TIL MODELLARI VA CHUQUR O‘RGANISH(DEEP LEARNING) USULLARI:UMUMIY TAVSIF

Sobirova Zarnigor G‘anijon qizi,
tayanch doktorant
sobirovazarnigor1996@gmail.com
ToshDO‘TAU

Annotatsiya. Hozirda sun’iy intellekt, xususan chuqur o‘rganish (deep learning) yutuqlari asosida yaratilgan til modellari inson tilini tushunish, qayta ishlash va yaratish borasida katta yutuqlarga erishmoqda. Til modellari – bu kompyuterlar yordamida inson tilini tushunish va undan samarali foydalanishga imkon beruvchi algoritmlar bo‘lib, ular matnlarni tahlil qilish, tarjima qilish, chatbot yaratish va boshqa ko‘plab vazifalarda ishlatiladi. Ushbu maqolada til modellarining NLPdagi o‘rni, ularning tarixiy rivojlanishi, texnik arxitekturasi, qo‘llanilish sohalari va ayniqsa, o‘zbek tiliga oid tadqiqotlar tahlil qilinadi.

Annotation. Nowadays, language models developed based on artificial intelligence, particularly deep learning advancements, have achieved significant progress in understanding, processing, and generating human language. Language models are algorithms that enable computers to comprehend and effectively utilize human language. They are applied in various tasks such as text analysis, translation, chatbot creation, and many others. This article analyzes the role of language models in NLP, their historical development, technical architecture, areas of application, and especially research related to the Uzbek language.

Аннотация. В настоящее время языковые модели, разработанные на основе достижений искусственного интеллекта, особенно глубинного обучения (deep learning), достигли больших успехов в понимании, обработке и генерации человеческой речи. Языковые модели — это алгоритмы, которые позволяют компьютерам понимать человеческий язык и эффективно его использовать. Они применяются в таких задачах, как анализ текста, перевод, создание чат-ботов и многих других. В данной статье рассматриваются роль языковых моделей в области обработки естественного языка (NLP), их историческое развитие, техническая архитектура, области применения, а также исследования, посвящённые узбекскому языку.

Kalit so‘zlar: *NLP, til modellari, statistik yondashuvlar, RNN, LSTM, transformer, UzBERT, BERTbek, ASR*

Kirish. So‘nggi yillarda sun’iy intellekt va chuqur o‘rganish texnologiyalaridagi yutuqlar natijasida tabiiy tilni qayta ishlash (Natural Language Processing, NLP)

sohasi jadal rivojlanmoqda. NLPning asosiy maqsadi kompyuterga inson tilini tushunish, tahlil qilish va yaratish qobiliyatini singdirishdir. Bu maqsadga erishishda asosiy vositalardan biri – til modellari (language models) hisoblanadi. Til modellari matnni tushunish va yangi matn yaratish uchun mo‘ljallangan neyron tarmoqlar bo‘lib, ular turli arxitekturalarda – rekurrent neyron tarmoqlar (RNN), ularning takomillashtirilgan turlari LSTM va GRU, hamda eng yangi Transformer arxitekturasi kabi modellar ko‘rinishida namoyon bo‘ladi.

Til modeli (Language Model) nima?

Til modellari inson yozgan yoki aytgan matnlar asosida **til qonuniyatlarini o‘rganadigan** va yangi mazmunli matnlar yaratishga qodir algoritmik tizimlardir. Ular matndagi so‘zlar ketma-ketligiga asoslanib keyingi so‘zni bashorat qilish, gapning ma’nosini aniqlash, tarjima qilish, savollarga javob berish kabi ko‘plab murakkab vazifalarni bajaradi.

Til modellari dastlab oddiy **statistik yondashuvlar** (masalan, N-gram modellar) asosida qurilgan bo‘lsa, keyinchalik **neyron tarmoqlarga** asoslangan chuqur modellar — **RNN**, **LSTM**, va ayniqsa **Transformer** arxitekturasi asosidagi yirik modellar (masalan, BERT, GPT, T5) orqali ulkan yutuqlarga erishildi.

