LINGVISTIK ANNOTATSIYA TURLARI VA UNING TIL KORPUSLARIDAGI AHAMIYATI

Xudayarova Sabura Shuxrat qizi
xudayorovasabura@navoiy-uni.uz
ToshDO‘TAU magistranti

Annotatsiya. Til va kompyuter, ushbu tushunchalar bugungi kunda tilshunoslikning zamonaviy yo‘nalishlarini shakillantirishi bilan birga sohada ko‘plab amaliy natijalarga erishmoqda. Ko‘plab til korpuslarining yaratilishi hamda til birliklarining korpuslarda annotatsiyalanishi erishilgan yutuqlar natijasidir. Ushbu maqolada lingvistik annotatsiya va uning turlari hamda bugungi kunga qadar yaratilgan til korpuslari, turli til korpuslarida til birliklarining annotatsiyalanishi haqida fikr yuritilgan.

Abstract. Language and computers, these concepts shape the modern directions of linguistics today and achieve many practical results in the field. The creation of many language corpora and the annotation of language units in corpora are the result of the achievements. This article discusses linguistic annotation and its types, language corpora created to date, annotation of language units in different language corpora.

Аннотация. Язык и компьютеры — эти концепции сегодня формируют современные направления лингвистики и достигают многих практических результатов в этой области. Результатом достижений является создание множества языковых корпусов и аннотирование языковых единиц в корпусах. В данной статье рассматриваются лингвистическая аннотация и ее виды, созданные к настоящему времени языковые корпуса, аннотация языковых единиц в разных языковых корпусах.

Kalit so‘zlar. *Obyekt, modellashtirish, sun’iy intelekt, xalaro korpuslar, NLP, algoritm, matn, token, lemma, annotatsiyalash.*

Shaxsning hujjat va matn ma’lumotlarini ifodalovchi belgilar bilan o‘zaro aloqasi Tabiiy tilni qayta ishlash (NLP) sun’iy intellektdagi eng yirik ishlanmalardan biridir. Chatbotlar, nutqni avtomatik aniqlash va his-tuyg‘ularni tahlil qilish dasturlari kabi ko‘plab NLP yechimlari butun dunyo bo‘ylab son-sanoqsiz biznes samaradorligini va mahsuldorligini oshirmoqda. NLP sohasidagi so‘nggi yutuqlar, hatto nutqida nuqsoni bo‘lgan odamlarga avtomatik nutqni aniqlash qurilmalari va ularning atrofidagi odamlar bilan erkin muloqot qilish imkoniyatini ko‘rsatdi. Biroq, bu ajoyib texnologiyalarning hech biri matn annotatsiyasisiz amalga oshirilmaydi.

NLP algoritmlari uchun izohli matn ma’lumotlarining katta to‘plami talab qilinadi. Matn izohlarining beshta keng tarqalgan turiga qisqacha to‘xtalamiz:

1. Entity Annotation Entity annotatsiyasi chatbot trening ma’lumotlar to‘plami va boshqa NLP o‘quv ma’lumotlarini yaratishda eng muhim jarayonlardan biridir. Bu matndagi obyektlarni aniqlash, ajratib olish va etiketlash harakati.

Obyekt izohi turlariga quyidagilar kiradi:

• Nomlangan obyektni tanib olish (NER): obyektlarning tegishli nomlari bilan izohlash.

• Kalit iboralarni belgilash: matn ma’lumotlaridagi kalit so‘zlar yoki kalit iboralarning joylashuvi va teglari.

• Nutq qismini (POS) teglash: tanib olish va izohlash [Core, Ishizaki, Moore, Nakatani, Reithinger, Traum, Tutiya, 1998: 65]. Nutqning funksional elementlari (sifatlar, otlar, qo‘shimchalar, fe’llar va boshqalar). Shaxsiy izohlar NLP modelini matndagi nutq qismlarini, nomlangan shaxslarni va asosiy iboralarni aniqlashni o‘rgatadi. Ushbu vazifada izohlovchilar matnni diqqat bilan o‘qiydilar, maqsadli obyektlarni topadilar, ularni annotatsiya platformasida ajratib ko‘rsatishadi va oldindan belgilangan ro‘yxatdan teglarni tanlaydilar. NLP modellariga nom berilgan obyektlar haqida ko‘proq ma’lumot olishga yordam berish uchun obyekt izohi ko‘pincha obyektni bog‘lash bilan birlashtiriladi.

