TRANSFORMER ARXITEKTURASIGA ASOSLANGAN TIL MODELLARI YORDAMIDA O‘ZBEK TILI MATNLARINI SEMANTIK TAHLIL QILISH

Mo‘minov Bahodir Boltayevich[1], Allaberganova Nasiba Muradovna[2]

[1]Toshkent davlat iqtisodiyot universiteti t.f.d., professor
E-mail: mbbahodir@tsue.uz
[2]Toshkent axborot texnologiyalari universiteti, tayanch doktorant
E-mail: abisan_1990@mail.ru

K E Y W O R D S

Transformer arxitektura, sun’iy intellekt, semantik tahlil, tabiiy tilni qayta ishlash, til modellari, BERT, GPT, RoBERTa, XLNet, T5, kontekst.

A B S T R A C T

Maqolada sun’iy intellekt texnologiyalariga asoslangan holatda matnlarning semantik tahlil qilish jarayonlari, zamonaviy yondashuvlar va mavjud muammolarga yondashuvlar o‘rganilgan. Tadqiqotda Transformer arxitekturasiga asoslangan til modellari (BERT, GPT, RoBERTa) tahlil qilinib, ularning imkoniyatlari va cheklovlari o‘zbek tilidagi matnlar misolida baholangan. Shuningdek, xalqaro miqyosda keng tarqalmagan milliy yoki mintaqaviy tillar(lokal tillar)da semantik tahlilni rivojlantirish bo‘yicha tavsiyalar ishlab chiqilgan.

KIRISH

So‘nggi yillarda sun’iy intellekt texnologiyalarining rivojlanishi tabiiy tilga ishlov berish jarayonlari (NLP) yo‘nalishida sezilarli yutuqlarga olib keldi. Semantik tahlil - bu matn mazmunini tahlil qilish, so‘zlar va gaplar o‘rtasidagi ma’no bog‘liqliklarini aniqlash jarayonidir. Bu NLPning eng muhim yo‘nalishlaridan biri hisoblanadi[1].

Sun’iy intellekt modellarining semantikani tushunish darajasi ancha rivojlangan bo‘lsada, ular hanuzgacha kontekst asosida ma’no chiqarish, kinoya ma’nosida aytilgan so‘zlarni va tilga xos xususiyatlar(tilning fonetik, leksik, grammatik va pragmatik xususiyatlari)ni chuqur anglashda qiyinchiliklarga duch kelmoqda.

Semantik tahlil dastlab statistik yondashuvlarga asoslanib TF-IDF, LSA va LDA kabi modellarda matnning statistik xususiyatlari tahlil qiladi. Bu usullar so‘zlar orasidagi kontekstual bog‘liqlikni chuqur anglashda cheklovlarga ega. Neyron tarmoqlarga asoslangan modellar Word2Vec, GloVe semantik munosabatlarni vektor shaklda ifodalash imkonini berdi. Biroq bu modellar bir tomonlama kontekstni hisobga oladi, ya’ni so‘z ma’nosini to‘liq anglashda yetarli emas.

2017 - yilda Transformer arxitekturasi Google kompaniyasining sakkiz nafar tadqiqotchisi tomonidan taklif qilinib, NLPda inqilob yasadi. Ushbu yondashuv asosida yaratilgan BERT, GPT, RoBERTa, XLNet, T5 kabi modellar tilni chuqur tushunishda yangi bosqichni boshlab berdi. Xususan, BERT modeli ikki tomonlama kontekstni hisobga olish xususiyati bilan ajralib turadi[2].

