \begin{cases} \sigma\left({v'_{predak_j}}^\top v_{w_I}\right) & \text{, ako je idući čvor lievi} \\ \sigma\left(-1 \cdot {v'_{predak_j}}^\top v_{w_I}\right) & \text{, ako je idući čvor desni} \end{cases}$$

gdje je $w.kod$ dvojčana nizanica, uznaka (kod) rieči w, a pojedine nam sastavnice govore kamo treba skrenuti u svakom čvoru da dođemo do w. Ako skrećemo u lievo to možemo označiti s `1`, tako nam `10` može značiti da moramo skrenuti lievo u korienu, a zatim desno da bismo došli do trenutačnog čvora. j teče po svim predcima rieči w (ima ih kolika je duljina koda od w), tj. svim čvorovima na putu do w.

Dakle, vjerojatnost rieči w računamo tako da množimo vjerojatnosti pojedinih skretanja (odabira lievog ili desnog djeteta) na putu do te rieči. Vjerojatnost pojedine odluke računa se kao logistička funkcija primienjena na unutrašnji umnožak rječnog predstavka naše ulazne rieči w_I i rječnog predstavka toga unutrašnjeg čvora stabla. Dodatno, kako bismo imali izpravnu vjerojatnost $p(1|\ldots) + p(0|\ldots) = 1$ množimo taj umnožak s $+1$ ili -1 jer je $\sigma(x) + \sigma(-x) = 1$.

Za razliku od standardnog skip-grama kod kojeg imamo po dva vektora v_w i v'_w za svaku rieč w, ovdje imamo jedan prikaz v_w za svaku rieč w i jedan prikaz v'_n za svaki unutrašnji čvor dvojčanog stabla (uzporedi to se posljednjim modelom u odjeljku 3.4 gdje smo isto imali oddvojene rječne predstavke za ulazne rieči i za čvorove stabla).

Iz ovoga sliedi da je trošak računanja $\log p(w_O|w_I)$ i $\nabla p(w_O|w_I)$ razmjeran $L(w_O)$, koji u prosjeku nije veći od $\log W$.

Za izgradnju dvojčanog stabla rabi se Huffmanova uznaka.

O još jednoj metodi ubrzanja izračunavanja nazvanu *negativno uzorkovanje* čitatelj može pročitati u (Mikolov et al., 2013b), a ovdje pogledajmo jednu jednostavnu metodu.

[11] Ovdje iznosim svoje shvaćanje, u navedenom je radu opis prilično štur pa se oslanjam na svoje razumievanje.

3.6.4. Poduzorkovanje čestih rieči

U velikim se korpusima najčešće rieči pojavaljuju stotinama milijuna puta, ali ono što je često nosi manje obaviesti od nečeg riedkog. Tako se naprimjer u englezkom određeni član *the* supojavljuje sa skoro svakom riečju, tj. ne nosi skoro nikakvu obaviest. Možemo predpostaviti da se vektori čestih rieči neće značajno promieniti nakon učenja nad milijunima primjera. Stoga kako bi se uravnotežio nesrazmjer između čestih i riedkih rieči rabimo poduzorkovanje, svaku rieč w_i u skupu za učenje odbacujemo s vjerojatnosti:

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

gdje je $f(w_i)$ čestota rieči w_i, a t odabrani prag (obično $\sim 10^{-5}$).

Ovim se postupkom ubrzava učenje i dobivaju se kakvotniji vektori za riedke rieči.

3.6.5. Pseudokod CBOW-a

Kako bismo bolje razumieli ove modele pokažimo kako izgleda postupnik učenja.[12] Pogledajmo na primjer CBOW, u kojem na temelju susjednih rieči predviđamo trenutačnu.

1. Stvori matricu `syn0` u koju ćemo pohranjivati rječne vektore i popuni ju slučajnim brojevima. Njezine su protege $|V| \times layer1\_size$, gdje je `layer1_size` veličina skrivenog sloja, tj. protežnost vektora.

2. Stvori matricu `syn1` jednakih protega kao predhodna u koju ćemo pohranjivati rječne vektore za unutrašnje čvorove stabla i popuni ju ništicama.

3. Učitaj skup za učenje. Izgradi Huffmanovo stablo. Sada za svaku rieč znamo put od koriena do nje prikazan dvojčanom nizanicom kako je predhodno opisano (kôd/uznaka rieči), tj. znamo tko su joj roditelji.

4. Za svaku `rieč`, odredimo njezine susjede kako je gore opisano i čini:

5. ulazni → skriveni: zbroji rječne vektore za rieči susjede = `neu1`

6. skriveni → izlazni: Za `d` od `0` do `duljina(rieč.code)`:
   - `f = dot(neu1,syn1[predak_d])`
   - `f = 1/(1 + exp(f))`

[12] U navedenim radovima nije dan nikakav postupnik niti obrazice osim gore navedenih. Ovaj je opis nastao proučavanjem izvorištnog koda ovoga modela (vidi kasnije word2vec).

- `g = stopa_učenja * (1 - rieč.code[d] - f)` // `rieč.code` odgovara *w.kod* u gornjem opisu, ovaj bi se gradijent trebao moći dobiti logaritmiranjem pa deriviranjem gore navedene obrazice za vjerojatnost.[13]
- greška iz izlaznog → skriveni: `neu1e += g * syn1[predak_d]`
- nauči težine skriveni → izlazni: `syn1[predak_d] += g * neu1`

7. skriveni → ulazni: za svaku rieč susjeda `w_s` od `rieč`: `syn0[w_s] += neu1e`
8. Ako ima još rieči, vrati se na korak 4, tj. 5. Inače je `syn0` matrica rječnih predstavaka koju tražimo.

3.6.6. Kakvoća dobivenih prikaza

U (Mikolov et al., 2013a) pokazanu je da se ovim vektorima postižu najbolji poznati rezultati na zadatcima skladnjane i značbene sličnosti. Neke od tih pokusa ponovit ćemo za hrvatski, a ostale pokuse čitatelj može potražiti u navedenom radu.

[13] Primietimo prvo da je $\sigma(-x) = \exp(-x)/(1+\exp(-x))$, sada možemo dva slučaja u obrazici za vjerojatnost zapisati sažeto kao $\sigma(umnožak) = \exp((1-w.kod[d])\cdot umnožak)/(1+\exp(umnožak))$.

4. Naputačno ostvarenje

U sklopu rada izrađen je skup naputaka (programa) za provedbu pokusa opisanih u sljedećem poglavlju. Potrebni naputci izrađeni su u naputnom jeziku Python inačice 2.7. Python je općenamjenski, tumačen naputni jezik visoke razine koji podržava višestruke naputbene paradigme, uključujući zapovjednu, predmetno usmjerenu i, u manjoj mjeri, funkcijsku.

Kao pomoć pri radu s naučenim modelima rabi se Pythonova knjižnica Gensim. GENSIM[1] je otvorenokodna Pythonova knjižnica za obradu prirodnog jezika i ponalazbu obaviesti. Sastoji se od učinkovitih ostvarenja postupnika poput pritajene značbene razglobe, pritajene Dirichletove dodjele, slučajnih uzmeta, supodrednog Dirichletova procesa i dubokog učenja word2vecom.[2]

Za učenje jezičnih modela i rječnih predstavaka rabimo postojeće naputke.

Oruđnica za PMJM

Oruđnica za PMJM[3] (engl. RNNLM Toolkit) je skup oruđa koji implementira povratnomrežni jezični model kako je opisano u predhodnim poglavljima.

Rabimo ju za učenje povratnomrežnih jezičnih modela.

SRILM

SRILM[4] je oruđnica za izgradnju i primjenu uzorkoslovnih jezičnih modela.

U ovom se radu rabi za učenje n-rječnih modela.

[1] `http://radimrehurek.com/gensim/index.html`

[2] Englezki su nazivi redom: latent semantic analysis, LSA/LSI; latent D. allocation, LDA; random projections, RP; hierachical D. proces, HDP; word2vec deep learning.

[3] `http://rnnlm.org/`

Inačica uvećana značajkama `https://research.microsoft.com/en-us/projects/rnn/`

[4] `http://www.speech.sri.com/projects/srilm/`

word2vec

WORD2VEC[5] je oruđe koje pruža učinkovito ostvarenje arhitekturâ neprekinute vreće rieči te skip-grama za računanje vektorskih prikaza riečî.

Rabimo ga za učenje rječnih predstavaka.

Naputak ima mnogo parametara koji se proslieđuju pri pozivu. Popis parametara može se dobiti pozivanjem programa bez parametara.

Za provedbu pokusa 5.3.5 rabimo sustav CroNER.

CroNER

CRONER je sustav za prepoznavanje i razredbu imenovanih sućaka u hrvatskom jeziku temeljen na nadziranom obilježavanju sliedova pomoću uvjetnih nasumičnih polja (conditional random fields, CRF). Prepoznavanje imenovanih sućaka zadatak je pronalaženja i razredbe svih imena, vremenskih i brojevnih izraza u orječju. CRONER rabi bogat skup rječnih (sama rieč, njezina natuknica, završetak rieči, svi padežni oblici, oblikoskladnjani opisnici itd.) i imeničkih[6] značajki. Trenutačno je najbolji postojeći sustav za slavenske jezike. Potanje je opisan u (Karan et al., 2013)

Rječni su predstavci dodani kao dodatne rječne značajke tako da svaku protegu vektora dodamo kao novu značajku plus jedna dvojčana značajka koja je jednaka jedan ako za tu rieč postoji vektor, inače je nula.

Zbog prevelikih memorijskih zahtjeva nije bilo moguće dodati i značajke za nekoliko predhodnih i nekoliko sljedećih rieči u odnosu na trenutačnu rieč. Niti je bilo moguće provesti pokuse za prevelike vektore.

4.1. Važniji naputci za provedbu pokusa

Svi su ovi naputci samorazumljivi, a ovdje se donosi kratak opis njihovog funkcioniranja koji je manje razumljiv od samih naputaka.

4.1.1. `synonyms.py`

Ovaj naputak služi za provedbu pokusa odabira suznačnica 5.3.2.

[5] `https://code.google.com/p/word2vec/`

[6] IMENICI (engl. gazetteers) popisi su imena ljudi (osobna imena i prezimena), organizacija, ulica, gradova, država i drugih rieči koje odgovaraju rabljenim razredima.

Naputak pri pozivu prima putanju do rječnih vektora i skupa nad kojim se izpituju. Putanja u oba slučaja može biti do jedne datnice (datoteke) ili do mape koja sadrži više datnica s vektorima ili skupom.

Naputak učitava jednu po jednu datnicu s rječnim vektorima pomoću Gensimove funkcije `load_word2vec_format` koja kao parametar prima podatak je li model zapisan dvojčano ili orječno, potom prolazi po svim datnicama za izpitivanje, i zatim za svaki redak računa sličnost između ciljne rieči i ostalih (vidi opis pokusa) pomoću Gensimove funkcije `similarity` koja računa kosinusnu sličnost dvaju vektora. Rieči se poslože po sličnosti od najveće i računa se broj točnih slučajeva (izpravna rieč je na prvom mjestu) i srednji obratni rang. Ako za ciljnu rieč ne postoji vektor taj se redak preskače, a ako za neki od ponuđenih suznačnica ne postoji vektor, sličnost se postavlja na nulu. Uzput se bilježe rieči kojih nema u rječniku (OOV).

4.1.2. `relatedness.py`

Ovaj naputak služi za provedbu pokusa 5.3.3. Isto kao i gornji naputak, učitava modele i pitanja, potom prolazi po svim modelima, pa po svim pitanjima i računa sličnost između dviju rieči u redku Gensimovom funkcijom `similarity` i izpisuje ju na zaslon. Ako neka rieč ne postoji u rječniku, sličnost se postavlja na nulu.

4.1.3. `comparatives.py`

Ovaj naputak služi za provedbu pokusa 5.3.4.

Kao i gornji naputci učitava modele i pitanja. Potom prolazi po svim modelima i pitanjima i poziva Gensimovu funkciju `accuracy`. Ona prima popis pitanja, prolazi po njima, računa s vektorima kako je opisano kasnije u pokusu, prolazi po svim riečima tražeći najsličniju. Ako je najsličnija ona koja treba biti, tada je to točan slučaj. Na kraju se izpisuje broj točnih i netočnih slučajeva.

5. Pokusno vrednovanje

5.1. Skupovi dataka

5.1.1. fHrWaC

HRWAC je hrvatska spletna građara prikupljena s vršnog područja `.hr`. Trenutačna inačica (v1.0) sadrži 910 milijuna pojavnica. Podrobnije na `http://nlp.ffzg.hr/resources/corpora/hrwac/`

FHRWAC je procieđena (filtrirana) inačica starije inačice hrWaC-a. Uklonjen je veći dio neorječnog sadržaja (npr. dielovi uobličnog kôda), uznačne pogreške i stranojezični sadržaj. Podrobnije u radu (Šnajder et al., 2013). Građara sadrži 50.940.598 rečenica (jedna rečenica po redku, opojavničena[1]) i 1.232.632.208 (1.2G) pojavnica. Duljina prosječne rečenice iznosi 24,1974 pojavnica.

Građaru se može preuzeti s `http://takelab.fer.hr/data/fhrwac/`.

5.1.2. Skup podataka za odabir suznačnica

Skup podataka za zadatak odabira suznačnica (sinonima) u hrvatskom jeziku. Skup se sastoji od tri datnice, po jedna za imenice, pridjeve i glagole. Svaka datnica sadrži 1000 pitanja, jedno po redku, a svako se pitanje o suznačnicama sastoji od ciljne rieči i četiriju odgovora od kojih je samo jedan suznačnica ciljne rieči, a preostala tri odvraćaju s točnoga:

```
ciljnaRieč:odgovor1:odgovor2:odgovor3:odgovor4:IDtočnogOdgovora
```

Skup je samodjelno (automatski) proizveden iz strojnočitljiva rječnika, potanje u (Karan et al., 2012) i (Šnajder et al., 2013). Može ga se preuzeti s `http://takelab.fer.hr/data/crosyn/`. Primjer diela skupa:

[1] Za objašnjenje nekih osnovnih pojmova vidi dodatak A.

```
autodidakt:primopredajnik:samouk:prestup:pripovijetka:2
konformizam:prilagodljivost:blento:komedija:čok:1
divan:nat:sobarica:ljutnja:otoman:4
```

5.1.3. CroSemRel450

CROSEMREL450 je skup podataka za zadatke značbene povezanosti (engl. semantic relatedness) u hrvatskom jeziku. Dostupan je na `http://takelab.fer.hr/data/crosemrel450/`

Skup se sastoji od 450 parova rieči i prosječne ocjene značbene povezanosti koje su dodielili ljudski ocjenivači. Ocjene se kreću od 1 do 5 (najjača povezanost). Prva datnica (`CroSemRel450-12.txt`) sadrži prosječne ocjene dvanaestero ocjenjivača, a druga (`CroSemRel450-6.txt`) prosječne ocjene šestero ocjenjivača (od početnih 12) s najvećim uzajamnim slaganjem u ocjeni.

Pojam značbene povezanosti ovdje obuhvaća kako paradigmene značbene odnose (protuznačnost, suznačnost, podređenost, supodređenost i meronimiju (čestnost)) tako i sintagmene odnose u ustaljenim izrazima (ustaljenice, višerječni izrazi), ali i tvorbene odnose i šire asocijativne odnose. Podrobnije u (Janković et al., 2011).

Primjer:

politika	politički	4.91666666667
momčad	tim	4.91666666667
istaknuti	isticati	4.91666666667
igrač	igrati	4.91666666667
država	državni	4.91666666667
reći	kazati	4.83333333333
reći	izjaviti	4.83333333333
reći	govoriti	4.83333333333
nov	star	4.75
film	utorak	1.16666666667
domaći	vidjeti	1.16666666667

5.1.4. Skup poredbenika pridjevâ

Za potrebe ovog rada izgrađen je skup pitanja sljedećeg oblika:

$\texttt{osnovnik\_pridjeva}_1$ $\texttt{poredbenik\_pridjeva}_1$ $\texttt{osnovnik\_pridjeva}_2$ $\texttt{poredbenik\_pridjeva}_2$

Dakle, u prva se dva stupca nalaze osnovni i poredbeni stupnjevi nekoga pridjeva, a u trećem i četvrtom nekog drugog pridjeva.

Skup je izgrađen tako da je na temelju podatka o učestalosti u korpusu odabrano 50 pridjeva čiji su poredbenici česti. Zatim je od tih 50 odabrano osobnom procjenom 10 najčešćih ili zanimljivih pridjeva. Potom se za svaki od tih 10 pridjeva nasumično bira 35 od preostalih 49 pridjeva. Skup se dobiva tako da se 35 puta zapišu parovi osnovnik-poredbenik za svaki od odabranih 10, a pored njih, u trećem i četvrtom stupcu, osnovnici i poredbenici njima pridruženih 35 pridjeva. Ukupno ima, dakle, 350 dvoparova (četvorki). Jasnije će biti na primjeru skupa:

```
bogat bogatiji debeo deblji
bogat bogatiji važan važniji
...
bogat bogatiji poznat poznatiji
brz brži malen manji
brz brži blag blaži
...
brz brži poznat poznatiji
dobar bolji normalan normalniji
dobar bolji potreban potrebniji
...
```

Ideja vodilja bila je da će tih 10 najčešćih parova osnovnik-poredbenik dobro uhvatiti "ideju" poredbenog stupnja (vidi pokus 5.3.4).

5.1.5. Skup najčešćih država i njihovih glavnih gradova

Skup se sastoji od 506 pitanja oblika:

$$\texttt{glavni\_grad}_1\ \texttt{država}_1\ \texttt{glavni\_grad}_2\ \texttt{država}_2$$

Skup je dobiven prevođenjem englezke inačice skupa izgrađene u (Mikolov et al., 2013a), koji je dostupan na `https://code.google.com/p/word2vec/source/browse/trunk/questions-words.txt` (prvi odjeljak).