2. Obyektni bog‘lash. Obyekt izohi matndagi muayyan obyektlarni tartibga solish va izohlash bo‘lsa, obyektni bog‘lash bu obyektlarni ular haqidagi ma’lumotlarning kattaroq omborlariga ulash jarayonidir.

3. **Matnni tasniflash**, shuningdek, matnni turkumlashtirish yoki hujjat tasnifi sifatida ham tanilgan, matnni tasniflash matnni yoki matnning qisqa satrlarini o‘qish bilan izohlovchi vazifalarni bajaradi. Annotatorlar tarkibni tahlil qilishlari, undagi mavzu, niyat va his-tuyg‘ularni aniqlashlari va oldindan belgilangan toifalar ro‘yxati asosida tasniflashlari kerak. Obyekt annotatsiyasi alohida so‘zlar yoki iboralarni belgilash bo‘lsa, matn tasnifi butun matn yoki matn qatoriga bitta yorliq bilan izoh berish jarayonidir.

4. **Hissiy intellekt** - bu mashinani o‘rganishning eng qiyin sohalaridan biri. Ba’zan hatto odamlar uchun matnli xabar yoki elektron pochta ortidagi haqiqiy tuyg‘uni taxmin qilish qiyin. Mashina uchun istehzoli, aqlli yoki boshqa tasodifiy aloqa shakllaridan foydalanadigan matnlarda yashirin konnotatsiyalarni aniqlash eksponent jihatdan qiyinroq. Mashinani o‘rganish modellariga matn ichidagi his-tuyg‘ularni tushunishga yordam berish uchun modellar his-tuyg‘ularga izohli matn ma’lumotlari bilan o‘rgatiladi. Kengroq ma’noda his-tuyg‘ularni tahlil qilish yoki fikrni tahlil qilish deb ataladi, hissiyot izohi - bu tuyg‘u, fikr yoki his-tuyg‘ularning yorlig‘i. Annotatorlarga tahlil qilish uchun matnlar beriladi va matn ichidagi his-tuyg‘u yoki fikrni qaysi belgi eng yaxshi ifodalashini tanlashi kerak.

5. **Lingvistik annotatsiya**, shuningdek, korpus annotatsiyasi deb ataladi, lingvistik annotatsiya shunchaki matn yoki audio yozuvlardagi til ma’lumotlarini belgilash jarayonini tasvirlaydi. Lingvistik annotatsiya bilan annotatorlarga matn yoki audio ma’lumotlardagi grammatik, semantik yoki fonetik elementlarni aniqlash

va belgilash vazifasi yuklanadi. Lingvistik annotatsiyaning turlariga quyidagilar kiradi [Bunt H, 2010. 29–46]:

• **Diskurs annotatsiyasi**: anafor va kataforalarning o‘z oldingi yoki postsedentlari bilan bog‘lanishi. Nutqning bir qismi (POS) teglari: matndagi turli funksiyali so‘zlarga izoh beradi.

• **Fonetik annotatsiya**: nutqdagi intonatsiya, urg‘u va tabiiy pauzalarni belgilaydi.

• **Semantik annotatsiya**: ta’riflarga izoh beradi. Lingvistik annotatsiya chatbotlar, virtual yordamchilar, qidiruv tizimlari, mashina tarjimasi va boshqalar kabi turli NLP yechimlari uchun sun’iy intellekt bo‘yicha o‘quv ma’lumotlar to‘plamini yaratishda qo‘llaniladi. Bular bugungi kunda mashinani o‘rganishda keng qo‘llaniladigan beshta turdagi matn izohlaridir. Korpusdagi til namunasining xususiyatlari. Bunday izoh tokenizatsiya deb ataladigan dastlabki segmentatsiya jarayonini talab qiladi, bu esa korpusdagi birliklarni belgilaydi – so‘zlar, raqamlar, tinish belgilari va boshqalar. Ba’zi hollarda bu qo‘shimcha qadamni o‘z ichiga oladi.

