Uporaba modela BERT pri postavljanju vejic v slovenskem jeziku

SDJT JOTA

Martin Božič, Ljubljana 2021
Mentor pri diplomskem delu: prof. dr. Marko Robnik Šikonja
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Motivacija

- Postavljanje vejic je najpogostejša napaka v slovenščini
- Avtomatski sistem bi olajšal pisno komunikacijo.
- Slovenščina je morfološko bogat jezik.
- Vejica je skladenjska, mnogo je tudi izjem.

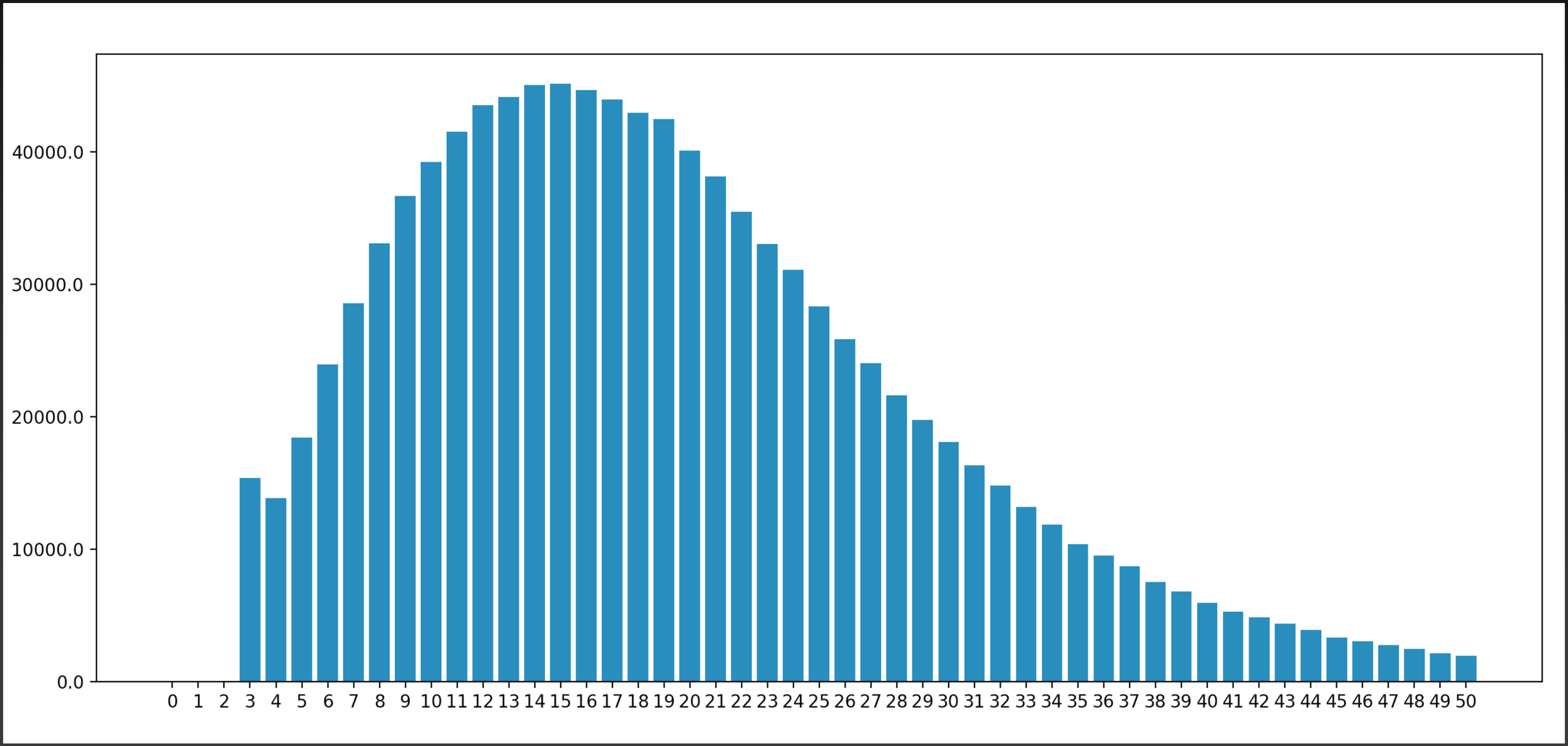
Pristopi

- Uporaba nevronskih mrež na ravni znakov.
- Uporaba dvostrane nevronske mreže GRU.
- Uporaba in prilagoditev vnaprej naučenih modelov tipa BERT.

