Generiranje imena naselja pomocu LSTM mreže

Antonio Čogelja Morena Granić Fran Lubina Iva Jurković Jakov Juvančić Matej Logarušić

***Sažetak*—Cilj projekta je LSTM rekurzivna neuronska mreža na razini znakova koja generira realistična imena hrvatskih naselja. Fokus projekta je treniranje i razvijanje neuronske mreže za generiranje realističnih imena hrvatskih naselja. Korištenjem LSTM mreže, koja je prilagođena za analizu sekvencijskih podataka, cilj je razviti model sposoban za učenje jezičnih obrazaca i struktura iz postojećih imena naselja. Svrha mreže je generiranje novih imena temeljenih na tim naučenim obrascima, pri čemu se zadržavaju jezične i strukturne zakonitosti specifične za taj kontekst. Željena točnost modela η je $\lim_{\tau\to 0}\eta = 0.5$**

***Index Terms*—Naselje, LSTM, rekurzivne mreže, neuronske mreže-**

I. Uvod

Ishod projekta je LSTM rekurzivna neuronska mreža na razini znakova koja generira realistična imena hrvatskih naselja. Mreža radi sa vektorima koji predstavljaju slova hrvatske abecede proširene specijalnim znakovima $\Sigma = \{\text{hrv. abeceda}\} \cup \{\langle start\rangle, \text{"} \setminus 0\text{"}\}$.
Ulaz mreže je one-hot vektor $\mathbf{x}^{(t)}$ dimezije $|\Sigma| = 30 + 2$.

$$\mathbf{x}_i^{(t)} = \begin{cases} 1, & \text{ako } i = j \\ 0, & \text{inače} \end{cases} \tag{1}$$

Izlaz dobiven na kraju pojedinog vremenskog koraka t je vektor vjerojatnosti pojave pojednog znaka abecende.

$$\hat{\mathbf{y}}^{(t)} = \begin{bmatrix} p(c_0) \\ p(c_1|c_0) \\ \vdots \\ p(c_{|\Sigma|-1}|\bigcap_{i=0}^{|\Sigma|-2} c_i) \end{bmatrix} \quad \text{Gdje} \quad c \in \Sigma \tag{2}$$

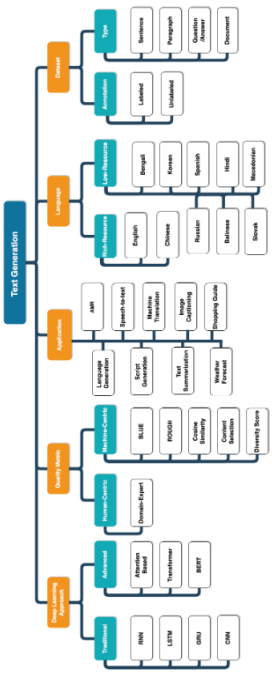
Vjerojatnosti su dobivene softmax funkcijom parametriziranom hiperparametrom temperature τ.
Na temelju tih vjerojatnosti se uzorkuje konačni izlazni vektor $\mathbf{y}^{(t)}$, odnosno t-ti znak u imenu naselja.

$$\mathbf{y}^{(t)} \sim \hat{\mathbf{y}}^{(t)} = \sigma_\tau(f(\mathbf{x}^{(t)}; \boldsymbol{\theta})) \tag{3}$$

$f(\mathbf{x}; \boldsymbol{\theta})$ predstavlja ukupno djelovanje ćelija modela nad njenim ulazom parametrizirano hiperparametrima modela $\boldsymbol{\theta} = \begin{bmatrix} |\mathbf{a}| & \mu & \tau \end{bmatrix}$
(opisani u poglavlju III-B) Temperaturno uzorkovanje je izabrano, jer omogućava eksperimentiranje i generiranje zanimljivih toponima. Izlaz mreže je niz znakova $\{\mathbf{y}^{(t)}\}\Big|_{t=0}^{T-1}$, odnosno ime naselja. Željena točnost modela η je $\lim_{\tau\to 0}\eta = 0.5$

II. Pregled literature

U ovom poglavlju dajemo kratki pregled postojeće literature na način kako je to učinjeno u [1]. Svrstavanjem radova na temelju kriterija: pristupa dubokom učenju, funkciji pogreške, primjeni, jeziku i skupu podataka.
Dobivamo taksonomiju na slici 1.
Nas samo zanimaju radovi na jezicima spomenutim u cjelini I na razini riječi.
U konačnici odabiremo karakteristike našeg rješenja, navedene u tablici **??**.

