LINGVOTEK: SO‘Z VA SUN’IY INTELLEKT

ISRAILOVA DILDORA ATXAMOVNA

Toshkent irrigatsiya va qishloq xo‘jaligini mexanizatsiyalash muhandislari instituti Milliy tadqiqot universiteti, Ingliz tili kafedrasi, o’qituvchi

O‘ZBEK TILIDAGI KORPUSLARDA SO‘Z BOYLIGINI SUN’IY INTELLEKT YORDAMIDA AVTOMATIK ANIQLASH: MUAMMOLAR VA YECHIMLAR

Annotatsiya: Ushbu maqolada o‘zbek tilidagi elektron korpuslarda so‘z boyligini sun’iy intellekt yordamida avtomatik aniqlashning nazariy va amaliy jihatlari tahlil qilinadi. Maqolaning dolzarbligi mavjud korpuslarda leksik birliklarni avtomatik tahlil qilish mexanizmlarining yetarli darajada rivojlanmaganligi bilan belgilanadi. Maqolaning maqsadi – o‘zbek tili morfologik xususiyatlariga moslashtirilgan neyrotarmoqli modellar asosida so‘z boyligini aniqlash samaradorligini oshirish yo‘llarini taklif etishdir. Sun’iy intellekt vositalari o‘zbek tili korpuslarining leksikografik salohiyatini sezilarli darajada kengaytirishi mumkin.
Kalit so‘zlar: o‘zbek tili korpusi, so‘z boyligi, sun’iy intellekt, neyrotarmoqlar, avtomatik aniqlash, raqamli filologiya, morfologik tahlil, kompyuter lingvistikasi.

Abstract: This paper examines the theoretical and practical aspects of automatic vocabulary extraction in Uzbek electronic corpora using artificial intelligence. The relevance of this study stems from the underdeveloped state of automatic lexical analysis mechanisms in existing corpora. The aim is to propose methods for improving the efficiency of vocabulary identification based on neural network models tailored to the morphological characteristics of the Uzbek language. Artificial intelligence tools can substantially enhance the lexicographic potential of Uzbek language corpora
Keywords: Uzbek language corpus, lexical richness, artificial intelligence, neural networks, digital philology, morphological analysis, computational linguistics.

KIRISH.O‘zbek tilining milliy korpuslarini yaratish va rivojlantirish so‘nggi o‘n yillikda filologiyaning muhim yo‘nalishlaridan biriga aylandi. Bugungi kunda “O‘zbekiston Milliy korpusi” (uzbekcorpus.uz) va bir qator mintaqaviy korpuslarda millionlab so‘z birliklari jamlangan. Biroq mavjud korpuslarning aksariyati so‘z boyligini avtomatik tahlil qilish imkoniyatlariga ega emas. So‘z boyligi – tilning lug‘at tarkibining ranginligi, xilma-xilligi va miqdoriy jihatdan ifodalanishi bo‘lib, uni qo‘lda aniqlash vaqt va mehnat talab etadi.

Sun’iy intellekt (SI), xususan, chuqur o‘rganish (deep learning) va neyrotarmoqli modellar tabiiy tilni qayta ishlashda (NLP) yuqori natijalarni ko‘rsatmoqda. Shunga qaramay, o‘zbek tili kabi agglyutinativ tillar uchun SI modellarini moslashtirish muammosi haligacha yetarlicha o‘rganilmagan. Ushbu maqolaning asosiy muammosi – o‘zbek tili korpuslarida so‘z boyligini avtomatik aniqlashning mavjud usullarining past samaradorligi va bu kamchiliklarni bartaraf etishda SI imkoniyatlarini o‘rganish.

TADQIQOT METODOLOGIYASI

Tadqiqotda uch asosiy metod qo‘llanildi:

1. Korpus lingvistikasi metodi – O‘zbekiston Milliy korpusining 500 ming so‘zli vakillik namunasi asosida leksik birliklarni ajratib olish.
2. Kompyuter modellashtirish – Python dasturlash tilida TensorFlow va spaCy kutubxonalari yordamida neyrotarmoqli model yaratish.
3. Eksperimental sinov – Taklif etilayotgan modelni mavjud statistik (TF-IDF) va qoida-asosidagi (rule-based) usullar bilan aniqlik (precision), to‘liqlik (recall) va F1-mezon bo‘yicha solishtirish.

Modelni o‘qitish uchun korpusdan 80% (400 ming so‘z) – trening, 10% (50 ming so‘z) – validatsiya, 10% (50 ming so‘z) – test tanlanmasi sifatida ajratildi.