Podatkovna množica

- Podatkovno množico smo zgradili z uporabo korpusa Gigafida.
- Vsebovala je 907.879 povedi, v katerih je bilo 1.645.120 vejic.
- V povprečju vsaka poved vsebuje skoraj 2 vejici.
- Pri učenju smo podatkovno množico razdelili na učno (90%), razvojno (5%) in testno (5%).

Število povedi glede na vsebovano število besed.



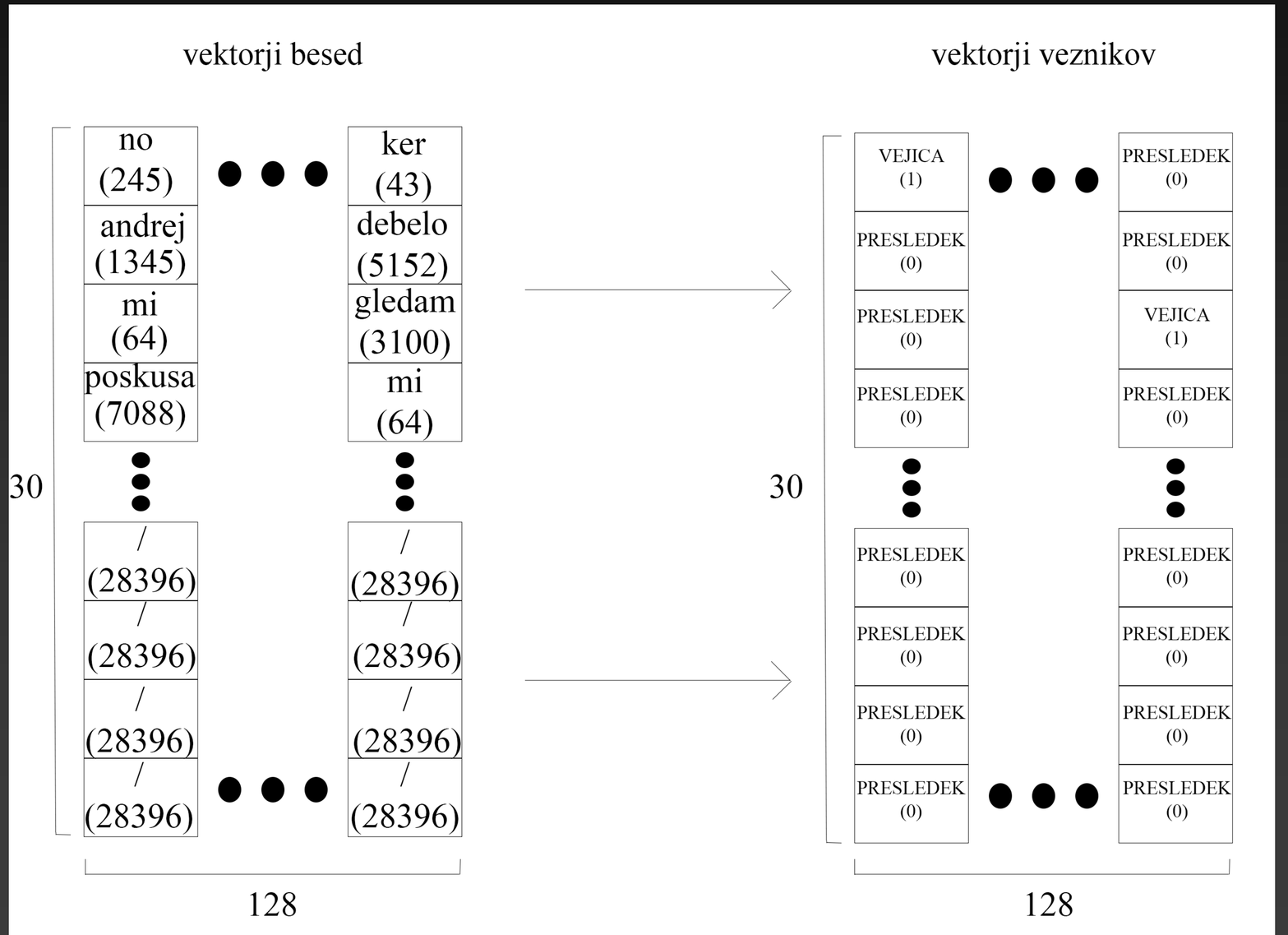
Rešitev z nevronsko mrežo GRU

- Uporabili smo dvosmerno rekurenčno GRU nevronsko mrežo.
- Za kodiranje besed smo uporabili fastText vložitve.
- Po vsaki iteraciji učenja smo model validirali in po dveh neizboljšanih validacijah modela učenje zaključili.

Primer vektorja vhoda in vektorja izhoda pri učenju nevronske mreže GRU

besede	ločilo
Ko	PRESLEDEK
je	PRESLEDEK
ura	PRESLEDEK
prišla	,VEJICA
sem	PRESLEDEK
še	PRESLEDEK
vedno	PRESLEDEK
sanjal	PRESLEDEK
o	PRESLEDEK
domu	.PIKA
0	0
0	0

Primer vektorja vhoda in vektorja izhoda pri učenju nevronske mreže GRU



Evalvacija rezultatov pri uporabi GRU nevronske mreže

- Na izhodu smo izbrali najverjetnejše napovedano ločilo.
- Od 111.607 vejic v besedilu jih je mreža pravilno predvidela 86.810, kar je 77.78%.
- Opazili smo, da je mreža pri večih napačno klasificiranih primerih vejico izpustila, kot jo postavila na napačno mesto.

