O’ZBEK TILIDAGI GAPLARNI PERIFRAZ QILISH AXBOROT TIZIMINING FUNKSIONAL JARAYONLARINI LOYIHALASHTIRISH

Axmedova Xusniya Xusanovna

Muhammad al- Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti
khusniya150586@gmail.com

K E Y W O R D S

biznes-jarayon, modellashtirish, perifraz, axborot tizimi, semantik analizator

A B S T R A C T

Hozirgi kunda biznes-jarayonlarni modellashtirish metodologiyalari va vositalari bir vaqtning o‘zida ilmiy tadqiqot yo‘nalishi va dasturiy ta’minot (DT) bozorining rivojlanayotgan sektoriga aylangan. Ushbu usullar axborot tizimlarining zaif tomonlarini aniqlash va ularni bartaraf etish uchun zarur choralarni taklif qilish uchun ishlatilishi mumkin. Ushbu maqolada o‘zbek tilida gaplarni perefraz qiluvchi axborot tizimining biznes-jarayonlari BPMN metadologiyasi yordamida shakllantirildi. Semantik analizator axborot tizimining umumiy biznes-jarayonidagi har bir jarayonda alohida to‘xtalib o‘tilgan. Tabiiy tildagi matnlarni perefraz qilishda foydalaniladigan yondashuvlar ko‘rib chiqilgan.

Kirish.

Jahonda ta’limni boshqarish axborot tizimlarini ishlab chiqish, modellashtirish, ma’lumotlar bazasini loyihalashtirishga hamda algoritmlarni ishlab chiqishga yo‘naltirilgan ilmiy tadqiqotlar olib borilmoqda. Bu borada, jumladan, oliy ta’lim muassasasi o‘quv jarayoni boshqaruvi axborot tizimi biznes jarayonlarining Business Process Model and Notation (BPMN) metodologiyasi va relyatsion algebra asosida shakllantiriladigan Entity-Relationship Diagrams (ERD) modellari, algoritmlari hamda dasturiy ta’minotlarni SMART (Social, Mobile, Access, Regulated, Technology) texnologiyalar asosida yaratish muhim vazifalardan biri hisoblanadi. Shu bilan birga, ma’lumotlarining mantiqiy va jismoniy modellarini E.Koddning 1,2,3-normal formalari asosida optimallashtirish hamda taqsimlangan axborot resurslarida ma’lumotlarni tavsiflash formatlari va axborot izlash algoritmlarini ishlab chiqish zarur hisoblanadi [1].

Adabiyotlar tahlili

Iboralarni qayta ifodalash bugungi kunda dolzarb masaladir va ko‘plab tadqiqotchilar bu borada chuqur o‘rganish metodlari hamda millionlab parametrli nutq modellari arxitekturalaridan foydalanmoqda. Biroq, bu muammoni hal qilishga birinchi urinishlar ikki so‘z orasidagi ma’nolarning farqini WordNet ma’lumotlar bazasidagi masofa orqali aniqlashga asoslangan sodda usullardan boshlangan [2]. Shuningdek, WordNet lug‘atidagi so‘zlarning ma’nolari o‘rtasidagi bog‘liqlik darajasini hisobga oladigan yondashuv ham taklif etilgan [3]. Mazkur usul so‘zlar juftligini ularning grammatik xususiyatlariga qarab solishtirish tamoyiliga asoslangan.

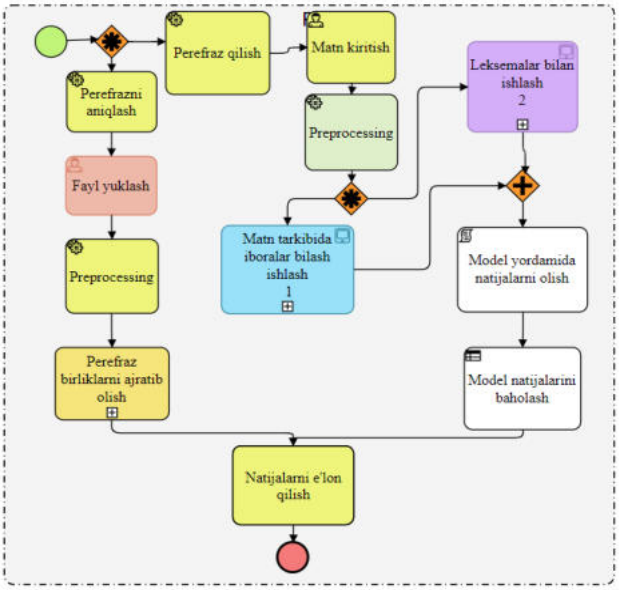
Ba’zi tadqiqotlarda so‘zlarning semantik o‘xshashligi sinsetlar o‘rtasidagi eng qisqa yo‘l uzunligiga bog‘liq ekanligi muhokama qilingan [4]. Tabiiy tillarda so‘z o‘xshashligi berilgan terminlarga mos keluvchi so‘z ta’riflari to‘plamlarining kesishmasi sifatida aniqlanadi [5]. Boshqa ilmiy manbaga ko‘ra – katta korpusda umumiy ajdod bilan uchrashish ehtimolligiga, umumiy ajdod va taqqoslanayotgan ikki sinset bilan uchrashish ehtimolligiga bog‘liq [6-9].

