KORPUS MATNLARINI TEGLASH MUAMMOLARI

Yodgorov Umid Saydilla o‘g‘li
yodgorov@navoiy-uni.uz,
ToshDO‘TAU o‘qituvchisi

Annotatsiya. Korpus matnlarini teglash chatbotlar, virtual yordamchilar, qidiruv tizimlari va mashina tarjimasi kabi tabiiy tilni qayta ishlash (NLP) ilovalarida asosiy amal hisoblanib, u sun’iy intellekt modellariga inson tili va muloqotni yanada samaraliroq tushunishga yordam beradi. NLP ilovalaridagi o‘quv ma’lumotlarini teglashdagi noaniqliklar AI modellarini real jahonda mavjud ilovalarida ishlamay qolishiga olib kelishi mumkin. Gartner hisobotiga ko‘ra, 8% noto‘g‘ri ta’lim ma’lumotlarini kiritish sun’iy intellektning aniqligini 75%ga kamaytirishi mumkin. Korpus matnlarini annotatsiyalash yoki teglash orqali so‘z, ibora, so‘z birikmasi shu kabi leksik birliklarni izohlash kabi matnga teg (izoh) qo‘shish va ularni belgilashdan iborat. Ushbu maqolada til korpusi matnlarini teglash muammolari ko‘rib chiqiladi.

Abstract. Corpus text tagging is a key practice in natural language processing (NLP) applications such as chatbots, virtual assistants, search engines, and machine translation, where it helps AI models understand human language and communication more effectively. Inaccuracies in the labeling of training data in NLP applications can cause AI models to fail in their existing real-time applications. According to a Gartner report, entering 8% of the wrong training data can reduce the accuracy of artificial intelligence by 75%. Annotating or tagging corpus texts consists of adding a tag (explanation) to the text and defining them, such as explaining words, phrases, phrases, and similar lexical units. This article examines the problems of tagging language corpus texts.

Аннотация. Разметка корпуса текста - это ключевая практика в приложениях обработки естественного языка (NLP), таких как чат-боты, виртуальные помощники, поисковые системы и машинный перевод, где они помогают моделям ИИ более эффективно понимать человеческий язык и общение. Неточности в маркировке обучающих данных в приложениях НЛП могут привести к сбою моделей ИИ в существующих приложениях реального времени. Согласно отчету Gartner, ввод 8% неправильных данных обучения может снизить точность искусственного интеллекта на 75%. Аннотирование или маркировка корпусных текстов заключается в добавлении к тексту тега (пояснения) и его определении, например, пояснении слов, словосочетаний, словосочетаний и подобных лексических единиц. В данной статье рассматриваются проблемы разметки текстов языкового корпуса.

Kalit so‘zlar: *Til korpusi, matnni teglash, mashinali o‘qitish modellari, korpus turlari, parallel korpus, teglangan korpuslar, POS teg.*

Kirish

Korpus shakllantirilganidan so‘ng undan NLP vazifalarini bajarishda foydalanish uchun qayta ishlash lozim. Til korpusidan tegishli ma’lumotlarni olish uchun *tizim, vosita, metod* va *dasturlarni* ishlab chiqish talab etiladi. Korpusni qayta ishlash nafaqat asosiy lingvistik tadqiqot va ishlanmalar uchun, balki til texnologiyasi ishlarida ham ajralmas jarayon hisoblanadi.

Bugungi kunda korpusni qayta ishlashda *statistik tahlil, konkordans, leksik kollokatsiya, kalit so‘zlarni qidirish, so‘zlarni lokal guruhlash, lemmalash, morfologik ishlov berish va generatsiya, qismlarga ajratish, so‘zlarni qayta ishlash, so‘z turkumlarini belgilash, izohlash, tahlil qilish* kabi amallar bajariladi [Garside, Leech, Mcenery, 2020; Hovy, Lavid, 2010]. Ba’zi hollarda korpusni qayta ishlash evaziga olingan natijalar til va uning xususiyatlari haqidagi xulosalarga zid kelishi kuzatiladi.