Izpopolnjevanje modela BERT

- Uporabili smo več že vnaprej naučenih modelov BERT, ki smo jih še doučili in prilagodili problemu postavljanja postavljanja vejic.
- Učenje modela poteka na dva načina in sicer z zamenjavo 15% vhodov z maskiranimi žetoni in napovedovanjem zaporednosti stavkov.
- Model uporablja lastne vektorske vložitve.

originalen stavek	Ko se je obrnila, je bila pretresena.
vektorske vložitve	Ko se je obrnil ##a , je bila pretres ##ena .

Izpopolnjevanje modela BERT

- Dodali smo dodatno preslikovalno plast
- Uporabili smo plast BertForMaskedLM iz zbirke transformers.
- Uporabili smo tehniko maskiranja besed in model doučili.
- Izpopolnili smo CroSloEngual in bert-base-multilingual-cased model.
- Za učenje smo uporabili platformo Google Colab Pro.

Priprava vhodov

- Na vhodu smo modelu podali vektor indentifikatorjev žetonov, vektor preslikanih ločil in vektor pozornosti.
- Vse vektorje smo podaljšali ali porezali na dolžino 128.
- Za podaljševanje vektorja žetonov model uporablja poseben žeton [PAD].
- Mesta z besedami smo v vektorju preslikanih ločil označili z vrednostjo -100. Ta modelu pri učenju pove, da za obravnavani žeton ni potrebno iskati preslikave.

