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. 1155

Duboko učenje vektorskih reprezentacija riječi za modele označavanja tekstova na hrvatskome jeziku

Goran Gašić

Zagreb, srpanj 2015.

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

Zagreb, 6. ožujka 2015.

Predmet: **Strojno učenje**

DIPLOMSKI ZADATAK br. 1155

Pristupnik: **Goran Gašić (0036454355)**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Duboko učenje vektorskih reprezentacija riječi za modele označavanja tekstova na hrvatskome jeziku**

Opis zadatka:

Semantičke reprezentacije riječi svaku riječ prikazuju niskodimenzijskim vektorom tako da semantički slične riječi imaju slične vektore. Takve su se reprezentacije pokazale iznimno korisnima u nizu zadataka obrade prirodnog jezika, uključivo leksičkosemantičkim zadatcima i ekstrakciji informacija. U posljednje vrijeme osobito su se uspješnima pokazale semantičke reprezentacije generirane nadziranim modelima strojnog učenja (engl. word embeddings), posebice neuronskim mrežama i modelima dubokoga strojnog učenja.
U okviru diplomskoga rada potrebno je proučiti semantičke reprezentacije generirane neuronskim mrežama i modelima dubokoga učenja. Izgraditi semantičke reprezentacije za riječi hrvatskoga jezika korištenjem javno dostupnih korpusa. Proučiti pristupe integriranja semantičkih reprezentacija u modele za polunadzirano slijedno označavanje tekstova, uključivo modele višezadaćnog učenja. Razviti odgovarajuću programsku implementaciju modela za označavanje vrsta riječi i označavanje imenovanih entiteta u tekstovima na hrvatskome jeziku. Provesti iscrpno eksperimentalno vrednovanje modela na ispitnim skupovima podataka te analizu pogrešaka. Radu priložiti izvorni i izvršni kod razvijenog sustava, skupove podataka i programsku dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 13. ožujka 2015.
Rok za predaju rada: 30. lipnja 2015.

Mentor:

Doc. dr. sc. Jan Šnajder

Djelovođa:

Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za diplomski rad profila:

Prof. dr. sc. Siniša Srbljić

Zahvaljujem doc. dr. sc. Janu Šnajderu na mentorstvu tijekom studija te roditeljima na neizmjernoj podršci.

SADRŽAJ

1. Uvod

Duboko učenje kao grana strojnog učenja u posljednjih nekoliko godina obara sve vrhunske rezultate u obradi prirodnog jezika (engl. *natural language processing*, NLP). Konj za trku strojnog učenja su umjetne neuronske mreže, koje istražujemo u ovom radu. U području obrade prirodnog jezika, njima su nadopunjujuće vektorske reprezentacije riječi, koje su aktivan predmet opsežnog istraživanja budući da nose bogate sintaktičke i semantičke odnose između riječi. Nije iznenađujuće da je daleko najviše istraživanja u smjeru dubokog učenja provedeno na zadacima obrade teksta na engleskom jeziku. Cilj ovog rada jest izgradnja modela za niz zadataka označavanja tekstova na hrvatskome jeziku koji se zasniva na vrhunskim metodama dubokog učenja, uspješno primijenjenim na engleskom jeziku. Našim radom dajemo tri bitna znanstvena doprinosa obradi prirodnog jezika za hrvatski jezik, pojašnjena u nastavku:

- Dajemo preglednu usporedbu četiri tipa vektorskih reprezentacija u svrhu označavanja tekstova na hrvatskome jeziku. Prethodno svaki od njih uvjerljivo podiže vrhunske rezultate na zadacima gdje je primijenjen. Držimo da će naša dokumentirana evaluacija biti od neizmjerne pomoći u istraživanju njihove primjene na hrvatskom jeziku. Sve reprezentacije javno objavljujemo;
- Kombiniramo te činimo javno raspoloživima sve javno dostupne označene skupove podataka za niz zadataka označavanja teksta na hrvatskome jeziku u svrhu poticanja istraživanja na ovom području;
- Naši modeli po prvi put dubokim učenjem postižu vrhunske ili njima bliske rezultate na nizu zadataka označavanja teksta na hrvatskome jeziku. Iscrpno dokumentiramo implementaciju u svrhu dalje prilagodbe i usavršavanja.

Struktura rada je sljedeća. U drugom poglavlju dajemo pregled teorijske pozadine rada. Definiramo duboko učenje te opisujemo kako je ova nekoć zapuštena grana strojnog učenja dosegla današnju popularnost zahvaljujući umjetnim neuronskim mrežama. Dalje predstavljamo koncept vektorskih reprezentacija riječi, nužnih za uspješnu primjenu dubokog učenja na zadatke obrade prirodnog jezika. Dajemo pregled algoritama

za efikasno učenje vektorskih reprezentacija riječi na ogromnim neoznačenim skupovima podataka. Nadalje, predstavljamo važne zadatke obrade prirodnog jezika u čiju svrhu razvijamo naše modele.

U trećem poglavlju dajemo pregled srodnih radova u literaturi. Predstavljamo radove koji istražuju primjenu dubokog učenja za obradu prirodnog teksta na engleskom jeziku te time motiviraju naš rad na hrvatskom jeziku. Dalje dajemo pregled prethodnog rada na nizu zadataka obrade prirodnog jezika na hrvatskom jeziku.

U četvrtom poglavlju predstavljamo naš pristup izgradnji modela za niz zadataka kojima se bavimo. Ugrubo ga možemo podijeliti na dva koraka; nenadzirano učenje vektorskih reprezentacija te nadzirano učenje modela označavanja tekstova.

U petom poglavlju podrobno opisujemo implementaciju prethodno definiranog pristupa. Ovdje se mahom oslanjamo na postojeće alate te Pylearn2 biblioteku za duboko učenje koji su slobodni za korištenje (engl. *open source*). Korištenjem navedene biblioteke izgrađujemo konačne modele te navodimo sve parametre učenja.

U šestom poglavlju opisujemo metodologiju evaluacije izgrađenih modela u svrhu pouzdane usporedbe različitih vektorskih reprezentacija za naše zadatke obrade prirodnog jezika. Dajemo detaljan pregled korištenih skupova podataka na svim zadacima; odlične performanse modela zahvaljujemo upravo kombiniranju svih javno dostupnih skupova podataka. Dajemo potpun pregled rezultata unakrsne provjere te konačnih rezultata na ispitnim skupovima.

2. Teorijska pozadina

2.1. Duboko učenje

Duboko učenje grana je strojnog učenja koja razvija modele složene od više nelinearnih transformacija, što omogućava učenje reprezentacija podataka na više razina apstrakcije. Najpopularniji takav model je umjetna neuronska mreža s više skrivenih slojeva, inspirirana neokorteksom čovjeka. Premda je duboko učenje započeto u 80-im godinama prošlog stoljeća, tek nakon što Hinton et al. (2006) značajno optimiziraju algoritam učenja, umjetne neuronske mreže stječu široku popularnost. Uspjeh umjetnih neuronskih mreža najviše možemo zahvaliti Mooreovom zakonu (Schaller, 1997), koji predviđa udvostručavanje gustoće tranzistora u mikroprocesoru svake dvije godine te još uvijek vrijedi nakon 50 godina. Eksponencijalan rast procesorske moći te prodor pristupačnog računarstva u oblaku (engl. *cloud computing*) omogućio je učenje modela koji savladavaju zadatke obrade slika, zvuka i teksta na neviđenim količinama podataka. Nadalje, njihovom uspjehu doprinosi razvoj računanja na grafičkim karticama (engl. *GPU computing*). Ciresan et al. (2011) razvijaju GPU implementaciju umjetne neuronske mreže koja poražava prethodne rezultate klasifikacije slika na standardnom skupu podataka MNIST.

2.1.1. Umjetna neuronska mreža

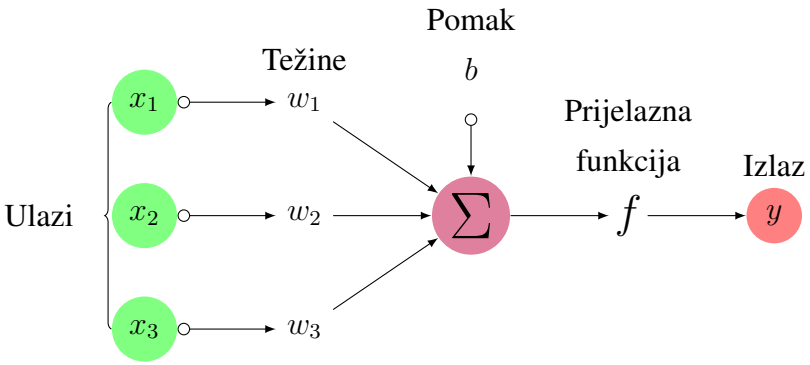
Umjetna neuronska mreža statistički je model inspiriran biološkim neuronskim mrežama. Ovakve mreže naročito su izražajne te imaju sposobnost aproksimacije proizvoljnih funkcija. Umjetnu neuronsku mrežu možemo promatrati kao graf gdje svaki čvor predstavlja jedan neuron. Svakom neuronu pridružujemo niz parametara koje prilagođavamo tijekom učenja. Na slici 2.1 prikazujemo općenit primjer neurona. Niz neurona $[x_1, x_2, x_3]^T$ s lijeva daje ulaz neuronu u središtu. Svaki ulaz množimo težinom na odgovarajućem bridu te njihovoj sumi dodajemo vrijednost pomaka neurona. Pomak možemo promatrati kao izglednost da izlaz čvora bude jednak 1, čime modeli-

ramo sposobnost aktivacije biološkog neurona. Konačno, vrijednost izlaza y dobivamo po jednadžbi 2.1.

$$y = f(\sum_i w_i x_i + b) \tag{2.1}$$

Potpunu neuronsku mrežu izgrađujemo nizanjem slojeva neurona kao na slici 2.2. Ovdje čvorovi na ulaznom sloju daju ulaz za čvorove u skrivenom sloju, čije parametre optimiramo učenjem modela.

