Generiranje neologizama iz njihovog opisa na hrvatskom jeziku

Ovaj rad istražuje primjenu dubokog učenja za generiranje neologizama — novih riječi koje sažimaju određeni koncept ili definiciju. Problem se tretira kao zadatak nadzirane sumarizacije gdje je ulaz definicija pojma, a izlaz jedna, novostvorena riječ. U projektu je implementiran i dotreniran model **ByT5-small**, koji radi na razini znakova, što ga čini pogodnim za morfološke manipulacije potrebne pri stvaranju novih riječi.

1. Uvod

Stvaranje novih riječi (neologizama) ključno je u brendiranju, marketingu i umjetnosti, ali zahtijeva duboko poznavanje pravila tvorbe riječi. Tradicionalne metode često se oslanjaju na ručno definirana pravila , dok ovaj projekt koristi pristup temeljen na prijenosu učenja (transfer learning) kako bi model sam naučio implicitna pravila kombiniranja morfema.

2. Prijašnji rad

Ovaj rad istražuje primjenu dubokih predtreniranih modela za generiranje neologizama (novih riječi). Autori problemu pristupaju kao zadatku **nadzirane sumarizacije** (supervised summarization) u kojem je ulaz definicija pojma, a očekivani izlaz nova, sažeta riječ koja predstavlja taj koncept.

Ključne ideje i metodologija:

- **Modeli:** U radu se analiziraju i uspoređuju dva tipa "encoder-decoder" modela (seq2seq): **T5-base** (koji radi na razini podriječi i ima 220 milijuna parametara) te **ByT5-small** (koji radi na razini bajtova/karaktera i ima 300 milijuna parametara).
- **Podaci:** Za treniranje je korišten engleski rječnik **Kaikki**, koji sadrži preko milijun riječi s pripadajućim definicijama. Podaci su pročišćeni kako bi se zadržali smisleni termini (riječi duže od 5 znakova, bez brojki i sl.).
- **Cilj istraživanja:** Cilj je bio ispitati mogu li modeli induktivno naučiti osnovna pravila tvorbe riječi (poput spajanja osnova i sufiksa) te generirati originalne i kreativne neologizme koji su razumljivi ljudima.

Rezultati i doprinos: Rad demonstrira uspješnost modela u usvajanju pravila tvorbe riječi. Budući da standardne metrike za ovaj specifičan zadatak ne postoje, provedena je kvalitativna analiza putem ljudske evaluacije na nasumičnom uzorku generiranih riječi.

3. Opis modela

U ovom radu za zadatak generiranja neologizama korišten je model **ByT5-small** (varijanta *google/Byt5-small*), koji predstavlja arhitekturu temeljenu na transformeru, optimiziranu za rad na razini bajtova. Za razliku od standardnih T5 modela koji koriste rječnik podriječi, ByT5 obrađuje tekst kao niz bajtova, što ga čini iznimno robusnim za zadatke kreativnog generiranja riječi gdje morfološka

struktura i sastavljanje novih nizova znakova igraju ključnu ulogu, naročito za morfološki bogate jezike poput hrvatskog jezika.

Proces treniranja i evaluacije proveden je na sljedeći način:

- **Arhitektura i podaci:** Model je konfiguriran kao *sequence-to-sequence* (seq2seq) model gdje je ulazni podatak opis (stupac *Description*), a ciljani izlaz sažeta nova riječ (stupac *Word*). Podaci su podijeljeni u skupove za trening, razvoj i testiranje, pri čemu je skup za treniranje sadržavao 30.909 uzoraka.
- **Hiperparametri:** Trening je proveden kroz 5 epoha uz stopu učenja (engl. *learning rate*) od 0.00005 koristeći **AdamW** optimizator. *Batch size* postavljen je na 8, uz maksimalnu duljinu ulaznog niza od 64 te izlaznog od 16 znakova.
- **Metrike evaluacije:** Kvaliteta generiranih neologizama mjerena je kroz tri različite dimenzije:
    1. **n-gram F1 rezultat:** Mjeri preklapanje n-grama znakova (n=3) između ciljane i generirane riječi.
    2. **Semantička sličnost:** Korišten je model *distiluse-base-multilingual-cased-v2* za izračun kosinusne sličnosti između vektorskih prikaza (engl. *embeddings*) riječi kako bi se utvrdilo koliko generirana riječ po značenju odgovara originalu.
    3. **Cross-Entropy Loss:** Korišten je za mjerenje vjerojatnosti kojom model predviđa točnu ciljanu riječ s obzirom na opis.
- **Inferencija:** Prilikom generiranja riječi korištena je metoda uzorkovanja (engl. *sampling*) s parametrima top_k=10, top_p=0.95 i temperaturom od 0.7 kako bi se osigurala ravnoteža između kreativnosti i koherentnosti izlaza.