Asosiy qism

O‘zbek tilida gaplarni perefraz qiluvchi axborot tizimining biznes-jarayonlari BPMN metadologiyasi yordamida shakllantirildi. Axborot tizimi quyidagi 1-rasmdagi umumiy jarayonlarni ketma-ket amalga oshiradi hamda keltirilgan har bir murakkab jarayon o‘z navbatida

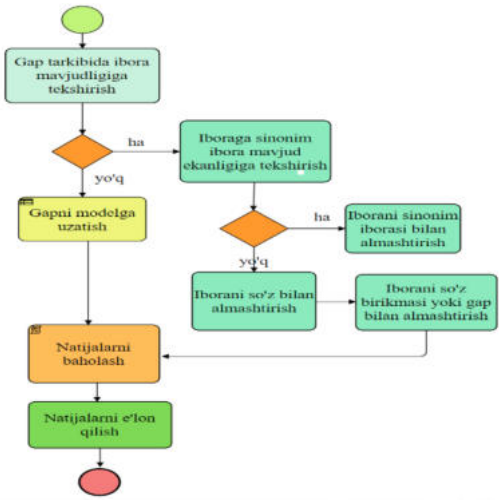
© Xusniya Axmedova

bir nechta jarayonlarni o‘z ichiga oladi. Semantik analizator axborot tizimining umumiy biznes-jarayonidagi har bir jarayonda alohida to‘xtalib o‘tamiz. Dastlab foydalanuvchi tomonidan kerakli funksiya tanlanadi, ya’ni perefraz qilish yoki perefrazni aniqlash funksiyalaridan biri tanlanadi. Tanlangan funksiya qarab jarayonlar ketma-ketligi bajariladi. Aytaylik foydalanuvchi perefraz qilish funksiyasini tanlagan bo’lsin. Ushbu funksiyani tanlaganda foydalanuvchi tomonidan matn kiritish talab qilinadi. Kiritilgan gap tokenlarga ajratiladi. Kiritilgan matn ustida dastlabki ishlov berish jarayonlari olib boriladi. Ya’ni gapdagi barcha so’zlar kichik harflar bilan boshlanadi. Ortiqcha belgilardan tozalash, emojilar ustida ishlash kabi amallar bajariladi.



1-rasm: Axborot tizimining umumiy jarayonlarining BPMN modeli.

1-rasmdan 1-jarayonni tahlil qilsak. Foydalanuvchi tomonidan kiritilgan gap tarkibida ibora mavjudmi deya tekshiriladi. Agar gap tarkibida ibora mavjud bo’lsa, iboralar bilash ishlash bandiga o’tiladi.