Slično kao i kod poredbenikâ pridjevâ, po 22 se puta ponavlja sadržaj prvih dvaju stupaca:

```
Atena Grčka Bagdad Irak
Atena Grčka Bangkok Tajland
...
Atena Grčka Tokio Japan
Bagdad Irak Bangkok Tajland
...
Bagdad Irak Atena Grčka
...
```

5.2. Mjere

5.2.1. Kosinusna sličnost

Najpopularniji način mjerenja sličnosti dvaju vektora jest kosinus kuta koji zatvaraju:

$$\cos(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Dakle, kosinus kuta dvaju vektora jest njihov unutarnji umnožak pošto su normalizirani na jediničnu duljinu. Uočimo da su duljine vektora nevažne, važan je samo kut između njih.

Kosinus poprima vriednosti od -1 (kut 180°) kada vektori pokazuju u suprotnim smjerovima, preko 0 (kut 90°) kada su vektori okomiti do $+1$ (kut 0°) kada pokazuju u istom smjeru.

5.2.2. Zamršenost (perpleksnost)

ZAMRŠENOST (engl. perplexity) je obvjestnozorbena mjera koja mjeri koliko dobro vjerojatnostna razdioba ili vjerojatnostni model predviđa uzorak. Rabi se za poredbu vjerojatnostnih modela.

Zamršenost razlučne (diskretne) vjerojatnostne razdiobe p definirana je kao:

$$2^{H(p)} = 2^{-\sum_x p(x)\log_2 p(x)}$$

gdje je $H(p)$ entropija razdiobe, a x teče po događajima.

Zamršenost slučajne promjenljivice X definira se kao zamršenost razdiobe po svim mogućim x-evima. Perpleksnost se može shvatiti i kao inverz vjerojatnosti izpitnog skupa (prema dotičnom JM-u), normalizirano geometrijskim prosjekom po broju rieči:

$$PPL = \sqrt[\kappa]{\prod_{i=1}^{\kappa} \frac{1}{P(w_i|w_1,\ldots,w_{i-1})}}$$

Ugrubo rečeno, zamršenost jest mjera veličine skupa rieči iz kojeg se bira sljedeća rieč ako smo prehodno opazili tu i tu poviest.

Važno je napomenuti da perpleknost ne ovisi samo o kakvoći modela, nego i o podatcima za učenje i izpitivanje.

Što je zamršenost niža, to je model bolji.

5.2.3. Srednji obratni rang

Srednji obratni rang (engl. mean reciprocal rank) je uzorkoslovna mjera za vrednovanje bilo kojeg procesa koji proizvodi spisak mogućih odgovora na neki upit, poredanih po izpravnosti. Obratni rang odziva na upit je množitbeni obrat ranga prvog točnog (relevantnog) odgovora. Dakle, obratni rang je 1 ako je izpravan odgovor dohvaćen na prvom mjestu, $0,5$ ako je na drugom mjestu itd. Kada se uprosječi po svim upitima Q dobivamo srednji obratni rang:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rang_i}$$

5.2.4. Točnost

Točnost (engl. accuracy) je obćenito omjer broja točnih slučajeva i ukupnog broja slučajeva:

$$T = \frac{točno}{ukupno} \times 100\%$$

5.2.5. Koeficijenti korelacije

PEARSONOV KOEFICIJENT KORELACIJE (sučinilac suodnosa) je mjera *linearne* korelacije (crtovnog suodnosa) između dviju promjenljivica X i Y, koja poprima vriednosti između $+1$ i -1, gdje je $+1$ podpuna pozitivna korelacija, 0 bez korelacije i -1 podpuna negativna korelacija.

Računa se kao omjer su(s)mjenljivosti (engl. covariance) promjenljivica X i Y i umnožka standardnih odstupanja: $\frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$, odnosno računano nad nekim uzorkom (x_i-evi i y_i-evi):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

gdje je $\bar{x}$ srednja vriednost promjenljivice X.

SPEARMANOV KOEFICIJENT KORELACIJE je mjera statističke zavisnosti dviju promjenljivica koja ocjenjuje koliko se dobro njihov odnos može opisati *monotonom* (jednolikom) funkcijom. Također poprima vriednosti između $+1$ i -1.

Spearmanov se korelacijski koeficijent definira kao Pearsonov korelacijski koeficijent između uporedkovanih (rangiranih) vriednosti.

$$\rho = \frac{\sum_i (rg(x_i) - \overline{rg}_x)(rg(y_i) - \overline{rg}_y)}{\sqrt{\sum_i (rg(x_i) - \overline{rg}_x)^2} \sqrt{\sum_i (rg(y_i) - \overline{rg}_y)^2}}$$

gdje je $rg(x_i)$ rang od x_i.

Rang se određuje tako da se podatci svrstaju po veličini, potom se najmanjoj vriednosti pridruži rang 1, sljedećoj rang 2 itd. U slučaju istih vriednosti uzimamo prosjek odgovarajućih rangova.

5.3. Pokusi

Provedeno je nekoliko pokusa u kojima je naglasak stavljen na izpitivanje kakvoće rječnih predstavaka, a jednim je pokusom uzpoređena perpleksnost ŽMJM-a i n-rječnog JM-a.

5.3.1. Zamršenost jezičnog modela

Pomoću oruđnice za PMJM naučena su dva modela:

1. Model naučen nad prvih 300.000 redaka korpusa FHRWAC, skriveni sloj duljine 500 (`-hidden 500`), s 500 razreda (`-classes 500`), 2 koraka vremenskog odmotavanja (`-bptt 2`) u blokovima po 10 (`-bptt-block 10`), ostali su parametri ostavljeni na zadanim vriednostima. Učenje je trajalo 10 iteracija, a za validaciju je rabljeno sljedećih 75.000 redaka korpusa.

2. Model naučen nad zadnjih 10.188.120 redaka (20%) korpusa FHRWAC, skriveni sloj duljine 50, s 10 razreda, 2 koraka vremenskog odmotavanja u blokovima po 10, ostali su parametri ostavljeni na zadanim vriednostima. Učenje je trajalo 10 iteracija, a za validaciju je rabljeno prvih 300.000 redaka korpusa.

Zbog velikog vremenskog troška izračunavanja odabrani su manji dielovi korpusa ili manji skriveni slojevi.

Pomoću SRILM-a naučena su dva n-rječna jezična modela:

1. Dvorječni model naučen nad prvih 80% korpusa (40M rečenica) s Kneser-Neyevom zagladom.

2. Trorječni model naučen nad prvih 80% korpusa (40M rečenica) s Kneser-Neyevom zagladom.

Zamršenost modelâ računata je nad skupom od 1M rečenica koje nisu rabljene pri učenju.

Križnica 5.1: Zamršenost jezičnog modela

Model	Zamršenost
RNNLM-1	888,465
RNNLM-2	462,560
Dvorječni	513,762
Trorječni	**351,442**

5.3.2. Prepoznavanje suznačnica

Prepoznavanje suznačnica važno je u brojnim zadatcima na području obrade jezika i ponalazbe obaviesti poput razgraničbe višeznačnosti, proširivanja upita, prepričavanja, proizvedbe jezika, prikupa WordNeta ili pojednostave (uprostbe) orječja.

Modele vrednujemo rječničkim izpitom sličnosti, tj. nad skupom pitanja o suznačnicama koja se sastoje od ciljne rieči i četiri odgovora od kojih je samo jedan suznačnica ciljne rieči, a preostala tri odvraćaju s točnoga. Rabi se skup podataka 5.1.2.

Pokus je obavljen tako da se između vektorâ (rječnih predstavaka) ciljne rieči i svake od ponuđenih suznačnica izračuna kosinusna sličnost (u slučaju da neka rieč nije u rječniku sličnost je 0), potom se suznačnice poredaju po sličnosti od najsličnije.

Mjere koje se rabe su točnost i srednji obratni rang (MRR). Točan je slučaj onaj u kojem je izpravna suznačnica na prvom mjestu suznačnicâ poredanih po sličnosti.

Modeli s kojima uzpoređujemo naše su najbolji model iz (Karan et al., 2012) izgrađen pomoću pritajene značbene razglobe (engl. latent semantic analysis, LSA) s 500 protega i odlomcima (P) kao surječje te model razdiobne memorije iz (Šnajder et al., 2013).

Posljedci pokusa prikazani su u križnici 5.2. U prvom se stupcu nahodi rabljeni model, u drugom, trećem i četvrtom točnost, a u prestalima srednji obratni rang (MRR) za imenice, pridjeve i glagole redom. MRR dodatno dielimo s brojem primjera (1000) kako bismo dobili manje strogu suvrst točnosti.

Vektori LSA500D dobiveni su u (Karan et al., 2012) LSA-om s dokumentom kao kontekstom i njih smo rabili za provedbu pokusa jer nam vektori LSA500P koji postižu bolje rezultate nisu bili dostupni.

Sve rječne uložbe učene su word2vecom nad cielim korpusom FHRWAC. Parametri su zapisani ovako: vrsta-veličina_vektora-veličina_prozora. Ostali parametri imaju nenavodne (engl. default) vriednosti.

Vrste su *skip* za skip-gramski, a *cbow* za model neprekinute vreće rieči.

Križnica 5.2: Posljedci prepoznavanja suznačnica

Model	I	P	G	I-MRR	P-MRR	G-MRR
LSA500P (Karan et al., 2012)	68,7	68,2	61,6	–	–	–
LSA500D	60,0	60,8	50,7	75,2	75,8	69,6
Dm.Hr (Šnajder et al., 2013)	70,0	66,3	63,2	–	–	–
skip_100_5	71,9	69,9	71,3	82,8	81,5	82,9
skip_200_5	73,4	71,9	74,1	83,8	82,5	84,5
skip_200_10	75,6	72,6	70,1	84,7	82,8	82,4
skip_500_5	75,5	**73,0**	**75,8**	84,9	**83,1**	**85,6**
skip_1000_10	**76,8**	72,7	72,2	**85,6**	**83,1**	83,4
cbow_100_5	61,7	69,3	69,0	76,7	80,9	81,5
cbow_100_10	62,5	67,3	64,9	76,7	79,7	79,0
cbow_200_5	66,2	70,6	72,1	79,1	81,7	83,1
cbow_200_10	64,7	67,8	68,6	78,4	80,0	81,1
cbow_500_5	66,9	70,3	72,8	79,5	81,5	83,7
cbow_1000_5	66,6	70,3	72,1	79,3	81,4	83,2
cbow_1000_10	29,8	25,9	27,6	55,2	51,3	54,1

5.3.3. Ocjena značbene povezanosti i sličnosti

Rieči mogu ulaziti u razne značbene odnose, a ovdje ćemo izpitati odnos značbene sličnosti (engl. similarity) i odnos značbene povezanosti (engl. relatedness).

Slične su naprimjer rieči *automobil* i *kamion*, a primjer rieči koje nisu slične, ali su povezane su *automobil* i *vozač*. Možemo vidjeti da je povezanost općenitija kategorija od sličnosti, svi su slični pojmovi povezani, ali nisu svi povezani pojmovi slični. Rieči koje su značbeno slične imaju slična susjedstva, ali se obično ne supojavljuju, dok se povezane rieči često supojavljuju.

Mjerenje značbene povezanosti rieči temeljni je problem u obradi prirodnog jezika koji ima mnoge koristne primjene, poput zaključivanja u orječju, razgraničenja višeznačnosti, ponalazbe obaviesti itd.

Sličnost, tj. povezanost mjerimo nad skupom 5.1.3 kosinusnom sličnošću između rječnih vektora. U slučaju da neka rieč nije u rječniku sličnost je 0.

Za uzporedbu posljedaka rabljeni su Pearsonov i Spearmanov koeficijent korelacije između izračunate sličnosti i ručnih ocjena.

Posljedci pokusa prikazani su u križnici 5.3. Oznake 12 i 6 upućuju na ocjene 12-ero i 6-ero ocjenjivača kako je opisano u 5.1.3.

Modeli su isti kao u predhodnom pokusu.

Križnica 5.3: Ocjena značbene povezanosti

Model	Pearson-12	Spearman-12	Pearson-6	Spearman-6
LSA500D	0,468	0,240	0,438	0,225
skip_100_5	0,666	0,573	0,670	0,575
skip_200_5	0,671	0,569	0,665	0,600
skip_200_10	**0,677**	0,590	**0,677**	0,591
skip_500_5	0,666	0,595	0,673	0,573
skip_1000_10	0,651	**0,619**	0,649	**0,623**
cbow_100_5	0,522	0,429	0,533	0,438
cbow_100_10	0,492	0,421	0,501	0,432
cbow_200_5	0,565	0,455	0,570	0,468
cbow_200_10	0,532	0,440	0,537	0,453
cbow_500_5	0,574	0,493	0,576	0,504
cbow_1000_5	0,561	0,479	0,560	0,490
cbow_1000_10	0,467	0,338	0,466	0,351

5.3.4. Skladnjane i značbene nalike

Po uzoru na (Mikolov et al., 2013a) kakvoću vektora rječnih značajki izpitujemo nad zadatcima skladnjanih i značbenih nalika (engl. syntactic and semantic analogies).

Rječni predstavci mogu uhvatiti mnoge međurječne sličnosti kao što je naprimjer to da se rieč *velik* odnosi prema *veći* isto kao *malen* prema *manji*. Ali također i vrlo suptilne značbene odnose, kao što je to odnos grada i države u kojoj se nalazi, npr. Zagreb je za Hrvatsku, što je Berlin za Njemačku. Takvi značbeno bogati vektori mogu se uporabiti za poboljšanje mnogih postojećih primjena obrade jezika, poput strojnog prevođenja, ponalazbe obaviesti, odgovaranja na pitanja itd.

Takve odnose uobličujemo u pitanja, npr. “Koja se rieč odnosi prema *malen* kao što se *veći* odnosi prema *velik*?”

Mikolov je pokazao da se na ta pitanja može odgovoriti jednostavnim slovnoračunskim djelatbama (engl. algebraic operations) nad vektorskim predstavcima rieči. Da bismo odgovorili na pitanje koja se rieč odnosi prema *malen* kao što *veći* odnosi prema *velik*, dovoljno je da izračunamo vektor $X = vektor(\text{“}veći\text{”}) - vektor(\text{“}velik\text{”}) + vektor(\text{“}malen\text{”})$. Potom tražimo u vektorskom prostoru rieč koja je najbliža, tj. najsličnija X-u mjereno kosinusnom sličnošću i proglasimo ju odgovorom na pitanje (zanemarujemo ulazne rieči).

Skladnjane nalike izpitujemo nad poredbenicima pridjeva: *velik* se prema *veći* odnosi kao *malen* prema čemu? Rabi se skup 5.1.4.

Značbene nalike izpitujemo nad državama i glavnim gradovima. Rabi se skup podataka 5.1.5.

Posljedci pokusa prikazani su križnicom 5.4. Mjera je u oba slučaja točnost.

Modeli su kao u prehodnim pokusima. Za LSA nismo računali poredbenike jer su ti vektori naučeni za natuknice (leme) pa ne postoje vektori za poredbenike.

Križnica 5.4: Skladnjane i značbene nalike

dodel	Poredbenici	Glavni gradovi
LSA500D	–	6,72
skip_100_5	36,6	13,83
skip_200_5	47,1	18,38
skip_200_10	**48,3**	28,46
skip_500_5	42,0	23,72
skip_1000_10	34,0	**33,79**
cbow_100_5	30,3	8,30
cbow_100_10	24,6	8,70
cbow_200_5	31,4	7,91
cbow_200_10	28,9	8,70
cbow_500_5	31,1	10,08
cbow_1000_5	23,4	10,87
cbow_1000_10	0	0

5.3.5. Prepoznavanje imenovanih sućaka

Prepoznavanje i razredba imenovanih sućaka (engl. named entity recognition and classification, NERC) zadatak je u obradi prirodnog jezika i izlučivanju obaviesti koji nastoji izlučiti i razrediti sva imena, vremenske i brojevne izraze u prirodnojezičnom orječima.

Imenovane se sućke obično razređuje u imena ljudi, organizacija, lokacija, pa u vremenske izraze, nadnevke, novčane izraze itd.

U ovome pokusu uvršćujemo rječne vektore u sustav CroNER tako da je svaka protega vektora trenutačne rieči dodatna rječna značajka (vidi i odjeljak 4). Posljedci su prikazani u križnici 5.5. Za objašnjenje mjera vidi (Karan et al., 2013), a ovdje recimo da za svaku vriedi da je više bolje (P – preciznost/natankost, R – odziv).