4. Podaci

Proces izgradnje korpusa za ovaj rad sastojao se od sustavnog prikupljanja podataka iz relevantnih jezičnih izvora te njihove naknadne predobrade kako bi se osigurala kvaliteta ulaza za model. Podaci su prikupljeni iz dva glavna izvora:

- **Hrvatski jezični portal (HJP):** Korištenjem biblioteke *BeautifulSoup* za automatizirano strugaanje weba (engl. *web scraping*), prikupljene su definicije i pripadajuće riječi iz baze HJP-a. Ovaj izvor poslužio je kao temelj za učenje standardnih morfoloških i semantičkih pravila hrvatskog jezika.
- **Baza neologizama:** Uz standardne riječi, u korpus su uključeni i specifični primjeri neologizama kako bi se model potaknuo na kreativno generiranje.

Predobrada i čišćenje podataka: Sirovi podaci prošli su kroz proces čišćenja koji je uključivao uklanjanje HTML oznaka, normalizaciju teksta (pretvaranje u mala slova, uklanjanje suvišnih razmaka) te filtriranje prekratkih ili nerelevantnih unosa. Posebna pažnja posvećena je uklanjanju posebnih znakova i brojeva koji bi mogli unijeti šum u proces učenja.

Podjela skupa podataka: Konačni skup podataka (korpus) podijeljen je u tri dijela kako bi se omogućio ispravan razvoj i evaluacija modela:

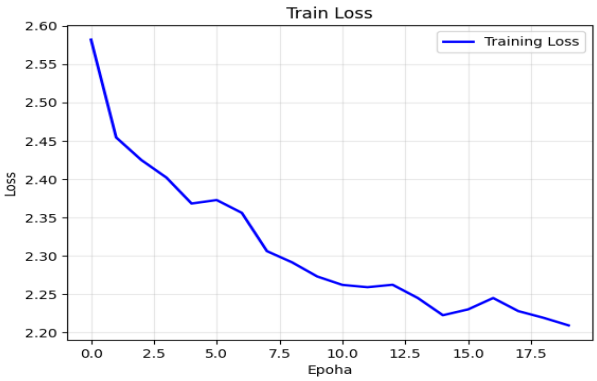
1. **Skup za treniranje (Train):** Najveći dio podataka (80%) korišten za prilagodbu parametara modela.
2. **Skup za razvoj (Dev):** Korišten za ugađanje hiperparametara i praćenje gubitka (engl. *loss*) tijekom treninga.
3. **Skup za testiranje (Test):** Zadržan kao neovisni skup za konačnu evaluaciju performansi i sposobnosti generalizacije modela na neviđenim primjerima.

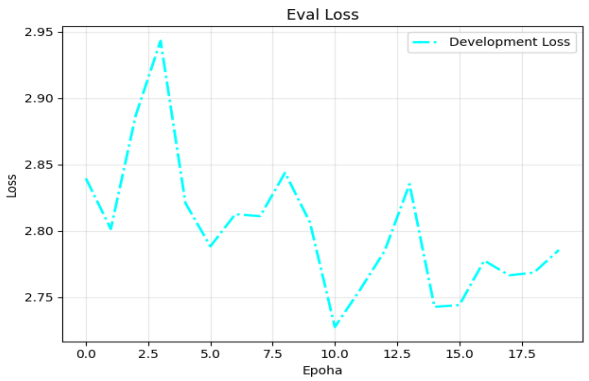
Ukupni broj uzoraka u konačnom korpusu iznosi preko 30.000, što osigurava dovoljnu raznolikost za stabilan trening modela poput ByT5.