O‘zbek tilida NLP bo‘yicha olib borilgan tadqiqot ishlarini yetarli deya emas. Ko‘p hollarda mavjud modellar ingliz yoki rus tilida o‘rgatilgan bo‘lib, ularni o‘zbek tiliga to‘g‘ridan-to‘g‘ri tatbiq etish natijalar sifatsizligiga olib kelmoqda. Transformer arxitekturasi semantik tahlil, mashinali tarjima, matn tasnifi kabi vazifalarni bajarishda an’anaviy RNN va LSTM modellarini ortda qoldirdi[3],[4],[5]. Transformer asosidagi BERT, RoBERTa, XLM-R modellarining imkoniyatlarini va tuzilishini tahlil qilamiz. Ushbu modellar ko‘p tilli matnlar ustida o‘qitilgan bo‘lsada, o‘zbek tili uchun chuqur semantik tahlilni amalga oshira olmaydi. Shuningdek, o‘zbek tilida NLPga oid ishlanmalar mavjud bo‘lsada, ular semantik tahlilning chuqur sintezi, ayniqsa transformer asosida ishlab chiqilmagan.

Transformer arxitekturasiga asoslangan umumiy til modellari (BERT, RoBERTa, XLM-R, GPT) uchun matematik model quyida keltirilgan[6].

(cc) BY ©Mo‘minov B., Allaberganova N.

1. Kirish ma'lumotlarini formallashtirish:

X matn quyidagi so'zlar ketma-ketligi sifatida tasvirlanadi:

$$X = \{x_1, x_2, \dots, x_n\}, \quad x_i \in v$$

X — matn ketma-ketligi, ya'ni tahlil qilinadigan kirish jumlasi yoki hujjat, x_i *— i-chi so'z (so'z yoki belgilar birligi),* v *— so'zlar lug'ati.*

2. So'zlar Embedding(matnni vektorga aylantirish) funksiyasi yordamida vektorlarga o'tkaziladi[7]:

$$E = \{e_1, e_2, \dots, e_n\}, \quad e_i = f_{embed}(x_i), \quad e_i \in R^d$$

f_{embed} *— Embedding funksiyasi, ya'ni har bir tokenni* d *— o'lchamli real sonlar vektoriga o'zgartiradi,* R^d *— bu d-o'lchamli haqiqiy sonlar fazosi degani. Ya'ni, embedding vektori d ta haqiqiy (real) sondan iborat vektor sifatida tasvirlanadi;* d *—Embedding vektorining o'lchami, ya'ni vektor nechta koordinatadan tashkil topganligini bildiradi (masalan, d=768 yoki 1024).*

Ilk yirik transformer modellarining (BERT, GPT) o'zgarishlari asosida 768 va 1024 embedding o'lchamlari keng qo'llanib, standartga aylangan. Ko'pgina tadqiqotlar va amaliy ishlarda shu o'lchamlar eng samarali deb topilgan.

3.Transformer modellarida ketma-ketlik tartibini saqlash uchun pozitsion kodlash amalga oshiriladi[8]:

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10000^{2i/d}}\right), \quad PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{2i/d}}\right)$$

$PE_{(pos,k)}$ *— pos pozitsiyadagi k-o'lchamli pozitsion kodlash komponentasi,* i *— komponentaning indeks raqami (0 dan boshlab),* d *— Embedding vektorining o'lchami.*

Faraz qilaylik, pozitsiya: pos=3, Embedding o'lchami: d=8 (soddaroq misol bo'lishi uchun shunday oldik). Transformer modelidagi pozitsion kodlash vektori qiymatlari (pos=3, d=8) uchun komponentalarning sinus-kosinus hisoblari quyidagi 1-jadvalda keltirilgan.

1-Jadval.