Samo učenje umjetne neuronske mreže tipično se provodi stohastičkim gradijentnim spustom, koji najprije zahtijeva definiranje funkcije pogreške kao što je kvadratna funkcija pogreške $E = (t - p)^2$ gdje je t točna vrijednost izlaza, dok je p predviđena vrijednost. Zatim je potrebna računski učinkovita metoda izračuna gradijenta. Najčešće koristimo propagaciju pogreške unatrag (engl. *backpropagation*) (Rumelhart et al., 1988).



Slika 2.1: Umjetni neuron

2.2. Vektorske reprezentacije riječi

Vektorske reprezentacije riječi (engl *word embeddings*) su prikaz riječi u prirodnom jeziku pomoću niskodimenzijskih vektora realnih brojeva kojima pokušavamo predstaviti sintaktičke i semantičke odnose između riječi. Takve reprezentacije dobivene nenadziranim učenjem na ogromnim količinama teksta u proteklih su nekoliko godina korištene u raznim sustavima za obradu prirodnog jezika (engl. *natural language processing*, NLP) te uspješno primijenjene za poboljšanje rezultata na raznim zadacima. U nastavku opisujemo četiri korištena tipa vektorskih reprezentacija; reprezentacije dobivene Brownovim grupiranjem, modelom kontinuirane vreće riječi, Skip-gram modelom te Collobert-Weston modelom.

2.2.1. Brownovo grupiranje

Brownovo grupiranje (engl. *Brown clustering*) je metoda hijerarhijskog grupiranja riječi na osnovu konteksta u kojima se pojavljuju u tekstu. Grupiranje se provodi "odozdo prema gore" (engl. *bottom-up*), odnosno od svake riječi u svojoj grupi. Algoritam iterativno gradi binarno stablo spajajući grupe riječi na osnovu logaritamske vjerojatnosti teksta uz zadano grupiranje. Logaritamsku vjerojatnost teksta dobivamo predstavljanjem zajedničke vjerojatnosti $P(w_1, \ldots, w_n)$ umnoškom uvjetnih vjerojatnosti $P(w_1)P(w_2|w_1)\ldots P(w_n|w_{n-1}\ldots w_1)$. Zatim ovu vjerojatnost aproksimiramo bigramima, tj. kao kontekst uzimamo tek prethodnu riječ, te logaritmiramo radi lakšeg računa. Algoritam 1 prikazuje naivnu implementaciju Brownovog grupiranja. $Q(C)$ predstavlja kvalitetu grupiranja C dobivenu jednadžbom 2.2.

Algoritam 1 Brownovo grupiranje

Ulaz: riječi $(w_1, ..., w_n) \in V$, broj grupa $K \in \mathbb{N}$

Izlaz: binarno stablo gdje listovi odgovaraju riječima

za sve $v_i \in V$ **radi**

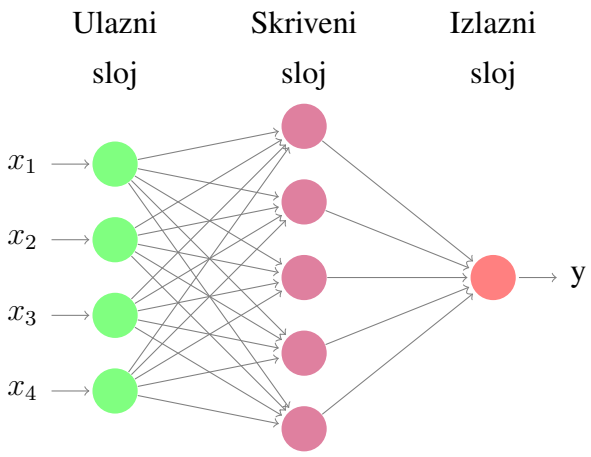
$C_i := i$

završi

dok postoje dvije različite grupe **radi**

spoji dvije grupe koje maksimiziraju $Q(C)$

završi



Slika 2.2: Umjetna neuronska mreža s jednim skrivenim slojem

$$
\begin{aligned}
Q(C) &= \frac{1}{n} log P(w_1, ..., w_n) \\
&= \frac{1}{n} log P(w_1, ..., w_n, C(w_1), ..., C(w_n)) \\
&= \frac{1}{n} log \prod_{i=1}^{n} P(C(w_i)|C(w_{i-1}) P(w_i|C(w_i))
\end{aligned} \tag{2.2}
$$

Prikazanu implementaciju nazivamo naivnom zbog velike vremenske složenosti, koja iznosi $O(|V|^5)$, gdje je V skup riječi u tekstu. Provodimo $O(|V|)$ iteracija, u svakoj razmatramo svake dvije grupe u složenosti $O(|V|^2)$ te računamo kvalitetu grupiranja nastalog njihovim spajanjem u složenosti $O(|V|^2)$. Brown et al. (1992) spuštaju složenost na $O(|V|^3)$ održavanjem matrice promjene kvalitete grupiranja uz svako od $O(|V|^2)$ spajanja. Za odabir najboljeg spajanja dovoljno je pronaći polje s najvećom vrijednosti u složenosti $O(|V|^2)$. Nakon spajanja matricu je također moguće osvježiti u složenosti $O(|V|^2)$. Liang (2005) dalje spušta složenost implementacije algoritma na $O(K^2|V|)$, ograničavanjem broja grupa na K u svakom koraku algoritma.

Turian et al. (2009) izgrađuju vektorske reprezentacije riječi iz Brownovih grupa odabirom čvorova na zadanoj dubini u stablu. Svaki takav čvor odgovara nizu riječi u rječniku, kojoj pridjeljuju jednaku vektorsku reprezentaciju iz uniformne distribucije $\mathcal{U}(-1, 1)$. Ne ograničavaju se na jednu dubinu, već odabiru dubine 4, 6, 10, 20 te svakoj riječi pridjeljuju više vektorskih reprezentacija koje potom spajaju. Primijetimo da je nužno dalje učiti ovakve reprezentacije u NLP sustavu gdje su primijenjene jer su konkretne vrijednosti samih vektora u početku tek šum. Njihova izražajnost dolazi iz Brownovog grupiranja, odnosno mapiranja riječi u indekse u dijeljenoj matrici vektorskih reprezentacija.

2.2.2. Mikolov at al.

Mikolov et al. (2013a) predstavljaju dvije arhitekture modela za računanje vektorskih reprezentacija riječi na ogromnim skupovima podataka te pokazuju da nadmašuju prethodno najbolje rezultate na zadatku sličnosti riječi uz daleko manju vremensku složenost računanja. Najprije opišimo motivaciju za ovaj model. Bengio et al. (2003) prvotno predlažu arhitekturu modela zasnovanu na neuronskoj mreži kojom istovremeno uče jezični model (engl *neural network language model*, NNLM) te vektorske reprezentacije riječi, time savladavajući "prokletstvo dimenzionalnosti" (engl. *curse of dimensionality*). Njihov model najpopularnija je arhitektura dubokog učenja za obradu prirodnog jezika:

1. Svakoj različitoj riječi koja se pojavljuje u skupu podataka pridijelimo vektorsku reprezentaciju, vektor realnih vrijednosti u $\mathbb{R}^n$,

2. Izrazimo zajedničku vjerojatnost nizova riječi u tekstu kao funkciju njihovih vektorskih reprezentacija,

3. Istovremeno učimo vektorske reprezentacije riječi te parametre funkcije zajedničke vjerojatnosti modelom umjetne neuronske mreže.

Ovaj model čine četiri sloja:

- Ulazni sloj koji prima indekse prethodnih N riječi u rječniku veličine V,
- Projekcijski sloj koji svaki indeks pridružuje vektoru u tablici širine D,
- Skriveni sloj širine H,
- Izlazni sloj širine V.

Složenost učenja na jednom primjeru Q stoga iznosi

$$Q = ND + NDH + HV \tag{2.3}$$

dok je ukupna složenost učenja O uz broj epoha E te broj primjera T jednaka

$$O = ETQ \tag{2.4}$$

Dominantan izraz HV Mikolov et al. (2013a) zanemaruju korištenjem hijerarhijske softmax aktivacijske funkcije (Morin i Bengio, 2005) te zapisujući rječnik kao Huffmanovo stablo prema frekvencijama riječi u tekstu, što smanjuje veličinu izlaznog sloja na $O(\log V)$ te dominantan izraz postaje NDH. Mikolov et al. (2013a) dalje istražuju u smjeru jednostavnijih modela bez skrivenog sloja, koji učenje na ogromnim količinama podataka čini nepraktičnim. Domingos (2012) potvrđuje motivaciju za iskorištavanje što većih skupova podataka na jednostavnijim modelima u svrhu nadmašivanja najboljih rezultata na zadacima strojnog učenja. Primijetit ćete da je ovo jednostavno pravilo poslužilo kao nit vodilja u nastavku rada. Stoga, Mikolov et al. (2013a) razvijaju arhitekture koje uče NNLM u dva koraka. Najprije uče vektorske reprezentacije riječi na jednostavnom modelu neuronske mreže. Zatim uče NNLM na N-gramima iskorištavajući prethodno naučene reprezentacije, što nećemo dalje opisivati jer nije predmet našeg rada. Opišimo samo obje arhitekture jednostavnih modela neuronske mreže korištenih za računanje vektorskih reprezentacija riječi u nastavku.

Kontinuirana vreća riječi

Kontinuirana vreća riječi (engl *continuous bag-of-words*, CBOW) prvi je od dva modela. Arhitektura prati standardni NNLM, uz sljedeće razlike:

– Uklonjen je skriveni sloj;

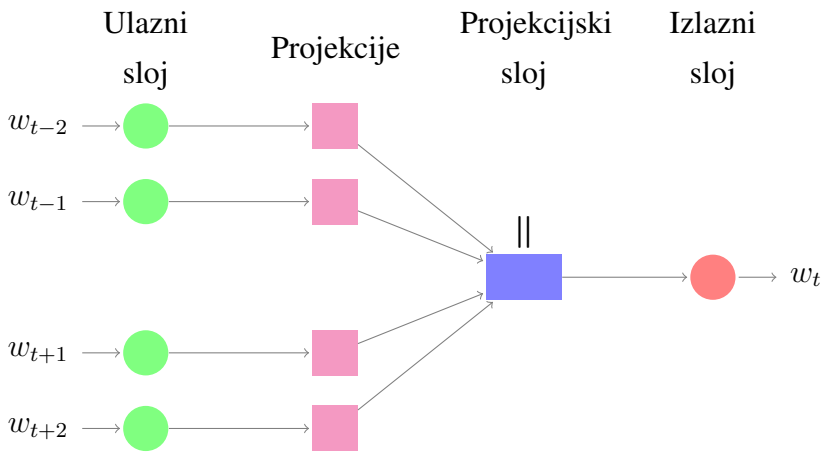
– Projekcijski sloj širine D dijele sve riječi na ulazu. Preciznije, njihove vektorske reprezentacije uprosječene su u projekcijskom sloju. Kao posljedica, redoslijed riječi u ulazu je nebitan, što objašnjava vreću riječi u nazivu modela;

– Na ulaz dovodimo ne samo riječi koje prethode trenutačnu, već i riječi koje slijede. Kriterij učenja je točna klasifikacija riječi u središtu prozora.

