ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O'ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 3 (6)

www.compling.tsuull.uz

MUNDARIJA

Mavjuda Alimbekova
Abdurauf Fitrat mualliflik korpusini yaratishning ahamiyati.....6

Madinabonu Qodirova, Shahlo Hamroyeva
Zamonaviy dunyoda mashina tarjimasi tadriji:
tahlillar va natijalar.....22

Noila Matyakubova
"Aligner" dasturiy vositasi uchun o‘zbek-ingliz tilida sifat va uning
darajalarining morfologik tahlili.....41

Mohiyaxon Uzoqova, Mansurbek Narzullayev
Sinonimayzer dasturida RoBERTaForMaskedLM modelidan leksik
sinonimlarni aniqlash uchun foydalanish.....54

Dlafruz Xudoyqulova
O‘zbek-ingliz farmatsevtika terminlari korpusli lingvistik
ta'minotining milliy-madaniy asoslari.....69

Ruhillo Alayev, Gulshaxnoz Maxmudjonova
O‘zbek tilidagi matnli hujjatlarda izlashni amalga
oshirishni takomillashtirish.....78

Sanjarbek Baxodirov
Tabiiy tilni qayta ishlashda matnni tozalash tizimini
ishlab chiqish.....91

Azizaxon Raxmanova
Sun’iy intelekt yordamida o‘zbek va ingliz tili lingvistik asoslarini
o‘qitishning zamonaviy uslublari.....106

ISSN 2181-922X 54-68

Sinonimayzer dasturida RoBERTaForMaskedLM modelidan leksik sinonimlarni aniqlash uchun foydalanish

Mohiyaxon Uzoqova[1]
Mansurbek Narzullayev[2]

Annotatsiya. Roberta turkumiga kiruvchi RobertaForMaskedLM modelini leksik sinonimlarni aniqlash uchun ishlatish mumkinligini ilgarigi tadqiqotlarimizda qayd etgan edik. Biroq mashq uchun to‘plangan ma’lumotning hajmi kichikligi va sifatining yuqori bo‘lmaganligi bois qoniqarli natija olinmaganligini xabar bergan edik. Ushbu tadqiqot davomida esa mashq uchun mo‘ljallangan hujjatlar hajmi hamda sifatini oshirgan holda bir necha bosqichda mashq jarayonini qayta amalga oshirdik. Buning natijasida esa yangi va ishonarli javoblarni qo‘lga kiritdik, natijalar tahlilini amalga oshirdik va jadvallarda aks ettirdik. Shuningdek, modelni o‘zbekcha sinonimayzer dasturiga muvaffaqiyatli tarzda integratsiya qildik.

Kalit so‘zlar: *sinonim, sinonimayzer, dataset, Roberta, tokenayzer, training loss, validation loss.*

Kirish

RoBERTa (Robustly Optimized BERT Pretraining Approach – BERT modelini ilgaridan o‘qitishga asoslangan puxta optimallashtirilgan model) ForMaskedLM modeli BERT modelining [Jacob va boshqalar, 2019. 4171] kuchaytirilgan varianti bo‘lib [Liu et al. 2019], neyron tarmoq bilan ishlovchi zamonaviy model hisoblanishi haqida ilgarigi tadqiqotimizda to‘xtalgan edik [Uzoqova, 2023. 328]. Bu model 2019-yilda FacebookAI guruhi tomonidan ishlab chiqilgan va turli maqsadlarda: mashina tarjimasida, tabiiy tilni qayta tushunishda va matn tasnifida ishlatilishi mumkin [Huggingface, 2023].

Modelning asosiy vazifasi matndan maqsadli tushirib qoldirilgan tokenga (“masked”) mos tokenlarni taqdim etishdir. Biz mo-