Križnica 5.5: Posljedci prepoznavanja imenovanih sućaka – MUC

	Osoba			Lokacija			Organizacija			Narodnost			Nadnevak			Vremenski izraz			Valuta			Postotak		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
bez vektora	**91,08**	93,33	92,17	83,99	**83,08**	**83,48**	83,66	**73,42**	**78,19**	98,91	**85,30**	**91,6**	92,03	66,66	77,31	**81,35**	**80,33**	**80,67**	**100,00**	**50,66**	**66,97**	**100,00**	**93,54**	**96,64**
cbow_100_5	90,97	93,64	92,27	84,09	82,39	83,20	85,96	69,58	76,89	**100,00**	62,33	76,79	**93,83**	**72,51**	81,79	**81,35**	**80,33**	**80,67**	**100,00**	50,56	66,84	99,70	91,92	95,59
cbow_100_10	90,97	**93,71**	92,31	**84,19**	82,39	83,25	85,83	69,58	76,84	**100,00**	62,47	76,90	93,80	72,13	81,53	**81,35**	**80,33**	**80,67**	**100,00**	50,56	66,84	99,70	91,92	95,59
cbow_200_5	90,43	93,26	91,81	83,27	80,31	81,72	**86,36**	66,24	74,96	**100,00**	46,22	63,09	93,82	72,39	**81,70**	80,47	**80,33**	80,20	**100,00**	49,11	65,64	99,18	91,13	94,90
cbow_200_10	90,43	93,26	91,81	83,48	80,31	81,82	86,29	66,26	74,95	**100,00**	47,29	64,12	93,82	72,39	**81,70**	80,47	**80,33**	80,20	**100,00**	49,11	65,64	99,18	91,13	94,93
skip_100_5	90,96	93,61	**92,25**	84,06	82,50	83,24	86,07	69,59	76,95	**100,00**	62,47	76,9	93,79	72,00	81,45	**81,35**	**80,33**	**80,67**	**100,00**	50,56	66,84	99,70	91,92	95,59
skip_200_5	90,41	93,26	91,80	83,33	80,31	81,75	86,23	66,25	74,92	**100,00**	46,22	63,09	93,82	72,39	**81,70**	80,47	**80,33**	80,20	**100,00**	48,07	64,66	99,18	91,13	94,93
skip_200_10	90,41	93,26	91,80	83,48	80,31	81,82	86,21	66,13	74,83	**100,00**	47,29	64,12	93,82	72,39	**81,70**	80,47	**80,33**	80,20	**100,00**	49,11	65,64	99,18	91,13	94,93

Križnica 5.6: Posljedci prepoznavanja imenovanih sućaka – egzaktno

	Osoba			Lokacija			Organizacija			Narodnost			Nadnevak			Vremenski izraz			Valuta			Postotak		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
bez vektora	89,10	91,29	90,16	81,49	**80,58**	80,99	72,17	63,34	67,45	98,91	**85,30**	**91,6**	84,08	60,88	70,61	62,54	61,8	62,04	94,60	47,87	63,3	91,53	85,56	88,42
cbow_100_5	88,52	91,11	89,78	81,75	80,06	80,87	75,66	61,23	67,68	**100,00**	62,33	76,79	84,94	65,63	74,04	60,05	59,46	59,63	94,84	47,97	63,4	94,27	86,71	90,27
cbow_100_10	88,52	91,18	89,82	81,84	80,06	80,92	75,56	61,24	67,64	**100,00**	62,47	76,90	85,03	65,37	73,90	60,05	59,46	59,63	94,84	47,97	63,4	94,27	86,71	90,27
cbow_200_5	87,81	90,54	89,14	80,45	77,56	78,94	75,79	58,12	65,78	**100,00**	46,22	63,09	84,17	64,93	73,29	59,38	59,46	59,28	94,71	46,52	62,17	93,72	85,93	89,60
cbow_200_10	87,81	90,54	89,14	80,65	77,56	79,04	75,59	58,05	65,66	**100,00**	47,29	64,12	84,17	64,93	73,29	59,38	59,46	59,28	94,71	46,52	62,17	93,72	85,93	89,60
skip_100_5	88,51	91,08	89,77	81,62	80,08	80,82	75,76	61,24	67,72	**100,00**	62,47	76,90	85,00	65,25	73,81	60,05	59,46	59,63	94,84	47,97	63,40	94,27	86,71	90,27
skip_200_5	**90,41**	**93,26**	**91,80**	**83,33**	80,31	**81,75**	**86,23**	**66,25**	**74,92**	100,00	46,22	63,09	**93,82**	**72,39**	**81,70**	**80,47**	**80,33**	**80,20**	**100,00**	**48,07**	**64,66**	**99,18**	**91,13**	**94,93**
skip_200_10	87,79	90,54	89,13	80,65	77,56	79,04	75,68	58,05	65,69	**100,00**	47,29	64,12	84,17	64,93	73,29	59,38	59,46	59,28	94,71	46,52	62,17	93,72	85,93	89,60

5.4. Razprava posljedaka

5.4.1. Zamršenost jezičnog modela

Zbog visoke vremenske složenost težko je raditi s PMJM-ovima nad velikim skupovima podataka. Malen broj modela onemogućava tumačenje učinka pojedinih parametara modela, ali možemo vidjeti da je model RNNLM-2 s vektorima duljine 50 (što je jako malo, u englezkom se kreće od 300 do 1000 protega) naučen nad 10M rečenica nadmašio dvorječni model naučen nad 40M rečenica. Pogledamo li RNNLM-1, model s prilično velikim skrivenim slojem (500) i s velikim brojem razreda (500) uvještban na svega 300k rečenica, vidimo da je podbacio, vjerojatno zbog premale količine podataka za učenje. To je i bilo očekivano, što model ima više parametara to je više podataka za učenje potrebno.

5.4.2. Prepoznavanje suznačnica

Iz križnice 5.2 odmah uočavamo da naučeni vektori uvelike povećavaju točnost prepoznavanja suznačnica. Poboljšanje je najveće za glagole 12,6%, sliede imenice s 6,8% te pridjevi s poboljšanjem od 6,7%.

Glede na vrstu rieči, imenice imaju najveću točnost, sliede glagoli (!) te na kraju pridjevi. Kod vektora naučenih u ovome radu za svaki (jednorječni) oblik rieč postoji zaseban vektor, pa tako i jedan za natuknicu (lemu). A u ovome pokusu radimo s lemama. Moguće je da pridjevi postižu nižu točnost jer se pojavljuju u množtvo oblika, pa onda i rjeđe u natukničnom obliku. Za budući rad bi bilo zanimljivo za svaku rieč pronaći vektore za njezine oblike te ih uprosječiti s nadom da će se time dobiti bolji predstavak te rieči.

Vidimo da je skip-gramski model bolji od modela CBOW. To je u skladu s rezultatima za englezki gdje skip-gramski model pokazuje znatno veću točnost na značbenim zadatcima (Mikolov et al., 2013a). U englezkom se pokazuje da skip-gram bolje modelira rjeđe rieči, a hrvatske rieči zbog množtva oblika možemo smatrati riedkima (hrvatska se imenica može pojavljivati u 10-ak oblika nasprem englezke koja u najviše tri).

Glede na veličinu vektora, možemo opaziti da je više bolje. Ako pak pogledamo veličinu prozora, ne može se donieti zaključak, osim što imamo neobičnu situaciju s 1000-protežnim CBOW-vektorima kod kojih je s povećanjem prozora s 5 na 10 točnost izrazito opala čak i preko 40%. Težko je reći što je uzrok tomu.

Bacimo li oko na MRR vriednosti, vidimo da svi naši modeli (osim cbow_1000_10) nadmašuju LSA-vektore.

5.4.3. Ocjena značbene povezanosti i sličnosti

Možemo vidjeti veliko poboljšanje u odnosu na LSA-vektore, posebice u Spearmanovu koeficijentu koji mjeri koliko se odnos dviju promjenljivica može opisati monotonom funkcijom.

Čini se da su bolji vektori naučeni s manjim prozorima. S povećanjem broja protega rezultati su bolji do neke točke. Za veće vektore možda dolazi do izražaja premalog broja podataka u odnosu na broj parametara modela.

5.4.4. Skladnjane i značbene nalike

Naše rezultate ne možemo uzporediti s LSA-vektorima koje smo rabili jer su oni naučeni za leme, a mi radimo s poredbenicima pridjeva. Ali možemo reći da je pronalaženje izpravne rieči koja je poredbeni stupanj nekog pridjeva u rječniku od 3M rieči s točnošću od 48,3% zadovoljavajući rezultat.

Skip-gramski se modeli opet pokazuju boljima. Najbolja se točnost od 48,3% postiže za model s 200 protega i prozorom veličine 10. S nižim i većim brojem protega točnost opada, možda zbog premalih vektora, odnosno premalog skupa za učenje. Najgori se rezultat postiže za CBOW-model s 1000 protega i prozorom veličine 10, točnost je 0, za svaki primjer model vrati pogrešnu rieč, što je možda posljedica podnaučenost.

Što se tiče glavnih gradova skip-gramski su modeli uvjerljivo bolji od CBOW-a. Čini se da širi prozor doprinosi točnosti. Najbolji se rezultat od 33,79% postiže za skip-gramski model s 1000 protega i prozorom širine 10.

5.4.5. Prepoznavanje imenovanih sućaka

Zbog visokih memorijskih zahtjeva nije bilo moguće prevesti pokuse za vektore većih protega ili uvrstiti i vektore za okolne rieči. Ovi su pokusi provođeni nad manjim skupom podataka nego pokusi u (Karan et al., 2013) pa zato rezultati nisu uzporedivi, odnosno identični za slučaj bez vektora.

Gledajući MUC vidimo da vektori uglavnom odmažu, jedino je zamjetno poboljšanje u F1-mjeri kod nadnevaka (81,70% nasprem 77,31%). Nema neke uočljive razlike između CBOW-a i skip-grama.

Kod egzaktne ocjene, dobiveno je malo poboljšanje sa skip-gramskim modelom s 200 protega i prozorom veličine 5. Najviše za vremenske izraze (18%), a do gubitka točnosti je došlo kod narodnosti (-28.51%). Težko je utvrditi zašto dolazi do toga.

Moguće je da bi dodavanjem rječnih predstavaka za susjedne rieči dobili bolje rezultate.

6. Zaključak

Jezični modeli svakom sliedu rieči pridružuju vjerojatnost pripadanja u neki jezik i jedan su od osnovnih alata u obradi prirodnog jezika. Najviše se rabe u strojnome prevođenju, samodjelnom prepoznavanju govora, prepoznavanju pismena i provjeri pravopisa. Tradicionalni se modeli uče na temelju statistike o pojavljivanju sliedova od n rieči (n-rječja) u građi. No taj pristup pati od nekoliko problema: ograničen je na n-ove do najviše 5, što zbog goleme količine potrebnih podataka za kakvotnu procjenu modela, što zbog same računske složenosti, zbog toga ne mogu ni dobro modelirati odnose između udaljenih rieči. Dodatno, nemaju koncept sličnosti između rieči ili sliedova rieči, tj. ne mogu učiti što rieč znači, tj. kako se upotrebljava pa ne mogu poobćavati na neviđene kombinacije rieči analogne već viđenima, npr. model je možda vidio *mačka leži u sobi*, no nije vidio *pas leži u sobi* i premda su *mačka* i *pas* slične rieči n-rječni model ne može prihvatiti tu rečenicu. Ti se problemi djelomično rješavaju zaglađivanjem.

Kao alternativa n-rječnim modelima nametnuli su se modeli temeljeni na umjetnim živčanim mrežama. Ti modeli rješavaju većinu navedenih problema učenjem vektorskih prikaza rieči koji sadrže množtvo značbenih i inih obaviesti o samoj rieči. Živčanomrežni jezični modeli uče istodobno vektorske prikaze i vjerojatnosti sliedova rieči prikazanih tim vektorima. Takvi modeli mogu naučiti da su *pas* i *mačka* slične rieči (imat će blizke vektore) te će i rečenice sa *psom* moći prepoznati kao dobre. Usredotočujući se samo na učenje vektorski prikaza živčanim mrežama nedavno su razviena dva modela koja iznimno brzo uče visokokakvotne vektorske prikaze rieči koji obiluju mnogim zanimljivim svojstvima. U ovom smo se radu usredotičili na te modele, tj. na učenje vektorskih prikaza rieči. Dobivene vektore primienili smo u hrvatskom jeziku na zadatke prepoznavanja suznačnica, na ocjenjivanje značbene povezanosti i sličnosti rieči, potom na zadatke traženja poredbenika pridjeva i povezivanja država i glavnih gradova analogijskim zaključivanjem te na prepoznavanje imenovanih sućaka. U svim je tim zadatcima (osim prepoznavanja sućaka) došlo do značajnog poboljšanja točnosti primjenom vektorskih prikaza.

N-rječni modeli postaju stvar prošlosti, živčanomrežni modeli daleko su kakvotniji i robustniji. Vektorski prikazi naučeni živčanim mrežama nose iznimno mnogo podataka o nekoj rieči, o nekom konceptu, pravo je pitanje koje sve podatke nose, kako protumačiti

te brojeve kao značenje, kao padeže, ... i kako ih najbolje primieniti. Kako kombinirati vektore rieči da se dobiju vektori rečenica i cielih dokumenata? Jesu li pojmovi u našem mozgu pohranjeni na sličan način preko jakosti međuneuronskih veza? Nema sumnje da je pred vektorskim prikazima rieči svietla budućnost u mnogim područjima obrade prirodnog jezika, posebice strojnome prevođenju. Ali isto tako i pred živčanim mrežama općenito.

UPOTREBLJENA GRAĐA

Yoshua Bengio. Neural net language models. Scholarpedia, 3(1):3881, 2008. URL http://www.scholarpedia.org/article/Neural_net_language_models.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, i Christian Jauvin. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137–1155, 2003.

Stanley F Chen i Joshua Goodman. An empirical study of smoothing techniques for language modeling. U Proceedings of the 34th annual meeting on Association for Computational Linguistics, stranice 310–318. Association for Computational Linguistics, 1996.

Yoav Goldberg i Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722, 2014.

Joshua Goodman. Classes for fast maximum entropy training. U Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on, svezak 1, stranice 561–564. IEEE, 2001a.

Joshua T Goodman. A bit of progress in language modeling. Computer Speech & Language, 15(4):403–434, 2001b.

Martin T. Hagan, Howard B. Demuth, i Mark Beale. Neural Network Design. PWS Publishing Co., Boston, MA, USA, 1996. ISBN 0-534-94332-2.

Vedrana Janković, Jan Šnajder, i Bojana Dalbelo Bašić. Random indexing distributional semantic models for croatian language. U Text, Speech and Dialogue, stranice 411–418. Springer, 2011.

Mladen Karan, Jan Šnajder, i Bojana Dalbelo Bašić. Distributional semantics approach to detecting synonyms in croatian language. Information Society, stranice 111–116, 2012.

Mladen Karan, Goran Glavaš, Frane Šarić, Jan Šnajder, Jure Mijić, Artur Šilić, i Bojana Dalbelo Bašić. Croner: Recognizing named entities in croatian using conditional random fields. Informatica (Slovenia), 37(2):165–172, 2013.

Bulcsú László. Neka pitanja strojnoga razumijevanja prirodnoga jezika. U Slavko Tkalac i Miroslav Tuđman, urednici, Informacijske znanosti i znanje. Zavod za informacijske studije, Zagreb, 1990.

Bulcsú László i Damir Boras. Tuđinština u jeziku hrvatskome. Studia lexicographica, (1): 27–52, 2007.

Tomáš Mikolov. Statistical language models based on neural networks. Doktorska disertacija, Ph. D. thesis, Brno University of Technology, 2012.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, i Sanjeev Khudanpur. Recurrent neural network based language model. U INTERSPEECH, stranice 1045–1048, 2010.

Tomáš Mikolov, Stefan Kombrink, Lukas Burget, JH Cernocky, i Sanjeev Khudanpur. Extensions of recurrent neural network language model. U Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, stranice 5528–5531. IEEE, 2011.

Tomáš Mikolov, Kai Chen, Greg Corrado, i Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013a.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, i Jeff Dean. Distributed representations of words and phrases and their compositionality. U C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, i K.Q. Weinberger, urednici, Advances in Neural Information Processing Systems 26, stranice 3111–3119. Curran Associates, Inc., 2013b.

Tomáč Mikolov i Geoffrey Zweig. Context dependent recurrent neural network language model. U SLT, stranice 234–239, 2012.

Tomáč Mikolov, Quoc V Le, i Ilya Sutskever. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168, 2013c.

Frederic Morin i Yoshua Bengio. Hierarchical probabilistic neural network language model. U AISTATS, svezak 5, stranice 246–252. Citeseer, 2005.

Nikolaos Pappas i Thomas Meyer. A survey on language modeling using neural networks. Technical report, Idiap, 2012.

David E Rumelhart, Geoffrey E Hinton, i Ronald J Williams. Learning representations by back-propagating errors. Cognitive modeling, 1988.

Jan Šnajder, Sebastian Padó, i Željko Agić. Building and evaluating a distributional memory for croatian. U 51st Annual Meeting of the Association for Computational Linguistics, stranice 784–789, 2013.

Joseph Turian, Lev Ratinov, i Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. U Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, stranice 384–394. Association for Computational Linguistics, 2010.

Wei Xu i Alexander I. Rudnicky. Can artificial neural networks learn language models? 2000.

Dodatak A
Osnovni pojmovi

U ovom se poglavlju definiraju osnovni pojmovi koji se rabe u radu.

JEZIK je sustav znakova koji služi priobćavanju. SUSTAV je cjelokupnost jedinicâ i odnosâ među njima. ZNAK je predmet koji stoji umjesto kojega drugoga predmeta. PRIOBĆAVANJE je slanje viesti od pošiljatelja k primatelju. Viest se ostvaruje u vremenu nizanjem jezičnih znakova. Jezični je znak izraz (oblik) kojemu je pridružen sadržaj (značenje). Jezik može biti prirodan (naravan) ili umjetan. Umjetni je jezik za razliku od prirodnoga onaj znakovni sustav koji nastaje tako da mu opis predhodi porabi. Prirodni su jezici glavno sredstvo ljudske[1] suobćitbe (László, 1990).

U području *obrade prirodnog jezika* primjenjujemo stroj, datkovni rednik nad (stvarnim) jezičnim (po)datcima.

JEZIČNA IZVORIŠTA su strojnočitljivi skupovi jezičnih dataka koji se rabe za izgradnju, poboljšavanje i vrednovanje sustava koji rade s prirodnim jezikom. To su obično zbirke orječne građe, tzv. građare, i rječnici.[2]

GRAĐARA (KORPUS) je zbirka kusova[3] jezika koji su odabrani i uređeni prema izričitim jezikoslovnim sudilima s namjenom da služe kao uzorak jezika.[4] Obično se (smatra da se) radi o zbirci orječjâ pa možemo govoriti o orječnim građarama ili kraće ORJEČNICAMA (litavski *tekstynas*, esperantski *tekstaro*).

Jeftin način prikupa građe jest pobiranje po spletu. Svesvjetski splet (WWW) je neizcrpan izvor vjerodostojnih prirodnojezičnih dataka za iztraživače na području jezikoslovlja, obrade prirodnoga jezika, umjetne inteligencije i mnogih drugih. Taj se pristup naziva

[1] Vjerojatno 'ljudske' ovdje nije najbolja rieč jer time odmah izključujemo možebitne visokointeligentne neljude (izvanzemaljce) koji suobćavaju znakovnim sustavima koji podpadaju pod definiciju jezika.

[2] Definicija pojednostavljena s `http://www.elra.info/Definition.html`.

[3] Rabi se neobvezujuća rieč 'kus' (piece), a ne 'orječje' jer je to pitanje tehnike uzorkovanja, ako su na primjer svi uzorci jednake veličine tada ne mogu svi biti orječja u strogom smislu, već fragmenti orječja proizvoljno oddvojeni od sadržaja (EAGLES).

[4] Ovo je popularna definicija EAGLES-a (Expert Advisory Group on Language Engineering Standards, Stručna savjetodavna skupina za mjerila/uzore jezičnog mjerništva/ustrojništva/sastrojništva) `http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html`

“splet kao građara” (engl. web as corpus), a njime se dobiva ‘spletna građara’.