Složenost učenja na jednom primjeru Q jest

$$Q = ND + D \log V \tag{2.5}$$

Arhitektura modela prikazana je na slici 2.3.



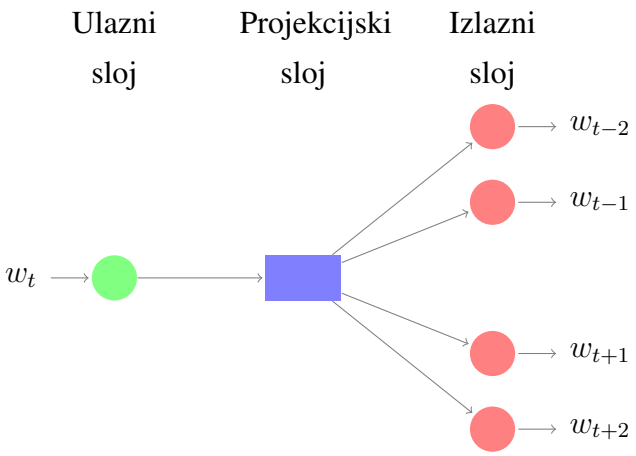
Slika 2.3: Arhitektura modela CBOW

Skip-gram

Skip-gram je model dobiven obratom kontinuirane vreće riječi. Preciznije, na ulaz dovodimo trenutačnu riječ te učimo predvidjeti ostale riječi u prozoru. Budući da je u izlazu preskočena trenutačna riječ, model se naziva skip-gram (od engl. *skip*, preskočiti). Složenost učenja na jednom primjeru Q jest

$$Q = C(D + D \log V) \tag{2.6}$$

gdje je C najveća udaljenost riječi u izlazu od trenutačne riječi. U svakom koraku učenja širina prozora riječi u izlazu nasumično je odabrana te nije veća od $2C$. Na slici 2.4 prikazana je arhitektura modela.



Slika 2.4: Arhitektura modela Skip-gram

2.2.3. Collobert-Weston

Collobert i Weston (2008) razvijaju model koji uči razlikovati nizove riječi u tekstu od iskvarenih nizova riječi. Funkcija pogreške ovisi o razlici izlaza mreže za iskvaren te ispravan niz riječi. Preciznije, neka je dan niz riječi $S = [w_{t-n}, ..., w_t, ..., w_{t+n}]$ u korpusu T. Iskvaren niz riječi S' dobivamo zamjenom srednje riječi w_t slučajnom riječi u rječniku $\tilde{w}_t$. Umjetna neuronska mreža predstavlja funkciju f koja pridjeljuje rezultat svakom nizu riječi. Konačno, parametre modela osvježavamo po funkciji gubitka zglobnice:

$$J(T) = \frac{1}{|T|} \sum_{t \in T} |1 - f(S') + f(S)| \tag{2.7}$$

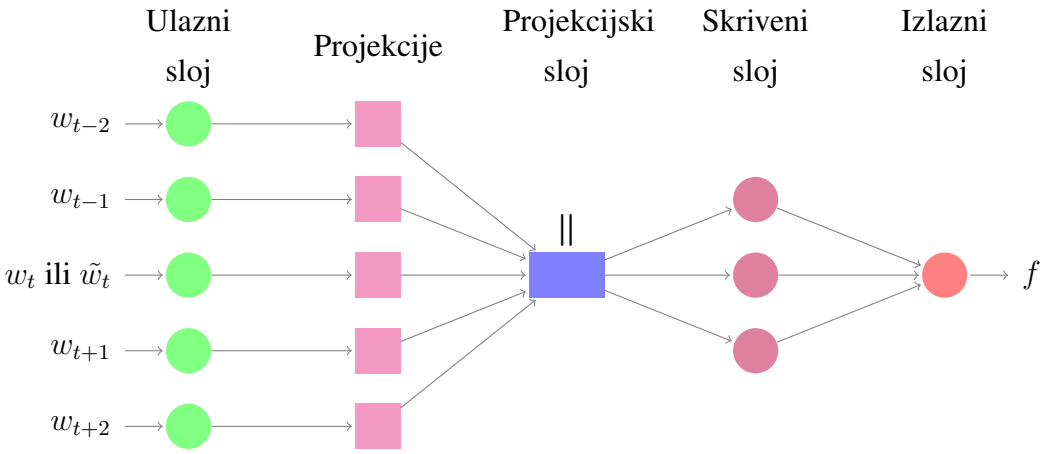
Na slici 2.5 prikazan je model umjetne neuronske mreže koji računa rezultat uz širinu prozora n jednaku 5. Svakoj riječi u ulazu najprije je pridijeljen indeks u rječniku V. Zatim njime indeksiramo vektorsku reprezentaciju riječi u dijeljenoj matrici C dimenzija $|V| \times M$, gdje su dimenzije vektorske reprezentacije svake riječi $1 \times M$. Potom sve vektorske reprezentacije spajamo u projekcijski sloj P veličine $2(n+1)*M$, što označavamo simbolom ||. Projekcijski sloj daje ulaz za skriveni sloj veličine $|H|$, gdje je aktivacijska funkcija A jednaka

$$A = \tanh(W_1 P + b_1) \tag{2.8}$$

gdje su W_1 i b_1 težine te pomak skrivenog sloja, tim redoslijedom. Rezultat niza riječi dobivamo linearnom kombinacijom aktivacija A u skrivenome sloju:

$$f(P) = W_2 A + b_2 \tag{2.9}$$

gdje su W_2 te b_2 težine te pomak izlaznog sloja. Dakle, svi parametri koje učimo su W_1, W_2, b_1, b_2, C, dok je ukupan broj parametara $(2n+1) * M * H + H + H + 1 + |V| * M \approx M * (nH + |V|)$.



Slika 2.5: Arhitektura Collobert-Westonova modela

2.3. Obrada prirodnog jezika

U neuropsihologiji i lingvistici prirodnim ili običnim jezikom nazivamo svaki jezik nesvjesno razvijen u mozgu skupine ljudi. Ovakve jezike ljudi koriste za sve oblike komunikacije; komunikaciju govorom, znakovno, dodirom ili pismom. Razlikujemo ih od konstruiranih jezika koje ljudi svjesno razvijaju, te formalnih jezika, korištenih za programiranje računala ili proučavanje područja logike. Obrada prirodnog jezika je područje računarske znanosti, umjetne inteligencije te računalne lingvistike koje proučava interakciju računala te prirodnog jezika, blisko povezano s područjem interakcije čovjeka i računala (engl. *human-computer interaction*, HCI). U nastavku podrobnije opisujemo zadatke obrade prirodnog jezika koje istražujemo u radu: označavanje vrsta riječi (engl. *part-of-speech tagging, POS tagging*), označavanje morfosintaktičkih

opisa (engl. *morpho-syntactic descriptors, MSD tagging*) te prepoznavanje imenovanih entiteta (engl. *named entity recognition*, NER).

2.3.1. POS označavanje

Označavanje vrsta riječi ili POS označavanje je proces pridruživanja vrste riječi svakoj pojavnici (engl. *token*) u tekstu, npr. glagol, imenica, pridjev i sl. Vrsta riječi najbitnija je u razrješavanju homonima u tekstu u svrhu dalje obrade. Algoritme za POS označavanje možemo podijeliti u dvije kategorije:

- Algoritmi zasnovani na pravilima poput prvoga algoritma za POS označavanje teksta na engleskom jeziku (Brill, 1992),
- Stohastički algoritmi zasnovani na statističkoj analizi teksta te vjerojatnosti. Najpopularniji primjer su skriveni Markovljevi modeli (engl. *Hidden Markov models*, HMMs)

2.3.2. MSD označavanje

Označavanje morfosintaktičkih opisa ili MSD označavanje primjenjuje se na jezike visoke fleksije. Fleksija je prilagodba oblika riječi u svrhu izražavanja različitih gramatičkih kategorija kao što su vrijeme, brojnost, rod, padež i sl. MSD oznaka opisuje riječi po gramatičkim kategorijama. Hrvatski jezik ima iznimno visok stupanj fleksije:

- Imenice deklliniramo po sedam padeža u jednini i množini;
- Glagole konjugiramo po vremenu, licu, broju i rodu;
- Pridjeve deklliniramo s imenicama u sedam padeža te različitoj brojnosti, rodu, stupnju i određenosti.

Poput POS označavanja, na zadatku su najuspješniji skriveni Markovljevi modeli.

2.3.3. NER označavanje

Prepoznavanje imenovanih entiteta ili NER označavanje je zadatak pronalaženja uzastopnih pojavnica u tekstu koje čine imenovane entitete te njihovo klasificiranje u različite kategorije. Najčešće su korištene osnovne tri kategorije: nazivi osoba, lokacija ili organizacija. Osim njih, ovdje spadaju kategorije poput vremenskih izraza kao što su datumi ili numeričkih izraza za novac, postotke i sl. Primijetite da posljednje dvije kategorije intuitivno ne svrstavamo u imenovane entitete. Usprkos tome, korištene su

ponajprije iz praktičnih razloga; one nose vrijednu jezičnu informaciju u obradi prirodnog jezika. Pristupe NER označavanju možemo podijeliti u dvije kategorije:

- Algoritmi zasnovani na ručno osmišljenim jezičnim gramatikama pružaju visoku preciznost uz manji odziv. Osmišljavanje takvih algoritama zahtijeva duboko poznavanje računalne lingvistike;
- Označavanje statističkim modelima naučenim algoritmima strojnog učenja. Najpopularniji takav klasifikator su uvjetna slučajna polja (engl. *conditional random fields*, CRF) (Lafferty et al., 2001).