[1]*Uzoqova Mohiyaxon Tuyg‘un qizi* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi mutaxassisligi 2-kurs magistranti
E-pochta: uzoqovamohiyaxon@navoiy-uni.uz
ORCID: 0000-0001-7102-0824
[2]*Narzullayev Mansurbek Nurali o‘g‘li* – Muhammad Al-Xorazmiy nomidagi Toshkent davlat axborot texnologiyalari universiteti Samarqand filliali Dasturiy Injiniring mutaxassisligi 2-kurs magistranti
E-pochta: mansurbeknarzullayev@mail.ru
ORCID: 0000-0002-8160-6631

delning ushbu vazifasidan kelib chiqqan holda uni matndagi leksik sinonimlarni topish uchun yo‘naltirishga qaror qilgan edik. Ya’ni *RoBERTa modeli bilan gapdagi bir so‘zning sinonimlari shu gapga (kontekstga) mos kelishini tekshirgan edik* [Uzoqova, 2023. 328]. Bu tekshiruv dastlab juda kichkina hajmdagi va nisbatan kattaroq hajmdagi (taqriban 300MB) data to‘plam bilan amalga oshirilgan edi. Ammo samaradorlik yuqori bo‘lmagani bois biz data to‘plamning hajmini, sifatini hamda mashq qilish uchun resurslarni kuchaytirgan holda yangi tajriba o‘tkazdik va sezilarli darajada ijobiy natija oldik. Jarayon quyidagi ketma-ketlikda amalga oshirildi.

Asosiy qism

I bosqich. RobertaForMaskedLM modeli va tokenayzer uchun dataset tayyorlash.

Tadqiqotimiz davomida shunga amin bo‘ldikki, o‘zbekcha matnlarni qayta ishlashda asosiy ish quroli bo‘lib xizmat qiluvchi datasetlar ko‘p emas. Shu sababli modellarni mashq qildirish maqsadida oldindan tayyorlab qo‘yilgan tayyor datasetlarni topish jarayoni qiyin kechdi.

Ma’lumki, dataset tayyorlashning bir qancha sinalgan mashhur yo‘llari mavjud [Dullin, 2023]. Masalan, a) datasetlar saqlanuvchi tadqiqotchilarga mo‘ljallangan hugging face [Huggingface, datasets 2023], kaggle [Kaggle, datasets 2023], amazon [Amazon, opendata, 2023], google [Google, dataresearch 2023], subreddit [Reddit, datasets 2023] kabi platformalardagi maxsus sahifalardan foydalanish; b) GitHubdan izlash; d) datani tortib olish (crawling) yo‘lidan foydalanish; e) ilmiy ishlarda e’lon qilingan ilovalardan muallifning ruxsati bilan foydalanish va b. Biz o‘z tadqiqotimizda yuqorida aytilgan holat tufayli quyidagi yo‘llar orqali datasetni shakllantirishga harakat qildik:

1. Veb saytlarni “crawl” qilish orqali ulardagi ma’lumotlarni ajratib olish.

Axborotning ishonchliligi, hajmi va sifatini inobatga olgan holda quyidagi veb-saytlardan data “crawl” qilindi:

– Kun.uz;

– Daryo.uz;

– Uza.uz;

– Lex.uz.

2. O‘zbek tili uchun yaratilgan tayyor datasetlarni yuqorida eslangan platformalardan topish.

Quyidagi datasetlardagi ma’lumotlar olindi:

– tahrirchi/uz-crawl [Huggingface, tahrirchi (a) 2023];

– tahrirchi/uz-books [Huggingface, tahrirchi (b) 2023];

– murodbek/uz-text-classification [Huggingface, murodbek 2023];

– elmurod1202/uzbek-sentiment-analysis [Huggingface, elmurod 2023];

– CC100 Dataset [Metatext, cc100 2023].

3. Ko‘ptilli datasetlardan (multilingual datasets) o‘zbek tiliga oid qismlarini ajratib olish.

Quyidagi ko‘ptilli datasetlardan o‘zbekcha qismi ajratib olindi:

– oscar [Huggingface, oscar 2023];

– mc4 (huggingface) [Huggingface, mc4 2023];

– wikimedia/wikipedia [Huggingface, wikimedia/wikipedia 2023].

Ushbu usullardan foydalanib 12.7 GB parquet format hajmidagi umumiy dataset to‘plandi va huggingface platformasidagi sahifamizga yuklandi [Huggingface, sinonimayzer 2024] .

Dataset bir qancha algoritmlar bilan kichik-kichik qatorlarga ajratilgan. So‘ng datasetdagi faqat lotincha ma’lumotlarni qoldirish va nisbatan katta qatorlarni hamda faqat 1 ta tokendan iborat qatorlarni tushirib qoldirish uchun bir qancha filterlar ishlatilgan, bular:

1. Matnda 50% dan ko‘p kirillcha belgilar bo‘lmasligi kerak.

2. Matn kamida 2 ta tokendan iborat bo‘lishi kerak (bunda bitta so‘zdan iborat datasetlarni hisobga olinmadi).

3. Matn ko‘pi bilan 32768 ta tokendan iborat bo‘lishi kerak (datasetni huggingface platformasida saqlashimiz uchun undagi har bir qator uzunligi, taxminan, 40000 ta tokendan oshmasligi kerak, shu sababdan 2^15 ta token maksimal uzunlik deb olingan).