Language Modeling

1. Statistical Language Models

- Linguistically motivated
- Adaptive
- Exponential
- Continuous Space
- Decision Tree
- N-Gram

2. Neural Language Models

- Speech Recognition
- Machine Translation
- Sentiment Analysis
- Text Suggestions
- Parsing Tools

3. Pre-trained Language Models

- BERT
- DeBERTa
- biLSTM
- T5
- Switch Transformers

4. Large Language Models

- ChatGPT
- GPT-3
- GPT-4 API
- PaLM
- Galactica
- Minerva
- GLAM
- AlexaTM
- LLaMA
- PanGu-Σ
- Cerebras-GPT
- BloombergGPT
- Falcon

Bugungi kunda til modellari **mashinaviy tarjima, chatbotlar, nutqni matnga aylantirish, matn generatsiyasi, xulosa chiqarish** kabi bir qator amaliy masalalarda asosiy yechim sifatida xizmat qilmoqda.

1. Chatbotlar

- Misollar: ChatGPT, LaMDA
- Tavsif: Tabiiy dialog uchun o‘rgatilgan model

2. Mashinaviy tarjima

- Misol: Google Translate
- Tavsif: Transformer asosida ishlaydi

3. Matn yaratish

- Misollar: GPT-3, GPT-4
- Tavsif: Hikoya, maqola, kod yozish

4. Nutqni matnga aylantirish

- Misollar: Whisper, ASR
- Tavsif: Audio kirishini matnga o‘zgartiradi

5. Savol-javob tizimi

- Misollar: BERT, T5
- Tavsif: Matn asosida savollarga javob beradi

6. Tilni avtomatik aniqlash

- Model: LID

- Tavsif: Matn qaysi tilda yozilganini aniqlaydi

Til modellari bugungi kunda quyidagi sohalarda keng qo‘llanilmoqda:

- **Tibbiyot:** klinik yozuvlarni avtomatik tahlil qilish;
- **Huquq:** hujjatlarni avtomatik tahlil va saralash;
- **Talim:** avtomatik tarjima, esse tekshirish, chat-botlar;
- **Dasturlash:** kod yozuvchi AI tizimlar (masalan, GitHub Copilot).

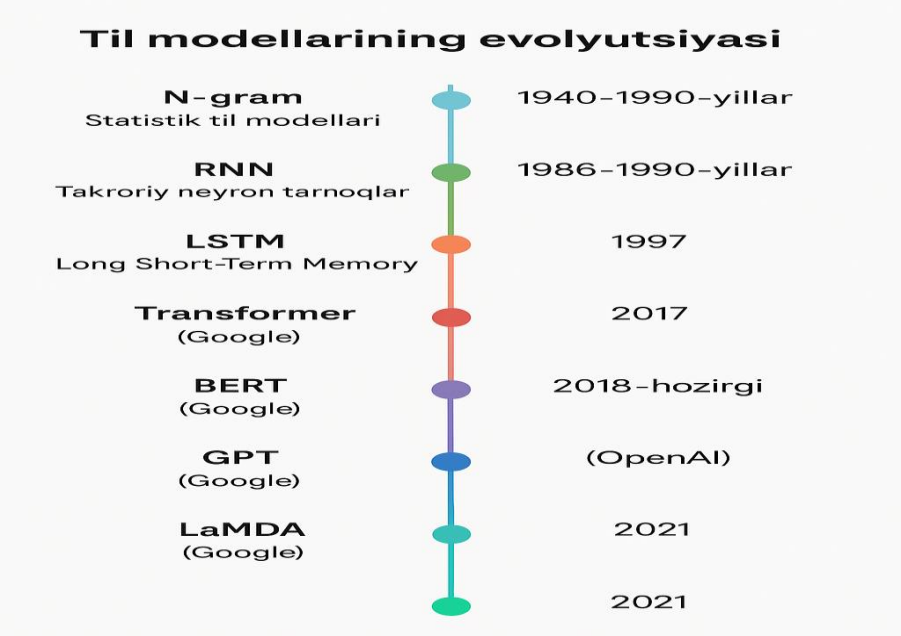
Soha	Model misoli	Izoh
Mashinaviy tarjima	Google Translate	Transformer asosida ishlaydi
Chatbotlar	ChatGPT, LaMDA	Tabiiy dialog uchun o‘rgatilgan model
Matn yaratish	GPT-3, GPT-4	Hikoya, maqola, kod yozish
Nutqni matnga aylantirish	Whisper, ASR	Audio kirishni matnga o‘zgartiradi
Savol-javob tizimi	BERT, T5	Matn asosida savollarga javob beradi
Tilni avtomatik aniqlash	LID model	Matn qaysi tilda yozilganini aniqlaydi

Til modellari tarixi va evolyutsiyasi

Til modellarining rivojlanishi dastlabki statistik modellardan boshlangan. 1980 –1990-yillarda **N-gram** modellar keng qo‘llanilgan bo‘lib, ular berilgan so‘zlar ketma-ketligiga asoslangan ehtimolliklarni hisoblaydi. Biroq bu yondashuvlar uzoq kontekstni hisobga ololmas edi. Bugunga kelib an’anaviy n-gram modellardan boshlab, chuqur neyron tarmoqlarga asoslangan yondashuvlargacha til modellari sezilarli darajada rivojlandi.

