KOGNITIV-ENSIKLOPEDIK MODELLARNI TABIIY TILNI QAYTA ISHLASH TIZIMLARIGA JORIY ETISHNING LINGVISTIK TAMOYILLARI

Abduqaxxorov Sardorbek Akramjon o‘g‘li
[ilmiy darajasi, lavozimi], Andijon davlat chet tillari instituti
E-mail: sardorbek.akramovich@gmail.com

Annotatsiya. Tezisda ensiklopedik bilim modelini til modeli yoki qidiruvga asoslangan generatsiya tizimiga ulashdan oldin lingvistik jihatdan baholash g‘oyasi ilgari suriladi. Shakl va ma’no ajratilishi haqidagi bahs, bilim grafi ta’riflari hamda ma’lumotlar hujjatlashtirilishi bo‘yicha xalqaro tajriba asosida beshta tamoyil taklif etiladi. Ular konseptual moslik, granulyarlik, madaniy moslashuv, dalil shaffofligi va bosqichma-bosqichlikdir. Tamoyillar o‘zbek tilidagi resurslar uchun joriy etish oldi sinovi sifatida qo‘llanishi mumkin.

Kalit so‘zlar: kognitiv-ensiklopedik model; tabiiy tilni qayta ishlash; bilim grafi; qidiruvga asoslangan generatsiya; lingvistik tamoyil; madaniy moslashuv.

Katta til modellari matn hosil qilishda yuqori natijalarga erishgan bo‘lsa-da, ularning tilni “tushunishi” haqidagi da’volar lingvistlar tomonidan jiddiy shubha ostiga olinmoqda. E. Bender va A. Koller fikricha, ma’no shakl bilan tildan tashqaridagi narsa o‘rtasidagi munosabat bo‘lgani uchun matnning faqat shakliy tomonidan o‘rgangan model ma’noga o‘z-o‘zidan yetib bora olmaydi [1]. T. Brown va hammualliflarining 175 milliard parametrli GPT-3 modeli haqidagi maqolasida ham bunday modellarning video yoki real jismoniy o‘zaro ta’sir kabi tajriba sohalariga tayanmasligi cheklov sifatida qayd etilgan [2]. Demak, tashqi bilim manbai til modeli uchun ixtiyoriy qo‘shimcha emas, balki ma’no bilan bog‘lanishning amaliy yo‘lidir.

Bunday bog‘lanishning keng tarqalgan texnik shakli qidiruvga asoslangan generatsiyadir. P. Lewis va hammualliflari parametrli xotirani (oldindan o‘qitilgan generator) Vikipediyaning zich vektor indeksidan iborat noparametrik xotira bilan birlashtirgan va bilim talab qiladigan vazifalarda natija yaxshilanganini ko‘rsatgan [3]. Mualliflar qarorlar uchun manba ko‘rsatish va dunyo haqidagi bilimni yangilash ochiq tadqiqot muammolari bo‘lib qolayotganini ham tan oladi [3]. Ammo tashqi manbaning o‘zi lingvistik jihatdan qanday bo‘lishi kerakligi texnik ishlarda kam muhokama qilinadi. Ensiklopedik model tizimga ulanganda uning tuzilishidagi har bir nuqson generatsiya natijasiga to‘g‘ridan to‘g‘ri o‘tadi.

Shu sababli kognitiv-ensiklopedik modelni ulashdan oldin beshta lingvistik tamoyil asosida baholashni taklif qilamiz. Birinchisi konseptual moslik tamoyilidir. Til birligi va bilim bazasidagi tugun o‘rtasidagi bog‘lanish leksema darajasida emas, ma’no darajasida o‘rnatilishi lozim. Bilim grafi haqidagi ta’riflarni tahlil qilgan L. Ehrlinger va V. Wöß bilim grafini axborotni ontologiyaga yig‘ib integratsiya qiluvchi

va yangi bilim chiqarish uchun mantiqiy xulosalash vositasini qo‘llovchi tizim deb ta’riflaydi [4]. Agar ko‘p ma’noli so‘zning har bir ma’nosi alohida tugunga bog‘lanmasa, xulosalash vositasi noto‘g‘ri asosdan to‘g‘ri ko‘ringan, lekin xato natija chiqaradi.

Ikkinchisi granulyarlik tamoyilidir. Bilim bloki so‘rovning tabiiy hajmiga mos bo‘lishi kerak. Juda yirik blok (butun ensiklopedik maqola) generatorga ortiqcha va bir-biriga zid ma’lumot beradi, juda mayda blok (alohida fakt) esa kontekstni yo‘qotadi. Shartli misol sifatida “Navro‘z” konseptini olaylik. Sana haqidagi so‘rovga bitta faktli blok yetarli, bayram marosimlari haqidagi so‘rovga esa voqea ssenariysini qamrab olgan, ishtirokchilar, harakatlar va predmetlar bog‘langan freymli blok kerak bo‘ladi. Demak, model bir konsept uchun bir nechta o‘lchamdagi blokni ta’minlashi lozim. Bloklar o‘rtasida “qism va butun” munosabati aniq ko‘rsatilsa, qidiruv tizimi so‘rov turiga qarab kerakli darajani tanlay oladi va bir javobda bir xil faktning takrorlanishi kamayadi.