Dataset filterlangach, uni 2 ta qismga ajratdik: “train” – modelni mashqi qildirish uchun va “validation” – modelni testlash uchun. Shundan umumiy datasetning 90%i modellarni mashq qildirish uchun, 10% qismi modelni baholash uchun uchun ishlatildi.

Datasetlar alohida alohida huggingface sahifasiga yuklanadi. Buning uchun datasets kutubxonasidan foydalanilgan. Datasetni yuklashda har bir fayl hajmi maksimal 500MB qilib belgilangan.

II bosqich. Tokenizer modelini train qilish.

1-qadam. HuggingFace jamiyati tomonidan ishlab chiqilgan *tokenizers* kutubxonasi orqali avvaldan mashq qildirilgan tokenayzer chaqirildi (import qilindi) va dataset uchun sozlamalar moslandi

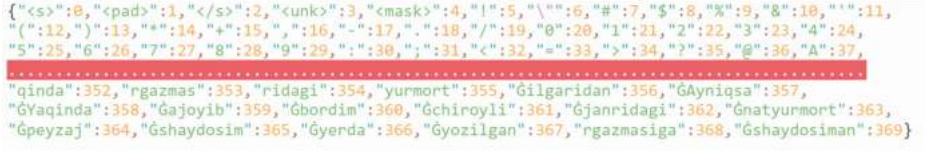
(Qarang: 1-rasm).

```
DATASET_ID      = "sinonimayzer/mixed-data"
VOCAB_SIZE      = 52000
NUM_PROC        = 20
def trainTokenizer(dataset):
    def batch_iterator(batch_size=100 * NUM_PROC):
        for i in range(0, len(dataset), batch_size):
            yield dataset[i : i + batch_size]["text"]
    printLine("Train tokenizer")
    tokenizer = ByteLevelBPETokenizer()
    tokenizer.train_from_iterator(
        batch_iterator(),
        vocab_size=VOCAB_SIZE,
        min_frequency=2,
        show_progress=True,
        special_tokens=[
            "<s>",
            "<pad>",
            "</s>",
            "<unk>",
            "<mask>",
        ],
    )
```

1-rasm. Tokenayzerni chaqirish va berilgan matn uchun sozlamalarni moslash

Berilgan matn 1: *Men ilgaridan go‘zal rasmlar shaydosiman. Yaqinda ajoyib rasmlar ko‘rgazmasiga bordim. U yerda chiroyli rasmlar ko‘p edi. Ayniqsa, peyzaj va natyurmort janridagi rasmlar go‘zal yozilgan edi.*

2-qadam. Tokenayzer matndan (datasetdan) o‘zi uchun unikal tokenlarni yasadi va ularga tartib raqami (ID) berdi. Sozlamalarda yasama tokenlarning maksimal hajmi (“vocab_size”, qarang: 3.2.5-rasm) standart holat uchun 52 000 gacha deb belgilangan. Bizning holatimizda (berilgan matn 1) unikal yasama tokenlar miqdori 369 taga yetdi (2-rasm).



2-rasm. Berilgan matnning tokenlarga ajratilishi va raqamlanishi

Umumiy dataset bo‘yicha esa 52000 tagacha token yasalgan (Qarang: 3-rasm):

3-rasm. Berilgan matnning tokenlarga ajratilishi va raqamlanishi

To‘liq ro‘yxatni bizning sahifamizda [Huggingface, sinonimayzer/UzRoBerta 2024] ko‘rish mumkin.

3-qadam. Hosil qilingan tokenlar saqlanadigan jildning manzili ko‘rsatildi. Model ko‘rsatilgan jildda ikkita fayl hosil qilindi: vocab.json va merges.json. Bu fayllarda Tokenayzerning vazifasi bu jarayonda yakuniga yetdi.

III bosqich. RoBERTa modelini mashq qildirish.