2.- rasm: Iboralar bilash jarayonlarining BPM modeli

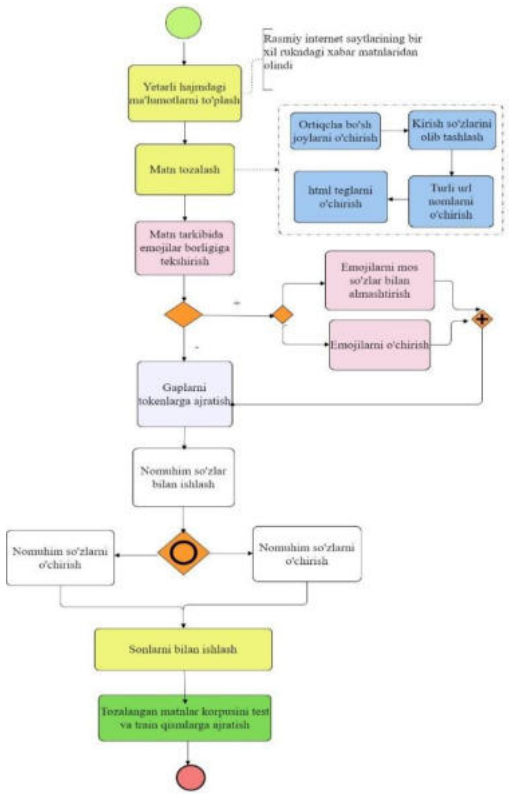
Ushbu jarayonda foydalanuvchi tomonidan kiritilgan gap tarkibida ibora mavjud bo’lsa ular ustida ishlash haqida bu rasm. Ma’lumki o’zbek tili iboraga boy til, og’zaki nutqda ham yozma nutqda ham juda ko’p uchraydi. Demak foydalanuvchi tomonidan kiritilgan gap dastlab ibora mavjudligiga tekshiramiz. Buning uchun iboralar ishtirok etgan gaplar jamlandi va iboralar [ph] tegi bilan teglab chiqildi. Ibora qatnashgan gaplar soni 40 000 tani tashkil etadi. Ma’lumki iboralar o’z ma’nosida ham gaplar tarkibida uchrashi mumkin. Buni farqlash uchun qo’lda teglangan gaplar jamlandi. Jamlangan dataset ustida Yashirin Markov hisoblashlari olib borilib, gapdagi ibora ajratib olinadi. O’zbek tilidagi iboralar ro’yhatini shakllantirishda “O’zbek tili frazeologik lug’ati” dan foydalanildi. Ma’lumotlar bazasiga jami 3500 ta ibora jamlangan. Ma’lumki, iboralarni bir so’z bilan ham, so’z birikmasi bilan ham, gap bilan ham, ibora bilan ham almashtirish mumkin. Tadqiqotimiz davomida iboralarning bu xususiyatlarini ham inobatga oldik. Ibora o’rnini bosuvchi so’zlar, so’z birikmasi, hamda gaplardan bittadan belgilab qo’yildi. Iboralarning hammasini ham bitta so’z bilan almashtirib bo’lmaydi, xuddi shunday hammasini sinonim iboralari mavjud emas. “O‘zbek tili frazeologik lug‘ati” kitobidan jami 200 ta iboraning sinonim iboralari mavjud ekanligini guvohi bo’ldik va ularni ajratib oldik. Ibora ishtirok etgan gaplarni ham asosan ushbu kitobdan oldik.

Gap tarkibidagi iboralarni aniqlab ularni sinonimlari bilan almashtirish natijasida hosil bo’lgan gapni to’g’ridan- to’g’ri foydalanuvchiga taqdim etdik, faqat ularni so’z bilan

(cc) BY © Xusniya Axmedova

almashtirilgandagina modeldan foydalandik. Shu o'rinda aytish kerakki, iborani o'zining sinonim iborasi bilan almashtirishda qiyinchilikka duch kelindi. Bir iboraning bir nechta sinonim iborasi mavjud bo'lsa, aynan qaysi iborani tanlab olish masalasini yechimi ko'ndalang bo'ldi. Buni yechishda yana qo'shimcha dataset zarur bo'ldi. Ya'ni iboralarning turli sinonimlari ishtirok etgan gaplar jamlandi. 200 ta iboraning har bir sinoninim ishtirokidagi jami 1000 ta gap jamlandi va GloVe algoritmi yordamida eng yuqori ehtimollikka ega bo'lgan sinonim ibora tanlab olindi.

Navbatdagi jarayon iborani gap bilan almashtirish jarayoni bo'lib, hamma iboralar ham bir gap bilan ifodalanmaydi. Biz axborot tizimi ma'lumotlar bazasida jami 250 ta ibora o'rnini bosuvchi iboralarni belgilab qo'ydik. Ibora mavjud bo'lmagan gaplarni perefraz qilishda modeldan foydalanildi. Navbatdagi BPM jarayonida model hosil qilish jarayonlarini ko'rishimiz mumkin.



3-rasm: Model uchun dataset hosil qilish jarayonlari

3-rasmdagi jarayonlarni tahlil qilsak.