Neobrađeno je orječje u redniku prikazano kao dugačka nizanica pismenâ (engl. character string). Takav prikaz nije pogodan za većinu postupnika strojnoga učenja, stoga orječje podvrgavamo predobradi. Predobrada uključuje nekoliko koraka odabir kojih ovisi o našim potrebama, a najčešći su: opojavničenje, razdioba na rečenice, lematizacija itd.

OPOJAVNIČENJE (engl. tokenization) orječja možemo jednostavno definirati kao postupak kojim se orječje iz jedne dugačke nizanice pretvara u spisak pojavnicâ.

POJAVNICA (engl. token) je nizanica koja odgovara onome što shvaćamo pod pojmom ‘rieč’. Pojavnicom obično smatramo sve što se nalazi između dvaju pismena koja služe kao graničnici (najčešće su to bjeline/razmaci).[5] Dakle, pojavnica je svaka pojedina ‘rieč’ koja se pojavljuje u orječju, to je ono što mislimo kad kažemo da neko orječje ima 5000 rieči.

RAZLIČNICE (engl. types) su različne rieči (engl. distinct words), rieči koje se razlikuju jedna od druge; možemo ih shvatiti kao skupove istovjetnih pojavnicâ. To je ono što mislimo kad kažemo da je beba izrekla dvie rieči, želimo reći da je izrekla dvie različite rieči, a ne dvaput istu.

Pogledajmo na primjeru kako rabimo te pojmove, kažemo da rečenica `"žena vidi mnogo žena"` sadrži četiri pojavnice, a tri različnice (dvaput imamo nizanicu `"žena"`).[6]

Dodatni korak koji možemo provesti je lematizacija.

LEMATIZACIJA (ONÁTUKA, UZPRAVA) (< onatučiti < o-natučica-iti) jest postupak svođenja rieči (pojavnice) na njezinu natučicu. NÀTUČICA (natuknica, lema) osnovni je izraz rieči kako se pojavljuje u rječnicima. Onatučimo li, primjerice, pojavnicu `"pjevaj"` dobivamo `"pjevati"`, tj. neodređeni oblik toga glagola, za imenice bismo dobili nazovnik jednine itd.

Nakon što smo od nizanice pismena dobili spisak rieči u podobnu obliku možemo pristupiti uznaci riečî, tj. odabiru načina na koji će rieč biti predstavljena iliti prikazana. O tome vidi u radu odječak 2.5.

[5] Ponekad poteškoću mogu predstavljati višerječne jedinice koje ponekad želimo shvatiti kao jednu pojavnicu, npr. željeli bismo da se `"Novi Zagreb"` ne opojavniči na `["Novi", "Zagreb"]` kao da se radi o `"novom kaputu"`, nego na `["Novi Zagreb"]`.

[6] Razlika između *type* i *token* mnogo je obćenitija, radi se, ugrubo rečeno, o bitoslovnoj (ontološkoj) razlici između obće vrste, koncepta neke stvari i njezinih pojedinačnih zbiljnih primjeraka, oprimjerenjâ. Ako želimo prenieti nazive na ovo obćenitije viđenje, možemo reći da su *različnice* ‘međusobno različne vrste stvarî’, a *pojavnice* ‘pojavne oprimjerbe’ (engl. instantiations) tih mislenih koncepata.

Podrobno izlaganje o razlikovanju type-token zaniman (engl. interested) čitatelj može pronaći primjerice na: `http://plato.stanford.edu/entries/types-tokens/`

Dodatak B
O jeziku rada

U radu susrećemo neke neobične rieči (značbeni, osvještaj, ...), a i neke poznate zapisane su na neobičan način (rieč, jednačba, ...) – o čemu se tu radi?

Pitanje sloga (stila) je pitanje izbora. U ovom je radu izabrano prevoditi strane rieči i pisati sustavom boljim od onog nametnutog. Ovakav se stil ~~obično~~ uviek dovodi u vezu s piščevim svjetonazorom, smatra se odrazom nacionalizma (narodničtva), nazadnjačtva, tuđomrzstva, stranobojazni itd. Nažalost, obično i je više ili manje tako. No to što je Hitler gradio autoceste ne znači da su svi koji grade autoceste nacisti. Tako je i s jezikom.

Ovim bih se poglavljem htio u podpunosti distancirati od takvog svjetonazora i pokazati da nema ničeg demonskog u stvaranju novih rieči i pisanju ovakvim sustavom.

Važna napomena: ovo je poglavlje nabrzaka napisano, pa je nepodpuno ili loše organizirano i sl.

Kako bih pokazao da moji motivi nemaju veze s nacionalizmom, tuđomrzstvom itd. prvo ću iznieti kako doživljavam ovaj naš sviet.

Pogledajmo sliku B.1. Poslušajmo kako ju je doživio Carl Sagan s čijem se viđenjem slažem:[1]

> Zbog odraza Sunca s letjelice čini se da Zemlja leži na zraci svjetlosti, kao da je ovaj maleni sviet nekako posebno važan, ali to je samo igra geometrije i optike. Na slici nema traga ljudima, nema naše prerade Zemljine površine, nema naših strojeva, nema nas. S tog udaljenog motrišta, nema traga našoj obsjednutost nacionalizmom. U mjerilu svjetova, ljudi su nevažni, tek tanka prevlaka života na zabitnoj i samotnoj grudi stienja i metala.
>
> S tog udaljenog motrišta, Zemlja bi mogla izgledati ne posebice zanimljiva. No za nas je drugačije. Razmotrite još jednom tu pjegu. To je ovdje. To je naš dom. To smo mi. Svi koje volite, svi koje poznajete, svi za koje ste ikada čuli, svako ljudsko biće koje je ikada postojalo proživjeli su svoje živote na njoj.

[1] Izvorno u knjizi *Pale Blue Dot*. Na Youtubeu su dostupni brojni videouradci s ovim orječjem u izvedbi Carla Sagana, a ovdje donosimo vlastiti prievod jednog od njih.

Sveukupnost naše radosti i patnje, tisuće samouvjerenih religija, ideologija i gospodarskih učenja, svaki lovac i sakupljač hrane, svaki junak i kukavica, svaki tvorac i razaratelj civilizacije, svaki kralj i kmet, svaki zaljubljeni par, svaka majka i otac, diete puno nade, izumitelj i iztraživač, svaki učitelj morala, svaki korumpirani političar, svaka "superzviezda", svaki "vrhovni vođa", svaki svetac i grešnik u poviesti naše vrste živješe ovdje – na trunki prašine ovješenoj na zraci Sunca.

Zemlja je vrlo mala pozornica u ogromnoj svemirskoj areni. Sjetite se rieka krvi koje su prolili svi ti generali i carevi da bi mogli u slavi i trijumfu postati trenutni gospodari djelića ove točke. Sjetite se bezkrajnih okrutnosti stanovnika jednog kutka ove piknje prema stanovnicima nekog drugog kutka koji se od njih jedva razlikuju. Kako li su česti njihovi nesporazumi, kako li su revni pobiti se, kako li su gorljive njihove mržnje.

Našu umišljenost, našu zamišljenu važnost, naše zavaravanje da zauzimamo neki povlašteni položaj u svemiru osporava ova točka bliede svjetlosti. Naš je planet usamljena mrlja u velikom svemirskom mraku koji je okružuje. Nema nikakve naznake da će u našoj zabitosti, u svom tom prostranstvu odnekuda doći pomoć koja bi nas spasila od nas samih.

Zemlja je jedini sviet za koji se zasad zna da gaji život. Nema drugog mjesta, barem ne u blizkoj budućnosti, na koje bi se naša vrsta mogla preseliti. Posjetiti ga, da. Nastaniti, još ne. Svidjelo se to nama ili ne, trenutačno je Zemlja mjesto gdje igramo našu predstavu. Kažu da je astronomija izkustvo koje uči skromnosti i izgrađuje karakter. Vjerojatno ne postoji bolji pokaz blesavosti ljudske taštine nego što je ta udaljena slika našeg malešnog svieta.

Za mene ona naglašava našu odgovornost da budemo prijazniji jedni prema drugima i da sačuvamo i njegujemo bliedu modru točku, jedini dom koji smo ikada poznavali.

Slika B.1: Blieda modra točka (Pale Blue Dot) – svjetlopis obhodnice Zemlje snimljen 1990. sa svemirske iztraživalice *Voyager 1* (putovatelj) na udaljenosti od oko 6 milijardi kilometara od Zemlje. Sva ljudska poviest odigrala se na ovoj sitnoj piknjici (ovdje prikazanoj u plavom krugu) koja je naš jedini dom.

Sada kad samo razkrstili s nacionalizmom i tuđomrzstvom, pogledajmo prave razloge.

B.1. Zagovor domaćica

Pogledajmo za početak tablicu (križnicu, skrižaljku) B.1. U sadašnjem jeziku rieči iz prvoga stupca služe kao pridjevi rieči u drugome stupcu. Ono što je očito i nekome tko ne zna hrvatski jest to da ne postoji neka vidljiva veza između tih riječi; one nisu izvedene jedne iz drugih. Ali, uzporedimo li prvi i četvrti stupac[2] , uvidjet ćemo da između tih rieči postoji neka veza; one jesu nastale jedne od drugih.

Križnica B.1: Rieči i njihovi pridjevi

"hrvatski"	osnovna rieč	hrvatski	"osnovna rieč'
solarni	Sunce	sunčani, Sunčev	Sol
lunarni	Mjesec	mjesečni, Mjesečev	Luna
digitalni	znam(en)ka; prst	znam(en)čani; prstni	digit(us)
nuklearni	jezgra	jezgreni	nucleus
auditivni, audio-	sluh	slušni	auditus, audire
vizualni	vid	vidni	visus
audiovizualni	sluh, vid	slušno-vidni	auditus, visus
komercijalni	trgovina	trgovački, trgovinski	commercium
kromatski	boja	bojni	khroma
heterokroman	razni, boja	raznobojan	heteros, khroma
linearni	crta, pravac, linija	crtovni	linea (>linija)
modalni	način	načinovni, načinski	modus
binarni	dva	dvojčani, dvojni, ...	bini, bis
seksagezimalni	šestdeset	šestdesetični, šestdesetni	sexaginta
somatski	tielo	tjelesni	soma
kardiovaskularni	srce, žila	srčano-žilni	kardia, vasculum
binauralni	dva, uho	dvoušni	bini, auris
binokularni	dva, oko	dvoočni	bini, oculus

Problem je u tome što nama u hrvatskom četvrti stupac nije poznat, mi ne kažemo *sol* nego kažem *sunce*, ne kažemo *digit(us)* nego *znamenka* itd. Rieči u četvrtom stupcu su rieči drugih jezika: latinskoga (englezkoga) i grčkoga i one su normalne svakodnevne

[2]Rieči u prvome stupcu pisane su onako kako se pojavljuju u hrvatskom, a rieči u četvrtome stupcu su izvorne latinske ili grčke rieči. Grčki je ulatiničen po načelu slovo po slovo, nije se pazilo na standardna pravila prieslova ili priepisa pa je moguće da su neke grčke rieči pogrešno ulatiničene.

rieči za pojmove koje mi označujemo riečima iz drugoga stupca. A rieči u prvome stupcu normalni su svakodnevni pridjevi od tih rieči (ovdje pohrvaćeni), isto kao što mi kažemo *svinjski* kada se nešto odnosi na *svinju*, tako bi i na latinskom rekli *lunaris* za nešto što se odnosi na *Lunu* (Mjesec).

U trećem su stupcu navedeni pridjevi rieči drugoga stupca načinjeni unutarsustavno, u hrvatskom. Veza drugoga i trećega stupcu vidljiva je kako na izraznoj razini tako i na značbenoj, što je mnogo važnije. Upravo je u tome stvar, pridjevi iz prvoga stupca nisu na očigledan način povezani s riečima drugoga stupca (kao što su povezani s riečima četvrtoga) i tu smo na gubitku, mi ne osjećamo vezu tih rieči kao što ju osjećamo između drugoga i trećega stupca. Mi ne osjećamo da *nuklearni* ima veze s *jezgrom*, ne osjećamo da *modalni* ima veze s *načinom* itd. Smatram da je podpuno glupo rabiti rieči iz prvoga stupca umjesto rieči trećega stupca jer smo na gubitku, rieči koje trebaju biti povezane (kao što i jesu u izvornom jeziku) u hrvatskome nisu povezane na našu štetu. **Ovo nema nikave veze s tuđomrzstvom nego sa zdravim razumom, posve je normalno i koristno tvoriti pridjeve unutar sustava, oni su samorazumljivi i povezani na očit način sa svojim imenicama, kao što je to i slučaj u izvornom jeziku iz kojeg smo ih pokrali.**

Ovdje nema mjesta ni sporu ni daljnjem zboru o tome da su hrvatske rieči bolje!

Valja još nešto reći. Rieč *boja* nije baš hrvatska, dolazi iz turskog, a povezana je sa značenjem ‘ukras’, neke od hrvatskih rieči za boju bile su *mast* (uzp. *mastnica* ‘obojeno mjesto na koži’, i *premazan svim mastima*) i *šara*. Problem je s posuđenicama što često povuku sa sobom i svoju rodbinu, pa je tako *boja*, predpostavljam, povukla sa sobom i *bojadisati* (bojiti), rieč koja nije nastala unutar sustava za razliku od *bojiti*, pa je onda teže razumljiva i unosi anomaliju u sustav. Dakle, **ako već uzimamo tuđe rieči, nemojmo onda skupa s njima preuzeti i njihove izvedenice, nego napravimo izvedenice u svom jeziku, tj. postojećim tvorbenim načinima.** Dakle, uzeli smo *boju* i onda kažimo *bojiti*, a ne *bojadisati*. Recimo da nemamo rieč *jezgra* i da prevladava rieč *nukleus*, tada valja napraviti pridjev od nje unutar sustava čime dobivamo *nukleusni*, a ne preuzeti *nuklearni*. Isto tako *bakterijski*, a ne *bakterijalni*, *planetni*, ne *planetarni*, *plazmeni*, ne *plazmatski* itd.

Znače li te hrvatske rieči baš isto što i strane?

Ukratko, da. Iz nekoliko razloga, jedan je taj da u standardnom jeziku rieč znači ono što je propisano da znači; standardni je jezik umjetan, nastaje odabiranjem i propisivanjem. Ali, to nije pravi razlog, pogledajmo na par primjera druge razloge zašto su to dobre rieči.

Pogledajmo prvo *jezgreni* u značenju *nuklearni*. Netko će tvrditi da ‘nuklearni’ ne može bit ‘jezgreni’ jer je ‘nuklearni’ nešto više, nešto drugačije itd. No tu se često radi o nerazumievanju samog pojma. Ljudi ne razumiju što zapravo rieč znači. Na primjer,

meni je kad sam bio mlađi i neobrazovaniji ‘nuklearno’ bilo nešta povezano s nuklearnom bombom ili nuklearnom elektranom, i onda kad se susretneš s npr. *jakom nuklearnom silom* (osnovnom silom koja povezuje između ostalog protone i neutrone u jezgri) ostaneš malo u nedoumici kakve to sad ima veze s atomskom bombom. Problem je u tome što nisam razumio što zapravo ‘nuklearni’ znači, nisam znao definiciju toga pojma, imao sam tek nekakvo djelomično, zapravo pogrešno, shvaćenje pojma ‘nuklearni’. ‘Nuklearni’ znači ‘jegreni’, koji se odnosi na (atomsku) jezgru, pa je onda jasno da je ‘nuklearna sila’ ona koja djeluje u atomskoj jezgri, a sve druge uporabe poput ‘nuklearne elektrane’ ili ‘nuklearne medicine’ dolaze od tog osnovnog značenja koje je uviek pristuno u njima. Tako da je u redu reći *jezgreni rat* za *nuklearni rat* jer je to rat koji se vodi *jezgrenim oružjem*, a to oružje rabi *jezgrenu reakciju*, a u toj reakciji sudjeluje atomska *jezgra*. Imamo: *jezgra* > *jezgrena reakcija* > *jezgreno oružje* > *jezgreni rat* > *jezgrena zima* itd. Dakle, uviek si možemo odigrati u glavi tu evoluciju uporabe neke rieči, nema veze ako se i djelomično proširilo značenje, hrvatska rieč može odmah značiti sve što i strana rieč, dovoljno je samo shvatiti kako se uporaba ili značenje širilo. Primietimo kako nas ta rieč *nuklearni* onemogućava da shvatimo o čemu se zapravo govori, ta rieč nema očite veze s *jezgrom* i zapravo ne razumijemo što ona znači, pa joj pripisujemo svakakva značenja koja su zapravo samo odrazi stvarnog značenja.

Pogledajmo još *znamčani* u značenju *digitalni*. Prošle je godine u Srbiji održano nekakvo na(d)tjecanje u nekoj ustanovi gdje se tražio najbolji prievod za rieč *digitalizacija*.[3] Pobjedila je rieč *ubrojčavanje*. Odmah su usliedile burne reakcije, spominjalo se Ustaše, tuđomrzstvo i tako to kako to već ide i kod nas, možda i jače, no preskočimo sad te irelevantne kritike i osvrnimo se na jednu važniju koja je došla od tamošnjih jezičnih stručnjaka (i ove za ustaštvo su došle i od nekih stručnjaka), a ta je da *digitalni* ne znači *brojčani*, da je to samo jedno značenje itd. No to je zapravo bedastoća, kao i kod ‘nuklearnog’ radi se o nerazumievanju pojma, a tvrdim da tome najviše doprinose nerazumljive tuđice. Gledano tvorbeno, *digital* je u englezkom pridjev od *digit* (znamenka, brojka; prst), ako to prevedemo dobivamo *znamenčani*, *brojčani*, *prstni*. Ali ajmo prvo razumieti što znači *digitalni*. Digitalno je oprječno analognom. Radi se o tome kako se predstavljaju veličine. U digitalnim se napravama veličine predstavljaju razlučnim (diskretnim) brojevima, a u analognim napravama neprekidnim (kontinuiranim) spektrom vriednosti u obliku neke mjerljive fizičke veličine (napon, kut kazaljke, ...). U analognim su napravama veličine analogne fizikalnoj informaciji koju predstavljaju. Ovdje *analogne* znači *razmjerne*, *proporcionalne*. Na primjer sat na kazaljke, kod njega je mjera kuta zakreta male kazaljke jednaka mjeri dvostrukog pripadnog luka što ga Sunce prevali po putanji na nebu. Dakle, kut kazaljke je razmjeran, proporcionalan, analogan luku Sunca. S druge strane, kod

[3] `http://www.021.rs/Novi-Sad/Vesti/Ubrojcavanje-srpska-rec-za-digitalizaciju.html`

digitalnih naprava veličine se pretvaraju u brojeve (znamenke), takav je prikaz razlučan (diskretan), ne može poprimiti sve vriednosti nego je ograničen brojem upotrebljenih znamenaka. U digitalnim se uređajima radi s brojevima, brojevi se zapisuju u dvojčanom sustavu, a te se dvie znamenke onda mogu predstaviti kao npr. viši i niži napon. A u analognim bi uređajima recimo visina napona bila razmjerna veličini koju predstavlja.