1-qadam. Model uchun sozlamalar yozildi va bu sozlamalar asosida RobertaForMaskedLM modeli hosil qilindi (Qarang: 4-rasm):

```
data_collator = DataCollatorForLanguageModeling(
    tokenizer=tokenizer,
    mlm=True,
    mlm_probability=0.15
)
training_args = TrainingArguments(
    per_device_train_batch_size=BATCH_SIZE,
    per_device_eval_batch_size=BATCH_SIZE,
    dataloader_num_workers=NUM_PROC,
    logging_steps=LOGGING_STEPS,
    save_total_limit=SAVE_LIMIT,
    hub_model_id=HUB_MODEL_ID,
    output_dir=MODEL_FOLDER,
    logging_dir=LOG_FOLDER,
    save_steps=SAVE_STEPS,
    eval_steps=EVAL_STEPS,
    hub_token=HUB_TOKEN,
    max_steps=MAX_STEPS,
    evaluation_strategy="steps",
    save_strategy="steps",
    gradient_checkpointing=True,
    overwrite_output_dir=True,
    disable_tqdm=False,
    push_to_hub=True,
    data_seed=True,
    report_to=["tensorboard"]
)

config = RobertaConfig(
    vocab_size=VOCAB_SIZE,
    type_vocab_size=1,
)

model = RobertaForMaskedLM(config=config)
```

4-rasm. RoBERTa modeli uchun sozlamalarni moslash

2-qadam. Datasetni model uchun moslash (5-rasm):

```
def genDataset(dataset, tokenizer, block_size):

    def tokenize_function(examples):
        return tokenizer(
            examples["text"]
        )

    def group_texts(examples):
        concatenated_examples = {k: sum(examples[k], []) for k in examples.keys()}
        total_length = len(concatenated_examples[list(examples.keys())[0]])
        total_length = (total_length // block_size) * block_size
        result = {
            k: [t[i : i + block_size] for i in range(0, total_length, block_size)]
            for k, t in concatenated_examples.items()
        }
        result["labels"] = result["input_ids"].copy()
        return result

    printLine("Tokenize dataset")
    tokenized_dataset = dataset.map(
        tokenize_function,
        batched=True,
        batch_size=NUM_PROC * NUM_PROC,
        num_proc=NUM_PROC,
        remove_columns=["text"]
    )

    printLine("Group dataset map")
    ready_dataset = tokenized_dataset.map(
        group_texts,
        batched=True,
        batch_size=NUM_PROC * NUM_PROC,
        num_proc=NUM_PROC,
    )

    return ready_dataset
```

5-rasm. To‘plangan datasetning model uchun moslashtirish jarayoni

3-qadam. Berilgan matn avvaldan hosil qilingan tokenayzer yordamida raqamli ko‘rinishga o‘tkazildi (Qarang: 6-rasm).

<s>	Men	ilgaridan	go	‘	zal	rasmlar	shaydosiman	.	Yaqinda
0	348	356	283	269	285	270	369	18	358
ajoyib	rasmlar	ko	‘	rgazmasiga	bordim	.	U	yerda	chiroyli
359	270	284	269	368	360	18	328	366	361
rasmlar	ko	‘	p	edi	.	Ayniqsa	,	peyzaj	va
270	284	269	84	282	18	357	16	364	339
natyurmort	janridagi	rasmlar	go	‘	zal	yozilgan	edi	.	</s>
363	362	270	283	269	285	367	282	18	2

6-rasm. Berilgan matnni tokenayzer natijasiga muvofiq raqamlashning jadval ko‘rinishida aks etishi

4-qadam. Model mashq qildirildi va argumentda ko‘rsatilgan joyga saqlandi (Qarang: 7-rasm).

```
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
model.to(device)

trainerRoberta = Trainer(
    model=model,
    args=training_args,
    tokenizer=tokenizer,
    data_collator=data_collator,
    train_dataset=train_dataset,
    eval_dataset=eval_dataset
)

trainerRoberta.train()
trainerRoberta.save_model(MODEL_FOLDER)
```

7-rasm. Berilgan matnni mashq qildirish jarayoni kodlarda

III bosqich. Modeldan foydalanish.