1. Dastlabki jarayon yetarli hajmdagi ma'lumotlarni jamlash. Ushbu jarayonda O'zbek tilidagi rasmiy internet saytlaridagi matnlar jamlab olindi. Ya'ni, bilamizki har kungi yangiliklar barcha internet saytlarida namoyish etiladi, har biridagi matn mazmunan bir xil, ammo turlicha so'zlardan foydalanilgan holda foydalanuvchilargan namoyish etiladi. Biz ham xuddi shunday qildik. Ya'ni *kun.uz, daryo.uz, stadion.uz, qalampir.uz, effect.uz* kabi jami 12 ta rasmiy saytlar tanlab olindi. Saytlardagi kundalik yangiliklar matnlarini olib birlashtirdik. Internet saytlaridagi matnlarni yuklab olishda Python dasturlash tili va *Scrapy* frameworkidan, murakkab tuzilishga ega va interaktiv buttonlarga ega bo'lgan internet saytlari uchun *brouzerni avtomatlashtirish* uchun mo'ljallangan ochiq kodli dasturiy ta'minot *(Selenium Browser Automation Project)* dan foydalanildi.

2. Matn tozalash jarayoni. Ma'lumki internetdan matnlarni ko'chirib olganimizda unda har xil ortiqcha belgilar, bo'sh joylar, keraksiz internet url manzillari uchraydi. Matnlardan foydalanish uchun albatta ular yaroqli bo'lishi kerak. Imlo xatolaridan holi bolishi maqsadga muvofiq. Bu jarayonni ham avtomatik amalga ochirdik. Internet manbalarini quyidagilardan tozalandi:

1. Ortiqcha bo'sh joylar
2. html teglar
3. Internet manbalarining url manzillari
4. Gap bo'lmagan birikmalar

Elektron pochta nomlarini o'chirish orqali matn ortiqcha belgilardan tozalanadi. Bu bilan matn to'liq foydalanish uchun yaroqli deya olmaymiz. Matn tarkibida gaplarni alohida satrlarga joylash masalasi ham matnni normallashtirishga kiradi. Matnlarni gaplarga ajratishni ham dastur yordamida amalga oshirdik. Ya'ni o'zbek tilida gapni oxirini ifodalovchi tinish belgilar mavjud, ular nuqta (.), undov (!), so'roq belgisi (?), uchta nuqta (…), undan tashqari ritorik so'roq (?!) belgisi ham mavjud bo'lib, ushbu belgilar yodamida matn tarkibidagi gaplarni alohida ajratib olish mumkin.

3. Emojilar bilan ishlash. Internet matnlari tarkibida emojilarning uchrashi tabiiy albatta. Emojilar matndagi hissiyotlarni ifodalaydi. Sentimental analizda ularni so'zlar bilan almashtirish maqsadga muvofiq. Gaplarni perefraz qilishda emojilarni so'z bilan almashtirish yoki o'chirish mumkin. Ikkalasidan bittasini tanlash uchun emojilarni gapdagi o'rnini aniqlash kerak.

(cc) BY © Xusniya Axmedova

Masalan:

Bugungi uchrashuv natijalar kutilganidan ham 👍 bo’ldi.

Ushbu gapda emojilarni so’z yoki so’z birikmasi bilan almashtirish kerak, chunki emoji bu gapda natijani qanday bo’lganligini aniqlab kelgan. Shu kabi gaplarda ularni so’zlar bilan almashtiriladi. Tadqiqotimiz favomida 10 xil turdagi emojilarni so’z bilan almashtirib chiqdik. Ushbu emojilar haqida quyidagi jadvalda ma’lumotlar berilgan.

1-jadval:

Emojilar va ularning ekvivalent so’zlari

№	Emoji	Emojilarga mos ekvivalent so’z yoki so’z birikmalari
1	❌	Taqiqlanadi, mumkin emas
2	⚡	Juda muhim
3	✅	Roziman
4	👨‍💼	Erkak
5	⏰	Davomiyligi
6	⚖️	Adolat
7	🤝	Kelishdik
8	👍	Zo’r, ajoyib
9	😄	Xursand bo’lmoq
10	😟	Hafa bo’lmoq
…	…	…