Keyinchalik neyron tarmoqlarga asoslangan modellar, xususan **rekurrent neyron tarmoqlar (RNN)**, **Long Short-Term Memory (LSTM)** va **Gated Recurrent Unit (GRU)** kabi modellar paydo bo‘ldi. Ular ketma-ketlikdagi so‘zlar orasidagi murakkab bog‘liqliklarni aniqlash imkonini berdi.

2017-yilda **Transformer** modeli[1] taqdim etilishi bilan yangi bosqich boshlandi. Transformer arxitekturasi parallel hisoblash imkoniyati, self-attention mexanizmi va samaradorligi bilan til modellarining rivojiga inqilobiy turtki berdi.



Til modellar turlari

A) **N-gram modeli (Statistik model):** so‘zlar ketma-ketligi asosida keyingi so‘zni ehtimol asosida aniqlaydi. Masalan, trigram model: $P(w_3 \mid w_1, w_2)$ Faqat uzoq kontekstni tushunmaydi.

B) ***RNN (Recurrent Neural Network):*** *Har bir so‘z ketma-ketlikda o‘rganiladi , har bir qadamda* ***hidden state*** *yangilanadi. Uzoq kontekstni eslab qolishda qiyinchilik bor (*vanishing gradient *muammosi).*

C) ***LSTM (Long Short-Term Memory):*** *RNN asosida yaratilgan.* ***Forget gate, Input gate, Output gate*** *kabi eshiklar bilan uzoq kontekstni eslab qoladi. Matn tarjima, nutq tanish kabi sohalarda ishlatilgan.*

D) ***Transformer :*** *Parallel ishlaydi, self-attention mexanizmi bilan har bir so‘z boshqa so‘zlarga “qaraydi”. Asosiy komponentlar:*

- **Self-attention**
- **Multi-head attention**
- **Position encoding**

GPT, BERT, T5 kabi modellar shu asosda yaratilgan.

Til modellarining asosiy komponentlari:

- **Plan** – Reja tuzish bosqichi.

- **Do** – Amalga oshirish bosqichi.
- **Check** – Tekshirish yoki nazorat qilish bosqichi.
- **Act** – Harakat qilish yoki takomillashtirish bosqichi.

O‘zbek tili uchun tabiiy tilni qayta ishlash (NLP) yo‘nalishida so‘nggi yillarda jadal izlanishlar boshlanib, bir qator resurs va modellar yaratilmoqda.Shu bilan birga, *O‘zbek tili hozircha resurslar jihatidan “low-resource” (resursi cheklangan) til* deb qaraladi – ya’ni kattaroq korpuslar, belgilangan (annotated) ma’lumotlar bazalari va yirik o‘qitilgan modellar kamligi sababli til texnologiyalari sohasida hali mavjud bo‘shliqlar bor[2].

Zero, ilgari O‘zbek tili uchun katta hajmli matn korpuslari va belgilangan ma’lumotlar to‘plamlarini yaratish ustida yetarlicha ish olib borilmagani ta’kidlanadi [3].

Biroq, ayni paytda bu bo‘shliqni to‘ldirish yo‘lida dastlabki qadamlar qo‘yilmoqda – masalan, 2022-yilda O‘zbek tilining morfologik va sintaktik jihatdan belgilangan korpusini yaratish bo‘yicha loyiha doirasida maxsus *POS (so‘z turkumi) va sintaksis* anotatsiya tizimi ishlab chiqilib, birlamchi belgilangan korpus tuzila boshlandi. Shuningdek, O‘zbek tilining Universal Dependencies formatidagi birinchi sintaksis daraxtkorpusi ham yaratilgan bo‘lib, unda 500 ga yaqin jumla sintaktik tahlil bilan belgilangan (ushbu resurs hozircha kichik hajmda bo‘lsa-da,kelgusida kengaytirilmoqda).