Model saqlangan joydan qayta yuklab olindi va unga test uchun gap berildi. Test sifatida berilayotgan gapdan bir so‘z *<mask>* tokeni ostida maqsadli ravishda tushirib qoldirildi. Namuna uchun *“U yerda chiroyli rasmlar ko‘p edi”* gapi olindi (Qarang: 8-rasm).

```
from transformers import pipeline
unmasker = pipeline('fill-mask', model='sinonimayzer/UzRoBERTa-v3')
unmasker("U yerda <mask> rasmlar ko'p edi")
```

8-rasm. Model yordamida tushib qolgan so‘zni topish

Quyidagi natija olindi:

```
C:\Disc\Projects\uz-synonimizer\test>python tester_roberta.py
{'score': 0.0001581029937369749, 'token': 285, 'token_str': 'zal', 'sequence': 'U yerdazal rasmlar ko'p edi'}
{'score': 0.0001485073708156123 8, 'token': 7835, 'token_str': '', 'sequence': 'U yerda rasmlar ko'p edi'}
{'score': 0.0001331435058015343 9, 'token': 43643, 'token_str': '', 'sequence': 'U yerda rasmlar ko'p edi'}
{'score': 0.0001282269950024783 6, 'token': 26899, 'token_str': '', 'sequence': 'U yerda rasmlar ko'p edi'}
{'score': 0.0001239131524926051 5, 'token': 13471, 'token_str': '', 'sequence': 'U yerda rasmlar ko'p edi'}
```

9-rasm. Sinov uchun berilgan gap asosida olingan eng yuqori beshta natija

Natijadan ko‘rinib turibdiki, birorta token ma’noga ega emas. Faqatgina birinchi token “go‘zal” so‘zining ikkinchi bo‘g‘inini takrorlagani bois “chiroyli” so‘ziga shartli ravishda yaqin deb qarash mumkin. Modelning bunday natija qaytarishiga asosiy sabab datasetning hajmi kichkinaligidir. Shuning uchun biz Word2Vec modelida sinaganimiz kabi kattaroq dataset [Kaggle, mohiyaxonuzokova 2023] bilan natija olishga harakat qildik (Qarang: 10-rasm):

```
[{'score': 0.11681189388036728, 'token': 378, 'token_str': ' ham', 'sequence': "U yerda ham rasmlar ko'p edi"},
 {'score': 0.06719416379928589, 'token': 419, 'token_str': ' bu', 'sequence': "U yerda bu rasmlar ko'p edi"},
 {'score': 0.05429129028511047 85, 'token': 4662, 'token_str': ' ajoyib', 'sequence': "U yerda ajoyib rasmlar ko'p edi"},
 {'score': 0.04048766940832138, 'token': 710, 'token_str': ' esa', 'sequence': "U yerda esa rasmlar ko'p edi"},
 {'score': 0.03192049637436867, 'token': 508, 'token_str': ' boshqa', 'sequence': "U yerda boshqa rasmlar ko'p edi"}]
```

10-rasm. Sinov uchun berilgan gap bo‘yicha yangi mashq-matn orqali olingan eng yuqori beshta natija

Biroq bu safar ham “chiroyli” so‘zi o‘rniga *“G”, “,”, “A”, “E”, “K”* kabi ma’nosiz natijalarni oldik. Buning asosiy sababi test sifatida olingan gapdagi so‘zlar butun hujjat davomida juda kam o‘zaro aloqaga kirishgani deb qaraldi. Modelni boshqa namuna bilan sinashda davom etdik.

Model turli xil parametrlarda mashq qildirildi.

Modelning kichik dataset bilan mashq qildirilgan **dastlabki versiyasida** [Uzoqova, 2023. 328] ***“Bugun Toshkent <mask> anjuman bo‘lib o‘tdi”*** gap uchun quyidagicha natijalar olindi (Qarang: 1-jadval):

score	token	token_str	Sequence
0.06776456534862518	1417	shahrida	Bugun Toshkent **shahrida** anjuman bo‘lib o‘tdi
0.011984631419181824	774	viloyati	Bugun Toshkent **viloyati** anjuman bo‘lib o‘tdi
0.0024553691036999226	264	da	Bugun Toshkent**da** anjuman bo‘lib o‘tdi
0.000220525631448254	2191	hududida	Bugun Toshkent **hududida** anjuman bo‘lib o‘tdi
0.00011574733798624948	501	larda	Bugun Toshkent**larda** anjuman bo‘lib o‘tdi

1-jadval. Sinov uchun berilgan gap asosida modelning kichik dataset bilan mashq qildirilgan dastlabki versiyasidan olingan eng yuqori beshta natija