Pozitsion kodlash komponentalari (sinus-kosinus)

Komponent(k)	**Formula**	**Natija**
$PE_{(3,0)}$	$\sin\left(\frac{3}{10000^{0/8}}\right) = \sin(3)$	0.141120
$PE_{(3,1)}$	$\cos\left(\frac{3}{10000^{0/8}}\right) = cos\ (3)$	-0.989992
$PE_{(3,2)}$	$\sin\left(\frac{3}{10000^{2/8}}\right) = \sin(3/100^{0.5})$	0.811094
$PE_{(3,3)}$	$\cos\left(\frac{3}{10000^{2/8}}\right) = cos(3/100^{0.5})$	0.584925
…	…	…
$PE_{(3,7)}$	$\cos\left(\frac{3}{10000^{6/8}}\right)$	0.900659

4. Yakuniy kirish ma'lumotlari (Embedding +Pozitsiyali kodlash):

$$Z = E + PE$$

Z - Har bir token uchun ma'no va tartibni o'z ichiga olgan yakuniy kirish vektorini beradi, E - Embedding vektori (so'zlarni vektorlash natijasi) so'zning ma'nosini ifodalaydi , PE - har bir pozitsiya uchun sinus-kosinus kodlashidan olingan vektor. Bu vektor so'zning joylashuvini bildiradi. Ushbu Z vektor self-attention qatlamlariga uzatiladi. Ushbu qatlamni "Ko'p funksiyali e'tibor mexanizmi" (Multi-Head self Attention) deymiz. Har bir bosqichda quyidagi operatsiyalar bajariladi:

$$Q = ZW^Q, \quad K = ZW^K, \quad V = ZW^V$$

Z - kirish vektori (embedding + pozitsion kodlash), W^Q, W^K, W^V *- mos ravishda query, key, value vazifasini bajaruvchi o'rgatiladigan og'irlik*

©Mo'minov B., Allaberganova N.

(weight) matritsalari, Q, K, V - e'tibor mexanizmida ishlatiladigan asosiy vektorlar.

E'tibor mexanizmi:

$$Attention(K,Q,V) = soft\,max\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

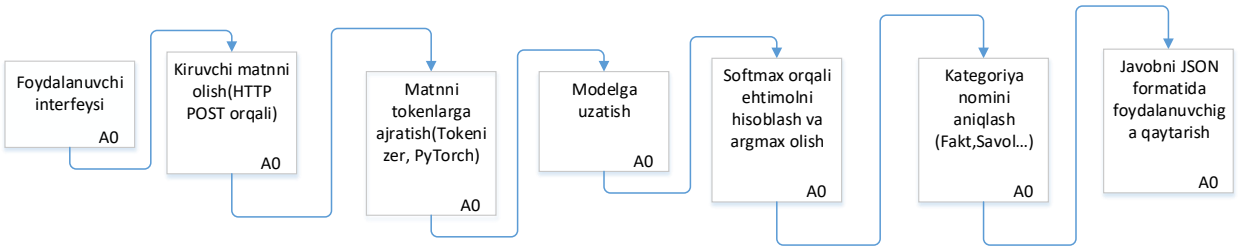
$\sqrt{d_K}$ — softmaxda gradient barqarorligini ta'minlash uchun normalizatsiya. Bu formula orqali model har bir tokenning boshqa tokenlarga e'tibor darajasini hisoblaydi. Natijada, har bir token o'zining global kontekstini hisobga olgan yangi vektorga ega bo'ladi.

5.Semantik ma'no chiqish bosqichi. Yakuniy vektorlar semantik tasniflovchi qatlamga uzatiladi[9]:

$$y = soft\,max(W_f \cdot h + b)$$

Bu yerda: h — so'nggi qatlamdagi semantik vakillik (hidden state), W_f — chiqish vazifasi uchun o'rgatiladigan og'irliklar, b — bias vektori, y – semantik toifalar: fakt, savol, buyruq, ijobiy/inkor, hissiy, neytral va boshqalar.

Yuqoridagi matematik modellarga asoslangan holda ishlar ketma-ketligini quyidagi funksional tuzilma ko'rinishida ifodalanadi.



1-rasm. Dasturiy vositaning funksional tuzilmasi.