Dodatno, *digitalni* može znači da rabi znamenke ili da ima veze s znamenkama, recimo one budilice s padajućim listićima s brojevima, koje ne moraju biti digitalne u gornjem smislu ili npr. 7-odsječni *digital displays* (znamčani prikaznici). Ali i tu podpuno odgovora *znamčani*.

Sva druga značenja *digitalnog* poput *digitalne televizije* dolaze od osnovnog, a zapravo se i ne radi o drugim značenjima. Kao i kod *nuklearnog* možemo promotriti evoluciju uporabe rieči: *znamčana televizija* je ona koja radi s *znamčanim signalima*, a to su upravo oni koji su prikazani *znamčano*. Dakle, opravdano je u svim kontekstima zamieniti *digitalni* s *znam(en)čani*, a *analogni* s *razmjerni*. Zasigurno nam čudno zvuči govoriti *razmjerni* umjesto *analogni* (u ovom značenju), ali to to znači i tako je valjda i Grcima, njima je *analogan* razumljivo kao i nama *razmjeran*. I mislim da je blesavo zazirati od razumljivosti.

Nije hrvatski jedini jezik koji prevodi, digitalan je u francuzkom *numérique*, u češkom *číslicový* (číslice = znamenka), u poljskom *cyfrowy* (cifra), u ruskom *cifrovój*, u grčkom *psifiakós* itd. To želim iztaknuti: i drugi jezici prevode strane rieči, neko više, neki manje. Poslije ćemo pogledati kako se mnogo prevodi u grčkom.

Može se dogoditi da dugom porabom nerazumljiva rieč prikupi više značenja koja su težko dovediva u vezu s polazištnim. To smo imali upravo s digitima, to su zapravo prsti, a tek kasnije je to poprimilo značenje znamenke, i kad prevodimo digitalni, onda to naravno nećemo uviek prevoditi s prstni, nego i sa znamčani jer je to ono značenje koje ta rieč ima u tom kontekstu. Dakle, imamo slučajeve u kojima možemo prevesti polazištnu rieč (nuklearni – jezgreni) i onda furati tu rieč u svim uporabama, a imamo i slučajeve gdje je polazištna rieč poprimila nova značenja koja mi izkazujemo drugim riečima (digit – prst, znamenka), onda naravno nećemo prevesti samo polazištnu rieč (prst) i gurati ju gdje joj nije mjesto. Zapravo se to svodi na to da kada se traži domaća zamjena prevođenjem strane rieči, da se onda prevodi značenje, smisao, a ne da se nuđno mora oponašati tvorba u izvornom jeziku, premda je to dobar početak.

Zaključujem da predložene hrvatske rieči uistinu mogu u većini slučajeva značiti sve što i njihovi strani odpovjednici.

Jasno, u gornjoj je tablici navedeno svega par primjera kojih sam se prvo sjetio, ali takvih je primjera koliko hoćeš.

Zašto birati domaće

Sažmimo što smo rekli kod pridjeva. Domaća je tvorenica bolja jer je povezana s polazišnom rieči, s kojom i treba biti povezana, kao što je i u izvornom jeziku. Ta se povezanost očituje na izraznoj razini i na značbenoj. Zbog toga se te rieči bolje uklapaju u jezik, odnosno u naš um. Zbog svoje unutarsustavne tvorbe od poznatih rieči, one su **samorazumljive** i lako protumačljive, ne treba ih posebno učiti. Bolje razumijemo što govorimo, tj. što mislimo.

To ne vriedi samo za pridjeve nego i za ostale vrste rieči.

Dosta je demonizacije stvaranja rieči. Stvaranje rieči je najnormalnija stvar. To je jedan oblik izražavanja. Ne znati stvarati rieči je kao da ne znaš napraviti upitnu rečenicu, zakinut si u izražavanju.

Samorazumljivost (ili kome je draže autokomprehensibilitet)

Razlikuju se dvie vrste značenja rieči:

1. tvorbeno značenje – značenje koje riječ ima zbog načina na koji je stvorena, npr. *računalo* je 'nešto čime se računa'

2. terminološko, rječničko značenje – to je propisano značenje, određeno značenje, tako je *računalo* 'stroj za obradu dataka '

Značenja su svih naziva na taj način propisana u standardnom jeziku. Rieč znači ono što je određeno da znači.

Koje je tvorbeno značenje sljedećih rieči: *šiljilo, ljepilo, olovka, snimalica, prevoditi,...* To nije bilo teško, zar ne?

A koje je tvorbeno značenje sljedećih rieči: *problem, kategorija, katalog, kritika, radijator, ventilator, transkranijalan, dijagnoza,...* Hm, ako ne znate latinski i grčki teško da ste to uspjeli odgonetnuti.

Upravo je u tome problem s tuđicama, one nemaju tvorbenoga značenja za govornika hrvatskoga, iz samog izraza ne može se ni približno shvatiti što ta rieč znači.

Tuđice govorniku hrvatskoga ništa ne govore same po sebi, a **stručno se nazivlje uglavnom bira tako da i sama riječ upućuje na ono označeno njome**. Kod tuđica nije jasna tvorbena motivacija pa one mogu značiti bilo što, primjerice *sustipletika* jest znanost o vodi, ribama, miševima, konjima, zvuku, nosu, prašini. Sve nam to zvuči prihvatljivo, a sama rieč ne znači ama baš ništa.

I to je glavni problem s tuđicama, one govorniku koji ne zna jezik iz kojeg su došle ništa ne znače same po sebi.

Da ju čujemo na televiziji dok se govori o nečemu novom/nepoznatom (’sustipletičari dokazali: ljudi si kao vrsta pridaju preveliku važnost’) prihvati bismo je bez problema.

Rieči se u izvornom jeziku biraju tako da budu razumljive, nitko ne bira rieči bezveze, i posve je normalno da sliedimo tu ideju i stvaramo rieči koje samo razumijemo. Kada su francuzi 1960-ih stvarali rieč *surjekcija*, znali su da *sur* znači *na*, a *jekcija* je *bacanje* kao*preslikavanje*, pa da bi nazvali *preslikavanje na* jednom rječju napravili su *surjekciju*. Ona je nama podpuno neprozirna i ništoznačna.

Uzobrazno (kulturno) dobro

Živimo u doba globalizacije (što je najčešće istoznačno s poameričenjem), sviet postaje manji i svi jezici trpe zbog toga, događa se izjednačivanje s englezkim, što premda je prirodno, ne znači da je i poželjno. To je loše, neki razlozi su spomenuti, no ima još nešto. To što se sviet sve više ujedinuje je dobro s jedne strane, možemo se samo nadati budućnosti bez rudimentarnih plemenskih osobina koje su ekvivalent obilježavanju teritorija mokraćom, no to ujedno stvara opasnost od gubitka uzobraznih dobara.

Nužno je razviti svijest o nužnosti očuvanja uzobraznih posebnosti, od kojih je jezik jedna. Te su “stvari” neponovljive, posljedkom su tisuća godina razvoja (zapravo i puno više, mogu slobodno reći posljedkom 14 milijardi godina razvoja), i sada izgubiti (ako od dva nastane jedan, to je gubitak) takvu posebnost zbog lienosti, zbog neznanja je najblaže rečeno blesavo. **Smatram da bi svaki jezik trebao raditi na očuvanju svoje posebnosti, uvijek pronalaziti svoje načine da izrazi nešto, a ne prihvaćati rieči koje mu ništa ne znače.** Prekrasan je pojam viđenja nas kao jednoga planeta, kao jedne vrste i jasno je da nam treba međunarodni jezik (dok još nemamo visokotočne zbiljnovremene prievodnike), pa neka je to i englezki (koji objektivno gledano sigurno nije najsretniji izbor), ali neprocjenjiv bi bio gubitak bilo kojega postojećega jezika. I nemojmo se zavaravati, možda jednu od deset rieči prevedemo, u visokoj znanosti sve vrvi latinštinom i engleštinom, a danas sve više ti nazivi ulaze u obći jezik, a **fonetizirani englezki ili latinski nije hrvatski**, to bi prije bio srpski.

Ovakav nemaran (i blesav jer zapravo ne razumijemo što govorimo) odnos prema jednom kulturnom dobru je još jedna stavka na popisu stvari kojima se sramotimo kao vrsta.

Iskreno, meni je bezveze ako se “učene” rieč kažu na jednak način u svim jezicima, izpada da svi jezici imaju neke svoje rieči za seosko gospodarstvo, ali čim počnemo o nečem učenijem pričati onda svi rabimo iste rieči. Čemu će nam onda svi ti silni jezici. Zašto bih ih učili kada su ionako nalik na sve ostale?

Ali ima i pozitivnih primjera. U islandskom je gotovo sve prevedeno, od naziva voća do kvantne fizike (mjeričnog/oliničnog naravoslovlja). A i kinezki više-manje sve prevodi.

A i grčki prevodi latinski (grčke rieči ne moraju prevoditi jer ih razumiju, kao što je nama jasno što je *svjetlopis*, tako je njima jasno što je *fotografija*). Tu je i finski, mađarski itd. i svi se oni normalno bave znanošću. Otvoriš wikipediju na islandskom i niti jedna rieč ti nije poznata (osim možda iz njemačkog neka) – divota!

Ali u nekim područjima i u hrvatskom se prevodi. Recimo u znanosti o živome , u živoslovlju iliti biologiji, ondje su mnogi (ako ne i svi) nazivi životinja prevedeni: bradnjaci (*Pogonophora*), opnokrilci (*Hymenoptera*), četinoljusci (*Chaethognata*), bodljikaši (*Echinodermata*)... Toga se svi sjećamo iz škole.

Jedna od najčešćih kritika jest da su nove rieči glupe ili smiešne.

Nove su rieči smiešne i glupe – zrakomlat

nogomet, košarka, zvučnik, kolodvor, kišobran, olovka, suncobran, neboder, sladoled, izlet, zrakoplov, sažetak, ledište, ...

Sve su to nove rieči, nastale u 19. ili 20.st., danas nikome ne smetaju jer smo ih primili s majčinim mlijekom, ali tada su nekima bile smiešne.

Smatram da ljudi doživljavaju nove rieči smiešnima uglavnom jer to žele, to je nešto što se kod nas očekuje, stvaranje rieči je demonizirano i valja se tomu izrugivati. Kada se želimo zabaviti svašta može biti smiešno, npr. komedija Željka Pervana ili Teorija velikog praska, ako se pak ne želimo zabaviti, onda se ni nećemo. Tako je i riečima.

Zapitajmo se zašto strane nove rieči nisu smiešne, a domaće jesu. Stalno prihvaćamo tuđe riječi, a nikad nam nisu smiešne. Osim ako čujemo neku slovensku riječ. A zašto su nam baš slovenske i nove hrvatske riječi smiešne? Čini mi se da uglavnom zato što su nam razumljive, smiešno nam je da se nešto može na taj način nazvati. Smiešno je zvati znanost o čovjeku *čovjekoslovlje*, premda se tako i u grčkom zove – antropologija. Grcima je to posve razumljiva rieč tvorena istovjetno kao i naša, ali im nije smiešna kao što ni nama nije smiešno *jezikoslovlje* jer smo to upili bez filtriranja po smiešnosti.

Isto nam tako nisu smiešne ni gore navedene rieči, nikome nije smiešno *izlet* (izlet(jeti), nastalo prevođenjem njemačkog *Ausflug*) ili *neboder* (nebo-der(ati), po englezkom *skyscraper*), ali da nemamo 'neboder' i da ga netko predloži kao zamjenu za *skyscraper* rekli bi da je to smiešno; vjerojatno bi uočili da se radi o doslovnom prievodu i smiehu ne bi bilo kraja.

Naravno ima i loših rieči. Izrazi poput *okopasno hlačodržalo* ili *vuneni okokućni travopas* nisu primjer valjanih novotvorenica, to su riječi/izrazi nastali kako bi se narugalo hrvatskom jeziku i valjanom stavu da stvari treba nazvati razumljivim (što za nas znači hrvatskim) riečima. Tu je i zloglasni *zrakomlat*, simbol smiešnosti i uzaludnosti stvaranja rieči. Ta je rieč nastala, koliko znam, kao šala ili poruga, a ne kao ozbiljan priedlog za helikopter. Rieč je nadahnuta *muhomlatom* i mogla je biti žargonski, zabavan naziv za

helikopter kao što je u englezkom *chopper* (sjekač) jer je sjekao rastlinje pri spuštanju u Korejskom ratu ili tako nešta. Ali čini se da je neki nadobudni domoljub upotriebio tu rieč na televiziji u Dnevniku u izjavi tipa:[4] *Predsjednik republike izišao je iz zrakomlata*. I odtada se valjda sprda s tom riečju. *Helikopter* dolazi od grčkog *helix* – zavojnica, spirala i *pteron* – pero, krilo, a ozbiljni hrvatski priedlozi bili su: *uzvrt* (*uz-* u značenju gore, kao uzletjeti, uzići; *vrt* od *vrtjeti*, nešto kao 'ono što se vrtnjom uzdiže') (ne ~~uvrt~~) ili u obliku *uzvrtnjak*, pa *vrtložnjak*, *vrtilet* itd., a lako se napravi još novih ili boljih kada bi bilo volje za to.

Zar su samo hrvatske rieči smiešne, a svi su drugi jezici "ozbiljni"? Naravno da ne, pogledajmo nekoliko odabranih "smiešnih" tuđica u tablici B.2.

Križnica B.2: Smiešne tuđice

tuđica	doslovno	hrvatski
banka	stol, klupa	novčara; pohranište, -nica ...
bankrot	slomljena klupa	stečaj; propast, slom
burza	vrećica, mošnja, kesa	tržara, trgovara
elektrana	jantarasta kuća	strujara, munjara
nuklearna elektrana	oraščićna jantarasta kuća	jezgrena strujara
juke-box	~zlobna kutija	sviralica, glasbena kutija
masa	tiesto	tvarina
izolirati	pootočiti, poostrviti	osamiti, izdvojiti, ...
lavabo	prat ću	umivaonik
investirati	odjenuti, zaodjenuti	uložiti
električan	jantarast	munjevni
elektron	jantarak	munjak
minuta	umanjena	časak
sekunda	druga (2.)	trenutak, časak
autor	povećavatelj	sročitelj, tvorac, stvaralac
tekst	tkanje	orječje
kontekst	sutkanje, *sutka	surječje
garderoba	čuvaj-robu	rušnica
eksplozija	iztjerati pljeskom	razbuk, buknuće, prasnuće, ...

Naravno da su te riječi raznim putovima kroz nekoliko međuznačenja došle do sudobnog značenja i da se danas ne pomišlja na doslovno značenje. Ali ono ju tu i mnogo

[4] Koga zanima može sam potražiti o čemu je točno bila rieč.

je smješnije i bezsmislenije od hrvatskih riječi kojima se ruga pozivajući se na doslovno značenje.

Pogledajmo još hrvatskih smiešnica: *mišić*, zapravo je to *mali miš*, nastalo prevođenjem latinskog *musculus* – mali *mus* jer su oblik i pokreti nekih mišića (biceps, dvoglavik) podsjećali na miša. Kako je to smiešno, treba odmah ukloniti tu rieč i zamieniti ju barem *muskulom* da ju ne razumijemo tako da se možemo usredotočiti. Ili *predsjednik*, onaj koji sjedi izpred, ha ha, prema lat. *praesidens*, ili *sadržaj/sudrž*, ono što se drži zajedno/skupa, ono što se sa-drži, isto prema latinskom *continere*, skupa držati, od čega je i *kontinent* (možemo hrv. reći *kopnina*), zemlja koja se drži skupa, tj. neprekinuta zemlja. Inače ima i rieč *usebina* (ono što nešto ima u sebi, valjda) za sadržaj, zapravo se u slovenskom rabi kao *vsebina*. Kako je to sve smiešno. *Promjer* prema *diametros*, haha, *tisak* (tiskati) prema *imprimere*, haha, *sljedba* od *secta*, haha, *sveučilište* od grč. *pandidakterion*, lol. Kako je sve to prevođenje smiešno, odakle im uobće ta ideja da se nerazumljive tuđice prevode tako da su razumljive.

Idemo dalje: *nebce*, tj. pisano nerazumljivo *nepce*, je upravo to, *malo nebo*; *koža* je zapravo pridjev koji danas glasi *kozja* od životinje *koze* jer je to bila *koža skora* (kozja koža). Pa imamo *kocku*, zapravo *kostku*, *malu kost* jer su kocke za igru bile napravljene od kosti, imamo i *tkivo* (staničje) od glagola *tkati*, haha skup stanica se zove tkanje itd. Kako li je to sve smiešno, mislim da je najbolje sve te rieči izbaciti i zamieniti ih nerazumljivim tuđicama.

Vidjeli smo da i strane rieči mogu biti smiešne, kao što to mogu biti i domaće dobro poznate. Nije *pticoslovlje* ništa smješnije od *ornitologije*, samo je razumljivije. Nekad su i tuđice mnogo smješnije od domaćica, npr. jantar, stalno govorimo o jantaru, imamo cijeli fakultet koji se bavi jantarom, jantar kruži oko jezgre atoma, jantar teče žicama, jantar mi omogućuje da ovo pišem; dakako, govorim o elektronima, elektronici, električnoj struji i sl. I to je kao posve prihvatljivo govoriti o jantaru (!!!), nimalo smiešno i glupo. Da razodkrijemo te rieči, pa kažemo jantarak, jantarkoslovlje to bi bilo strašno smiešno i glupo, kao što izvorne riječi elektron, elektronika, ...i jesu, a nitko se ne smije. Ali hrvatske riječi izvedene od rieči *munja*, valjda jedine električne pojave u prirodi poznate od pamtivieka, smiješne su i glupe (a jantar je posve prihvatljiv).[5]

Da zaključim, rieči su “smiešne” ako želimo da su smiešne, smiešne mogu biti nove rieči i postojeće rieči, bilo domaće bile tuđe. A vidjeli smo i da je prevođenje rieči česta pojava.