Modelning hozirgi to‘plangan dataset yordamida 200 ming qadamda mashq qildirilgan modelda [Huggingface, sinonimayzer/UzRoBerta-v1 2024] xuddi shu gap quyidagi natijalarni berdi (Qarang: 2-jadval):

score	token	token_str	Sequence
0.8364855051040649	1708	shahrida	Bugun Toshkent **shahrida** anjuman bo‘lib o‘tdi
0.04394292086362839	2342	viloyatida	Bugun Toshkent **viloyatida** anjuman bo‘lib o‘tdi
0.02987276390194893	1102	xalqaro	Bugun Toshkent **xalqaro** anjuman bo‘lib o‘tdi
0.026901429519057274	2988	shahridagi	Bugun Toshkent **shahridagi** anjuman bo‘lib o‘tdi
0.01506294310092926	4622	markazida	Bugun Toshkent **markazida** anjuman bo‘lib o‘tdi

2-jadval. Sinov uchun berilgan gap asosida modelning yirik dataset bilan 200 ming qadamda mashq qildirilgan dastlabki versiyasidan olingan eng yuqori beshta natija

Xuddi shu dataset orqali modelning ikkinchi versiyasi [Huggingface, sinonimayzer/UzRoBerta-v2 2024] 500 ming qadamda train qilinganida quyidagicha natijalar olindi (Qarang: 3-jadval):

score	token	token_str	Sequence
0.847356915473938	1708	shahrida	Bugun Toshkent **shahrida** anjuman bo‘lib o‘tdi
0.021089091897010803	1102	xalqaro	Bugun Toshkent **xalqaro** anjuman bo‘lib o‘tdi
0.0201409962028265	7817	universitetida	Bugun Toshkent **universitetida** anjuman bo‘lib o‘tdi
0.01779646798968315	2988	shahridagi	Bugun Toshkent **shahridagi** anjuman bo‘lib o‘tdi
0.01506294310092926	2342	viloyatida	Bugun Toshkent **viloyatida** anjuman bo‘lib o‘tdi

3-jadval. Sinov uchun berilgan gap bo‘yicha asosida modelning yirik dataset bilan 500 ming qadamda mashq qildirilgan dastlabki versiyasidan olingan eng yuqori beshta natija

Har bir versiyadan olingan natijalardan ko‘rinib turibdiki, bo‘shliq uchun taklif etilgan har bir tokenning (token_str) qiymati (score) avvalgi versiyaga qaraganda yuqoriroq va aniqroq.

Quyida modelning har 100 ming qadamda olingan “train” va “validation loss” qiymatlari ko‘rsatilgan (Qarang: 4-jadval):

Training loss	Epoch	Step	Validation loss
2.3673	0.25	100000	2.4588
2.0797	0.51	200000	2.1653
1.9369	0.76	300000	2.0265
1.8545	1.02	400000	1.9456
1.8133	1.27	500000	1.9101

4-jadval. Har 100,000 ta qadamda mashq qildirilgach olingan natija

Modelning ikkinchi versiyasidagi natijalarni tahlil qilib 500 ming qadamda model bor yog‘i 1.2 “epoch”ga yetishi va natijalarni yanada yaxshilash mumkinligi taxmin qilindi; modelning uchinchi versiyasi [Huggingface, sinonimayzer/UzRoBerta-v3 2024] 1 millionta qadam orqali mashq qildirildi va quyidagicha natijalar olindi (Qarang: 5-jadval):

score	token	token_str	Sequence
0.904608964920044	1708	shahrida	Bugun Toshkent **shahrida** anjuman bo‘lib o‘tdi
0.021837687119841576	2988	shahridagi	Bugun Toshkent **shahridagi** anjuman bo‘lib o‘tdi
0.011045671068131924	1102	xalqaro	Bugun Toshkent **xalqaro** anjuman bo‘lib o‘tdi
0.011010712012648582	7817	universitetida	Bugun Toshkent **universitetida** anjuman bo‘lib o‘tdi
0.004504858981817961	13072	aeroportida	Bugun Toshkent **aeroportida** anjuman bo‘lib o‘tdi

5-jadval. Sinov uchun berilgan gap asosida modelning yirik dataset bilan 1.000.000 ta qadamda mashq qildirilgan dastlabki versiyasidan olingan eng yuqori beshta natija

Uchinchi versiyadan ko‘rinib turibdiki, model datasetni anchagina yaxshi o‘rgangan va mos keladigan qiymatlar kattaroq “score”ga ega.