Matn korpusi mavjudligi chuqur o‘rganish modellari uchun nihoyatda muhim – chunki yirik **til modellari**ni sifatli o‘qitish uchun katta hajmdagi matn ma’lumotlari talab etiladi. O‘zbek tilida asosiy ochiq matn manbalaridan biri – **O‘zbekcha Vikipediya** bo‘lib, unda taxminan 124 ming maqola mavjudligi haqida xabar berilgan [4], biroq Vikipediya tarkibining bir qismi avtomatik tarjima qilingan yoki rasmiy ensiklopedik matnlardan to‘g‘ridan-to‘g‘ri ko‘chirilganligi tufayli til modeli o‘qitish uchun ma’lum cheklovlar mavjud – jumladan, maqolalarning uslubi qisqa va telegrafik bo‘lishi, yoki ayrim matnlarda kirill-lotin aralash yozuv belgilari uchrashi kabi muammolar aniqlangan. Shu sabab, tadqiqotchilar Vikipediya ma’lumotlarini tozalash va qo‘shimcha manbalar bilan boyitishga e’tibor qaratmoqdalar.

Xususan, yaqinda O‘zbek tilida keng qamrovli **yangiliklar korpusi** shakllantirildi – masalan, Daryo.uz saytining ~200 ming yangilik maqolasidan iborat korpus tuzilib, undagi jumlalar tuzilishi va til jihatdan Vikipediya ma’lumotiga qaraganda ancha xilma-xil va toza ekani qayd etildi. Bunday sifatli korpuslar yirik til modellari uchun asos bo‘lib xizmat qilmoqda.

UzBERT haqida

So‘nggi yillarda O‘zbek tilida ham **chuqur o‘rganishga asoslangan til modellari** paydo bo‘laboshladi. Dastlabki ishlardan biri – Mansurov B. va Mansurov A. tomonidan taqdim etilgan **UzBERT** modeli bo‘lib, u 2021-yilda e’lon qilingan[5].

UzBERT – **BERT arxitekturasi** asosida O‘zbek tilida oldindan o‘qitilgan ilk yirik model hisoblanadi. Tadqiqotchilar ushbu modelni O‘zbek tilidagi matnlarda (kirill va lotin yozuvlaridagi ma’lumotlar bilan) o‘qitib, natijada UzBERT **ko‘p tilli mBERT modeliga nisbatan** O‘zbek matnini tushunish vazifalarida ancha yuqori natijalarni ko‘rsatganini ma’lum qilganlar .Ya’ni, ayni paytgacha O‘zbek tilida alohida model bo‘lmagani sabab mBERT (100+ tilni qamragan ko‘p tilli BERT) ishlatilar edi, biroq monolingual UzBERT o‘sha vazifalarda aniq ustunlikka ega ekani empiric ravishda ko‘rindi.

UzBERT modeli ochiq manbada e’lon qilinib, MIT litsenziyasi ostida hamjamiyatga taqdim etilgan.

Yana bir yondashuv – 2023-yilda Elmurod Kuriyozov va hamkorlari tomonidan yaratilgan **BERTbek** modeli bo‘lib, bu ham O‘zbek tilidagi matnlarda pre-training qilingan Transformer arxitekturali modeldir. BERTbek modelini yaratish uchun tadqiqotchilar maxsus matn korpusi tuzdilar: O‘zbekcha Vikipediya maqolalarining tozalangan nusxasi va yangilik saytlaridan olingan 200 mingdan ortiq maqolalar jamlanmasida model **Masked Language Modeling** vazifasida oldindan o‘qitildi. So‘ngra BERTbek turli vazifalarga moslab fine-tuning qilinib, natijalari baholandi. Xususan, **O‘zbek tilidagi matnlarni toifalash (klassifikatsiya) bo‘yicha yangi ma’lumotlar to‘plami** ustida turli modellarning solishtirma tahlili o‘tkazildi Ushbu tahlilda an’anaviy **qoida asosida** ishlovchi usullar zamonaviy chuqur o‘rganish modellaridan ancha ortda qolishi ko‘rindi – **RNN** va **CNN** kabi neyron tarmoqlar bazaviy usullardan ancha yaxshiroq natija ko‘rsatdi. Ayniqsa, oldindan o‘qitilgan til modeli sifatida BERTbekni qo‘llagan yechim eng yuqori aniqlikka erishdi va boshqa modellardan ustun chiqdi. Bu natijalar O‘zbek tilida **pretraining+fine-tuning** yondashuvi samaradorligini ko‘rsatib, kelgusidagi tadqiqotlar uchun tayanch bo‘lib xizmat qilmoqda. Ta’kidlash joizki, O‘zbek tilidagi modellar faqat matn bilan cheklanmay, ovozli nutq uchun ham rivojlanmoqda: masalan, 2020-yillarda O‘zbek tili uchun bir necha soatlik nutq korpuslari tuzilib, **Automatic Speech Recognition (ASR)** tizimlarida *Deep Learning* modellari qo‘llanila boshlandi. Bunday dastlabki loyihalar O‘zbek nutqini matnga avtomatik o‘girishda ijobiy natijalar bermoqda. Umuman olganda, so‘nggi paytda O‘zbek tilini raqamli muhitda qo‘llashni osonlashtirish uchun **lug‘atlar**, **morfologik tahlilchilar**, **matn korpuslari** va **til modellari** borasida bir qator ochiq loyihalar paydo bo‘ldi. Bu jarayon hali boshida turgani bois, mavjud modellarni