[5]Hrvatski bi išlo: munjak – elektron, munjkoslovlje – elektronika, munjina/munjivo – elektricitet, munjevni – električni, munjara/strujara – elektrana (uzp. s mljekara, mjesto gdje se proizvodi mlijeko), munjevnica – baterija/akumulator, ...

Što nam to činu od jezika!?

Primjetba je s Omrežja (Interneta) na izmišljanje novih riječi. Začuđuje ta sljepoća, dakle, **u redu je primati tisuće i tisuće rieči na -tor i -cija koje nama uobće ništa ne znače, to je u redu, to je posve prihvatljivo i normalno, ali kada netko ponudi zamjenu za te rieči, bilo da doslovno prevede ili nekako drugačije, to je odmah glupo i smiešno**, svaki se priedlog dočekuje kao vic, moraš se smijati kada čuješ neki hrvatski prijedlog za neku tuđu riječ, to se očekuje.

Dodatno, tu se obično radi o tome da govornici ne shvaćaju kako je rieč tvorena, ne prepoznaju tvorbeni način kao izpravan, premda on to jest. To je posljedica izrazito površnog poznavanja hrvatske tvorbe, što je i za očekivati, to se ne može naučiti ako se na to ne okrene pozornost, to se neće usvojiti zajedno s ostatkom jezika jer to nije česta pojava itd. Ali onda je škola mjesto gdje bi to trebalo usvojiti.

Hrvatski ima vrlo dobra sredstva za stvaranje rieči uzporedivo s grčkim, samo je problem što su ljudi neupućeni u postojeće tvorbene načine. Nitko ne shvaća rieč *slučaj* kao izvedenicu od *slučiti*, osim ako ga netko ne uputi na to, pa mu onda može i biti nepoznati takav tvorbeni način.

I ne moramo se ograničiti na postojeće načine, tvorba rieči nije zacementirana, kako bi se dalo zaključiti iz djela nekih jezičara, u trenutku kad je završen prikup građe za knjigu *Tvorba riječi u hrvatskome književnom jeziku* Stjepana Babića. Pojavljuju se novi načini i nema ničeg lošeg u njima. Plaho se pojavljuje stapljanje ili *blending* rieči, pa imamo *prisavljotine* (bljuvotine+Prisavlje (HRT)) ili *kradarenost* (nadarenost za krađu) što sam na televiziji čuo od jedne djevojčice.

Otmjenost, učenost

Evo scijentifičke kroatolingvne akademske ekspresije definicije epistema.[6]

Epistem je instantno akcesibilna permanentno memorirana eficientno aplikabilna principijelno sistemitizirana faktička i proceduralna informacija.

Kako li samo učeno zvučim. Evo seljačkoga određaja:

Znanje je trenutačno dostupna trajno pohranjena učinkovito primjenljiva načeono ususta vljena činjenična i postupačna obavijest.

Sekventna je ekspresija definicije rase, naturalno iterno kroatolingvna:

Rasa je uniparijentalna ili biparijentalna propagacija homozigotnih individua.

[6] Primjeri prilagođeni iz (László, 1990; László i Boras, 2007).

Opet ispadam učen. A sada opet seljak:

Pasma je jednoroditeljan ili dvoroditeljan razmnožak istoplodnih jedinaka.

Sve je to pitanje ugleda i odgovora na *Što je cool*. Strani nam jezici ponekad izgledaju bolje od našeg, izgledaju nekako stabilno i sigurno za razliku od našeg u kojem smo često nesigurni. No to je uglavnom privid, svi su prirodni jezici po tome slični.

Kako drugi jezici prevode

Pogledajmo tablicu B.3. U prvom su stupcu navedeni neki latinizmi, pseudogrecizmi i anglizmi (latinština, lažigrčština i englezština), u drugome grčke rieči kojima su prevedene rieči iz prvog stupca, a u trećem moguće hrvatske rieči.

Želio sam pokazati da grčki kao jedan od dva osnovna jezika (drugi je latinski) za stvaranje nazivlja ne prihvaća latinizme nego ih prevodi. Razlog je uviek isti: rieči ili jedan dio nije grčki, tj. jedan je dio nerazumljiv, pa nisu blesavi da govore rieči koje ne razumiju ili koje su pogrešno tvorene. Zašto bi govorili *sociologija* kada im to zvuči kao *blablaslovlje*, zato uzmu svoju rieči za družtvo i načine *koinoniologija* itd.

Nisu ovdje navedeni, no grčki ima i vlastite razumljive nazive za velike brojeve za razliku od množtva jezika koji preuzimaju nerazumljive latinske nazive.

Važno je napomenuti da rieči iz prvog stupca nisu nuđno i rieči koje bi se uporabile u latinskom.

A ja znam, znam, znam sve, a najbolje padeže

Pogledajmo kako se zovu padeži na nekim jezicima.

Prevođenjem grčkog *ptosis* dobiveno je latinski *casus*, a to znači 'padanje, padež' (iz uzpravnog položaja). Prvo se odnosilo na 'kose' padeže, koji "padaju" iz uzpravnog (orthe, rectus) padeža, koji je kasnije preimenovan u onomastike.

- **Nominativus** je latinski prievod grčkog **onomastike**, padež imenovanja, nazivanja (nominatio) – **nazivnik/nazovnik**.
- *Genos* je 'rod, razred', a padež koji označava *rod* ili razred u koji neka stvar pripada je **genike**, hrvatski **rodnik**, što je pogrešno polatinjeno u **genitivus** što bi bio padež podrietla, postanka,...
- **Dotike**, **dativus** je od glagola 'dati', to je padež davanja – **datnik**.
- *Aitia* je uzrok, odgovarajući padež je **aitiatike** koji izražava uzrok neke radnje, padež koji *tvori* (čini) neku radnju – **tvornik**, padež koji se odnosi se na ono što je prouzročeno ili postignuto, kao padež učinka, padež stvari koja je izravno

Križnica B.3: Kako grčki prevodi strane rieči

latinski,"grčki", englezki	grčki	hrvatski
definicija	orismos	određaj, omeđaj,...
lingvistika	glossologia	jezikoslovlje
sociologija	koinoniologia	družtvoslovlje
medicina	iatrike	liečništvo
televizija	teleorase	dalekovidnica
terminologija	orologia	nazivoslovlje
signal	sema	dojavak
gravitacija	baryteta	teža
audiovizualan	optikoakoustikon	slušno-vidni
radioaktivan	radienergo	zračljiv, 'zračbodjelatan'
nuklearni	pyrenikos	jezgreni
procesor	epeksergastes	obradnik
kontinent	epeiros	kopnina, zemljočest
eksplozija	ekrekse	razbuk, buknuće, prasnuće
gen	gonidio	nasljedilo (kao osjetiti-osjetilo)
kompas	pyxida	sjevernica
sport	athlema	takma
informacija	pleroforia	ob(a)viest
datum	emeromenia	nadnevak
kalendar	emerologio	danovnik, danokaz
stipendija	ypotrofia	poduporka...
burza	khrematisterio	tržara
financirati	khrematodoto	podnovčiti
telegram	telegrafema	brzojav(ka)
automobil	autokineto	samovoz, samokret, samogib, kola
autobus	leoforeio ('pukonoša')	putnički/javni samovoz, javnovoz
roman	mythistorema (mithos + istoria)	?
bit	dyfio	dvojnica
interview	synenteuxi	subesjeda
kamion	fortego	teretnjak, teretni samovoz
weekend	sabbatokyriako ('subotonedjelja')	konac (tjedna)
banka	trapeza	novčara
paintball	kromatosphairise	šaroboj (<biti)
deterdžent	aporrypantiko	perilo, čistilo, ...
Internet	Diadiktuo	Omrežje, Svemrežje
sapun	sapouni	milo
flomaster (felt-tip pen)	markadoros	pustenka (pustena pisaljka)
garaža	gkaraz	kolnica
turizam	tourismos	putničarstvo

grčki	latinski	hrvatski	slovenski	poljski
ptosi	casus	padež	sklon	przypadek
onomastike	nominativus	nazovnik/nazivnik	imenovalnik	mianownik
genike	genitivus	rodnik	rodilnik	dopełniacz
dotike	dativus	datnik	dajalnik	celownik
aitiatike	accusativus	tvornik	tožilnik	biernik
kletike	vocativus	zovnik	zvalnik	wołacz
topike	locativus	mjestnik	mestnik	miejscownik
organike	instrumentalis	orudnik/oruđnik	orodnik	narzędnik

zahvaćena radnjom. U latinskom je pogrešno prevedeno u **accusativus** – padež obtuživanja.

- Padež zvanja (klese, vocatio), dozivanja je **kletike**, **vocativus**, **zovnik**.
- Padež koji izražava *mjesto* (topos, locus) gdje se nešto događa je **topike**, **locativus**, **mjestnik**.
- Padež koju izražava oruđe (organon[7], instrumentum), sredstvo kojim se nešto čini, dakle **organike**, **instrumentalis**, **orudnik/oruđnik**.

Zar su Grci odabrani narod?

Čime li se bave ove znanosti:

Potamologija, kraniologija, helmintologija, mirmekologija, oologija, ornitologija, psefologija, skatologija, veksilologija, zimologija.

A ove:

Rjekoslovlje, lubanjoslovlje, crvoslovlje, mravoslovlje, jajoslovlje, pticoslovlje, izboroslovlje, izmetoslovlje, zastavoslovlje, vrenjoslovlje.

Dakako istime se bave. Na ovom bi mjestu zapravo mogao biti kraj mojega dokazivanja da valja uzeti hrvatsku rieč, a ne tuđu. **Stvarno mi nije jasno zašto bi itko odabrao rieč iz prvog popisa umjesto iz drugoga. Što ne valja s hrvatskim riečima?** To što su samorazumljive, pa stoga gube dio svoje tajanstvenosti, nisu tako *cool* kao tuđe nerazumljive rieči. Zar neka znanost gubi na ugledu ako i najneobrazovaniji čovjek može razumjeti čime se bavi?

Ako možemo birati između niza glasova koji nam ništa ne znači (ornitologija, jer niti nam "orniti" pjevaju na granama, niti mi "logijamo") i niza koji nam znači

[7] Kakva je veza oruđa i tjelesnih organa? *Organon* je ‘ono čime se radi’ – *radilo*, što služi za rad (ergon), što služi za izvođenje nekog zadatka. Do značenja specijaliziranog diela tiela je došlo iz ovakvih rečenica “oko je organ (=oruđe) vida, ono čime se služi za vid”, dakle tjelesni organi su oruđa osjetilâ ili sposobnosti.

(pticoslovlje, jer nama 'ptice' pjevaju na granama, i mi 'slovimo', govorimo) zašto biramo ono što nam ništa ne znači? Mi nismo Grci, zašto onda govorimo grčki?

Ali prve su rieči međunarodne, omogućuje da se sporazumiju znanstvenici iz različitih zemalja. – reći će netko. Tu se miešaju neke stvari; recite zar naši znanstvenici s američkima razgovaraju na hrvatskome? Naravno da ne, obće na englezkome, najboljem, najuglednijem jeziku na svietu. Dakle znanje *engleskoga* omogućuje našim znanstvenicima da obće s drugima, a ne znanje hrvatskoga. A hrvatski sigurno nije odskočna daska za učenje englezkoga.

Sada smo došli do pravoga problema, lienost i kolonijalni mentalitet. Lieni smo, ne želimo pamtiti dva naziva, jedan hrvatski, drugi englezki za isti pojam, pa stoga "prevodimo" strani naziv na hrvatski fonetizacijom i tako u tren oka dobivamo hrvatsku riječ. Pa tako englezki *assertion* postane lažihrvatski *asercija*, a rieč je dakako o *tvrdnji* i sl. Koja je šteta? Pa šteta je u tome što *asercija* ne znači ama baš ništa, i onda se ta rieč rabi bez ikakvoga razumijevanja. Ili u geometriji *diamond* (romb) postane – wait for it – *dijamant*.

Uglavnom zbog lienosti i neznanja unosimo u jezik hrpu, bezkorisnih riječi koji se tvorbeno neprozirne i uobće nije jasno što označavaju, niti ih Hrvat koji ne zna englezki razumije ako ih nije usvojio odmalena, pa tako govorimo o monitoringu Hrvatske, o rejtingu Hrvatske, o biznisu, o kombajnu, o tramvaju kao da to zapravo išta znači. **Danas je uobće bezsmisleno pitati je li neka riječ hrvatska ili nije, danas je svaka latinština i engleština ujedno i hrvatski, ovakav je odnos** $grki \subset latinski \subset engleski \subset hrvatski$**.** A to neznalice i zlonamjernici pozdravljaju. Čak su nam i *false friends* postali pravi prijatelji, pa tako *eventually* (konačno, s vremenom) postane *eventualno* (možda, possibly, conceivably) što nikako nije isto, ili da čovjek ne povjeruje, *in a nutshell* (ukratko, sažeto) postane *u orahovoj ljusci*.

Kad netko to pogleda izvana zapita se što, pobogu, izvodimo. Nazivamo stvari imenima koja nam ništa ne znače, u svim područjima života rabimo nazive koji su nam posve neprozirni: ventilator, radijator, bojler, televizor, . . . Niti smo Rimljani, niti Grci, niti Englezi, zašto onda govorimo tim nerazumljivim jezicima, a imamo svoj?

Višestruke tuđice

Vidjeli smo kako se lako prisvajaju tuđe rieč, ali da bi sve bilo još absurdnije, ponekad ćemo rieči za istu stvar posuditi iz dva jezika.

Tako samo recimo iz *grčkog* uzeli *stih*, a onda iz *latinskog versifikaciju* za proučavanje građe stiha. Prikažimo to skrižaljkom B.4.

Uzput, *insekt* i *entomo* znače *urezanik*. haha

Križnica B.4: Višestruke posuđenice

grčki	latinski	hrvatski
stih – stihourgia(?)	versus – **versificatio**	vrstica – ?
entomo – **entomologija**	**insekt** – *insectiloquium	kukac – kukcoslovlje

Treba li se miešati u jezik

Savjetujem svakome tko se zanima za jezik da pročita knjigu *Čiji je jezik?* Mate Kapovića, dostupna je i zakonito na Internetu. U njoj razbija neke sveuzočne mitove o (standardnom) jeziku. No, u jednoj stvari grieši, osuđuje purizam kao negativnu pojavu jer ga promatra samo kao odraz nacionalizma. No purizam može biti i čisto zdravorazumski podhvat s ciljem povećanja razumljivosti i raznolikosti jezika.

Je li dobra ideja miešati se u prirodni razvoj? To je pitanje bezsmisleno, sve što činimo je prirodno jer mi jesmo priroda, čak ako i prirodno shvatimo kao 'bez ljudskog uplitanja', opet je bezsmisleno jer se "uplićemo u jezik" nesvjestno, ali ajmo se još distancirati i praviti da se jezik može događati "prirodno" bez ikakve ljudske intervencije, možemo li ga tada mienjati, smijemo li? Razmotrimo jednu drugu "prirodnu" situaciju, prirodno je recimo dobiti rak i umrieti, to se obično događa bez naše svjestne intervencije, ali nećemo stajati po strani i gledati kako se to događa (barem oni koji su prisebi). Tako da se i inače uplićemo u "prirodne" stvari kada god nama paše, bilo da liečimo bolesti, preusmjeravamo tokove rieka, biramo biljne i životinjske potomke s boljim karakteristikama za daljnje razmnožavanje ili pak kada gradimo kuće, stvaramo države itd. itd. Takoreći sve što činimo je petljanje u "prirodni" tiek stvari. Jezik je naše glavno sredstvo sporazumievanja, razvija se "prirodno", ali nije sve što nastane samo od sebe najbolje moguće, rak nastaje sam od sebe pa ne gledamo blagonaklono na njega.

Inače se polusvjestno događaju te promjene, zašto ne bismo poduzeli svjestne korake?

Ne kažem ja da trenutačni jezik, tj. govor nije dobar, on očito izpunjava svoju svrhu, ali to ne znači da ne može biti bolji. Postavlja se odmah pitanje što to znači "bolji"?

Ja mislim da je jezik 'bolji' ako nam ne treba rječnik nekog drugog jezika da bismo ga razumieli, tj. bolji je ako je samorazumljiviji, ako se njegove jedinice mogu razumieti pomoću postojećih unutarsustavnih jedinica.

To je i prirodna situacija u jeziku, npr. ako nam treba pridjev od rieči 'svinja', reći ćemo 'svinjski', a ne ne znam što.

Mislim da je glupo preuzimati tuđe nerazumljive tvorenice kada možemo napraviti svoje razumljive! Npr. u matematici imamo jednu matricu koja se zove *jakobijan*, to je očito osamostaljeni pridjev iz englezkog ili francuzkog koji nam ne znači puno, zašto Jakobijevu matricu ne bismo zvali *jakobijevka*, kao što je kapa *titovka*?

Kakve *citologije* i *histologije* kada možemo reći *stanicòslovlje* i *tkivòslovlje*. Hoćemo li govoriti *izomorfizam* ili *istolikost*, hoćemo li nekoga *interviewirati* ili ćemo *s nekime subesjediti*, hoćemo li govoriti *idempotencija* ili *istomoćnost*, *eufemizam* ili *blagorječje*, *eutanazija* ili *blagosmrće*,...

Na nama je, a smatram da je itekako koristno rabiti domaće razumljive rieči.

B.2. Pravopis

Pravopis ovoga rada jest inačica čitkotvornog pravopisa, pravopisa čitke tvorbe. To znači da je prirodno čitak, a prozirne tvorbe. Bilježenjem se prozirnosti tvorbe za pisatelja smanjuje broj pravila pri pisanju i istodobno povećava razbirljivost zapisa. Prirodna ili samodjelna čitkost znači da se za naravnoga govornika broj pravila pri čitanju ne povećava.