Obyektni tan olish, bu korpusdagi tegishli birliklarni aniqlashga xizmat qiladi. Lemmalar annotatsiyasi lemmatizatsiyaning eng asosiy turlaridan biri, korpusdagi har bir so‘zni asos (iqtibos yoki lug‘at) shakli bilan belgilash. Lemmatizatsiya mavjud shakl-lemma ma’lumotlar bazasi asosida amalga oshirilishi mumkin, (yarim) avtomatik yondashuv bo‘lib, unda so‘z shakllari kesiladi. Lemmaning umumiy ifodasiga kelish uchun belgilarni kesib tashlash yoki ushbu ikki strategiyaning ba’zi gibrid yondashuvlari, ular morfologik birlikni ham o‘z ichiga olishi mumkin. Sintaktik va morfologik jihatdan Til birliklarini teglash ham shular jumlasidandir. Til birliklarini teglash annotatsiyaning eng tez-tez uchraydigan va eng ko‘p ishlatiladigan turlaridan biridir. Annotatsiya ko‘plab korpus-lingvistik tadqiqotlarga tegishli semantik annotatsiya. Bu har bir tokenlashtirilgan so‘zga yorliq belgilashni o‘z ichiga oladi. Ba’zi grammatik kategoriya ma’lumotlari so‘zning nutq qismini minimal darajada aniqlaydi,

Annotatsiyaning keng tarqalgan turi - nutqning bir qismini teglash korpusda juda ko‘p turli xil annotatsiya turlari va formatlari qo‘llanilishi ajablanarli emas. Tilshunoslikda korpus lingvistikasining ko‘p qismida hali ham nisbatan kam sonli dastur turlari hukmronlik qilmoqda.

Lingvistik annotatsiya til korpuslarining ajralmas qisimidir. Korpusdagi har bir til birliklari lingvistik annotatsiyalangandagina til korpuslarini tashkil etadi. Lingvistik annotatsiyaning ahamiyati til korpuslarida namoyan bo‘ladi. Shunday ekan til korpuslari va uning turlari haqida fikr yuritish o‘rinli bo‘ladi. Korpus lingvistikasi amaliy tilshunoslikning lingvistik korpus (matnlar korpusi) qurish va ulardan foydalanishning umumiy tamoyillarini ishlab chiqish bilan shug‘ullanuvchi bo‘limidir. Turli xil matnlarni tahlil qilish asosida tadqiqotchini qiziqtiradigan lingvistik hodisa, masalan, grammatik tuzilmaning xatti-harakati, tilda ekspressiv vositalardan foydalanish va boshqalar haqida xulosa chiqarish mumkin. Kompyuter texnologiyalarining rivojlanishi ko‘plab matnlarning elektron shaklda paydo

bo‘lishiga yordam berdi. Bunday hajmdagi matnlar bilan ishlash, ulardan kerakli ma’lumotlarni olish uchun butun dunyoda lingvistik korpuslar yaratila boshlandi, ya’ni, maxsus tanlangan, turli lingvistik parametrlar bo‘yicha belgilangan va qidiruv tizimi bilan ta’minlangan matnlar to‘plami. Tadqiqot materiali hajmining ortishi lingvistik ma’lumotlarni tahlil qilishning yangi usullarini, jumladan, ularni statistik qayta ishlashni qo‘llashni taqozo etdi. Demak, korpus lingvistikasi ikki jihatni o‘z ichiga oladi: birinchidan, matn korpusini yaratish va belgilash (annotatsiya) va qidiruv vositalarini ishlab chiqish. ular uchun, ikkinchidan, lingvistikaning o‘zi - korpuslarga asoslangan eksperimental tadqiqotlar. Bu nisbatan yosh va faol rivojlanayotgan, hisoblash tilshunosligi bilan chambarchas bog‘liq va miqdoriy usullardan keng foydalaniladi. Elektron resurslardan foydalanish imkoniyati tilshunoslik tadqiqotlarida material to‘plash jarayonini ancha osonlashtirdi. Biroq, lingvistik ma’lumotlarning bunday mavjudligi lingvistik tadqiqotlarning dalillar bazasiga qo‘yiladigan talablarni tubdan o‘zgartirdi. Korpus lingvistikasi tilshunoslikning matn korpusini ishlab chiqish, yaratish va ishlatish bilan shug‘ullanuvchi bo‘limidir. Bu atama 1960-yillarda korpus yaratish amaliyotining rivojlanishi bilan bog‘liq bo‘lib, 1980-yillardan boshlab kompyuter texnologiyalarining rivojlanishi bilan bog‘liq holda qo‘llanila boshlandi. Korpus lingvistikasi tilni o‘rganadi, chunki bu til uning matn korpusida (ko‘plikda), uning “haqiqiy dunyo” matn korpusida ifodalanadi. Korpus tilshunosligi shuni ko‘rsatadiki, tilni ishonchli tahlil qilish eng kam tajriba aralashuvi bilan ushbu tilning tabiiy konteksti sohasida to‘plangan korpuslardan foydalanish mumkin. Katta matn to‘plamlari tilshunoslarga lingvistik tushunchalarning miqdoriy tahlilini o‘tkazishga imkon beradi. Lingvistik yoki til, matnlar korpusi - bu katta, taqdim etilgan. Mashinada o‘qilishi mumkin bo‘lgan formatda, birlashtirilgan, tuzilgan, belgilangan, aniq lingvistik muammolarni hal qilish uchun mo‘ljallangan filologik jihatdan malakali til ma’lumotlari to‘plami. Zamonaviy korpusning asosiy xususiyatlari - mashinada o‘qiladigan format, reprezentativlik va metalingvistik ma’lumotlarning mavjudligi. Reprezentativlikka matnlarni tanlashning maxsus tartibi yordamida erishiladi. Lingvistik korpus – ma’lum tamoyillarga muvofiq to‘plangan, ma’lum standart bo‘yicha belgilangan va ixtisoslashtirilgan qidiruv tizimi bilan ta’minlangan. Ba’zan korpus (“birinchi tartibli korpus”) oddiygina umumiy xususiyat (til, janr, muallif, matnlarni yaratish davri) bilan birlashtirilgan har qanday matnlar to‘plami deb ataladi. Matn korpusini yaratishning maqsadga muvofiqligi quyidagilar bilan izohlanadi: taqdimot, real kontekstdagi lingvistik ma’lumotlar; ma’lumotlarning yetarlicha katta vakili (korpusning katta hajmi bilan); turli lingvistik muammolarni hal qilish uchun bir marta yaratilgan korpusdan qayta foydalanish imkoniyati, masalan, grafik va leksikani amalga oshirish.