Biz olingan natijalarga asoslangan holda avvaldan shakllan-

tirilgan sinonimlar bazasiga modelni integratsiya qildik. Natijada, sinonimayzer dasturi quyidagicha ishlamoqda (Qarang: 11-rasm):

11-rasm. RoBertaForMaskedLM modelini sinonimayzer dasturiga integratsiya qilish orqali olingan natija

Xulosa

RobertaForMaskedML modeli, guvoh bo‘ldikki, datsetning hajmi qancha yirik bo‘lsa, shuncha sifatli ishlaydi. Masalan, **RoBERTa base** modelining o‘zini mashq qildirish uchun jami 160 GB hajmdagi matndan foydalanilgan [Huggingface, robertabase 2023]. Natijalar esa *ishonchli* va *yaroqli.*

Demak, yuqorida ta’kidlangandek, gapdagi muayyan so‘zning sinonimlari shu kontekstga mos kelishini tekshira olishimiz uchun bizga yirik hajmdagi sifatli va avvaldan o‘qitilgan dataset kerak [Baevski va boshqalar, 2019. 5360].

Foydalanilgan adabiyotlar

Amazon, opendata. 2023. https://registry.opendata.aws/

Baevski, Alexei; Edunov, Sergey; Liu, Yinhan; Zettlemoyer; Luke & Auli, Michael. 2019. Cloze-driven Pretraining of Self-attention Networks. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.* 5363-5372, Hong Kong, China. DOI: 10.18653/v1/D19-1539

DOI: 10.18653/v1/N19-1423.

Dullin, Theo. 2023. How to Create Datasets: strategies and examples https://kili-technology.com/data-labeling/machine-learning/create-dataset-for-machine-learning

Google, dataresearch. 2023. https://datasetsearch.research.google.com/

Huggingface, datasets. 2023. https://huggingface.co/docs/datasets/index

Huggingface, elmurod. 2023. https://huggingface.co/datasets/elmurod1202/uzbek-sentiment-analysis

Huggingface, mc4. 2023. https://huggingface.co/datasets/mc4

Huggingface, murodbek. 2023. https://huggingface.co/datasets/murodbek/uz-text-classification

Huggingface, oscar. 2023. https://huggingface.co/datasets/oscar

Huggingface, robertabase. 2023. https://huggingface.co/roberta-base

Huggingface, sinonimayzer. 2024. https://huggingface.co/datasets/sinonimayzer/mixed-data

Huggingface, sinonimayzer/UzRoBerta. 2024. https://huggingface.co/sinonimayzer/UzRoBERTa-v3/raw/main/vocab.json

Huggingface, sinonimayzer/UzRoBerta-v1. 2024. https://huggingface.co/sinonimayzer/UzRoBERTa-v1

Huggingface, sinonimayzer/UzRoBerta-v2. 2024. https://huggingface.co/sinonimayzer/UzRoBERTa-v2

Huggingface, sinonimayzer/UzRoBerta-v3. 2024. https://huggingface.co/sinonimayzer/UzRoBERTa-v3

Huggingface, tahrirchi (a). 2023. https://huggingface.co/datasets/tahrirchi/uz-crawl

Huggingface, tahrirchi (b). 2023. https://huggingface.co/datasets/tahrirchi/uz-books

Huggingface, wikimedia/wikipedia. 2023. https://huggingface.co/datasets/wikimedia/wikipedia

Huggingface. 2023. https://huggingface.co/roberta-base

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaggle, datasets. 2023. https://www.kaggle.com/datasets

Kaggle, mohiyaxonuzokova. 2023. https://www.kaggle.com/datasets/mohiyaxonuzokova/bigger-dataset-from-kunuz-and-daryouz

Metatext, cc100. 2023. https://metatext.io/redirect/cc100

Reddit, datasets. 2023. r/datasets

Uzoqova, Mohiyaxon. 2016. “Leksik sinonimlarni aniqlash uchun Word2Vec va RobertaForMaskedLM modellaridan foydalanish. Syn-roberta modeli haqida”. Kompyuter lingvistikasi: muammolar, yechim va istiqbollar. Shuhrat Sirojiddinov muharrirligida, 328 — 338. Toshkent. https://compling.navoiy-uni.uz/index.php/conferences/article/view/323

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692

Training the RoBERTaForMaskedLM model to identify lexical synonyms in the Synonymizer program

Mohiyaxon Uzoqova[1]
Mansurbek Narzullayev[2]

Abstract. We noted in our previous studies that the RobertaForMaskedLM model, which belongs to the Roberta family, can be used to identify lexical synonyms. However, we reported that a satisfactory result was not obtained due to the small amount and quality of the data collected for the training. During this study, we re-implemented the training process in several stages, increasing the size and quality of the documents intended for the ML training. As a result, we obtained new and convincing results, analyzed and presented them in tables. We also successfully integrated the model into the Uzbek synonymizer program.