3. Srodni radovi

Srodne radove podijelit ćemo u dvije kategorije radi preglednosti. Najprije dajemo pregled radova na engleskom jeziku koji su korištenjem vektorskih reprezentacija riječi za obradu prirodnog jezika motivirali pristup u našem radu. Zatim dajemo pregled prethodnog rada na konkretnim zadacima obrade prirodnog jezika kojima se bavimo. Ovdje ćemo se usredotočiti na radove na hrvatskom jeziku.

3.1. Vektorske reprezentacije riječi za NLP

Ovaj je rad izvorno motiviran nekolicinom prvobitnih radova koji istražuju primjenu vektorskih reprezentacija riječi na zadatke obrade teksta na engleskom jeziku.

Collobert et al. (2011) prvotno predlažu arhitekturu modela za NLP zadatke zasnovanu na umjetnoj neuronskoj mreži, koju koristimo u našem radu. Primarno žele njome izbjeći potrebu za pretjeranim traženjem značajki u svrhu poboljšanja rezultata modela. Uspješno postižu vrhunske ili njima bliske rezultate na nizu zadataka.

Ranije, Turian et al. (2010) istražuju korištenje vektorskih reprezentacija riječi u izgradnji modela za polunadzirano učenje te otkrivaju da svi tipovi evaluiranih reprezentacija poboljšavaju postojeće vrhunske rezultate na zadacima označavanja imenovanih entiteta te plitkog parsanja (engl. *chunking*). Od njih preuzimamo ideju za izgradnjom vektorskih reprezentacija iz Brownovih grupa te savjete o hiperparametrima izgradnje reprezentacija.

Konačno, Mikolov et al. (2013a) objavljuju dva algoritma učenja vektorskih reprezentacija – CBOW i Skip-gram – te uvjerljivo pokazuju bogatstvo sintaktičkih te semantičkih informacija koje nose tako izgrađene reprezentacije. Korištenjem njihove implementacije izgrađujemo oba tipa reprezentacija za hrvatski jezik.

3.2. Zadaci obrade prirodnog jezika

Vrhunske rezultate za POS i MSD zadatke na hrvatskom jeziku postiže Agić et al. (2013b) uz točnost od 97.13% za POS te 87.72% za MSD sustavom HunPos[1] (Halácsy et al., 2007), slobodnim za korištenje. Erjavec (2010) objavljuje MULTEXT-East 4 specifikaciju za MSD oznake, korištenu u prethodnim radovima. Označavamo po istoj specifikaciji.

Najbolje rezultate označavanja imenovanih entiteta u tekstovima na hrvatskom jeziku postižu sustav CroNER[2] (Karan et al., 2013) uz MUC F1 (Nadeau i Sekine, 2007) od 90.73% te Ljubešić et al. (2013) uz F1 od 89.86% uz manji skup oznaka na različitom skupu podataka. U našem radu koristimo jednak skup oznaka kao posljednji rad; organizacija, osoba, mjesto te ništa od navedenoga. Oba modela zasnivaju se na označavanju niza uvjetnim slučajnim poljima. Pritom je CroNER zatvoren, dok posljednji rad koristi Stanford NER[3] implementaciju slobodnu za korištenje (Finkel et al., 2005).

[1] `https://code.google.com/p/hunpos/`
[2] `http://takelab.fer.hr/croner/`
[3] `http://nlp.stanford.edu/software/CRF-NER.shtml`

4. Pristup problemu

U nastavku opisujemo pristup problemu izgradnje jedinstvenog modela za niz ranije navedenih zadataka označavanja tekstova na hrvatskome jeziku: POS i MSD označavanje te NER zadatak. Postupak izgradnje modela dijelimo na dva koraka. Najprije vršimo nenadzirano učenje vektorskih reprezentacija riječi na ogromnom neoznačenom skupu teksta na hrvatskome jeziku. Zatim iste vektorske reprezentacije koristimo kao ključan dio arhitekture modela za NLP zadatke u sljedećem koraku.

4.1. Nenadzirano učenje vektorskih reprezentacija

U području strojnog učenja nenadziranim učenjem nazivamo postupak pronalaženja skrivene strukture u neoznačenim podacima. U našem slučaju skrivena strukture koju želimo pronaći u ogromnom neoznačenom skupu teksta upravo su vektorske reprezentacije riječi, koje donose bitnu sintaktičku i semantičku informaciju o riječima. Popularni skupovi podataka za ovu svrhu u radovima na engleskom jeziku su engleska Wikipedia i Reuters (Collobert et al., 2011), koji ukupno sadrže 852 milijuna pojavnica. Karan et al. (2013) za hrvatski jezik koriste uzorak hrWaC 1.0 skupa (Ljubešić i Erjavec, 2011), koji sadrži ukupno 351 milijun pojavnica. U našem radu vektorske reprezentacije riječi gradimo iz daleko većeg skupa podataka hrWaC 2.0 (Ljubešic i Klubicka, 2014), filtriranog na prethodno predložen način (Šnajder et al., 2013). Naš skup podataka sadrži 1417 milijuna pojavnica u 64 milijuna rečenica. Ograničavamo rječnik za koji gradimo vektorske reprezentacije na 200000 najčešćih riječi u skupu.

Nakon pripreme opisanog neoznačenog skupa podataka, gradimo četiri različita tipa ranije predstavljenih vektorskih reprezentacija uz niz parametara:

- **Brown** su vektorske reprezentacije dobivene Brownovim grupiranjem,
- **CBOW** (od engl. *continuous bag of words*) su vektorske reprezentacije dobivene metodom kontinuirane vreće riječi,
- **Skip-gram** vektorske reprezentacije,

– **Collobert-Weston** vektorske reprezentacije.

Brown reprezentacije gradimo za različit broj grupa: 25, 50, 100, 250, 500, 1000, 2500, 5000, 10000 optimiziranim algoritmom (Liang, 2005). Duljina Brown vektorske reprezentacije je konstantna te iznosi 48. Zahvaljujući Mooreovom zakonu, u našem radu po prvi put uspješno provodimo Brownovo grupiranje u 10000 grupa, za što je utrošeno dva tjedna računanja na četvero-jezgrenom procesoru. Prethodni radovi na engleskom jeziku ograničavaju se na 3200 (Turian et al., 2010). Ostale vektorske reprezentacije gradimo za različitu duljinu vektora: 25, 50, 100, 250.

4.2. Nadzirano učenje modela označavanja tekstova

Nadziranim učenjem nazivamo strojno učenje modela na označenom skupu primjera, gdje svaki primjer sadrži vrijednosti ulaza te odgovarajućeg izlaza za zadatak u koju svrhu učimo model. Gradimo jedinstvenu arhitekturu modela zasnovanog na umjetnoj neuronskoj mreži u svrhu zadataka POS i MSD označavanja te NER zadatka, ranije predloženu u radu Collobert et al. (2011), opisanu u nastavku. Bitna razlika te poboljšanje modela potječe od odabranih značajki, koje, između ostaloga, iskorištavaju aspekt označavanja u kliznom prozoru za pamćenje povijesti oznaka.

4.2.1. Arhitektura modela za NLP zadatke

Na slici 4.1 dajemo pojednostavljen prikaz korištene arhitekture jedinstvenog modela za zadatke obrade prirodnog jezika, zasnovane na umjetnoj neuronskoj mreži. Formalno:

1. Na ulaz mreže dovodimo indekse riječi iz rječnika $w \in V^1$. Pritom je rječnik proširen posebnim pojavnicama koje označavaju početak ili kraj rečenice te nepoznatu riječ;

2. Riječ w predstavljamo pomoću K diskretnih značajki, što izražavamo kao $w \in V^1 \times \cdots \times V^K$, gdje je V^k rječnik za značajku k. Svakoj značajki pridružujemo projekcijsku tablicu $PT_{W^k}(.)$. Pritom je $W^k \in \mathbb{R}^{d_k \times |V^k|}$ gdje je d_k definiran hiperparametar duljine vektorske reprezentacije značajke k;

3. Potpunu vektorsku reprezentaciju riječi w dimenzija $d = \sum_k d_k$ stoga dobivamo konkatenacijom pojedinačnih vektorskih reprezentacija svake od K značajki.

$$PT_{W^1,\cdot,W^K}(w) = \begin{pmatrix} PT_{W^1}(w_1) \\ \vdots \\ PT_{W^K}(w_K) \end{pmatrix} \tag{4.1}$$

4. Konačno, vektorske reprezentacije svih pojavnica u kliznom prozoru $[w]_1^T$ spajamo u projekcijskom sloju.

$$PT_{W^1,\cdot,W^K}([w]_1^T) = \begin{pmatrix} PT_{W^1}([w_1]_1) & \cdots & PT_{W^1}([w_1]_T) \\ \vdots & & \vdots \\ PT_{W^K}([w_K]_1) & \cdots & PT_{W^K}([w_K]_T) \end{pmatrix} \tag{4.2}$$

Ključno je napomenuti da su sve vektorske reprezentacije parametri učenja, odnosno omogućavamo njihovu promjenu tijekom učenja povratnom propagacijom pogreške. Kao prijelazne funkcije koristimo $f_1 = \tanh$ za skriveni sloj (formula 4.3) te softmax funkciju $f_2 = \sigma(z)$ za izlazni sloj (formula 4.4). Odlučili smo broj neurona u skrivenom sloju postaviti na konstantnih 250 za sve naučene modele, vodeći se savjetom u radu (Collobert et al., 2011), gdje je uočeno da odabir hiperparametara poput veličine skrivenog sloja nema pretjeran učinak na performanse naučenog modela, dok god su dostatni za izražajnost neuronske mreže. Samo učenje provodimo stohastičkim gradijentnim spustom, uz mini-skupove (engl *mini-batches*) od 100 primjera.