Jahonda keng foydalaniladigan ingliz, fransuz, nemis va shunga o‘xshash tillar uchun korpuslarni qayta ishlovchi ko‘plab dasturlar mavjud. O‘zbek tili korpusi uchun yuqorida keltirilgan amallarning ba’zilari amaliyotga joriy etilgan bo‘lsa-da, samaradorligi qoniqarli emas. O‘zbek tilining tabiatini (o‘ziga xosligini) hisobga olgan holda o‘zbek tili korpusini qayta ishlash vositalarini loyihalash dolzarb. Shu sababli ingliz va jahonning boshqa tillarida faol qo‘llaniladigan ba’zi taniqli korpuslarni qayta ishlash usullari va vositalarini tahlil qilishga ehtiyoj tug‘iladi.

Korpus matnlarini teglash

Korpus matnlarini teglash – gaplarning xususiyatlarini aniqlash uchun matnga yoki uning turli elementlariga **teg**(lar) biriktirish (yozis/belgilash) jarayoni [Gries, Berez, 2017]. Bu jarayonda matn yoki audio yozuvlardagi *so‘z turkumlari (parts of speech, POS), grammatik kategoriya va grammatik ma’no, sintaktik xususiyarlar, kalit so‘zlar, iboralar, kinoya* va boshqa leksik birliklar ajratib olinadi hamda lisoniy ma’lumotlar qo‘shimcha yoki metama’lumotlar bilan belgilanadi [Elov va boshqalar, 2022; Elov va boshqalar, 2023].

Matnni teglash/annotatsiyalashda matn ma’lumotlarini turli xil tabiiy tillarni qayta ishlash vazifalari uchun tushunarli va foydalanishga imkon beradigan **yorliq(metka)lar** yoki **toifalarni belgilash** jarayonidir.

Shuni ta’kidlash kerakki, “matn annotatsiyasi”, “matnni teglash” atamalari ko‘pincha bir-birining o‘rnida ishlatiladi, chunki ular AI/ML modellarini o‘rgatish uchun matn ma’lumotlariga ma’lumot qo‘shishni o‘z ichiga oladi. Biroq matn annotatsiyasi – kengroq tushuncha, matnni teglash esa matn izohidagi o‘ziga xos kichik vazifadir (1-jadval).

1-jadval. Matn annotatsiyasi va matn teglash o‘rtasidagi farq

	Matn annotatsiyasi	matn teglash
Ta’rif	Matnga qo‘shimcha ma’lumot, *metama’lumotlar* yoki kontekst qo‘shishni o‘z ichiga olgan kengroq atama.	Matn ma’lumotlarini turli NLP vazifalari uchun tushunarli va foydalanishga yaroqli/qulay qilish uchun *teg* yoki *toifalarni belgilash* jarayoni.
Maqsad	AI/ML modellarini matnning turli qismlari orasidagi *kontekst, hissiyot* va *munosabatlarni* tushunish, o‘rgatishda yordam beradi.	Modellarni o‘rgatish va baholash uchun etiketli ma’lumotlar talab qilinadigan nazorat ostidagi mashinali o‘rgatish vazifalarida ishlatiladi, bu esa ularga teglangan ma’lumotlar asosida bashorat qilish, ma’lumot olish yoki boshqa NLP vazifalarini bajarish imkonini beradi.
Qo‘llanilishi	Aniqlik yoki qo‘shimcha kontekstni ta’minlash uchun matnga *sharh* yoki *izoh* qo‘shish, *rasm, video* yoki *matn fragmentlari* bilan *havolalar* kabi multimedia elementlarini bog‘lash.	Matnni tasniflash, NERlarni aniqlash, kalit so‘zlarni ajratib olish va hissiyotlarni tahlil qilish.
Metodologiya	Muayyan vazifaga qarab qo‘lda yoki avtomatlashtirilgan vositalar yordamida amalga oshirilishi mumkin.	Qo‘lda yoki avtomatlashtirilgan vositalar yordamida amalga oshirilishi mumkin, ammo bu matn izohidagi ma’lum bir kichik vazifa.