Internet manbalaridan jamlangan matnlar tarkibidagi emojilarni so’zlar bilan avtomatik almashtirish uchun o’zbek tilidagi gaplarni to’g’ri tuzilishga ega ekanligini tekshirib beruvchi tizim zarur. Undan tashqari gaplarni sintaktik tahlil qilish ham bu masalani yechishda muhim. Ya’ni gapning to’g’ri tuzilishini hamda undagi gap bo’laklarining yetarliligi aniqlanganidan keyingina gapdagi emojining o’rni aniqlash mumkin. Agar gapda emoji aniqlovchi yoki to’ldiruvchi yoki o’rnida kelgan bo’lsa ularni ekvivalent so’zlar bilan almashtirish mumkin bo’ladi. Agarda emojilarning gapdagi o’rni ma’lum bo’lmasa uni o’chirish zarur. Tadqiqotimizda sintaktik tahlil qilishni ko’rib chiqmaymiz, shu sababli ishning bu qismi ochiq qoldi. Matn yig’ish jarayonida uchragan emojilarni inson omili yordamida almashtirildi, o’chirildi.

4. Gaplarni tokenlarga ajratish. O’zbek tilidagi gaplarni so’z tokenlarga ajratishda ularni bo’sh joylardan foydalanish har doim ham to’g’ri bo’lmaydi. Chunki unda gapdayi tinish belgilari (vergul, tire, chiziqcha) so’zlarga qo’shilib qolishi mumkin. Undan tashqari qo’shma fe’llar, iboralar to’g’ri tokenlarga ajralmay qoladi. Bu muammoni ham biz ko’rib chiqmadik. Faqat gap tarkibida nomuhim so’zlarni aniqlashda ularni bo’sh joylar orqali tokenlarga ajratdik.

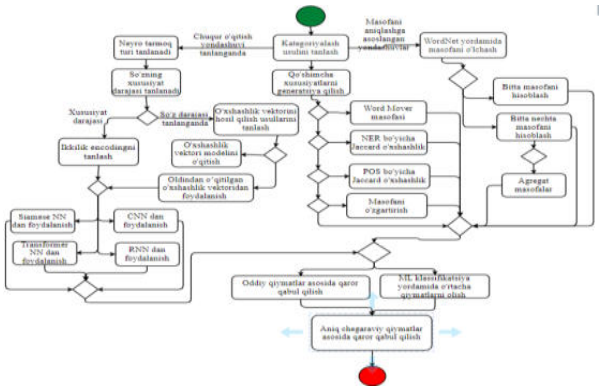
5. Nomuhim so’zlar bilan ishlash. Navbatdagi jarayon *Nomuhim so‘zlar bilan ishlash* jarayoni bo‘lib, unda gap tarkibidagi uning semantikasiga ta’sir qilmaydigan so‘zlarni olib tashlash nazarda tutilgan. O‘zbek tilidagi nomuhim so‘zlar (StopWords) bog‘lovchi, yuklama, olmosh, ravish kabi so‘z turkumlari orasida uchraydi. Shunday so‘z birikmalari ham borki, matn tarkibida uchraganda uning semantikasi uchun nomuhimligicha qoladi. Ammo ular ba’zi matnlar tarkibida muhim ahamiyatga ega. Bunday nomuhim so‘zlar ularning bigram va trigrammalari ro‘yxatini X.A.Madatov va Sh.Bekchanov [10]. Bu ro‘yxatga nazar solganimizda undagi barcha so‘zlar ham har doim nomuhim bo‘lmasdan, orasida omonim, polifunksional so‘zlar ham uchradi. Bundan kelib chiqadiki, bizga sof nomuhim so‘zlar, ya’ni har qanday gap tarkibida ham uning semantikasiga ta’sir qilmaydigan so‘zlar ro‘yxati kerak bo‘ladi. O’zbek tilidagi sof nomuhim so’zlarni B.B. Elov alohida ajratib ko‘rsatganlar [11]. Jami 182 ta sof nomuhim so‘zlarni yig‘ildi. Datasetni tozalash jarayonida o’zbek tilidagi nomuhim so’zlarni ham huddi emojilar kabi gapdan o’chirib tashlash yoki qoldirish masalasi ko’rib chiqildi. Agar nomuhim so’zlarning sinonimlari mavjud bo’lsa yoki ularni ekvivalenti mavjud bo’lsa ular gap tarkibida qoldirildi, aks holda o’chirib tashlandi.