$$f_1(x) = \tanh x = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{4.3}$$

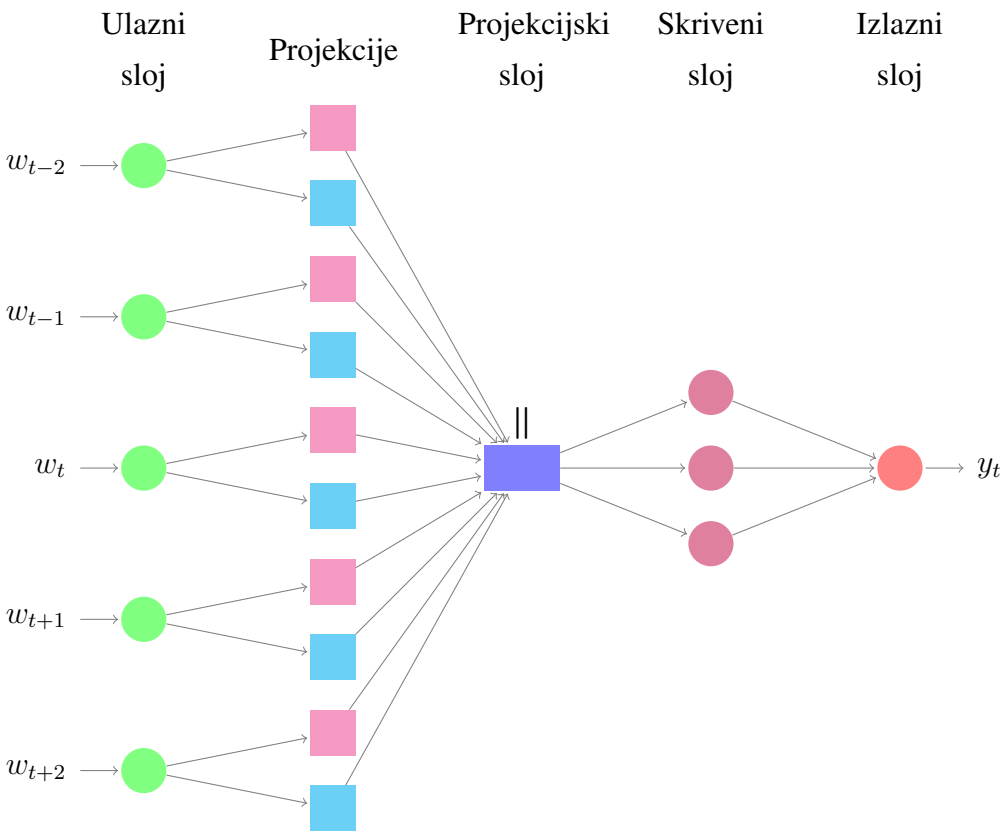
$$f_2(z) = \sigma(z)_j = \frac{e^{zj}}{\sum_{k=1}^{K} e^{zk}} \tag{4.4}$$

4.2.2. Odabir značajki

Koristimo sljedeći skup diskretnih značajki za sve naučene modele:

- **Oblik riječi** uz 4 različite oznake ovisno o tome je li riječ pisana malim slovima, velikim slovima, velikim početnim slovom ili ništa od navedenoga;
- **Sufiks riječi** gradimo koristeći javno dostupan alat za korjenovanje riječi (Ljubešić et al., 2007). Odbacivanjem dobivenog korijena riječi dobivamo jedan od 221 različita sufiksa (uključujući nepoznat sufiks) u koji preslikavamo riječ;

– **Prethodna oznaka** u kliznom prozoru. Budući da pojavnice u tekstu označavamo s lijeva na desno, slijedi da prethodno predviđene oznake možemo iskoristiti kao diskretne značajke za označavanje trenutačno promatrane pojavnice u središtu prozora. Pamćenje prethodno predviđenih oznaka u prozoru motivirano je radom Dietterich (2002), koji daju iscrpan pregled tehnika strojnog učenja na kliznom prozoru. Srodni radovi za obradu prirodnog jezika dubokim učenjem propuštaju iskoristiti ovu ključnu značajku u razvoju modela.



Slika 4.1: Arhitektura modela za NLP zadatke

5. Implementacija

Prateći prethodno poglavlje, opisujemo implementaciju oba koraka izgradnje modela. Najprije dajemo pregled alata iskorištenih za učenje vektorskih reprezentacija na skupu hrWaC 2.0. Zatim predstavljamo popularnu biblioteku za duboko učenje Pylearn2, koja nam omogućava da s lakoćom programski predstavimo složen model prethodno opisane arhitekture te provedemo učenje POS i MSD označavanja te NER zadatka.

5.1. Učenje vektorskih reprezentacija

Skup podataka hrWaC 2.0 [1] nakon preuzimanja najprije filtriramo alatom preuzetim iz rada Šnajder et al. (2013) [2], prevedenim u Python programski jezik. Na ovako priređenom skupu računamo frekvenciju pojavljivanja svake različite pojavnice. Gradimo rječnik od 200 000 najčešćih pojavnica, koji koristimo u nastavku. U tablici 5.1 prikazujemo pokrivenost skupa hrWaC 2.0 rječnikom ovisno o njegovoj veličini. Odabrana veličina pomiruje odličnu pokrivenost te razumnu veličinu rječnika, koja nam dozvoljava računanje vektorskih reprezentacija u razumnom vremenu.

5.1.1. Brownovo grupiranje

Brownovo grupiranje provodimo optimiziranom implementacijom u programskom jeziku C++ (Liang, 2005) [3], kojom računamo vektorske reprezentacije u složenosti $O(K^2|V|)$, gdje je K broj grupa, a $|V|$ veličina rječnika. Skup podataka prethodno izmjenjujemo umetanjem posebne pojavnice koja označava kraj rečenice. Nakon završetka izvršavanja programa, rezultantno binarno stablo s pripadajućim pojavnicama u svakom listu pohranjeno je u datoteci `paths`, na kojoj dalje izravno gradimo vektorske reprezentacije traženih parametara.

[1] `http://nlp.ffzg.hr/data/corpora/hrwac2.0.gz`
[2] `http://takelab.fer.hr/data/fhrwac/hrwac-filter.hs`
[3] `https://github.com/percyliang/brown-cluster`

Tablica 5.1: Pokrivenost rječnika na hrWaC 2.0 ovisna o veličini

Indeks	Pojavnica	Pokrivenost
1	,	5.23%
2	.	8.53%
5	u	17.25%
10	su	23.26%
25	iz	30.12%
50	bio	35.35%
100	dva	39.72%
1000	lijepo	55.97%
10000	Samsung	77.47%
100000	Jarca	94.25%
200000	silnoga	96.72%

5.1.2. CBOW

Za izgradnju vektorskih reprezentacija CBOW koristimo Googleov alat word2vec[4] (Mikolov et al., 2013a). Odabiremo prozor širine 5 po uzoru na ranije radove Collobert et al. (2011) te Turian et al. (2010). Nadalje, koristimo ranije opisani hijerarhijski softmax kao prijelaznu funkciju te poduzorkovanje (engl. *sub-sampling*) vrlo učestalih pojavnica kojime nasumično odbacujemo pojavnice koje čine više od 0.1% svih pojavnica u skupu. Učenje ubrzavamo odabirom 12 dretvi te traje tek nekoliko sati.

5.1.3. Skip-gram

Vektorske reprezentacije Skip-gram također gradimo alatom word2vec uz identične parametre kao za CBOW.

5.1.4. Collobert-Weston

Collobert-Westonove vektorske reprezentacije gradimo alatom Polyglot[5] (Al-Rfou et al., 2013) prevedenim uz biblioteku OpenBLAS[6] koja pruža ekstremno učinkovite implementacije matričnog računa te tehnika linearne algebre. Rezultat je učinkovito računa-

[4] `https://code.google.com/p/word2vec/`
[5] `https://github.com/aboSamoor/polyglot`
[6] `https://github.com/xianyi/OpenBLAS`

nje vektorskih reprezentacija na 4 jezgre procesora uz stopu učenja α jednaku 0.025, veličinu mini-skupa od 16 primjera te klizni prozor širine 5 pojavnica.

5.2. Učenje modela označavanja tekstova

Za učenje modela označavanja tekstova oslanjamo se na biblioteku Pylearn2, koju predstavljamo u nastavku. Dajemo pregled parametara modela u svrhu lake prilagodbe naše implementacije na novim zadacima.

5.2.1. Biblioteka Pylearn2

Pylearn2[7] (Goodfellow et al., 2013) je biblioteka za strojno učenje napisana u programskom jeziku Python. Pylearn2 oslanja se na biblioteku Theano (Bergstra et al., 2010) za definiranje te izvršavanje matematičkih izraza na matricama. Razvoj biblioteke Pylearn2 najviše je usredotočen na omogućavanje iterativnog razvijanja modela dubokog učenja, ponajviše umjetnih neuronskih mreža. Stoga odabirom upravo ove biblioteke za implementaciju našeg modela dobivamo niz postojećih implementacija komponenti modela te fleksibilne implementacije algoritama učenja.

Učenje ograničavamo na najviše 50 epoha. Za hiperparametre odabiremo stopu učenja $\alpha = 0.1$, koji linearno spuštamo tijekom učenja do 50. epohe. Konačna stopa učenja je $\alpha_{50} = 0.1\alpha_1$.

[7] `https://github.com/lisa-lab/pylearn2`

6. Evaluacija

Primarna svrha evaluacije jest pravedna ocjena performansi konačnih modela za označavanje tekstova. Kako bismo uspješno izgradili konačne modele, najprije za svaki zadatak provodimo unakrsnu provjeru u svrhu odabira najboljeg modela na skupu podataka namijenjenom za učenje, koji čini 80% ukupnog skupa podataka za svaki zadatak. Unakrsnu provjeru vršimo u 4 iteracije, čime ostvarujemo prihvatljiv kompromis između brzine učenja te statističke značajnosti rezultata. Prostor hiperparametara koji pretražujemo unakrsnom provjerom su prethodno navedene duljine vektorskih reprezentacija te broj Brownovih grupa. Konačno, za svaki zadatak učimo po jedan konačan model na cijelom skupu za učenje s hiperparametrima modela koji se pokazuje najuspješnijim nakon unakrsne provjere.

Najprije dajemo detaljan pregled korištenih skupova podataka. Zatim prikazujemo te raspravljamo o rezultatima.