Matnning grammatik tahlili. Birinchi yirik kompyuter korpusi 1960-yillarda Braun universitetida yaratilgan va har biri 2 ming so‘zdan iborat 500 ta matn bo‘laklarini o‘z ichiga olgan, 1961-yilda AQShda ingliz tilida nashr etilgan Brown

Corpus (BC) hisoblanadi. Natijada, u boshqa tillarda vakillik korpusini yaratish uchun 1 million hodisa standartini o‘rnatdi.

1970-yillarda Zasorina rus tilining chastotali lug‘ati yaratilgan bo‘lib, u matnlar korpusi asosida, shuningdek, 1 million so‘z hajmida va taxminan teng nisbatda ijtimoiy-siyosiy matnlarni, badiiy adabiyotlarni o‘z ichiga olgan, turli sohalardagi ilmiy va ilmiy-ommabop matnlar va drama. 1980-yillarda Shvetsiyaning Upsala universitetida yaratilgan rus Korpusi ham xuddi shunday model bo‘yicha qurilgan. Bir million so‘z hajmi faqat eng tez-tez uchraydigan so‘zlarning leksik-orfografik tavsifi uchun yetarli, chunki o‘rtacha chastotali so‘zlar va grammatik tuzilmalar har million so‘zga bir necha marta uchraydi (statistik nuqtayi nazardan, til noyob hodisalarning katta to‘plamidir, shuningdek, katta hajmdagi matnlar bilan ishlashga qodir kompyuter quvvatining o‘sishi tufayli 1980-yillarda butun dunyo bo‘ylab kattaroq korpuslarni yaratishga bir necha bor urinishlar qilingan. Buyuk Britaniyada bunday loyihalar Birmingem universitetidagi ingliz banki va Britaniya milliy korpusi (BNC) edi. Minglab matnlarni to‘plash, mualliflik huquqi muammolarini bartaraf etish, barcha matnlarni yagona shaklga keltirish, korpusni mavzular va janrlar bo‘yicha muvozanatlash juda ko‘p vaqtni oladi.