6. Sonlar bilan ishlash jarayoni. Tadqiqotda matn tarkibidagi sonlar o’zgartirilmaydi, ya’ni so’z bilan ifodalanmaydi. Chunki hamma sonlarni ham ekvivalentlari mavjud emas. Masalan 1 soni bitta, yagona kabi so’zlar bilan almashtirish mumkin, lekin hamma sonlarni ham ekvivalentlari mavjud emas. Shuning uchun matn tarkibidagi sonlar so’zlarga almashtirilmaydi.

(cc) BY © Xusniya Axmedova

7. Matnlarni test va training qismlarga ajratish jarayonida, matnlar ¼ ulushlarga ajratildi. Ya'ni datasetning 20% test va qolgan 80% training uchun ajratib qo'yildi hamda model o'qitildi.

Perefraz qilish quyidagi yondashuvlar yordamida aniqlash mumkin.



4-rasm: Perefraz qilish yondashuvlari

4-rasmda tasvirlangan tabiiy tildagi matnlarni perefraz qilishda foydalaniladigan yondashuvlar keltirilgan. Tabiiy tilning imkoniyatlarini nobatga olgan holda turli yondashuvlardan foydalanish mumkin. 3.2.4-rasmda tasvirlangandek Masofani aniqlashga asoslangan yondashuvdan tabiiy tilning WordNet tizimi yordamida foydalanish mumkin. Ya'ni WordNetdagi so'zlar iyerarsiyasida joylashuvi, undagi semantik maydonlar orasidagi masofalarni o'lchash orqali sinonim so'zlarni aniqlash mumkin. O'zbek tilining WordNet tizimi ishlab chiqilmaganligi sababli tadqiqodimizda bu yondashuv chetlab o'tildi. Qo'shimcha xususiyatlarni generatsiya qilish yondashuvidan ham foydalanish mumkin, bu yondashuv o'xshashlik algoritmlarini qo'llashni taqazo etadi. Bu yondashuv gaplar o'xshashligini hisoblash orqali ularni semantik yaqinlik darajasini aniqlashga yordam beradi. Ma'lumki, perefraz qilishda gaplarni o'xshashligini aniqlovchi algoritmlardan foydalanish degani aynan bir xil so'zlardan tarkib topgan gaplar bo'lsagina ularning o'xshashlik koeffitsiyenti yuqori ko'rsatkichga ega bo'lishi mumkin. Bu esa perefraz gaplarni emas, aynan bir xil so'zlardan iborat bo'lgan gaplarni aniqlashda qo'l keladi. Shularni inobatga olib tadqiqot davomida o'zbek tilidagi gaplarni perefraz qilish axborot tizimini ishlab chiqishda Chuqur o'qitish yondashuvlaridan foydalandik. Unga ko'ra *so'zlarning o'xshashlik darajasini aniqlashda* usulidan ham foydalanildi. Ushbu usuldan iboralarni sinonimlarini aniqlashda foydalanildi. Undan tashqari gaplar tarkibida omonim so'zlar uchraganda ularni sinonimlarini aniqlashda muammosiga duch keldik. Buni yechish uchun omonimiyani aniqlash uni qaysi so'z turkumiga tegishli ekanligi, yoki aynan qaysi ma'nosi bo'yicha sinonimlarini tanlab olish masalasi yechim talab qiladi. Shu sababli kelib chiqqan muammolarni bartaraf etishda GloVe algoritmidan foydalanildi. Glove algoritmi uchun o'qitish sifatida jamlangan umumiy datasetdan foydalanildi. Natijada so'zlar va iboralarning semantik ekvivalentlari (perefrazlari) aniqlandi.

Xususiyat darajasida aniqlashda oldindan tayyorlab qo'yilgan datasetni o'qitish orqali model hosil qilish tushuniladi. Model hosil qilishda biz zamonaviy transformerlardan foydalandik. Buning uchun Google tomonidan yaratilgan Gemma oilasiga mansub ochiq manbali (open-source) til modeli *google/gemma-3-4b-pt* tanlab olindi. Gemma modellarini ishlab chiqishda Google Gemini modellarida qo'llangan zamonaviy texnologiyalar va tadqiqotlar asosida yaratilgan. *google/gemma-3-4b-pt* nomini o'zidan quyidagi xususiyatlarni bilib olish mumkin.

–*google* — model yaratuvchisi.