6.1. Skupovi podataka

Skupove podataka za svaki zadatak odaberemo vodeći se najprije radom (Domingos, 2012), u kojem se savjetuje iscrpljivanje što većeg dostupnog skupa podataka za učenje modela te Ljubešić et al. (2013), koji upravo kombiniranjem postojećih skupova podataka za označavanje imenovanih entiteta postižu odlične rezultate na hrvatskom jeziku. Kategorizirajmo skupove podataka po zadacima.

POS označavanje

Kombiniramo tri javno dostupna skupa podataka, čiju strukturu prikazujemo u tablici 6.1: SETimes[1] (Tyers i Alperen, 2010), Student[2] (Filipić et al., 2012) te HOBS[3] (Hrvatsku ovisnosnu banku stabala) (Tadić, 2007).

[1] `http://nlp.ffzg.hr/resources/corpora/setimes/`

[2] Prethodno je skup Student neimenovan u srodnim radovima. Označili su ga studenti.

[3] `http://hnk.ffzg.hr/hobs/`

Tablica 6.1: Struktura skupova podataka za POS

Oznaka	Opis	Broj pojavnica				Primjeri
		SETimes	Student	HOBS	Ukupno	
Sve		89128	59143	117369	265640	
N	Imenica	33.37%	27.11%	29.89%	30.44%	godine Hrvatskoj Hrvatska
V	Glagol	14.51%	15.56%	14.67%	14.81%	je, su, će
Z	Interpunkcija	14.08%	12.07%	12.55%	12.96%	, . "
A	Pridjev	11.20%	11.17%	12.21%	11.64%	sve hrvatske hrvatski
S	Prijedlog	9.69%	9.14%	9.64%	9.54%	u, na, za
C	Veznik	5.24%	7.72%	7.05%	6.59%	i, da, a
P	Zamjenica	5.01%	7.41%	6.23%	6.08%	se, koji, što
R	Prilog	3.26%	5.47%	3.92%	4.05%	još, više, posto
M	Broj	2.56%	2.66%	1.84%	2.26%	jedan, tri, dva
Y	Kratica	0.00%	0.78%	1.01%	0.62%	BiH, dr, EU
Q	Čestica	0.42%	0.80%	0.47%	0.52%	ne, li, Ne
X	Strano	0.66%	0.10%	0.51%	0.47%	European Southeast Times

MSD označavanje

Za zadataka MSD označavanja koristimo tek skup SETimes, označen po standardu MULTEXT-East 4. Na žalost, premda je skup HOBS također označen za MSD, slijedi stariju verziju standarda broj 3. Stoga nismo mogli kombinirati sve skupove podataka. Detaljnu strukturu skupa SETimes označenog za MSD prikazujemo u tablici 6.2.

NER označavanje

Za označavanje imenovanih entiteta koristimo sve navedene skupove u radu Ljubešić et al. (2013); stoga su rezultati usporedivi. To su skupovi SETimes, Student te Vjesnik (Agic i Bekavac, 2013a). Prikazujemo njihovu strukturu u tablici 6.3.

Tablica 6.2: Struktura skupova podataka za MSD

Oznaka	Broj pojavnica	Primjeri
	SETimes	
Sve	89128	
Z	14.08%	, . "
Sl	3.75%	u, na, o
Vcr3s	3.35%	je, bude, jest
Cc	3.13%	i, a, ili
Npmsn	2.97%	Reuters, AFP, Erdogan
Sa	2.78%	za, u, na
Rgp	2.62%	također, još, Međutim
Ncmsg	2.39%	posto, tjedna, lipnja
Ncfsg	2.33%	godine, zemlje, stranke
Ncmsn	2.26%	ministar, predsjednik, premijer

Tablica 6.3: Struktura skupova podataka za NER

Oznaka	Broj pojavnica				Primjeri
	SETimes	Student	Vjesnik	Ukupno	
Sve	178981	59143	104134	342258	
Nazivi	11.30%	4.79%	8.63%	9.36%	
Organizacija	37.32%	44.35%	32.61%	36.62%	EU, za, Times
Osoba	30.61%	32.40%	39.17%	33.17%	Sanader, Kosor, Ivo
Mjesto	32.08%	23.25%	28.22%	30.22%	BiH, Hrvatska, Kosova

6.2. Rezultati

Ponovno, kategorizirajmo rezultate po zadacima. Konačne rezultate za sve zadatke radi preglednosti prikazujemo u tablici 6.7.

POS označavanje

Rezultati unakrsne provjere prikazani su u tablici 6.4. Na slici 6.1 grafički prikazujemo sve rezultate. Najuspješniji model koristi Collobert-Weston vektorske reprezentacije širine 250. Ostali modeli neznatno zaostaju. Učenjem modela s ovim hiperparametrima na cijelom skupu na učenje te evaluiranjem na ispitnom skupu dobivamo konačnu točnost od 96.40% (+0.64% u odnosu na prosječnu točnost prilikom unakrsne provjere).

MSD označavanje

Rezultati unakrsne provjere prikazani su u tablici 6.5. Na slici 6.2 grafički prikazujemo sve rezultate. Najuspješniji model koristi CBOW vektorske reprezentacije širine 250. Primjećujemo da Collobert-Weston daje nešto lošije rezultate, dok Brownove grupe daleko zaostaju. Na ispitnom skupu naš model postiže vrhunsku točnost od 89.95% (+2.04% u odnosu na unakrsnu provjeru).

NER označavanje

Rezultati unakrsne provjere prikazani su u tablici 6.6. Na slici 6.3 grafički prikazujemo sve rezultate. Daleko najuspješniji model koristi vektorske reprezentacije dobivene Brownovim grupiranjem riječi u 5000 grupa. Na ispitnom skupu naš model postiže točnost od 98.61% (+0.29% u odnosu na unakrsnu provjeru).

6.3. Rasprava

Trenutačno vrhunski model za POS označavanje na hrvatskom jeziku postiže točnost od 97.13% (Agić et al., 2013b), međutim na različitom ispitnom skupu koji čini 2.6% skupa SETimes. Za pravednu usporedbu modela predlažemo učenje te evaluaciju posljednjega na našem kombiniranom skupu podataka.

Iznimno smo zadovoljni odličnom točnosti od 89.95% koju postiže naš model za MSD označavanje. Rezultati ostvareni na ispitnom skupu daleko su bolji od prosječnih rezultata unakrsne provjere jer je ovaj model naučen na 33% većem skupu primjera.

Tablica 6.4: Rezultati unakrsne provjere za POS

(a) Brown

Broj grupa	Točnost
25	82.94
50	86.63
100	88.72
250	91.36
500	92.93
1000	93.93
2500	95.07
5000	95.37
10000	95.59

(b) Ostale vektorske reprezentacije

Duljina	Točnost		
	CBOW	Skip-gram	Collobert-Weston
25	94.92	92.34	95.38
50	95.10	94.04	95.53
100	95.26	94.65	95.69
250	95.44	95.04	**95.76**

Tablica 6.5: Rezultati unakrsne provjere za MSD

(a) Brown

Broj grupa	Točnost
25	57.65
50	62.35
100	65.92
250	71.03
500	74.72
1000	76.69
2500	78.09
5000	78.69
10000	80.14

(b) Ostale vektorske reprezentacije

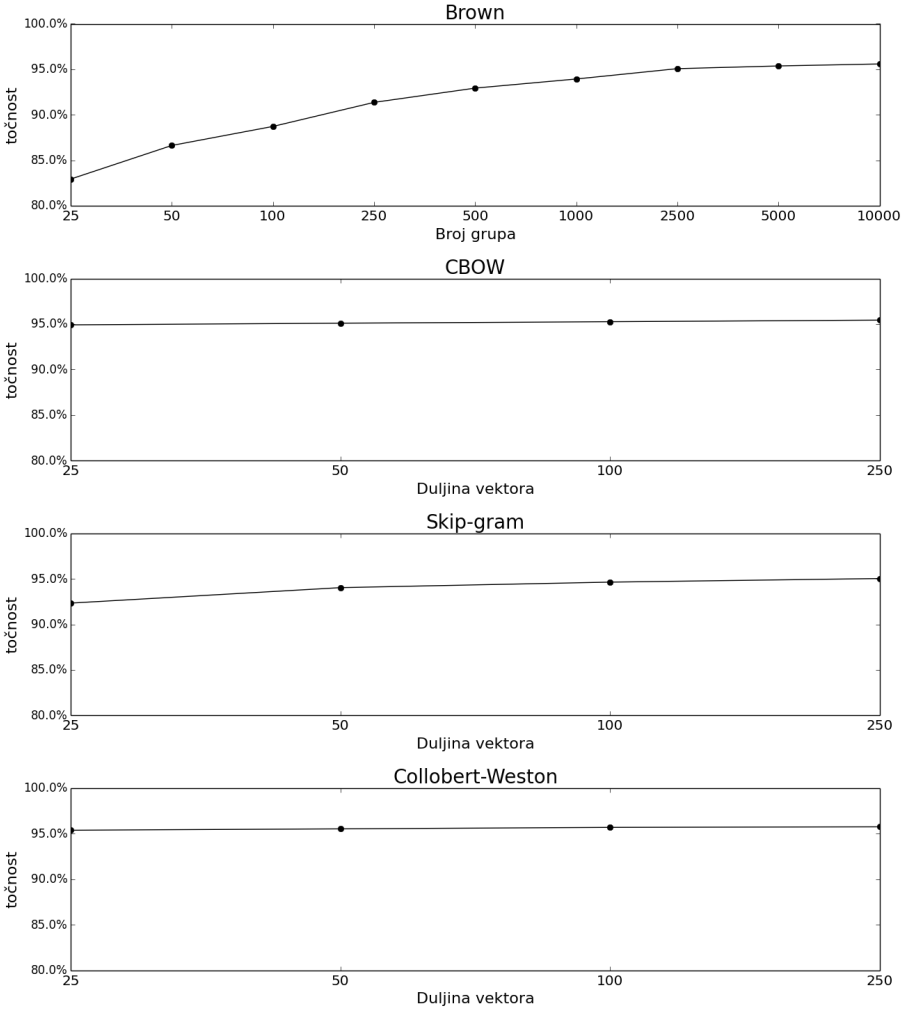
Duljina	Točnost		
	CBOW	Skip-gram	Collobert-Weston
25	85.82	83.73	81.80
50	86.88	85.72	82.80
100	87.41	86.60	83.10
250	**87.91**	87.46	83.33

Naime, naučen je na 80% cijelog skupa u usporedbi s $75\% * 80\%$ u jednoj iteraciji unakrsne provjere. Budući da za MSD imamo najmanji označen skup podataka, lako poboljšavamo rezultate dodatnim podacima. Stoga predlažemo istraživanje automatskog prijevoda oznaka skupa HOBS označenog po starijoj verziji standarda na novu verziju. Očekujemo točnost daleko iznad 90% uz model naučen na kombiniranim skupovima. Prethodno najbolji model postiže točnost od 87% (Agić et al., 2013b) na statistički manje značajnom ispitnom skupu koji ponovno čini 2.6% skupa SETimes, u

usporedbi s našim ispitnim skupom od 20%.