ASOSIY QISM

O‘zbek tilining agglyutinativ xususiyati va uning SI uchun ahamiyati

O‘zbek tili – agglyutinativ tillar guruhiga kiradi. Bu degani, so‘z o‘zagiga bir nechta affikslar ketma-ket qo‘shiladi. Masalan, “kel” (o‘zak) → “kela” → “keladi” → “kelmayapti” → “keltirilayotganlardan”. Bunday hosilalar soni nazariy jihatdan cheksizga yaqin. An’anaviy so‘z boyligini hisoblashda har bir so‘z shakli alohida birlik sifatida qaralsa, noto‘g‘ri natijalar yuzaga keladi. SI modellari uchun esa aynan shu morfologik xilma-xillikni bir o‘zakka qarashli variantlar sifatida tanib olish muhimdir.

Amaldagi uch usul tahlil qilindi:

TF-IDF – so‘z chastotasiga asoslangan, ammo affiksal o‘zgarishlarni hisobga olmaydi.

Rule-based – qo‘lda yozilgan morfologik qoidalar (masalan, -lar, -ga, -dan), lekin istisno va yangi so‘zlarga moslasha olmaydi.

Word2vec (skip-gram) – kontekstga asoslangan, biroq resurs intensiv va kam uchraydigan so‘z shakllarida xatoga yo‘l qo‘yadi.

Eksperimental natijalarga ko‘ra, ushbu usullarning o‘zbek tilidagi korpusda so‘z boyligini to‘g‘ri aniqlash bo‘yicha o‘rtacha F1-ko‘rsatkichi 67,3% ni tashkil etdi.

Taklif etilayotgan neyrotarmoqli model

Biz tomonimizdan ikki qavatli BiLSTM (Bidirectional Long Short-Term Memory) modeli ishlab chiqildi. Model quyidagi bosqichlarni o‘z ichiga oladi:

1. Tokenizatsiya – bo‘shliq va tinish belgilariga asoslangan boshlang‘ich ajratish.
2. Morfologik segmentatsiya – o‘zak va qo‘shimchalarni ajratuvchi yordamchi neyron tarmoq.
3. O‘zaklarni guruhlash – bir o‘zakka qarashli barcha so‘z shakllarini bitta leksemaga birlashtirish.

4. Chiqish – korpusdagi noyob o‘zaklar soni, uning umumiy so‘z shakllariga nisbati (type-token ratio) va leksik zichlik ko‘rsatkichi.

Model 50 epoch davomida o‘qitildi (batch size=64, learning rate=0.001). O‘qitish ma’lumotlarida 12 500 ta o‘zak va ularning 87 000 ga yaqin so‘z shakli belgilangan holda kiritildi.

TAHLIL VA NATIJALAR

Taklif etilgan BiLSTM modeli test tanlanmasida quyidagi natijalarni ko‘rsatdi:

Usul	**Aniqlik (Precision)**	**To‘liqlik (Recall)**	**F1-mezon**
TF-IDF	0,62	0,58	0,60
Rule-based	0,71	0,68	0,69
Word2vec	0,74	0,70	0,72
BiLSTM (taklif)	0,85	0,82	0,83

Taklif etilayotgan model F1-mezon bo‘yicha eng yaxshi mavjud usuldan (Word2vec) 0,11 ga (taxminan 12,4 foizga) yuqori natija qayd etdi. Model ayniqsa ko‘p ma’noli affikslar (masalan, -ma inkor va -ma ot yasovchi) va kam chastotali so‘z shakllarida sezilarli ustunlik ko‘rsatdi.

Shuningdek, modelning ishlash tezligi – 500 ming so‘zli korpusni qayta ishlash uchun o‘rtacha 4,2 soniya vaqt sarflandi (Intel Core i7, 16 GB RAM). Bu real vaqt rejimidagi tahlil uchun maqbul ko‘rsatkich hisoblanadi.

XULOSA

Ushbu tadqiqotda o‘zbek tilidagi korpuslarda so‘z boyligini avtomatik aniqlash uchun sun’iy intellektga asoslangan BiLSTM modeli taklif qilindi. Model o‘zbek tilining agglyutinativ xususiyatlarini inobatga olib, mavjud usullarga nisbatan 12,4% yuqori aniqlikni ta’minladi. Bu esa o‘zbek tilshunosligida keng qo‘llaniladigan korpus tahlillarini avtomatlashtirish, leksikografik ishlarni soddalashtirish va raqamli filologik resurslarni boyitish imkonini beradi.

Kelgusidagi tadqiqotlarda modelni dialektal korpuslar va tarixiy matnlar ustida sinovdan o‘tkazish, shuningdek, katta til modellari (LLM, masalan, LLaMA yoki GPT bilan) integratsiyasi rejalashtirilgan.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI

1. Abjalova, M. (2022). O‘zbek tilining milliy korpusi: yaratilish tamoyillari va imkoniyatlari. O‘zbek tili va adabiyoti, 3(4), 45–52.
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111–3119.
3. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.
4. Sirojiddinov, Sh., & Yusupov, O. (2021). Raqamli filologiya: nazariya va amaliyot. Toshkent: Mumtoz so‘z.
5. Iskandarova, D. (2023). O‘zbek tili korpuslarida leksik boylikni o‘lchash metodlari. Filologiyamasalalari, 2(1), 18–27.