Uchinchisi madaniy moslashuv tamoyilidir. Ensiklopedik bilim neytral ko‘rinsa-da, u qaysi til jamoasining nuqtai nazaridan yozilganiga bog‘liq. Sh. Safarov kognitiv tilshunoslikning markazida til orqali bilimning egallanishi, saqlanishi va uzatilishi masalasi turishini ko‘rsatadi [5]. Agar o‘zbekcha so‘rovga ingliz tilidagi manbadan tarjima qilingan blok qaytarilsa, konseptning milliy tarkibi tushib qolishi mumkin. Shartli misol: “hashar” konsepti tarjima orqali faqat “ko‘ngilli mehnat” sifatida berilsa, uning qo‘shnichilik, o‘zaro yordam burchi va jamoaviy marosim bilan bog‘liq qatlamlari yo‘qoladi. Shu bois blokda uning qaysi til va madaniyat materialiga tayanganini ko‘rsatuvchi belgi bo‘lishi kerak.

To‘rtinchisi dalil shaffofligi tamoyilidir. T. Berners-Lee, J. Hendler va O. Lassila semantik veb g‘oyasini axborotga aniq belgilangan ma’no berish va shu orqali kompyuter bilan inson hamkorligini ta’minlash sifatida bayon qilgan [6]. Bu yondashuvda har bir tasdiqning kelib chiqishini kuzatish imkoniyati asosiy shartlardan biri hisoblanadi. E. Bender va B. Friedman tabiiy tilni qayta ishlash uchun ma’lumotlar to‘plamiga til navi, so‘zlovchilar tavsifi va tanlov asoslari yozilgan “ma’lumotlar bayonnomasi”ni ilova qilishni taklif qilgan [7]. Shu talabni ensiklopedik blokka ko‘chirsak, har bir blok manba, sana, til navi va tahrir holati bilan birga uzatilishi kerak. Bunda generator javobida manbani ko‘rsata oladi va P. Lewis va hammualliflari qayd etgan manba muammosi qisman yechiladi.

Beshinchisi bosqichma-bosqichlik tamoyilidir. Bilim modeli yaxlit holda emas, sinalgan qismlar bilan ulanishi lozim. Avval eng barqaror va kam bahsli sohalar (geografik nomlar, sanalar, atamalar) ulanadi, keyin baholash turlicha bo‘lishi mumkin bo‘lgan tarixiy va madaniy konseptlarga o‘tiladi. Har bir bosqichda modelning javoblari mutaxassis tomonidan tekshiriladi va topilgan xatolar bilim bazasiga qaytariladi. Bunday yo‘l bilimni yangilash muammosiga ham mos keladi, chunki noparametrik xotirani modelni qayta o‘qitmasdan almashtirish mumkin [3].

Besh tamoyil birgalikda joriy etish oldi sinovini hosil qiladi. Tizim muhandisi ensiklopedik resursni ulashdan oldin beshta savolga javob berishi kerak. Har bir tugun ma’noga bog‘langanmi, blok o‘lchami so‘rov turiga mosmi, madaniy belgi bormi, manba va til navi ko‘rsatilganmi, ulanish bosqichlari belgilanganmi. Bu savollar texnik ko‘rsatkichlarni almashtirmaydi, ularni lingvistik mezonlar bilan to‘ldiradi. Sinovning afzalligi shundaki, u resurs qaysi texnologiya bilan ulanishidan qat’i nazar qo‘llanadi. Bilim grafi, vektor indeksi yoki oddiy matnli bloklar to‘plami uchun ham savollar o‘zgarmaydi, faqat ularni tekshirish usuli farq qiladi. Agar biror savolga salbiy javob olinsa, resurs ulanmaydi, avval tegishli qatlam tuzatiladi.

Xulosa qilib aytganda, kognitiv-ensiklopedik model tabiiy tilni qayta ishlash tizimida til modelining shakl bilan cheklangan imkoniyatini ma’no bilan to‘ldirishi mumkin, lekin buning uchun u oldindan lingvistik baholashdan o‘tishi lozim. Taklif etilgan beshta tamoyil o‘zbek tilidagi elektron ensiklopedik resurslarni zamonaviy generativ tizimlarga ulashda tilshunos va muhandis o‘rtasidagi umumiy mezonlar to‘plami vazifasini bajarishi mumkin. Keyingi ishlarda har bir tamoyil uchun miqdoriy baholash usullarini ishlab chiqish rejalashtirilgan.

ADABIYOTLAR

1. Bender E.M., Koller A. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – Online: ACL, 2020. – P. 5185–5198.
2. Brown T.B., Mann B., Ryder N. et al. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). – 2020. – P. 1877–1901.
3. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). – 2020. – P. 9459–9474.
4. Ehrlinger L., Wöß W. Towards a Definition of Knowledge Graphs // Joint Proceedings of the Posters and Demos Track of SEMANTiCS 2016 and SuCCESS’16. – Leipzig, 2016. – (CEUR Workshop Proceedings. Vol. 1695). – P. 1–4.
5. Safarov Sh. Kognitiv tilshunoslik. – Jizzax: Sangzor, 2006.
6. Berners-Lee T., Hendler J., Lassila O. The Semantic Web // Scientific American. – 2001. – Vol. 284, No. 5. – P. 34–43.
7. Bender E.M., Friedman B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science // Transactions of the Association for Computational Linguistics. – 2018. – Vol. 6. – P. 587–604.