–*gemma-3* — Gemma modelining 3-versiyasi.

–*4b* — modeldagi parametrlar soni 4 milliard.

–*pt* — pretrained ya'ni oldindan umumiy korpuslarda o'qitilgan, lekin aniq bir vazifaning yechimiga yo'naltirilmagan model.

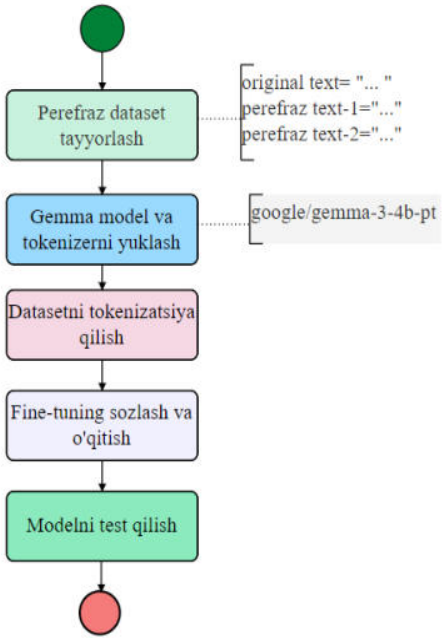
Modelning *gemma-3* versiyasi multimodel, ya'ni matn qabul qilib matn qaytarish imkoniyatiga ega. Undan tashqari model bir vaqtning o'zida 128 000 ta tokenni ko'rib chiqish imkoniyati ega. Model dunyoning 140 dan ortiq tillarini qo'llab quvvatlaydi, shu jumladan o'zbek tilini ham. Olding versiyalariga nisbatan kuchliroq va nisbatan ko'p variantlarda mavjud. Shu sababli gemma-3 versiyasidan foydalanildi. Model pretrained, ya'ni oldindan o'qitilgan va biror masalani yechishga moslashtirilmaganligi sababli, biz modeldan o'zbek tilidagi gaplarni perefraz qilishda foydalandik. Datasetimiz

(cc) BY © Xusniya Axmedova

[{"Original gap" = " "
"Perifraz gap-1" = " "
"Perifraz gap-2" = ""

},] ko'rinishda tayyorlandi. Uning umumiy hajmi 194MB, jami 375K gapdan iborat. Datasetni fine-tuning qilish jarayonida *4 ta RTX 4090 GPU* kerak bo'ldi. Model hosil qilish jarayon 11 kun davomida 5 epochda o'qitildi. Model qilish bosqichlarini 3.2.5-rasmda ko'rish mumkin. Ko'rinib turganidek dastlabki bosqich datasetni tayyorlash bosqichi bo'lib, bu haqida yuqorida batafsil ma'lumot berilgan. *Navbatdagi bosqich* Gemma model versiyasini va tokenizer kutubxonasini yuklab olish. Bu jarayon kodini quyida keltiramiz.

```
from transformers import AutoTokenizer, AutoModelForCausalLM
model_name = "google/gemma-3-4b-pt"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)
```



5-rasm: Model hosil qilish bosqichlari

Model va tokenizez yuklab olingandan keyin datasetni tokenizatsiya qilish bajariladi. Bu jarayonning python kodini quyida ko'rish mumkin.

```
from datasets import load_dataset
dataset = load_dataset("csv", data_files="dataset.csv")

def tokenize_function(example):
    return tokenizer(example["original_text"], text_target=example["paraphrase_text1", "paraphrase_text2"], truncation=True)

tokenized_datasets = dataset.map(tokenize_function, batched=True)
```

tokenizatsiya qilingan dataset uchun fine-tuning parametrlari belgilab olinadi va o'qitish amalga oshiriladi. Fine-tuning qilishda *transformers* bilan *Trainer* lardan foydalanildi.

```
from transformers import TrainingArguments, Trainer
training_args = TrainingArguments(
  output_dir="./gemma_paraphrase",
  evaluation_strategy="epoch",
  learning_rate=2e-5,
  per_device_train_batch_size=2,
  num_train_epochs=5,
  save_total_limit=2,
  fp16=True,
  gradient_accumulation_steps=16
)
trainer = Trainer(
  model=model,
  args=training_args,
  train_dataset=tokenized_datasets["train"],

eval_dataset=tokenized_datasets["validation"],
  tokenizer=tokenizer
)
trainer.train()
```

Fine-tuning sozlamalaridan ko'rish mumkinki, hosil bo'lgan perefraz modeli *output_dir="./gemma_paraphrase"* deb nomlanadi.