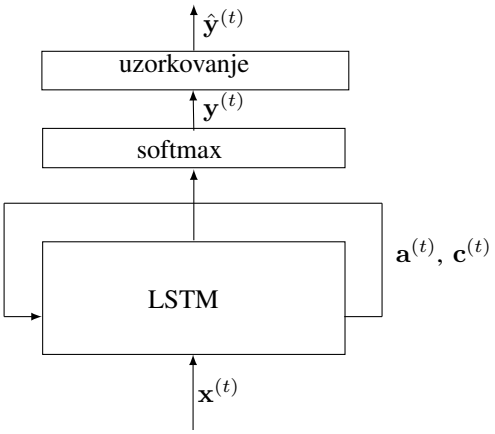


Slika 1. taksonomija rješenja za generiranje teksta

Nas samo zanimaju radovi na jezicima spomenutim u cjelini I na razini riječi sa primjenom generiranja jezičnih konstrukta.
U konačnici odabiremo karakteristike našeg rješenja, navedene u tablici I.

III. Opis implementirane LSTM mreže

Fokus projekta je treniranje i razvijanje neuronske mreže za generiranje realističnih imena hrvatskih naselja. Korištenjem



Slika 2. Arhitektura LSTM ćelije

LSTM mreže, koja je prilagođena za analizu sekvencijskih podataka, cilj je razviti model sposoban za učenje jezičnih obrazaca i struktura iz postojećih imena naselja. Svrha mreže je generiranje novih imena temeljenih na tim naučenim obrascima, pri čemu se zadržavaju jezične i strukturne zakonitosti specifične za taj kontekst. LSTM ćelija i mreža je implementirana u radnom okviru pyTorch.
Dizajn mreže i podešavanje hiperparametara se odvija paralelno sa implementacijom mreže u radnom okviru Keras.

A. Arhitektura

–

B. Hiperparametri

Tablica I
HIPERPARAMETRI NAŠE MREŽE

Hiperparametar	***Vrijednost***	***komentar***
broj jedinica po sloju	2	
stopa učenja (μ)		
temperatura (τ)		
dimenzija skrivenog stanja ($\|\mathbf{a}\|$)		
broj epoha	100	
aktivacijska funckija	tanh	zadano
povratna akt. funkcija	σ_τ	zadano
bias	da	zadano
inicijalizator kernela	glorot jednoliki	zadano
inicijalizator povratne veze	glorot jednoliki	zadano
bias inicijalizator	zeros	zadano
forget bias	da	zadano
regularizacija kernela		zadano
regularizacija kernela povratne veze		zadano
bias regularizacija		zadano
kernel ograničenje		zadano
povratno ograničenje		zadano
bias ograničenje		zadano
dropout	0	zadano
povratni dropout	0	zadano

C. Ćelija

Ćelija izgledda ovako.

Nadalje slika III-C prikazuje unutarnju shemu ćelije.

D. Treniranje

BPTT je korišten kao algoritam učenja. Kao funkcija gubitka koristi se unakrsna entropija.

$$L_i = -\sum_{t=0}^{T-1} \mathbf{x}_i^{(t)} \cdot log(\hat{\mathbf{y}}_i^{(t)}) \tag{4}$$

inačica BPTT koju mi koristimo je u biti propagirani stohastički gradijentni spust.

Pri treniranju koristimo parametre navedene u tablici II.