Nemis, polyak, chex, sloven, fin, zamonaviy yunon, arman, xitoy, yapon, bolgar va boshqa tillar uchun vakillik korpusi mavjud (yoki ishlab chiqilmoqda). Rossiya Fanlar akademiyasida yaratilgan rus tilining milliy korpusi; Hozirda 500 milliondan ortiq so‘z qo‘llanishlarini o‘z ichiga oladi [Core, Ishizaki, Moore, Nakatani, Reithinger, Traum, Tutiya, 1998:79]. Keng janrlar va funksional uslublarni qamrab oluvchi vakillik korpusi bilan bir qatorda lingvistik tadqiqotlar ko‘pincha opportunistik matnlar to‘plamidan, masalan, gazetalardan (ko‘pincha The Wall Street Journal va The New) foydalanadi. Ishning hajmi ham, tuzilishi ham reprezentativlik uchun muhimdir. Vakillik hajmi muammoga bog‘liq, chunki u o‘rganilayotgan hodisalar uchun qancha misol topish mumkinligi bilan belgilanadi.

Matn korpusi usuli har qanday tabiiy tilda yozilgan matnlardan ushbu tilni boshqaradigan mavhum qoidalar to‘plamini olish uchun foydalanadi. Ushbu natijalar ushbu maqsadli til va shunga o‘xshash tahlildan o‘tgan boshqa tillar o‘rtasidagi munosabatni o‘rganish uchun ishlatilishi mumkin. Birinchi bunday korpuslar manba matnlari asosida qo‘lda yaratilgan, ammo hozir bu ish avtomatlashtirilgan. Korpus nafaqat lingvistik tadqiqotlar uchun, balki lug‘atlarni [Carletta, 1996: 249–254] ham qayta tahrir qilish mumkin.

Tilshunoslikda va tabiiy tilni qayta ishlashda korpus yoki matn korpusi raqamli va raqamlilashtirilgan, izohli yoki izohsiz til resurslaridan iborat ma’lumotlar to‘plamidir. Ular korpus lingvistikasida statistik gipotezalarni tekshirish, hodisalarni tekshirish yoki ma’lum bir til hududida lingvistik qoidalarni tasdiqlash uchun ishlatilgan. Qidiruv texnologiyasida korpus bu qidirilayotgan hujjatlar to‘plamidir. Odatda, u manba nomi va u olingan butun matnga giperhavola bilan ta’minlanadi. Korpus menejeri korpus bo‘yicha har xil turdagi statistik ma’lumotlar haqida hisobot berishi mumkin. Masalan, ma’lum bir matn birligidan

foydalanish chastotasi lug‘atini tuzing. Bu ma’lumotlarning barchasi tilshunos tomonidan tilni bir butun sifatida tavsiflashda yoki alohida hodisani o‘rganishda qo‘llaniladi.

Bizning fikrimizcha, korpusning vakili bo‘lishi zaruriy shartdir, bu sizga turli xil matnlar to‘plamiga korpus maqomini berishga imkon beradi. Lingvistik tahlil qilish imkonini beruvchi matnlar ma’lum bir korpusning so‘zsiz vakilidir.

Xulosa qilib aytganda, ko‘plab til korpuslari yaratilgan bo‘lib, har bir tilning ichki imkoniyatlari o‘ziga xos xususiyatlarini aniq belgilab olgan holda uning muayyan bir shaklini ishlab chiqish bilan bir qatorda har bir til birligi lingvistik annotatsiyalanishi til korpusida amalga oshiriladi.

Foydalanilgan adabiyotlar:

1. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics, - 2008. – c 555–596
2. Bernsen, N.O., Dybkjær, L., Kolodnytsky, M.: The NITE Workbench. A Tool for Annotation of Natural Interactivity and Multimodal Data. In: Proceedings of the Third International Conference on Language Resources and Evaluation, - 2002. – C 214
3. Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., Liberman, M.: Atlas: A flexible and extensible architecture for linguistic annotation. In: Proceedings of the Second International Conference. – 2000
4. Bunt, H.: A methodology for designing semantic annotation languages exploiting semanticsyntactic isomorphisms. In: Proceedings of the Second International Conference on Global Interoperability for Language Resources, – 2010. – C 29-46.
5. Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, - 1996, - pp 249–254
6. Church, K.W.: A stochastic parts programm and noun phrase parser for unrestricted text. In: Proceedings of the Second Conference on Applied Natural Language Processing, ANLC ’88, pp.
7. Clear, J.H.: The british national corpus. In: G.P. Landow, P. Delany (eds.) The Digital Word, - 1993. pp. 163–187.
8. Core, M., Ishizaki, M., Moore, J., Nakatani, C., Reithinger, N., Traum, D., Tutiya, S.: The report of the third workshop of the discourse resource initiative. Tech. rep., Chiba University and Kazusa Academia Hall, - 1998.