Primer izgradnje vektorjev za izpopolnjevanje modela BERT.

vektor žetonov	[CLS] Ko [MASK] se [MASK] je [MASK] obrnil ##a [MASK] je [MASK] bila [MASK] pretres ##ena . [SEP] [PAD] [PAD]
vektor identifikatorjev žetonov	103 1561 105 1009 105 1001 105 11848 1002 105 1001 105 1108 105 14178 1579 47105 104 2 2
vektor preslikanih ločil	-100 -100 0 -100 0 -100 0 -100 -100 1 -100 0 -100 0 -100 -100 -100 -100 -100 -100
vektor pozornosti	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0

Izpopolnjevanje modela

- Model smo doučili v štirih iteracijah.
- Pri tem smo učno množico razdelili na pakete (batches) po 32 trojic vektorjev.
- Uporabili smo optimizator Adam iz knjižnice transformers.

Evalvacija modela BERT

- Model smo testirali s 'Corpus of comma placement Vejica 1.3 Petra Holozana'.
- Korpus vsebuje 104.185 stavkov, od katerih smo jih pri testiranju uporabili 78.472.
- Na vseh stavkih smo dosegli vrednost F-ocene: 0,9216
- Na stavkih iz korpusa Šolar smo dosegli vrednost F-ocene: 0,9415
- Na stavkih iz korpusa wiki smo dosegli vrednost F-ocene: 0,9265

Evalvacija modela BERT

- Na lastni testni množici smo pridobili točnost 95,3%, priklic 94,3% in F-oceno 94,8%.

Primerjava končnih rezultatov

	točnost	priklic	F-ocena
model GRU	83.6	78.8	80.6
CroSloEngual	95.3	94.3	94.8
bert-base-multilingual-cased	94.1	92.5	93.3

Primeri stavkov, ki jih je model BERT napovedal pravilno, a so bili v podatkovni množici označeni napačno.

postavljen stavek	Največji bedak med vsemi pa je bil neki moški in prav njega se je najbolj bala.
napačni stavek	Največji bedak med vsemi pa je bil neki moški, in prav njega se je najbolj bala.
postavljen stavek	Predolgo ga je mučila nemogoča izbira, naj se odloči za žensko, ki jo ljubi, ali za zvestobo svojemu klanu.
napačni stavek	Predolgo ga je mučila nemogoča izbira naj se odloči za žensko, ki jo ljubi, ali za zvestobo svojemu klanu.

Implementacija spletnega orodja

- Aplikacijo smo razvili pod okriljem Centra za jezikovne vire in tehnologije
- Uporabili smo programski jezik Python s knjižnico Flask.
- Uporabniški vmesnik smo napisali z uporabo jezikov HTML, CSS in javascript.

Implementacija spletnega orodja

- Celotno besedilo, ki ga uporabnik vnese, se posreduje na strežnik.
- Besedilo se na strežniku razdeli na posamezne povedi, ki jih program obdela in v njih postavi vejico.
- Iz stavka se najprej pobrišejo vse vejice. Program nato samostojno postavi vejice in na koncu primerja začetno in popravljeno različico besedila.
- Če je program vejico izbrisal, besedo pred njo obarva z modro barvo; če je vejico dodal, besedo obarva s sivo barvo.

To je vzorčno besedilo, s katerim želimo prikazati delovanje orodja. V tem besedilu so zato na določenih mestih vejice odstranjene, na drugih pa po nepotrebnem dodane. Orodje opozarja na manjkajoče vejice (s sivo barvo), hkrati pa tudi na odvečne (z modro barvo). Glede na teste program trenutno deluje 94 % uspešno.

Demonstracija spletnega orodja.

- Spletno orodje je dostopno na naslovu: https://orodja.cjvt.si/vejice/home

Načrti za prihodnost

- Od objave spletnega orodja smo zabeležili kar nekaj pripomb in pohval iz strani uporabnikov.
- Spremembe v povezavi s samim spletnim vmesnikom so lažje.
- Najtežje so spremembe pri delovanju samega modela, saj potrebujemo celoten model naučiti še enkrat, pri tem pa nimamo neke garancije, da bo model deloval boljše in da bomo omenjeno težavo sploh rešili.