Yuqoridagi matematik modellarga hamda amallar ketma ketligiga asoslangan holda dasturiy vositaning ishlash bosqichlari quyidagilardan tarkib topadi. Dasturning foydalanuvchi interfeysi qismida (matn kiritish, natijani ko'rish) hamma uchun qulay va sodda shaklda yaratiladi. Dasturning dasturiy vosita arxitekturasi qismida bo'lsa, PyTorch kutubxonasi asosida transformer modeli ishlatiladi. Transformer model o'zbek tilidagi katta hajmdagi matnlar Wikipedia va UzCorpora asosida o'qitildi. Bu modelga o'zbek tilining semantik qonuniyatlarini, matn tarkibini va kontekstual ma'nolarini o'zlashtirish imkonini berdi. Natijada, model real o'zbekcha matnlarni tahlil qilishda yuqori aniqlik bilan ishlashga erishdi. Model sinov qilinganida, u 89% hollarda to'g'ri aniqlashga erishgan (aniqlik), va 85% real javoblarni topa olgan. Ikkalasining o'rtacha muvozanatli bahosi - 87% (F1)ni tashkil qiladi. Bu model yuqori darajada ishonchli ishlaganini ko'rsatadi.

Dasturda model 10 000 ta o'zbekcha jumla asosida sinovdan o'tkazildi va sinov natijalari 2-jadvalda qayd etildi.

2-jadval.

Modelning matnlarni semantik toifalar bo'yicha tasniflash samaradorligi natijalari

Semantik toifa	Precision	Recall	F1-score
Fakt	0.91	0.88	0.89
Savol	0.87	0.84	0.85
Hissiy	0.88	0.83	0.85

Model XLM-R bilan solishtirilganda 6.3% yaxshilanish kuzatildi.

Transformer arxitekturasi asosidagi yondashuv milliy tilning lingvistik xususiyatlariga moslashtirildi. Natijalar mavjud ko'p tilli modellar bilan solishtirganda yuqori aniqlikni ko'rsatdi.

©Mo'minov B., Allaberganova N.

Kelgusida kontekstual tahlil, ko‘p modallilik[1] va real vaqtda ishlashga yo‘naltirilgan optimizatsiyalar amalga oshirilishi rejalashtirilgan. Tahlillar quyidagi natijalarni berdi:

Transformer modellar (xususan, BERT va GPT) o‘zbek tili uchun semantik tahlilda yuqori aniqlik ko‘rsatdi[10];

O‘zbek tilidagi matnlarda kontekst tushunilishi darajasi ingliz tiliga nisbatan sezilarli pastroq;

So‘z ko‘p ma’noliligi (polysemy) model uchun murakkablik tug‘diradi;

Mintaqaviy ya’ni lokal tillar uchun maxsus korpuslar va belgilangan (annotatsiyalangan) ma’lumotlar bazasi yetishmaydi;

Modellarning ishlash samaradorligi ularni o‘rgatishda foydalanilgan ma’lumotlar hajmi va sifati bilan bevosita bog‘liq.

XULOSA

Zamonaviy modellar chuqur neyron tarmoqlar asosida qurilgan bo‘lsada, ular hamon madaniy kontekst, metafora, tilga xos konnotatsiyalarni chuqur anglay olmaydi. Bu yerda madaniy kontekst deganda oddiy bir jumla turli madaniyatlarda turlicha ma’no kasb etishi mumkin. Masalan, “Men choy puli qoldirmadim” (I didn’t leave a tip) jumlasi AQShda xizmatdan qattiq norozilikni anglatsa, choy puli berish odatiy bo‘lmagan madaniyatlarda hech qanday qo‘shimcha ma’no anglatmaydi. Model bu madaniy normani bilmasa, jumlani noto‘g‘ri talqin qiladi. O‘zbek tilidagi NLP loyihalarida mavjud resurslar cheklanganligi sababli, universal modellar to‘liq moslashmaydi, bu esa natijalarning sifatsizligiga sabab bo‘lmoqda[11,12].