Tablica II
PARAMETRI PROCEDURE ZA TRENIRANJE

Parametar	***Vrijednost***
f-ja gubitka	kategorička unakrsna entropija
algoritam optimizacije	ADAM
metrike	točnost

IV. Opis eksperimentalnih rezultata

a) optimiranje hiperparametara: –

Tablica III
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

A. Usporedba rezultata

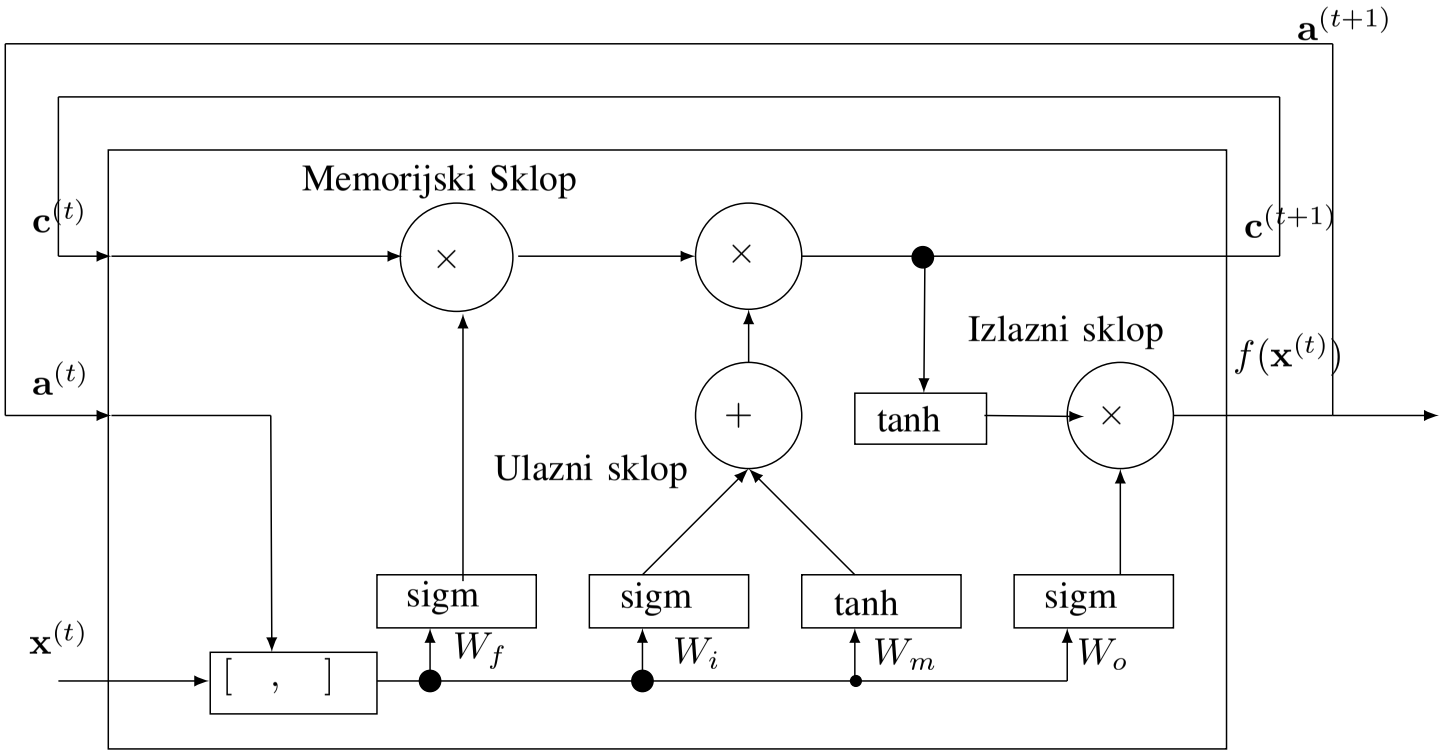
–

V. Zaključak

–

Literatura

[1] Fatima, N., Imran, A S., Kastrati, Z., Daudpota, S M., Soomro, A (2022) "A Systematic Literature Review on Text Generation Using Deep Neural Network Models" IEEE Access, 10: 53490-53503, https://doi.org/10.1109/ACCESS.2022.3174108



Slika 3. Unutarnja arhitektura LSTM ćelije