Dakle, ovo je jedna inačica za koju sam se odlučio, mogu se i drukčije neke stvari riešiti, sve je to stvar dogovora.

Pokazat ću da je ovaj pravopis bolji od postojećeg jer je bolji u svojoj osnovnoj funkciji zapisivanja značenja, jer ima mnogo manje pravila, tj. lakše ga je naučiti i obćenito je obavjestniji i koristniji.

Glavne su značajke da se dugi jat koji se inače bilježi s *ije* piše s *ie*, da se ne bilježe izpadanje glasova, jednačenja po zvučnosti i mjestu tvorbe. Time se dobive mnogo bolji pravopis što ću pokazati. Valjda napomenuti da se rieči čitaju kako se i inače čitaju, samo se drukčije zapisuju.

Vjerojatno u samome radu ima podosta pravopisnih pogrešaka, to je posljedica nenaviknutosti na ovakav način pisanja i brzine pisanja, a ne težine pravopisa.

Dakako, pitanje pravopisa je političko pitanje itd. Naravno posve se distanciram od toga.

Do pravopisa primienjenog u ovome radu dolazimo rješavanjem problema postojećeg pravopisa.

Pisanje glasa /ie/

Hrvatski standardni jezik ima 32 glasa (fonema, zvukova koji služe za razlikovanje značenja). To je tako odabrano, tj. propisano, premda dakako može biti i drugačije propisano ako želimo. Glasovi se diele na otvornike i zapornike.

Otvornika je šest: /a/, /e/, /i/, /o/, /u/ i dvoglas /ie/

Pitanje je kojim se slovima taj dvoglas /ie/ piše.

“Taj se dvoglas bilježi s ije. On je uviek dug i lako ga se razlikuje od kratkoga glasovnoga skupa je, koji se piše je.”

Sve bi to bilo liepo i krasno da nije sljedećega: nizom slova *ije* ne bilježi se samo glas /ie/ nego i niz triju glasova i-j-e: *dijeta* (i-j-e, 3 sloga) nasprem *dijete* (2 sloga), *nijedan* (i-j-e) nasprem *mlijeko* (mlie-ko) itd.

Dakle tri slova ***ije* stoje za dvie različite stvari: glas /ie/ i niz od tri glasa i-j-e**. To jest, na temelju slova čitatelj ne može znati koji su glasovi tima slovima zabilježeni, drugim riečima došlo je do gubitka obaviesti (informacije), ne postoji uzajamno jednoznačno (sujednoznačno) preslikavanje, supreslikavanje iliti bijekcija.

To je kao da se broj 1 i 2 bilježe istim znakom npr. 1, pa se 11 može pročitati na četiri načina: jedanaest (jedan, jedan), dvanaest (jedan, dva), dvadeset i jedan (dva, jedan) i dvadeset i dva (dva, dva). Mislim da svi uočavaju zašto je to problem, odnosno zašto taj sustav bilježenja nije dobar.

To onda potiči i pogrešan govor, ljudi vide napisano *mlijeko* i onda vodeći se po čitaj kako piše izgovore ga trosložno s j (mli-je-ko) umjesto dvosložno (mlie-ko) kako je pravilno u standardu.

Jedan je i od razloga zašto ljudi ne razlikuje *i(j)e* i *je* jer se ne mogu pouzdati u izgovor kada je *ije* jednom *i-j-e*, a drugi put *ie* pa to prieči uočavanje da se radi o dva glasa.

Toliko nam je inače važno zabilježiti svaku tančinu izgovora na štetu obaviestnosti, a ovdje gdje zaista postoji potreba da se pravilno zabilježuje to ne činimo.

Dodatno problem stvaraju pravila po kojima se glas /ie/ u nekim proizvoljnim situacijama piše sa *je*. Pa se ti osloni na uho pri pisanju, nije ni čudo da gotovo svi muče muku s i(j)e i je.

I da, u "ustaškom" se pravopisu tako bilježi, ali i u dovukovskom hrvatskom, iako je jedno i drugo podpuno nevažno za nas.

Izpadanje glasova

Pogledajmo što pravopis IHJJ-a (`pravopis.hr`) kaže o ovome (moj mastnopis):[8]

> Zbog jednostavnijega izgovora suglasnici ispadaju u nekim suglasničkim skupinama. To se ispadanje **katkad** bilježi u pismu.
>
> Ne pišu se:
>
> a) dva suglasnika nego jedan kad se dva ista suglasnika nađu jedan do drugoga: bezvučan (bez + zvučan), predvorje (pred + dvor + je)
>
> b) t ili d u skupinama stn, ždn u oblicima i izvedenicama od **domaćih** riječi: dvanaest – dvanaesnik; mastan – masna, masno, masni; most – mosni; nuždan (i nužan) – nužna, nužno, nužni; slastan – slasna, slasno, slasni; vjerojatnost – vjerojatnosni

[8] Zbog nedostatka vremena navodi nisu tiskopisno (tipografski) i pravopisno izpravno uobličeni. Preuzeto s `http://pravopis.hr/pravilo/ispadanje-glasova/8/`

Koliko li teksta, nije li vrieme za neku iznimku? Naravno da je!

> Kad imaju razlikovnu ulogu, t i d se pišu.
> grozdni (< grozd) | grozni (< grozan)
> (...)
> prstni (< prst) | prsni (< prsa)
> (...)
> U kontekstima u kojima je potrebno razlikovati pridjev od usta i pridjev od usna moguće je uspostaviti razliku ustni (< usta) i usni (< usna). **Gdje ta razlika nije potrebna**, i od usta se upotrebljava pridjev usni (usna šupljina). Slova t i d **zadržavaju se u pismu** i u tvorenicama koje bi se previše udaljile od osnovne riječi: mošt – moštni, brazda – brazdni, odmazda – odmazdni.
> (...)
> d) t u oblicima imenice otac i izvedenicama od te imenice: otac – oca, ocu... oci (uz očevi), očev, očinski; praotac – praoca, praocu... praoci, praočev.
> Piše se:
> a) t u skupini stn u izvedenicama od riječi **stranoga** podrijetla: aoristni, ametistni, azbestni, balastni, damastni, kontrastni, protestni, tekstni, testni, tvistni
> b) t ili d u oblicima imenica muškoga roda koje završavaju na -dak, -tak, -dac, -tac u kojima je samoglasnik a nepostojan: (...)
> predak – predci, sudac – sudci, svetac – svetci
> (...)
> f) j u superlativu pridjeva kojima komparativ počinje glasom j: najjači, najjednostavniji.
> Kad se dva ista suglasnika nađu jedan do drugoga, **u nekim se tvorenicama oba zapisuju**: dvadesettrećina, hiperrealističan, izvannastavni, naddržavni, nuzzarada, poddijalekt, preddušnični.

Naveli smo samo dio pravila, a već smo naišli na hrpu nelogičnosti i nepotrebne komplikaže:

(U daljnjem ću orječju nazivati pravopis kojim pišem “svojim” radi jednostavnosti, ali naravno da ga nisam ja u podpunosti izmislio nego se njime služim.)

1. Izpadanje se **kadkad** bilježi u pismu. Dakle, valja posebno naučiti kada se bilježi, a kada ne. Plus naravno, budući da je ovo hrvatski pravopis, u oba ćemo slučaja imati (proizvoljnih) iznimaka koje treba dodatno naučiti. Nasprem ovog zbunjenog pravopisa, u pravopisu za koji se ja zalažem izpadanje se **nikada** ne bilježi

u pismu (eventualno u jednom ili dva slučajeva, ali to ovisi za što se odlučimo). Dakle, **u "mom" pravopisu ne treba pamtiti sva ova pravila i množtvo njihovih proizvoljnih iznimaka**, zbog toga je lakše naučljiv.

2. Promotrimo sada točku b) u 'ne piše se' i točku a) u 'piše se': *t* u skupini *stn* izpast će ako se radi o domaćoj rieči (*bolesna*, ne *bolestna*), a neće izpast ako se radi o stranoj rieči (*protestni*, ne *protesni*), iako se u oba slučaja izgovora jednako! Ovo je podpuna bedastoća! Bedastoća je dielom u tome da su rieči prilagođenice ("rieči koje su prilagođene pravilima hrvatskoga jezika, ali se osjećaju kao strane") zapravo neprilagođene, s njima se drugačije postupa nego s domaćicama, a što je još gore s njima se bolje postupa, njih se ne unakažava, njih ne kastriramo, kod njih ne dolazi do gubitka obaviesti. Dodatno moramo stalno imati na umu podrietlo rieči. Prava je bedastoća da uobće postoji ovakvo pravilo koje opet treba posebno naučiti. **U mom pravopisu nijednog od ta dva pravila nema jer sve podpada pod jedno pravilo da slova ne izpadaju, ne treba učiti dodatna pravila i stalno ih imati na umu pri pisanju.**

3. Naravno, dolazi iznimka. U domaćim riečima ne dolazi do izpadanja slova (iako naravno u govoru i dalje izpadaju ti glasovi) u nekim slučajevima. Posebno je zanimljiv navedeni slučaj s *ustima* i *usnama*. Iz njega čitamo da ćemo nekad pisati *ustni*, a nekad *usni* s istim značenjem, dakle jedna te ista rieč ima dva pisanja. I naravno u nekim slučajevima ne izpadaju jer je to previše štetno. **Još dodatnih pravila!** Zanimljivost: po nekim drugim hrvatskim pravopisima piše se *dvobrazni* umjesto *dvobrazdni*, dakako s ogromnim gubitkom obaviesti, možemo samo pogađati od kojih se rieči sastoji.

4. Otac-svetac-sudac. Otvorimo li *Školsku gramatiku hrvatskoga jezika* Sande Ham pročitat ćemo da "Suglasnici *d, t* ispadaju i ne bilježe se ispred c samo u oblicima i izvedenicama od riječi *sudac, svetac, otac*". Po njoj se piše *suca, sveca, oca*. Ovaj navedeni pravopis se malo udosljedio pa se samo *otac* piše *oca* (d), a ovi ostali se pišu bez izpadanja (premda se dopušta i pisanje s izpadanjem) (b). Tko zna, možda ćemo jednog dana imati pravopis u kojem se i *otac* piše kako spada i **u kojem neće biti dodatnog pravila koje posebno treba naučiti i to pravila za jednu jedinu rieč**. Dakako, takav pravopis već postoji i normalno pišem *otca* kao i u svim drugim primjerima.

5. Pogledajmo još izpadanje istih glasova, tj. slova (a). Isti suglasnici jedni do drugih izpadaju u pismu, barem ponekad. Samo to izpadanje je jako štetno jer se gubi veza s polazištnim riečima. *Bezavjesni*, neki će možda morati pročitati naglas tu rieč

da shvate da je to *bezzavjesni*,*bezzastorni*. Ali zapravo neće se uviek ni izgovoriti s jednim /z/, nego će se često u rjeđim riečima čuti oba glasa. Dakle, pravopis ovdje nije zrcalo govora, iako tomu smjera, i dovodi do gubitka obaviesti: ako piše *odvojiti*, ne možemo znati da je to zapravo *oddvojiti*, a i *odvojiti* bi se možda moglo shvatiti kao *podvojiti* 'učiniti dvojnim'. Naravno, od toga se pravila odstupa kada se pravopisu prohtije (točka f i izpod). Sve to znači **još pravila, još iznimaka koje treba posebno naučiti**.

Da sažmemo malo, vidjeli smo hrpetinu pravila (nisu ni sva gore prikazana) koja nam govore kako pisati rieči u nekim situacijama. Naučili smo da se rieči pišu malo ovako, malo onako, hrvatske ćemo rieči koje zapravo nisu hrvatske pisati na jedan način, domaće ćemo rieči pisati malo na taj način malo na drugi način, neke ćemo istodobno pisati na oba načina, a naravno sve to vrvi množtvom iznimaka. Dakle, normalno je pisati *protes**t**ni*, *bole**sn**a*, ali *prs**t**ni* pa onda *us**t**ni* ili *usni*, pa onda pak *braz**d**ni*, pa u nekim riečima izpadaju isti suglasnici, a u nekima ne itd. Domaće rieči kastriramo, strane ne itd. Podpuno neujednačeno, ružno i komplicirano! Podsjetimo se što sav taj nered i proizvoljnost znači za korisnika: **sva ta pravila i iznimke treba posebno zapamtiti! Govornik stalno mora razmišljati o tome kada se nešto bilježi a kada ne.** Uzporedimo sva ta pravila i iznimke sa samo **jednim** pravilom: **Pri tvorbi rieči i oblika svi se glasovi pišu, ne bilježi se izpadanje do kojeg dolazi u govoru**.

JEDNO nasprem MNOGO proizvoljnih, bezkoristnih (dapače štetnih), težko naučljivih pravila i iznimaka.

Pogledajmo još nešto što piše u izpadanju glasova (izumljeno za ovaj pravopis):

> Nikad se ne pišu tri ista samoglasnika zaredom: od Joensuu (grad u Finskoj) dativ i lokativ nije Joensuuu nego Joensuu, od Waterloo instrumental nije Waterlooom nego Waterloom, od Yahoo instrumental nije Yahooom nego Yahoom, a posvojni pridjev nije Yahooov nego Yahoov.

Smatram da je ovo bedastoća. Napominjem da je ovo pravilo izmišljotina ovoga pravopisa (zapravo Jezičnog priručnika Coca-Cole istih autorica). Znamo da se *Yahoo* čita *ja-hu*, a kako bismo onda pročitali *Yahoov* nego *ja-huv*, a to nije kako govorimo: *ja-huov*. Dakle, ovakvo pisanje upućuje na posve pogrešan izgovor, a pravilo je kao i većina drugih bezkoristno i štetno. Sada nam pravopis stvara još jednu nepotrebnu nedoumicu, ima li izvorna rieč tri ista samoglasnika ili dva kako piše. Što da postoji neki *Yaho* čitano *ja-ho*, onda je pridjev normalno *Yahoov* (kao i za *Yahoo*) i sada, zbog bedastoća, više ne znamo o čemu govorimo.

Što ako imamo četiri ista samoglasnika zaredom? Koliko ih onda izpada? Što ako šest? Što ako devet? Onda ćemo izbaciti jedno od devet istih slova zato što... što, što dobivamo time, osim što zamućujemo stvari? Pogledajmo stvaran primjer. Recimo da govorimo o konju zvanom **Potooooooooo**[9] /potejtou/ (8 o), i sada hoćemo napraviti orudnik (*-om*), kako ćemo zapisati: *Potooooooooom* (9 kako je i normalno) ili *Potooooooooom* (8)? Zapravo, ako poobćimo pravilo to znači da moramo izbacivati po jedan samoglasnik dok ne dođemo do dva: *Potoom*!!!???

Pravilo je neprimjenljivo, bezkoristno, štetno, proizvoljno i nepotrebno (još jedno pravilo koje treba naučiti) kao i većina ostalih.

Jednačenja glasova

Neki se suglasnici diele po zvučnosti na zvučne (b, d, g, z, ž, dž, đ) i njihove bezzvučne parnjake redom (p, t, k, s, š, č, ć, c, h, f). Pri dodiru suglasnika različite zvučnosti prvi se suglasnik mienja u svoj parnjak tako da odgovara drugome po zvučnosti. Tako se primjerice govori *bespravni* od *bez +pravni* (z+p > sp).

Izdvojimo zanimljive slučaje kada se to zapisuje a kada ne:

> Jednačenje po zvučnosti provodi se: (...)
>
> – u većini riječi latinskoga podrijetla koje počinju s ab-, ob- i sub-: apsolvent, apsorbens, apsces, apscisa, apsolut, apsolutist; opservacija, opservatorij, opsesija, opskurnost, opstrukcija; supfebrilan, supskripcija, supstandard, supstantiv, supstitucija, supstrat
>
> Jednačenje po zvučnosti **ne zapisuje se** u riječima latinskoga podrijetla koje počinju sa sub- iza kojega slijedi p ili koje počinju s ad-: subpapilaran, subpolaran; adherencija, adhezija, adpozicija, adsorbens, adstrat.
>
> (...)
>
> Jednačenje po zvučnosti provodi se u izgovoru, **ali se ne zapisuje**:
>
> a) kad se d nađe ispred:
>
> – c: Gradac – Gradca, napredak – napredci, podcijeniti, podcrtati, redak – redci
>
> – č: mladac – mladče, nadčovjek, odčepiti, odčitati, podčiniti
>
> – ć: odćurlikati
>
> – s: brodski, gradski, podstanar, podsvijest, predsjednik, predstava, predstavnik, sredstvo, srodstvo, sudski, sudstvo
>
> – š: odškrinuti, podšišati

[9] https://en.wikipedia.org/wiki/Potooooooooo

b) u prefiksu ispod- i iznad-: ispodprosječan, iznadprosječan

c) u riječima **latinskoga podrijetla koje počinju s ad-**: adherencija, adhezija, adpozicija, adstrat

d) u riječima **latinskoga podrijetla koje počinju sa sub- iza kojega slijedi p**: subpapilaran, subpolaran

e) u **nekim riječima stranoga** podrijetla: bredpitovski, gangster, Habsburgovci

f) u **nekim** zemljopisnim imenima i njihovim tvorenicama: Josipdol, Križpolje, ivanićgradski.

Kad bi se izvedenica previše udaljila od osnovne riječi: pedesetdevetina (a ne pedesedevetina), podtočka (a ne pottočka ili potočka), predturski (a ne preturski), uzšetati se (a ne ušetati jer ušetati ima drugo značenje), do jednačenja po zvučnosti ne dolazi ni u izgovoru ni u pismu.

Opet množtvo proizvoljnih pravila:

1. Opet imamo situaciju da se nešto događa a ponekad se zapisuje a ponekad ne (= **pravila koja treba naučiti)**. Opet dielimo rieči po nacionalnosti, i sada vj. radi ustupka klasičnim filolozima svi moramo naučiti da se latinske rieči na ad- (**naučite koje su latinske rieči na ad-**) pišu s *ad-* prema se čitaju s *at-*. (**proizvoljno pravilo**).