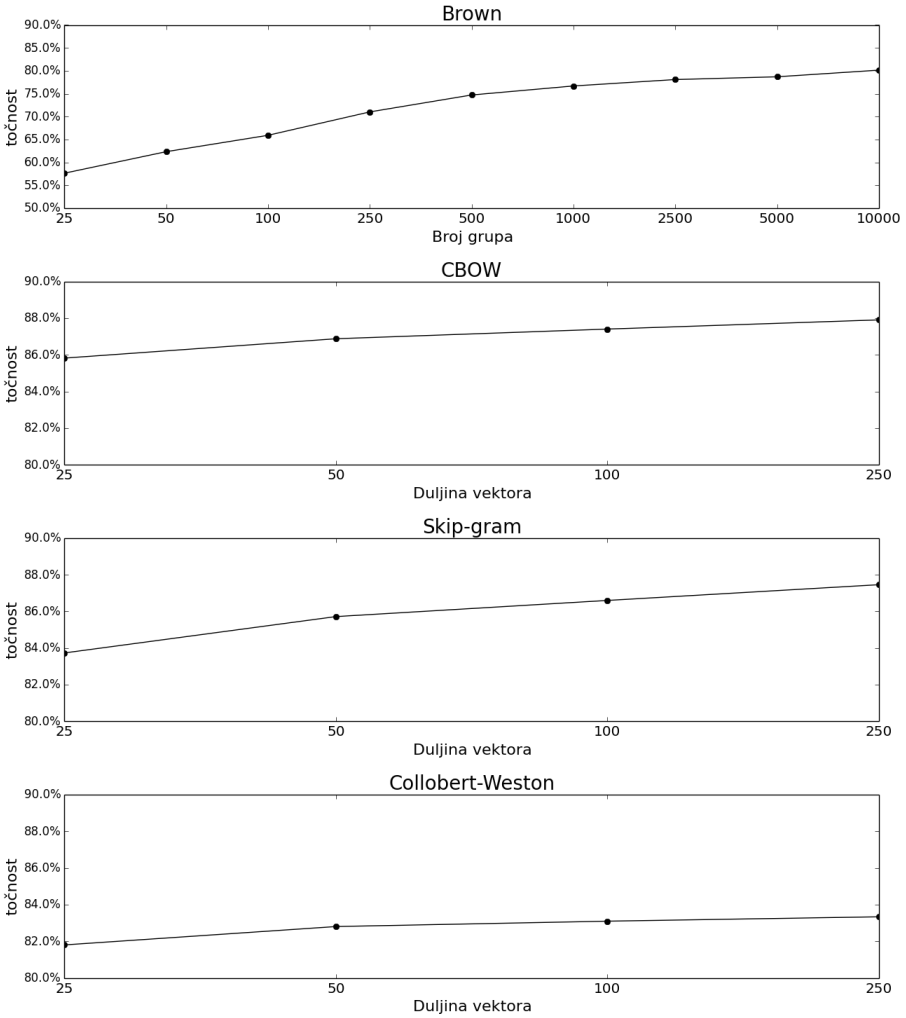
Rezultate NER označavanja na žalost ne možemo usporediti s trenutačno najboljim sustavom za NER na hrvatskom jeziku, CroNER (Karan et al., 2013) jer koristi veći skup oznaka. Premda postižemo vrhunsku točnost na statistički značajnom ispitnom skupu uz pouzdan model koji lako generalizira na neviđene primjere, smatramo da je potrebna dodatna evaluacija modela na identičnom skupu podataka korištenom za navedeni rad. Također, držimo da je za razvijanje boljih modela za NER neophodna otvorena podjela označenih skupova podataka, kojoj zahvaljujemo postignute rezultate.

Općenito, unakrsnom provjerom otkrivamo potrebu za širenjem prostora hiperparametara. Naime, vektorske reprezentacije koje daju najbolje modele za POS i MSD označavanje su najveće razmotrene duljine, 250. U oba slučaja postignuta točnost je zamjetno bolja od točnosti za jednake vektorske reprezentacije duljine 100. Stoga, dr-



Slika 6.1: *Rezultati unakrsne provjere za POS*

žimo da je vrijedno dalje istraživanje odabira hiperparametara našeg modela. Srodno tome, ograničili smo duljinu vektorskih reprezentacija dobivenih Brownovim grupiranjem na 48 po uzoru na prethodne radove. Promatrajući loše performanse modela koji koristi Brown reprezentacije za MSD označavanje uviđamo moguću nedostatnu sintaktičku i semantičku izražajnost kratkih reprezentacija.



Slika 6.2: *Rezultati unakrsne provjere za MSD*

Tablica 6.6: Rezultati unakrsne provjere za NER

(a) Brown

Broj grupa	Točnost
25	96.86
50	96.99
100	97.18
250	97.39
500	97.69
1000	98.10
2500	98.22
5000	**98.32**
10000	98.30

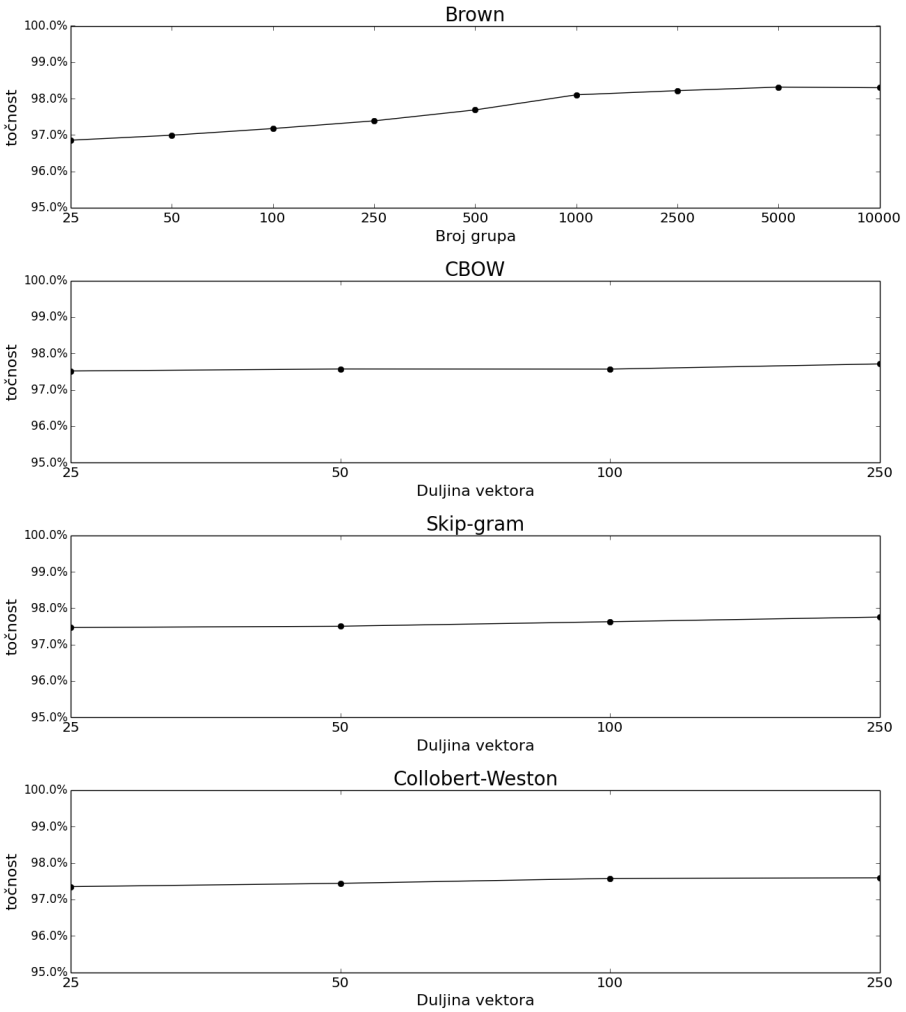
(b) Ostale vektorske reprezentacije

Duljina	Točnost		
	CBOW	Skip-gram	Collobert-Weston
25	97.52	97.47	97.35
50	97.57	97.50	97.44
100	97.57	97.63	97.58
250	97.71	97.76	97.59

Tablica 6.7: Rezultati na ispitnim skupovima za sve zadatke

Mjera	POS	MSD	NER
	Collobert-Weston (250)	CBOW (250)	Brown (5000)
Točnost [1]	96.40 (+0.64)	89.95 (+2.04)	98.61 (+0.29)

[1] U usporedbi s rezultatima unakrsne provjere.



Slika 6.3: *Rezultati unakrsne provjere za NER*

7. Zaključak

Cilj našeg rada bila je primjena dubokog učenja na izgradnju modela označavanja tekstova na hrvatskome jeziku. Naš rad izvorno je motiviran prvobitnim radovima Collobert et al. (2011) te Mikolov et al. (2013a), gdje je istražena primjena vektorskih reprezentacija riječi na zadatke obrade teksta na engleskom jeziku. Njima su redom poboljšani prethodno vrhunski rezultati na svim zadacima gdje su primijenjene. Dodatna motivacija je netom objavljen neoznačen skup teksta na hrvatskome jeziku hrWaC 2.0 od 1.4 milijardu pojavnica, koji obećava izgradnju sintaksno te semantički bogatih vektorskih reprezentacija.