5. Rezultati

5.1. Baseline model

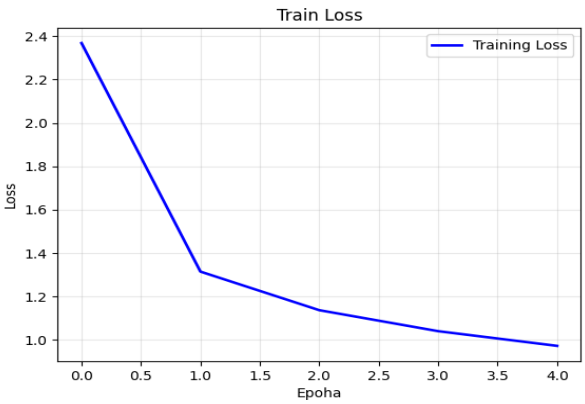
Baseline model korišten u ovom projektu je Seq2Seq model kojemu su i enkoder i dekoder LSTM. Model nije predtreniran, odnosno treniran je od nule na korpusu opisanom ranije u tekstu. Pri treniranju je korišteno 20 epoha, a stopa učenja je 0.001. Batch size je postavljen na 64, a LSTM ima 2 sloja.

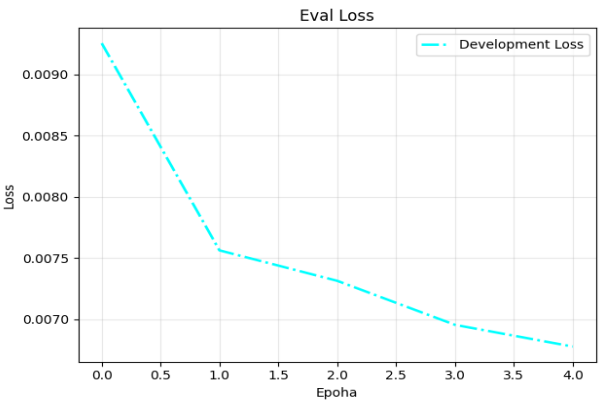




5.2. ByT5 model

Model je treniran na NVIDIA GPU (cuda). Tijekom treninga pratila se vrijednost funkcije gubitka kako bi se osiguralo da model uči pravila tvorbe riječi iz zadanih definicija.





5.3. Rezultati i usporedba modela

Baseline model nije predtreniran i ima puno jednostavniju strukturu, pa je time i rezultat osjetno lošiji. 3-gram score Nije poželjno da bude previsok, jer se u projektu teži generiranju neologizama, što znači da se riječi neće uvijek podudarati, ali bi trebale biti semantički slične, pa je bitno da semantic similarity bude što veći. Cross entropy mjeri koliko je model iznenađen

Model	Cross entropy	3-gram F1 score	Semantic similarity
Baseline	2.8086	0.0255	0.6460
ByT5	1.0090	0.2748	0.7503

Opis	Ciljana riječ	ByT5	Baseline
grafički prikaz koji pokazuje suodnos promjenjivih veličina	grafikon	obrađenje	proaanje
trava koja se skine s polja košnjom	otkos	košnjača	proaa
ne priznati komu što (o kakvu pravu)	pobiti	prikazivati	proviti
koji je napunjen perjem	pernat	peran	statan
glazbenica koja upravlja izvedbom glazbenoga djela	dirigentica	izvednica	proaa

6. Analiza i zaključak

Provedeno istraživanje potvrdilo je djelomičnu učinkovitost modela **ByT5-small** u zadatku generiranja neologizama na hrvatskom jeziku. Za razliku od **baseline LSTM modela**, koji je generirao fonetski neispravne nizove (poput *"proaa"*), ByT5 je uspješno usvojio morfološka pravila i mehanizme tvorbe riječi.

Iako model nije uvijek generirao identičnu riječ ciljanoj, visoka semantička sličnost (0.7503) i primjeri poput kreiranja riječi **"izvednica"** za opis dirigentice ukazuju na to da je model sposoban za kreativnu sintezu, a ne samo za puko reproduciranje rječnika. S obzirom na to da je rad proveden na jeziku s ograničenim resursima, dobiveni rezultati čine solidan temelj za daljnji razvoj automatiziranih sustava za brendiranje i kreativno pisanje.

Buduća istraživanja mogla bi se usmjeriti na:

- Korištenje većih varijanti modela (ByT5-base ili large) za postizanje još preciznije semantike.
- Proširenje korpusa specifičnim tehničkim i stručnim rječnicima.
- Uvođenje ljudske evaluacije (engl. *human-in-the-loop*) kako bi se preciznije izmjerila "kreativnost" i tržišna primjenjivost generiranih naziva.

Literatura

Gabriel Lencione, Neologism Generation from Descriptions, ICCC 2022.

google/byt5-small

Školski rječnik Hrvatskog jezika

GitHub repozitorij