2. **Posebno zapamtimo** kada se ne smijemo oslanjati na uho. Kada se recimo *d* nađe izpred toga i toga (**zapamtimo to**). Dodatno, zapamtimo predmetke ispod- i iznad-. Još jednom nas se podsjeća da su latinske prilagođenice zapravo neprilagođene i njih moramo posebno tretirati, **zapamtimo** koje su rieči latinskog podrietla s tim predmetcima. Sliedi proizvoljan popis nekih rieči koje moramo naučiti napamet. I naravno kada taj pravopis zakaže, onda se oslonimo na bolji pravopis.

Sve u svemu množtvo iznimaka koje treba naučiti napamet. No prije nego pokažemo što još ne valja s ovime, pogledajmo još jednu mjenu glasova – jednačenje po mjestu tvorbe.

Jednačenje po mjestu tvorbe kaže: “Suglasnici različiti po mjestu tvorbe zbog jednostavnijega se izgovora pri dodiru jednače u suglasničkim skupinama. Prvi se suglasnik skupine zamjenjuje suglasnikom koji je po mjestu tvorbe jednak drugomu suglasniku skupine.” *N* izpred *b* prelazi u *m*, a *s, z, h* izpred suglasnika *š, ž, č, ć, dž, đ, lj, nj, j* prelaze u suglasnike *š, ž* i *š*.

Tako se primjerice govori *iščupati* od *iz* i *čupati* ili *stambeni* prema *stan*.

Ta se promjena događa, ali su ne zapisuje u ovim slučajevima:

a) ako suglasnikom n ispred p i b završava prefiks: izvanbračni, izvanbrodski, izvanparnični

b) ako se suglasnik n ispred p ili b nalazi na granici dviju osnova: crvenperka, jedanput, stranputica, vodenbuba, vodenbuha

c) u riječima stranoga podrijetla koje su preuzete s n: nanbudo.

Dakle, opet neke **iznimke koje treba posebno naučiti**.

- Treba zapamtiti sve parove zvučno-bezzvučno i kada se koji mienja, a kada ne. A očito je da s tim mnogi imaju problema: Googleom sam odkrio da se na 27% (jedna četvrtina) hrvatskih stranica piše *podprogram* umjesto pravilnog *potprogram*, za srpske je stranice to 20%, za slovenske 99,99%. Što se tiče *podforuma* koji bi se trebao pisati *potforum* samo 1/1000% posto hrvatski stranica izpravno pišu, nasprem 6/1000% srbskih. Dodatno, na množtvo mjesta na Internetu ljudi se pitaju kako se pravilno piše. Ovo mnogo govori o tom pravopisu i pravo je pitanje, **zašto bismo uobće trebali razmišljati kako se piše**, kada može postojati samo jedno pravilo koje pokriva sve slučajeve?
- Pogledajmo rieč *redak*, ona se mienja po padežima ovako: *re**d**ak, re**t**ka, . . . , re**c**i/redci, . . . re**c**ima/redcima*. Oblici poput *retka* su bez veze, izgubljena je veza s osnovnim oblikom, on sad može biti *redak* ili *retak*, a pogotovo oblici sa *c – reci*, ne znaš je li to *redak*, *retak* ili *rec* ili *reći*. Zašto ne pisati informativnije?
- Dolazimo do smiešnih pojava *poTpoDnatuknica*. Uobće je smiešno da se malo piše po izgovoru, a malo ne: po**t**karpa**ts**ki. Stalno treba imati na umu kada ovako, kada onako.
- Dolazi do raznih zamućivanja značenja, kao npr. vještba (vidi kasnije). Dodatno vidi u (László i Boras, 2007).
- . . .

Še*z*desetše*st*godišnji po*st*diplomac

Hrvatski pravopis Babić-Mogušev iz 2011 kaže: “Kad bi se bezvučni šumnički skupovi st i št našli ispred zvučnih šumnika, također se jednače pošto im prethodno ispadne -t. Tako se piše vježba, izvlazben i dr. Od toga odstupaju složenice s brojevima na -st, kao šestgodišnji. (...) šestgodišnji (ne šezgodišnji)”

Ovo je tipičan oblik pravila u hrvatskim pravopisima: pravilo + (proizvoljna) iznimka. Ali da bi stvar bila hrvatskopravopisnija i iznimka mora imati iznimku, pa se tako ne odstupa od tog pravila uviek kod složenica s brojevima na -st jer moramo pisati *šezdeset* premda je tu uvjet pravila zadovoljen. Netko bi mogao prigovoriti da je ta rieč starija

od *šestgodišnji* ili *postdiplomac*, ali to nije važno jer je ta rieč tvorbena, jasno je kako se tvori, postoji sustav po kojem se tvore brojevi i *šestdeset* se liepo uklapa, možemo uviek iznova složiti tu rieč, to nam potvrđuju i pitanje piše li se *šestdeset* ili *šezdeset* koje možemo pronaći na Internetu. Postojanje takve nedoumice u govornika mnogo govori o samome pravopisnome sustavu.

Vježba je zapravo *vještba* = vještenje < vještiti se = postajati vještim u nečemu, ali zašto bi to itko znao, zašto bismo razumjeli kada možemo učiti napamet. *Vježba* može biti i glupo zapisano *vješba* = vješanje < vješati.

Nakazan *izvlazben* koji izaziva noćne more zapravo je *izvlastben* od *izvlastba* = izvlaštenje < izvlastiti, latinski ekspropriacija. Dakle, pod izgovorom znanstvenosti i inih samohvalnih izraza autori pravopisa sile nas da pišemo glupo, nerazumljivo po izgovoru i još je posebno naglašeno da se piše *izvlazben* da ne bi valjda netko slučajno napisao po razumu.

Pogledajmo kako se dakle prema sadašnjem pravopisu zapisuje kombinacija glasova s+t+d:

- šest + deset > šestdeset > šezdeset (tako se i izgovara)
- post + diplomski > postdiplomski (izgovara se pozdiplomski)

Dakle, malo pišemo po izgovoru, malo po razumu.

Tako dolazimo do *še**zd**esetše**stg**odišnji po**std**iplomac* (čitamo zd, zg, zd, ali /z/ tj. /zd/ pišemo sad ovako, sad onako)

Ovakav sustav je blesav, proizvoljan, težko učljiv i štetan!

Dodatno, ovaj je pravopis odporniji na nove glasovne promjene, dugoročno je primjenljiv. U tieku je jedna nova glasovna promjena koja nije baš iztražena, valjda je izmišljanje pravila i pseudologike pravi posao lingvista, a ne istraživanje jezika. Uglavno, možemo čuti na TV npr.: *drai gleatelji* umjesto *draGI gleDatelji*. U nekim sklopovima izpadaju neki suglasnici. Ta se promjena može proširiti, to nitko ne zna, a zar ćemo onda mienjati pravopis i pisati u inače *drag*, a u ovom obliku *drai* i tako dalje s promjenama u govoru, pa novi naraštaji neće moći čitati starije zapise. Bolja je opcija pravopis koji ne ovisi puno o izgovoru.

Ima i pravopisa gdje sve živo i neživo izpada, po nekima je posve normalno pisati *bici* umjesto *bitci*, pa si ti misli što je to, ili *počeci* ne znaš je li to *početak* ili *poček*, jesu li *lisci* mužke lisice (lisac) ili mali listovi *listci* (listak), posve se normalno piše *mlaci* bilo to *mladac*, *mlatac* ili *mlaka* itd. Takvo pisanje dodatno onemogućava "pogađanje" naglaska, npr. ako piše *početci* onda znam da je to vj. od *početak* i znam da je naglasak na drugom slogu, ako piše *počeci* znam da je naglasak na prvom slogu, ali ako se jedno i drugo isto piše onda ne znam. Nemoguće je iz samoga oblika rekonstruirati polazištnu rieč, a moglo bi biti moguće. Gubi se, dakle, veza s osnovnom rieči, moramo pogađati od čega dolazi,

a to ne mora uviek biti očito, recimo *snopovršče* od čega je to zovnik, od *snopovrzac* ili *snopovršac* ili *snopovrstac* ili *snopovržac*, tko će ga znati (=bezvezan gubitak obaviesti, neka si čitatelj misli što je pisac htio reći).

Pravopisi s izpadanjem glasova neriedko i pogrešno zapišu po uhu. *Zadatci* se može čitati [zadacci] s dugim c ili [zadaci]. Često se u ovakvim slučajevima čita s dugim c. Ali ako zapišemo *zadaci* više ni ne znamo o kojoj se rieči radi, a kamoli kako se čita, iako je to bila ideja.

Prednosti čitkotvornog pravopisa

1. Točniji zapis (recimo pisanje glasa /ie/).

2. Mnogo manje pravila i iznimaka, sva gore navedena pravila nestaju, što znači lakša (na)učljivost, mnogo manje toga treba zapamtiti.

3. Obavjestniji zapis, iz samoga je zapisa jasno kako je rieč nastala, njezina tvorba je prozirnija, time je i jasnije što rieč znači. Osim što je jasnije što rieč znači, jasnije i kako tvorbeni sustav u hrvatskome radi. Mnogi će reći "Težko je pisati po ovome pravopisu, to je samo za stručnjake…", ali nije tako. Samo treba razmisliti o tome što govorimo, o tome što mislimo, ovaj nas pravopis tjera da prestanemo gledati rieči kao neprozirne tvorevine i da uočavamo kako su rieči međusobno povezane, kako rastu jedne iz drugih, kako se jezik razvija. Time se odmah intimno upoznajemo s hrvatskom tvorbom rieči – rječotvorbom i time postajemo kadri i sami stvarati nove rieči i razumjeti nepoznate rieči, bilo nove bilo stare. Nekoliko se slučajeva koji nisu na prvu jasni lako nauči, ako jednom naučimo da je *vježba* zapravo *vještba*, tj. da dolazi od *vještiti se*, to nikada nećemo razboraviti za razliku od pravila o glasovnim promjenama ili pravila o ije i je. Ovo smatram najvažnijom prednošću ovoga pravopisa. Zamislimo da u školama ne trošimo bezbrojne sati nastojeći zapamtiti glasovne promjene koje nesvjestno ionako znamo i kada se one bilježe, a kada ne, da na kraju velik broj ljudi opet ne zna to, i da umjesto toga učimo kako se rieči tvore, kako jedna rieč nastaje od druge, da učimo kako se razvijao jezik, moglo bi se i pokazati kako se od staroslavenskog došlo do hrvatskog i kako se pretvorio u druge slavenske jezike i time stvoriti jaku podlogu za njihovo usvajanje, da shvatimo što zapravo govorimo kako bi iz škole izašli s izrazito solidnim razumievanjem vlastitog jezika i sposobni tvoriti nove rieči i razumieti nepoznate, a ne s recitiranjem glasovnih promjena i proizvoljnih pravila o tome kada se bilježe, a kada ne. **Svrha pravopisa ne smije biti da se ima što pitati na izpitu, svrha je pravopisa da zabilježi značenje i on mora biti što usvojljiviji i koristniji.**

4. Ne smeta izgovoru jer se po definiciji ne bilježe samo one glasovne promjene koje govornik jezika ionako već nesvjestno zna u svom ‘govornom aparatu’. Za razliku od ovog drugog pravopisa koji se kao temelji na *piši po uhu*, a zbog toga i svoje nedosljednosti često navodi na pogrešan izgovor.

Jedna je česta kritika da je ovaj pravopis **nenaučljiv**. Već sam gore kod prednosti opisao da baš i nije tako, ovaj pravopis ima mnogo manje pravila i iznimaka, a ona se dodatna obaviest o postanku rieči koja je potrebna u nekim slučajevima lako usvoji. Tvrdnja o nenaučljivosti je lažna ili barem neprovjerena. Ali ono što nije neprovjereno jest da je trenutačni pravopis nenaučljiv, akademski obrazovani građani ne znaju pisati po ovome pravopisu i to nema nikakve veze s raznolikošću pravopisa. O tome koliko će se naučiti ovisi o školi, englezki je pravopis najgori na svietu pa ga svi više ili manje nauče. Dodatno, slovenski se piše po sličnome pravopisu pa slovenci ne izlaze nepismeni iz š(k)ola. Možemo zaključiti da je početna tvrdnja neistinita i dodatno da je to jeftino podmetanje laži (ili barem neprovjerenosti) pod istinu.

Sažetak poglavlja

- Ja ovo ne radim zbog nacionalizma i sl.
- Sve se ovo može odnositi na bilo koji jezik. Smatram da svaki jezik treba raditi na očuvanju svoje posebnosti i povećanju raznolikosti, pomogućnosti vlastitim sredstvima, na korist sviju.
- Ne kažem ja da trenutačni jezik ne valja, nego da može biti bolji, a sigurno je bolji ako nam ne treba rječnik drugog jezika da bismo ga razumjeli.
- U redu je miešati se u jezik, kao što to radimo i sa svime ostalima kada imamo koristi od toga
- Prevođenjem rieči bolje razumijemo što govorimo, rieči su povezanije i lakše naučljive, pojmovi su nam u glavi povezaniji, imamo mogućnost stvaranja boljih asocijacija itd.
- Dosta je demonizacije stvaranja rieči, to je najnormalnija pojava i jedan oblik izražavanja
- Prevođenje stranih rieči nije glupo, nego je pametno, nerazumljive rieči zamjenjujemo razumljivima. Zašto odabirati nizove slogova koji nam ništa ne znače, koji ne upućuju na pojam (kao što to čine u izvornom jeziku), kada možemo odabrati rieči koje nam znače nešto, koje nasu upućuju na značenje

– Ne kažem ja da treba izbaciti, prognati “nepodbne” rieči, kažem samo da treba dati priliku domaćim riečima i neka ih rabi tko hoće, a ne osuđivati ih
– Ne kažem da treba uviek govoriti u hrvatskim riečima, niti taj pojam ima previše smisla, proizvoljan je; ono što hoću reći je: ako možeš od postojećih (po mogućnoti što hrvatskijih) rieči napraviti rieč za novi pojam, onda to i učini, tako da su rieči povezanije i razumljivije
– Kada se nazivaju novi pojmovi onda se odabiru protumačljive rieči za taj jezik, a normalno bi bilo da ostali jezici nazovu pojam svojim riečima kako bi razumjeli što govore
– Nove rieči nisu uglavnom smiešne, nego su samo razumljive i nove i mi se se moramo nasmijati jer se to očekuje. *Pticoslovlje* nije ništa smješnije od *ornitologije*.
– I drugi jezici prevode, a posebice je zanimljivo da grčki prevodi latinske rieči, on ne prihvaća nekakve mješane grčko-latinske barbarizme ili pogrešne tvorbe nego naziva stvari riečima koje razumije
– Ljudi su slabo upoznati s tvorbenim mogućnostima hrvatskoga što je krivica škole
– Sadašnji je pravopis nedosljedna, bezpotrebno složen, težko naučljiv, loš u svojoj osnovnoj funkciji zapisivanja značenja itd.
– Sadašnji nas pravopis tjera da temeljito usvojimo i svaki čas imamo na umu 80-ak pravila prirodnog glasoslovlja, kako se koji glas pred kojim, i u koje u nas pretvara, a to ionako ne možemo drugačije izgovoriti. Dakle, učimo pravila radi pravila, nastava je obremenjena obilatom, bezplodnom i nesvrhovitom građom (László i Boras, 2007).
– Sadašnji nas pravopis občarava tobožnjom znanstvenošću veoma velikog broja pravila za pisanje naravnoglasovnih promjena (László i Boras, 2007).
– Bolji je pravopis onaj koji traži učenje što manje pravila.
– Čitkotvorni pravopis je rješenje, ima mnogo manje pravila, što znači da ga je lakše naučiti, obavjestniji je, više smo obaviesti pohranili, razumijemo što rieči znače, kako su nastale, odmah se upoznajemo i s tvorbenim sustavom, bolje razumijemo razvoj jezika, odporniji je na nove glasovne promjene, …
– Važna napomena: rieči se čitaju kako se i inače čitaju, samo je zapis drukčiji
– Uviek imajmo na umu jedno od osnovnih pravila života: Stvari nisu istinite zato što netko kaže da jesu, tko god on bio. Stvari nisu istinite zato što ja želim da jesu.

JEZIČNI MODEL HRVATSKOGA JEZIKA ZASNOVAN NA POVRATNIM NEURONSKIM MREŽAMA

Sažetak

Jezični modeli procjenjuju vjerojatnost pripadanja nekog slieda rieči u neki jezik. Tradicionalni se modeli temelje na prebrojavanju pojavljivanja nizova od *n* rieči (n-rječja), no ti modeli pate od nedostatka podataka, nemogućnosti modeliranja udaljenih odnosa, nemogućnosti poobćavanja itd. Alternativa su modeli temeljeni na živčanim mrežama koji prikazivanjem rieči pomoću vektora rješavaju glavninu problema tradicionalnih modela. Ti rječni prikazi mogu se izkoristiti i u drugim zadatcima obrade prirodnog jezika. U radu su izrađeni rječni prikazi za hrvatske rieči pomoću naputka word2vec te su izpitani na zadatcima prepoznavanja suznačnica, ocjenjivanja značbene povezanosti i sličnosti, skladnjanih i značbenih nalika te na sustavu za prepoznavanje imenovanih sućaka u hrvatskom jeziku CroNER-u. U svim je slučajevima (osim kod CroNER-a) došlo do znatnog poboljšanja točnosti u odnosu na sustav bez vektorskih prikaza rieči.

Ključne riječi: obrada prirodnog jezika, jezični modeli, živčane mreže, rječni prikazi, hrvatski jezik

RECURRENT NEURAL NETWORK BASED MODEL OF CROATIAN LANGUAGE

Abstract

Language models are used to estimate the probability of a word sequence belonging to some language. Traditional models are based on counting the number of occurrences of n-word long sequences (n-grams). These models, however, suffer from insufficient data, inability to model distant relations or to generalize etc. An alternative approach are neural network based models which use vector representations of words to solve most of the traditional model's problems. These word representations can also be used in other natural language processing tasks. In this thesis we built word representations of Croatian words using word2vec software and tested them on the tasks of synonym detection, semantic similarity and relatedness judgment, syntax and semantic analogies and named entity recognition in Croatian using CroNER. In all the cases (except CroNER) we observed significant increase in accuracy as compared to the systems without vector representations of words.

Keywords: natural language processing, language models, neural networks, word representations, Croatian language