Odlučili smo izgraditi modele za zadatke označavanja vrsta riječi, morfosintaktičkih deskriptora te imenovanih entiteta. Najprije smo pretraživanjem prostora hiperparametara modela unakrsnom provjerom evaluirali različite vektorske reprezentacije riječi na svakom zadatku. Najbolje hiperparametre zatim smo odabrali za učenje modela na potpunom skupu za učenje te evaluaciju na ispitnom skupu za svaki zadatak. Uspješno smo izgradili modele koji postižu vrhunsku ili točnost blisku vrhunskoj za označavanje vrsta riječi (96.40%), morfosintaktičkih deskriptora (89.95%) te imenovanih entiteta (98.61%) na ispitnim skupovima. Objavljujemo izračunate vektorske reprezentacije kao poticaj daljem istraživanju obrade teksta na hrvatskom jeziku. Nusproizvod težnje k što većim označenim skupovima podataka u svrhu izgradnje boljih modela su kombinirani skupovi podataka za navedene zadatke koje također činimo javno dostupnima.

Otkrivamo važnu granu budućeg istraživanja koja vodi u smjeru izražajnijih modela s većim projekcijskim slojevima, ponajprije za vektorske reprezentacije dobivene Brownovim grupiranjem. Predlažemo proširenje skupa podataka označenog morfosintaktičkim deskriptorima istraživanjem automatske pretvorbe skupa podataka označenog po različitom standardu. Također, predlažemo primjenu većeg skupa podataka korištenog za izgradnju sustava CroNER (Karan et al., 2013) na učenje modela za označavanje imenovanih entiteta radi kvalitetne usporedbe pristupa te primjenu dodatnih značajki kao što su oznake vrsta riječi ili ograničeni rječnik imenovanih entiteta.

LITERATURA

Ţ Agić, Danijela Merkler, i Daša Berović. Parsing croatian and serbian by using croatian dependency treebanks. U *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, 2013a.

Zeljko Agic i Bozo Bekavac. Domain dependence of statistical named entity recognition and classification in croatian texts. U *Information Technology Interfaces (ITI), Proceedings of the ITI 2013 35th International Conference on*, stranice 277–282. IEEE, 2013a.

Zeljko Agic i Bozo Bekavac. Domain-aware evaluation of named entity recognition systems for croatian. *CIT. Journal of Computing and Information Technology*, 21 (3):195–209, 2013b.

Željko Agic i Zdravko Dovedan. Improving part-of-speech tagging accuracy for croatian by morphological analysis. *Informatica*, 32(4), 2008.

Željko Agić i Danijela Merkler. Three syntactic formalisms for data-driven dependency parsing of croatian. U *Text, Speech, and Dialogue*, stranice 560–567. Springer, 2013.

Željko Agić i Marko Tadić. Evaluating morphosyntactic tagging of croatian texts. U *The 5th International Conference on Language Resources and Evaluation (LREC 2006)*. Hrvatska znanstvena bibliografija i MZOS-Svibor, 2006.

Željko Agić, Nikola Ljubešić, i Danijela Merkler. Lemmatization and morphosyntactic tagging of croatian and serbian. U *Proceedings of ACL*, 2013b.

Rami Al-Rfou, Bryan Perozzi, i Steven Skiena. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*, 2013.

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

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, i Yoshua Bengio. Theano: a cpu and gpu math expression compiler. U *Proceedings of the Python for scientific computing conference (SciPy)*, svezak 4, stranica 3. Austin, TX, 2010.

Eric Brill. A simple rule-based part of speech tagger. U *Proceedings of the workshop on Speech and Natural Language*, stranice 112–116. Association for Computational Linguistics, 1992.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, i Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18 (4):467–479, 1992.

Dan C Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, i Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. U *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, svezak 22, stranica 1237, 2011.

Ronan Collobert i Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. U *Proceedings of the 25th international conference on Machine learning*, stranice 160–167. ACM, 2008.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, i Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

Thomas G Dietterich. Machine learning for sequential data: A review. U *Structural, syntactic, and statistical pattern recognition*, stranice 15–30. Springer, 2002.

Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, i Samy Bengio. Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11:625–660, 2010.

Tomaž Erjavec. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. U Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, i Daniel Tapias, urednici, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Lobel Filipić, Tereza Jurić, i Marija Stupar. Strojno prepoznavanje naziva u tekstovima pisanima hrvatskim jezikom, 2012.

Jenny Rose Finkel, Trond Grenager, i Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. U *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, stranice 363–370. Association for Computational Linguistics, 2005.

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

Ian J Goodfellow, David Warde-Farley, Pascal Lamblin, Vincent Dumoulin, Mehdi Mirza, Razvan Pascanu, James Bergstra, Frédéric Bastien, i Yoshua Bengio. Pylearn2: a machine learning research library. *arXiv preprint arXiv:1308.4214*, 2013.

Péter Halácsy, András Kornai, i Csaba Oravecz. Hunpos: an open source trigram tagger. U *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, stranice 209–212. Association for Computational Linguistics, 2007.

Geoffrey E Hinton, Simon Osindero, i Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

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*, 37(2), 2013.

John Lafferty, Andrew McCallum, i Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Percy Liang. *Semi-supervised learning for natural language*. Doktorska disertacija, Massachusetts Institute of Technology, 2005.

Nikola Ljubešić i Tomaž Erjavec. hrwac and slwac: Compiling web corpora for croatian and slovene. U *Text, Speech and Dialogue*, stranice 395–402. Springer, 2011.

Nikola Ljubešic i Filip Klubicka. {bs, hr, sr} wac: Web corpora of bosnian, croatian and serbian. U *Proceedings of the WAC-9 Workshop*, 2014.

Nikola Ljubešić, Damir Boras, i Ozren Kubelka. Retrieving information in croatian: Building a simple and efficient rule-based stemmer. *Digital information and heritage/Seljan, Sanja*, stranice 313–320, 2007.

Nikola Ljubešic, Marija Stupar, i Tereza Juric. Building named entity recognition models for croatian and slovene. U *Proceedings of the Eighth Information Society Language Technologies Conference*, stranice 117–122, 2012.

Nikola Ljubešić, Marija Stupar, Tereza Jurić, i Željko Agicć. Combining available datasets for building named entity recognition models of croatian and slovene. *Jezikovne tehnologije, Slovenščina*, 2(1):2, 2013.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, i Jeff Dean. Distributed representations of words and phrases and their compositionality. U *Advances in neural information processing systems*, stranice 3111–3119, 2013b.

Frederic Morin i Yoshua Bengio. Hierarchical probabilistic neural network language model. U *Proceedings of the international workshop on artificial intelligence and statistics*, stranice 246–252. Citeseer, 2005.

David Nadeau i Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Lev Ratinov i Dan Roth. Design challenges and misconceptions in named entity recognition. U *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, stranice 147–155. Association for Computational Linguistics, 2009.

Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

David E Rumelhart, Geoffrey E Hinton, i Ronald J Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5:3, 1988.

Robert R Schaller. Moore's law: past, present and future. *Spectrum, IEEE*, 34(6): 52–59, 1997.

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

Richard Socher, Yoshua Bengio, i Chris Manning. Deep learning for nlp. *Tutorial at Association of Computational Logistics (ACL), 2012, and North American Chapter of the Association of Computational Linguistics (NAACL)*, 2013.

Pontus Stenetorp, Hubert Soyer, Sampo Pyysalo, Sophia Ananiadou, i Takashi Chikayama. Size (and domain) matters: Evaluating semantic word space representations for biomedical text. U *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine*, 2012.

Marko Tadić. Building the croatian dependency treebank: the initial stages. *Suvremena lingvistika*, 63(1):85–92, 2007.

Joseph Turian, Lev Ratinov, Yoshua Bengio, i Dan Roth. A preliminary evaluation of word representations for named-entity recognition. U *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*, stranice 1–8, 2009.

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.

Francis M Tyers i Murat Serdar Alperen. South-east european times: A parallel corpus of balkan languages. U *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, stranice 49–53, 2010.

Kristina Vučković. Model parsera za hrvatski jezik. 2009.

Jason Weston, Frédéric Ratle, Hossein Mobahi, i Ronan Collobert. Deep learning via semi-supervised embedding. U *Neural Networks: Tricks of the Trade*, stranice 639–655. Springer, 2012.

Duboko učenje vektorskih reprezentacija riječi za modele označavanja tekstova na hrvatskome jeziku

Sažetak

Vektorske reprezentacije riječi prikazuju riječi niskodimenzijskim vektorima realnih vrijednosti u svrhu matematičkog zapisa sintaktičkih te semantičkih informacija. Nenadziranim učenjem na skupu teksta hrWaC od 1.4 milijardu pojavnica izgrađujemo 4 tipa reprezentacija za rječnik od 200 000 riječi. Pomoću njih dubokim učenjem izgrađujemo modele označavanja tekstova zasnovane na umjetnim neuronskim mrežama. Kombiniramo dostupne označene skupove podataka. Ostvarujemo vrhunsku ili točnost blisku vrhunskoj za označavanje vrsta riječi (96.40%), morfosintaktičkih deskriptora (89.95%) te imenovanih entiteta (98.61%) na ispitnim skupovima. Javno objavljujemo sve korištene skupove podataka.

Ključne riječi: obrada prirodnog jezika, duboko učenje, vektorske reprezentacije, neuronske mreže, označavanje vrsta riječi, morfosintaktički deskriptori, prepoznavanje imenovanih entiteta, POS, MSD, NER, NLP

Deep Learning of Word Embeddings for Tagging Models for Croatian Texts

Abstract

Word embeddings represent words using low-dimensional real-valued vectors to mathematically express their syntactic and semantic information. We use unsupervised learning on the hrWaC dataset containing 1.4 billion tokens to build 4 types of word embeddings for a dictionary of 200 000 words. Then we leverage deep learning to build tagging models based on artifical neural networks. We combine available labelled datasets. We achieve state-of-the-art or near state-of-the-art accuracy for part-of-speech tagging (96.40%), morphosyntactic tagging (89.95%) and named entity recognition (98.61%) on test datasets. We make all used datasets publicly available.

Keywords: natural language processing, deep learning, word embeddings, neural networks, part-of-speech tagging, morphosyntactic descriptors, named entity recognition, POS, MSD, NER, NLP