ENSIKLOPEDIK BILIMLAR BAZASI TUGUNINING IKKI YO‘NALISHLI TAVSIFI: O‘ZBEK TILIDA SEMANTIK QIDIRUV VA MASHINA TARJIMASI UCHUN LISONIY SHARTLAR

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

Annotatsiya. Tezisda elektron ensiklopedik bilimlar bazasidagi bitta tugun ham qidiruv so‘rovini kengaytirish, ham atoqli ot va madaniy terminni tarjima qilish uchun xizmat qila olishi asoslanadi. Buning uchun tugunga ikki xil belgi qo‘shish taklif etiladi, ya’ni munosabatlarning kengaytirish xavfi darajasi va tarjima siyosati belgisi. O‘zbek tili uchun zarur lisoniy shartlar sifatida lemma bilan bog‘lash, ikki yozuvdagi shakllar va omonim atoqli otlarni farqlash ko‘rsatiladi.

Kalit so‘zlar: semantik qidiruv; so‘rovni kengaytirish; mashina tarjimasi; atoqli ot; madaniy termin; bilimlar bazasi; o‘zbek tili.

Axborot qidiruvi va mashina tarjimasi odatda alohida yo‘nalish sifatida o‘rganiladi, biroq ikkalasi bir xil lisoniy muammoga duch keladi. Foydalanuvchi yozgan so‘z shakli u nazarda tutgan tushunchaga to‘g‘ri kelmasligi mumkin. Qidiruvda bu hol relevant hujjatning topilmay qolishiga, tarjimada esa atoqli otning buzilishi yoki madaniy terminning noto‘g‘ri talqin qilinishiga olib keladi. Elektron ensiklopedik bilimlar bazasi bu ikki vazifani yagona tushuncha tuguni orqali bog‘lashga imkon beradi. Lekin buning uchun tugun faqat ta’rif va havolalar to‘plami bo‘lib qolmasdan, uni ishlatadigan tizimga qaysi munosabatdan qanday foydalanish mumkinligini ham bildirishi kerak.

So‘rovni kengaytirish axborot qidiruvi nazariyasida yaxshi yoritilgan. R. Baeza-Yates va B. Ribeiro-Neto relevantlik bo‘yicha qayta aloqa va so‘rov kengaytirishni ikki yo‘l bilan tavsiflaydi. Global tahlil butun hujjatlar to‘plami asosida tuzilgan o‘xshashlik yoki statistik tezaurusdan foydalanadi, lokal tahlil esa so‘rovga javoban topilgan dastlabki hujjatlardan qo‘shimcha terminlar oladi [1]. Ensiklopedik bilimlar bazasi global tahlilning statistik tezaurusiga nisbatan boshqacha manba hisoblanadi, chunki undagi munosabatlar chastotadan emas, tushunchalar o‘rtasidagi aniq aloqadan kelib chiqadi. Shu sabab bazadagi har bir munosabat turi kengaytirishga turlicha xavf bilan qo‘shiladi.

Taklif etilayotgan tamoyil bo‘yicha munosabatlar kengaytirish xavfiga ko‘ra uch darajaga ajratiladi. Birinchi darajaga xavfsiz munosabatlar kiradi, ular so‘rov mazmunini deyarli o‘zgartirmaydi. Bu yerda lemma va uning so‘z shakllari, lotin va kirill yozuvidagi variantlar, aniq sinonimlar nazarda tutiladi (shartli misol “shifoxona” va “kasalxona”). O‘zbek tilida sinonimik qatorlar A. Hojiyev lug‘atida bosh so‘z atrofida jamlangan [2], bunday qatorlar bazaga ko‘chirilganda har bir a’zoning qo‘llanish doirasi ham saqlanishi lozim. Aks holda uslubiy chegaralangan so‘z so‘rovga qo‘shilib, natijalarni toraytirish o‘rniga chalg‘itadi.

Ikkinchi daraja shartli munosabatlarni qamrab oladi, bular gipernim va giponimlardir. Shartli misolda “o‘rik” so‘rovini “mevali daraxt” tuguniga ko‘tarish natijalar sonini oshiradi, lekin aniqlikni pasaytiradi. Shuning uchun gipernim bo‘yicha kengaytirish faqat dastlabki so‘rov kam natija berganda ishga tushirilishi va past vazn bilan qo‘shilishi maqsadga muvofiq. Gipernim nomzodlarini izohli lug‘at ta’riflaridan olish mumkin [3], ammo ta’rifdagi bosh so‘z har doim ham ontologik jins bo‘lavermaydi, shu sabab bunday nomzodlar mutaxassis tomonidan tekshirilishi kerak. Uchinchi daraja freym munosabatlaridir. Shartli misolda “paxta terimi” freymidagi vaqt, joy va ishtirokchi rollari so‘rovni hujjatlar tavsifi darajasida boyitadi. Ular foydalanuvchi niyati aniq bo‘lganda foydali, aks holda so‘rov mavzudan uzoqlashadi.

Bilimlar bazasini tashqi xotira sifatida ishlatish g‘oyasi zamonaviy til modellarida ham kuzatiladi. P. Lewis va hammualliflari taklif etgan qidiruvga tayangan generatsiya modelida parametrik bo‘lmagan xotira Wikipedia matnlarining zich vektorli indeksidan iborat bo‘lib, mualliflar faqat parametrlarga tayangan modelga qaraganda aniqroq va faktga ko‘proq mos matn olinganini qayd etadi [4]. Bu natija qidiruv bosqichi sifati butun tizim sifatini belgilashini ko‘rsatadi. Demak, o‘zbek tilidagi so‘rov noto‘g‘ri tugunga bog‘lansa, keyingi barcha bosqichlar xatoni takrorlaydi.