yanada takomillashtirish va katta resurslar yaratish ustida izlanishlar davom etmoqda.

Xulosa

Xulosa qilib aytganda, chuqur o‘rganishga asoslangan til modellari so‘nggi yillarda jadal rivojlanib, *rekurrent arxitekturalar*dan *attention*ga asoslangan Transformer arxitekturalariga o‘tish natijasida sifat va samaradorlik yangi bosqichga ko‘tarildi. Bu modellarning mashinaviy tarjima, dialoglashuv tizimlari (chatbotlar), matn yaratish, tilni avtomatik aniqlash kabi ko‘plab sohalarda muvaffaqiyatli qo‘llanilayotgani kuzatilmoqda. Ayniqsa, yirik *pretrained* modellarning kichik *fine-tuning* bilan moslashuvi turli topshiriqlarda inson darajasiga yaqin natijalarga erishish imkonini bermoqda.

O‘zbek tili kabi resurslari cheklangan tillar uchun ham chuqur o‘rganish asosida ilk yirik modellarning yaratilgani va ularning muloqot, klassifikatsiya kabi vazifalarda ijobiy natijalari mahalliy NLP rivoji uchun muhim poydevor bo‘lib xizmat qilmoqda. Albatta, hali oldinda O‘zbek tilida katta korpuslar tuzish, modellarning yanada takomillashtirilgan versiyalarini yaratish va keng qo‘llash vazifalari turibdi.

Modelni o‘rgatish va optimallashtirishning ilg‘or usullari – attention mexanizmlari, turli optimizatorlar, transfer learning yondashuvlari – esa bu jarayonda tadqiqotchilarga asosiy qurol bo‘lib qoladi.

Zamonaviy ilmiy tadqiqotlar shuni ko‘rsatmoqdaki, chuqur o‘rganishdagi yondashuvlar uyg‘unligi tufayli til modellari tez suratlarda kuchayib, nafaqat global tillar, balki o‘z tilimiz – O‘zbek tilida ham ajoyib natijalar sari siljimoqda. Til modellari tabiiy tilni kompyuter orqali qayta ishlashda muhim rol o‘ynaydi. Ularning arxitekturasi (RNN, Transformer), o‘qitish metodlari (pretraining, attention, optimizatorlar) va qo‘llanilish sohalari har tomonlama chuqur tadqiq etilmoqda. O‘zbek tili uchun ham ilk yirik modellar yaratilmoqda va resurslar boyitilmoqda. Kelgusida O‘zbek tilida ko‘p funksiyali, ochiq manbali va kuchli til modellarining paydo bo‘lishi kutilmoqda.

Foydalanilgan adabiyotlar

1. Vaswani, A., et al. (2017). *Attention is All You Need*.
2. Sanatbek Matlatipov, Hulkar Rahimboeva, Jaloliddin Rajabov, and Elmurod Kuriyozov. 2022. Uzbek sentiment analysis based on local restaurant reviews
3. Maksud Sharipov, Jamolbek Mattiev, Jasur Sobirov, and Rustam Baltayev. 2022. Creating a morphological and syntactic tagged corpus for the uzbek language. CEUR Workshop Proceedings, 3315:93 – 98.
4. https://dumps.wikimedia.org/uzwiki

5. B Mansurov and A Mansurov. 2021b. Uzbert: pretraining a bert model for uzbek. arXiv preprint arXiv:2108.09814.
6. Devlin, J., et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT.
7. Brown, T., et al. (2020). *Language Models are Few-Shot Learners*.
8. Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing*. Pearson.