Key words: *synonym, synonymizer, dataset, Roberta, tokenizer, training loss, validation loss.*

References

Amazon, opendata. 2023. https://registry.opendata.aws/

Baevski, Alexei; Edunov, Sergey; Liu, Yinhan; Zettlemoyer; Luke & Auli, Michael. 2019. Cloze-driven Pretraining of Self-attention Networks. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.* 5363-5372, Hong Kong, China. DOI: 10.18653/v1/D19-1539

DOI: 10.18653/v1/N19-1423.

Dullin, Theo. 2023. How to Create Datasets: strategies and examples https://kili-technology.com/data-labeling/machine-learning/create-dataset-for-machine-learning

Google, dataresearch. 2023. https://datasetsearch.research.google.com/

Huggingface, datasets. 2023. https://huggingface.co/docs/datasets/index

[1]*Uzoqova Mohiyaxon Tuyg'un qizi* –Master of degree. Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
E-mail: mohiyaxonuzokova@gmail.com
ORCID: 0000-0001-7102-0824
[2]*Narzullayev Mansurbek Nurali o'g'li* – Master of degree. Tashkent State University of Information Technologies named after Muhammad Al-Khorazmi, Samarkand branch.
E-mail: mansurbeknarzullayev@mail.ru
ORCID: 0000-0002-8160-6631

Huggingface, elmurod. 2023. https://huggingface.co/datasets/elmurod1202/uzbek-sentiment-analysis

Huggingface, mc4. 2023. https://huggingface.co/datasets/mc4

Huggingface, murodbek. 2023. https://huggingface.co/datasets/murodbek/uz-text-classification

Huggingface, oscar. 2023. https://huggingface.co/datasets/oscar

Huggingface, robertabase. 2023. https://huggingface.co/roberta-base

Huggingface, sinonimayzer. 2024. https://huggingface.co/datasets/sinonimayzer/mixed-data

Huggingface, sinonimayzer/UzRoBerta. 2024. https://huggingface.co/sinonimayzer/UzRoBERTa-v3/raw/main/vocab.json

Huggingface, sinonimayzer/UzRoBerta-v1. 2024. https://huggingface.co/sinonimayzer/UzRoBERTa-v1

Huggingface, sinonimayzer/UzRoBerta-v2. 2024. https://huggingface.co/sinonimayzer/UzRoBERTa-v2

Huggingface, sinonimayzer/UzRoBerta-v3. 2024. https://huggingface.co/sinonimayzer/UzRoBERTa-v3

Huggingface, tahrirchi (a). 2023. https://huggingface.co/datasets/tahrirchi/uz-crawl

Huggingface, tahrirchi (b). 2023. https://huggingface.co/datasets/tahrirchi/uz-books

Huggingface, wikimedia/wikipedia. 2023. https://huggingface.co/datasets/wikimedia/wikipedia

Huggingface. 2023. https://huggingface.co/roberta-base

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaggle, datasets. 2023. https://www.kaggle.com/datasets

Kaggle, mohiyaxonuzokova. 2023. https://www.kaggle.com/datasets/mohiyaxonuzokova/bigger-dataset-from-kunuz-and-daryouz

Metatext, cc100. 2023. https://metatext.io/redirect/cc100

Reddit, datasets. 2023. r/datasets

Uzoqova, Mohiyaxon. 2016. “Leksik sinonimlarni aniqlash uchun

Word2Vec va RobertaForMaskedLM modellaridan foydalanish. Syn-roberta modeli haqida". Kompyuter lingvistikasi: muammolar, yechim va istiqbollar. Shuhrat Sirojiddinov muharrirligida, 328 — 338. Toshkent. https://compling.navoiy-uni.uz/index.php/conferences/article/view/323

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692