Mashina tarjimasida xuddi shu tugun boshqa vazifani bajaradi. D. Bahdanau, K. Cho va Y. Bengio manba gapni bitta qat’iy uzunlikdagi vektorga siqish kodlovchi va dekodlovchi arxitektura uchun cheklov ekanini ko‘rsatib, har bir maqsad so‘z uchun manba gapning tegishli qismlarini yumshoq izlash mexanizmini taklif qilgan [5]. Bu mexanizm so‘zlarni moslashtiradi, lekin atoqli otning maqsad tilda qabul qilingan shaklini o‘zi bilmaydi. Qoidaga asoslangan Apertium platformasi esa bir tilli va ikki tilli lug‘atlarga hamda transfer qoidalariga tayanadi [6], shu bois unda bunday shakl lug‘at yozuvi sifatida aniq kiritilishi mumkin. Ensiklopedik tugun shu ikki yondashuv o‘rtasida ko‘prik bo‘la oladi. Neyron tizimda u tarjimadan oldin manba matndagi nomni belgilab beradi, qoidaga asoslangan tizimda esa ikki tilli lug‘atni yangi yozuvlar bilan to‘ldirish manbai bo‘ladi.

Buning uchun tugunga tarjima siyosati belgisi qo‘shish taklif etiladi. Bu belgi uch qiymatdan birini oladi. Birinchisi “qat’iy shakl” bo‘lib, maqsad tilda an’anaviy nomi mavjud joy va shaxs nomlariga beriladi (shartli misol “Termiz”, inglizcha matnda “Termez”). Ikkinchisi “transliteratsiya va izoh” bo‘lib, ekvivalenti yo‘q madaniy terminlarga qo‘llanadi (shartli misol “ko‘pkari”, ya’ni lotin yozuvida saqlangan shakl va qisqa izoh). Uchinchisi “tarjima qilinadi” qiymati bo‘lib, umumiy ot bilan ifodalangan muassasa va hodisa nomlariga tegishli. O‘zbek tilida atoqli ot bilan turdosh ot omonim bo‘lgan hollar alohida ehtiyotkorlik talab qiladi. Shartli misolda “Lola” ism sifatida transliteratsiya qilinishi, gul nomi sifatida esa tarjima qilinishi lozim, tanlov esa kontekstdagi tugun turiga qarab amalga oshiriladi.

Tarjima sifatini baholashda ham tugun darajasidagi nazorat kerak bo‘ladi. K. Papineni va hammualliflari taklif etgan BLEU metrikasi bir yoki bir nechta etalon tarjima bilan n-gramma moslikning cheklangan aniqligini hisoblaydi va qisqa tarjimalar uchun jarima koeffitsiyentini qo‘llaydi [7]. Metrika butun matn bo‘yicha umumiy ko‘rsatkich beradi, atoqli otdagi bitta xato esa umumiy ko‘rsatkichga sezilarli ta’sir qilmasligi mumkin. Shu sabab tarjima siyosati belgisi bor tugunlar uchun alohida tekshiruv o‘tkazish, ya’ni maqsad matnda qat’iy shakl saqlanganmi yoki yo‘qmi, shuni sanash maqsadga muvofiq.

Xulosa qilib aytganda, qidiruv va tarjima tizimlari ensiklopedik bazadan oladigan foyda bazadagi ma’lumot hajmidan ko‘ra tugunning qanday tavsiflanganiga ko‘proq bog‘liq. Munosabatlarning kengaytirish xavfi darajasi qidiruvda aniqlik bilan to‘liqlik o‘rtasidagi muvozanatni boshqaradi, tarjima siyosati belgisi esa atoqli ot va madaniy terminlarni izchil uzatishga xizmat qiladi. O‘zbek tili uchun bu tamoyil uch lisoniy shart bajarilganda ishlaydi. Matndagi so‘z shakli morfologik tahlil orqali lemmaga keltirilishi, har bir tugunda lotin va kirill yozuvidagi shakllar birga saqlanishi, omonim atoqli otlar esa turdosh otlardan alohida tugun sifatida ajratilishi kerak.

ADABIYOTLAR

1. Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval: The Concepts and Technology behind Search. – 2nd ed. – Harlow: Addison-Wesley, 2011.

2. Hojiyev A. O‘zbek tili sinonimlarining izohli lug‘ati. – Toshkent: O‘qituvchi, 1974.

3. O‘zbek tilining izohli lug‘ati: 5 jildli / A. Madvaliyev tahriri ostida. – Toshkent: O‘zbekiston milliy ensiklopediyasi Davlat ilmiy nashriyoti, 2006–2008.

4. 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.

5. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate // Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). – San Diego, 2015. – 15 p.

6. Forcada M.L., Ginestí-Rosell M., Nordfalk J. et al. Apertium: A Free/Open-Source Platform for Rule-Based Machine Translation // Machine Translation. – 2011. – Vol. 25, No. 2. – P. 127–144.

7. Papineni K., Roukos S., Ward T., Zhu W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. – Philadelphia, 2002. – P. 311–318.