Yechim sifatida quyidagi chora-tadbirlar keltirish mumkin:

- O‘zbek tiliga mos maxsus semantik korpuslar yaratish;
- Ko‘p tillik asosida tayyorlangan modellarni o‘zbek tiliga moslashtirish;

- Akademik hamda amaliy tadqiqotlar orqali til texnologiyalarini milliy darajada rivojlantirish.

Sun’iy intellekt asosidagi semantik tahlil tizimlari tilni chuqur tushunishga asoslangan avtomatlashtirilgan tizimlar yaratishda muhim vositadir. Global miqyosda BERT, GPT kabi modellar yuqori natijalarga erishgan bo‘lsada, ularni o‘zbek tilida samarali ishlatish uchun maxsus korpuslar, lingvistik resurslar va moslashtirilgan arxitekturalar zarur.

FOYDALANILGAN ADABIYOTLAR

1. Jurafsky, D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition / D. Jurafsky, J. H. Martin. – 2nd ed. – Upper Saddle River, NJ: Prentice Hall, 2009. –P. 988-1000
2. Raffel, C. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer / C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu // Journal of Machine Learning Research. – 2020. – Vol. 21, № 140. – P. 1–67
3. Vaswani, A. Attention is All You Need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008
4. Kakhorov A., Yodgorova D., Khayrullayev M. “AI-DRIVEN ENVIRONMENTAL

[1] O‘zbek tili semantik tahlili faqat matn bilan cheklanib qolmay, balki ovozli nutq (audio), yuz ifodalari (video), hatto rasmli kontekstlar bilan ham integratsiyalashgan holatda tahlil qilinishi ko‘zda tutilmoqda.

©Mo‘minov B., Allaberganova N.

COMMUNICATION: TRANSLATING SUSTAINABILITY REPORTS INTO UZBEK FOR GLOBAL AWARENESS” Экономика и социум. 2024. №6(121)-1. – P.292-300

5. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2019. – P. 4171–4186
6. Camacho-Collados, J. From Word to Sense Embeddings: A Survey on Vector Representations of Meaning / J. Camacho-Collados, M. T. Pilehvar // Computational Linguistics. – 2023. – Vol. 49, № 2. – P. 431–486
7. Sharipov, M. Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language / M. Sharipov, U. Salaev, S. Matlatipov, E. Kuriyozov // CEUR Workshop Proceedings. – 2023. – Vol. 3315. – P. 107–117
8. Allaberganova Nasiba Muradovna, Abdrimova Feruza Kamiljanovna. “His-tuyg‘ularni tahlil qilish usullari”, Scientific journal “Golden Brain”. Vol. 1, No 13. Uzbekistan – 2023. – B.260-267
9. Boltayevich, Mo‘minov Bahodir, Allaberganova Nasiba Muradovna, and Shermatov Boburjon Inoyat O‘g‘li. "NLPDA PYTHON DASTURLASH TILIDAGI BA’ZI JARAYONLAR". PROSPECTS A ND MAIN TRENDS IN MODERN SCIENCE 1.12 (2024): – B.88-97
10. Mo‘minov Bahodir Boltayevich, Allaberganova Nasiba Muradovna. “Семантик таҳлил дастури ва асосий элементлари”, “Development of science” Ilmiy jurnal (OAK). Buxoro – 2025. Volume 6, №6,ISSN 3030-3907. – B 68-79
11. Allaberganova Nasiba Muradovna, Asqarova Dildora Ulug‘bekovna. “Semantik tahlil” International scientific journal “Interpretation and researches”. Volume 1, Issue 10, 2023. ISSN: 2181-4163. – B.189-193
12. Allaberganova Nasiba Muradovna “Transformer avlodidagi modellar orqali o‘zbek tili uchun lemmatizatsiya va named entity recognition muammolarining yechimi”, “Sun’iy intellekt nazariyasi va amaliyoti: tajriba, muammolar va istiqbollar” mavzusidagi II Respublika ilmiy-amaliy anjumani materiallar to‘plami. Toshkent – 2025. – B.37-39

©Mo‘minov B., Allaberganova N.