Nihoyat so'ngi bosqich tayyor bo'lgan modelni test qilish bosqichi. Test qilishda biz yuqorida aytilda 20% dataset gaplaridan foydalandik. Test natijalari haqida keyingi bobda batafsil ma'lumot beriladi. Quyida test jarayonining python kodi keltiramiz [12].

```
input_text = load_dataset("csv", data_files="test.csv")
```

(cc) BY © Xusniya Axmedova

```
for i in input_text.original_text:
    inputs = tokenizer(i, return_tensors="pt").to("cuda")
    output = model.generate(**inputs, max_new_tokens=50)

    print(tokenizer.decode(output[0], skip_special_tokens=True))
```

Xulosa

Ushbu maqolada O'zbek tilida gaplarni perefraz qiluvchi axborot tizimining biznes-jarayonlari BPMN metadologiyasi batafsil ko'rib chiqilgan.

Ishlab chiqilgan modelni test qilishdan olingan natijalar va ularni statistik ko'rsatkichlari haqidagi ma'lumotlarni keying rejada ko'rib chiqamiz.

Foydalanilgan adabiyotlar

1. Pronoza E., Yagunova E., Kochetkova N. Sentence Paraphrase Graphs: Classification Based on Predictive Models or Annotators' Decisions?. In: Sidorov G., Herrera-Alcántara O. (eds) Advances in Computational Intelligence. MICAI 2016. Lecture Notes in Computer Science, vol 10061. Springer, Cham (2017).
2. Pronoza, E., Yagunova, E. Comparison of sentence similarity measures for Russian paraphrase identification. In Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), pp. 74–82, IEEE (2015a).
3. Pronoza, E., Yagunova, E. Low-Level Features for Paraphrase Identification. In: Sidorov, G., Galicia-Haro, Sofía N. (eds.) MICAI 2015. LNCS, vol. 9413, pp. 59–71. Springer, Cham (2015b).
4. Pronoza, E., Yagunova, E., Pronoza, A. Construction of a Russian Paraphrase Corpus: Unsupervised Paraphrase Extraction. Proceedings of the 9th Russian Summer School in Information Retrieval, August 24–28, 2015, Saint-Petersburg, Russia, (RuSSIR 2015, Young Scientist Conference), Springer CCIS (2015).
5. Androutsopoulos, I., Prodromos Malakasiotis, P. A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research, v. 38: 135–187 (2010).
6. Fernando, S., Stevenson, M. A semantic similarity approach to paraphrase detection. In Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics, pp. 45–52, (2008).
7. Pham, N., Bernardi, R., Zhang, Y. Z., Baroni, M. Sentence paraphrase detection: When determiners and word order make the difference. In Proceedings of the Towards a Formal Distributional Semantics Workshop at IWCS 2013, pp. 21–29 (2013).
8. Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., Blunsom, P. Reasoning about entailment with neural attention. arXiv preprint arXiv:1509.06664. (2015)
9. He, H., Gimpel, K., Lin, J. Multi-perspective sentence similarity modeling with convolutional neural networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1576–1586 (2015).
10. M. J. Kusner, Y. Sun, N. I. Kolkin, K. Q. Weinberger, From Word Embeddings to Document Distances, JMLR: W&CP 37 (2015) 957-966.
11. Y. Zhang, J. Baldridge, L. He, Zhang Y. PAWS: Paraphrase Adversaries from Word Scrambling, ArXiv, 2019. DOI: 10.48550/arXiv.1904.01130.
12. V.-A. Oliinyk, V. Vysotska, Y. Burov, K. Mykich, V. B. Fernandes, Propaganda Detection in Text Data Based on NLP and Machine Learning, CEUR workshop proceedings 2631 (2020) 132-144.
13. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, ArXiv, 2019. DOI: 10.48550/arXiv.1907.11692.
14. V. Vysotska, Y. Burov, V. Lytvyn, A. Demchuk, Defining Author's Style for Plagiarism Detection in Academic Environment, in: Proceedings of the International Conference on Data Stream Mining and Processing, DSMP, 2018, pp. 128-133. DOI: 10.1109/DSMP.2018.8478574.

(cc) BY © Xusniya Axmedova