Matn annotatsiyasi xatolarni aniqlash va qidiruv natijalarini yaxshilash uchun turli sohalarda qo‘llaniladi [Elov va boshqalar, 2023; Xusainova, 2023]. Quyida turli sohalarda yaxshiroq natijalarga erishish uchun matn annotatsiyasidan qanday foydalanish keltirilgan:

1. Sog‘liqni saqlash. Matn annotatsiyasi elektron sog‘liqni saqlash yozuvlari (electronic health records, EHR) va tibbiy adabiyotlarni tahlil qilishda ishlatiladi. Bu sog‘liqni saqlash xodimlariga bemor ma’lumotlaridagi *shablon, tendentsiya* va *korrelyatsiyalarni* aniqlashga yordam beradi. Tahlillar shuni ko‘rsatdiki, EHRlarni tahlil qilish uchun tabiiy tilni qayta ishlashdan foydalanish qayta davolashni 71,2% aniqlik bilan bashorat qilishi mumkin.

2. Chakana savdo. Matn annotatsiyasi mijozlarning *fikr-mulohazalari* va *ijtimoiy tarmoq sharhlarini* tahlil qilishda ishlatiladi. Bu esa chakana savdo sotuvchilarga tendensiya, imtiyoz va sohalar samaradorligini aniqlashga imkon beradi. IBM kompaniyasi tomonidan olib borilgan tadqiqot shuni ko‘rsatdiki, chakana savdo sotuvchilarning 62 foizi tahlil va ma’lumotlar tomonidan taqdim etilgan tushunchalar ularga raqobatdosh ustunlik berganini ta’kidlagan.

3. Moliya. Moliya institutlari *firibgarlikni aniqlash, mijozlarga xizmat ko‘rsatishni yaxshilash* va *biznes operatsiyalarni soddalashtirish* uchun mijoz *so‘rovlari, tranzaktsiyalari* va boshqa matnli ma’lumotlarni tahlil qilishda matnli annotatsiyadan foydalanadi. Misol uchun, Bank of America Erica deb nomlangan virtual yordamchini yaratish uchun tabiiy tilni qayta ishlash va bashoratli tahlildan foydalanadi. Bu esa mijozlarga yaqinlashib kelayotgan veksellar yoki tranzaktsiyalar tarixi haqidagi ma’lumotlarni ko‘rishga yordam beradi.

4. Mijozlarni qo‘llab-quvvatlash. Matn annotatsiyasi umumiy muammo, tendentsiya va soha samaradorligini yaxshilashda *chat transkriptlari* va *elektron pochta xabarlari* kabi mijozlarni qo‘llab-quvvatlash xizmatlari faoliyatini tahlil qilish uchun ishlatiladi. Korxonalar mijozlarning his-tuyg‘u va ehtiyojlarini tushunib, o‘z qo‘llab-quvvatlash xizmatlarini yaxshilashi va mijozlar talabiga javob berishi mumkin.

5. Inson resurslari (Human Resources, HR). Matn annotatsiyasi *ishga qabul qilish haqidagi ariza, rezyume* va *xodimlarning fikr-mulohazalarini* tahlil qilishda qo‘llanadi. Bu HR mutaxassislariga shablon, tendentsiya va potentsial muammolarni aniqlashga yordam beradi. Misol uchun, matn annotatsiyasidan ish joyidagi madaniyat yoki boshqaruv muammolari kabi sohalarni aniqlash va ularni hal qilishda tegishli choralarni ko‘rish uchun xodimlarning fikr-mulohazalarini tahlil qilishda amaliy ahamiyat kasb etdi.

Matn annotatsiyasidan foydalangan holda, yuqorida keltirilgan sohalardagi strukturlanmagan ma’lumotlarni qimmatli/zarur ma’lumotga aylantirish mumkin. Bu esa sohada muhim qarorlarni qabul qilish sifatini yaxshilash, mijozlar tajribasi va umumiy biznes natijalar samadorligining oshishiga olib keladi.

Matnni teglash(annotatsiyalash) muammolari

AI va ML kompaniyalari matnni annotatsiyalash jarayonlarida bir nechta qiyinchiliklarga duch kelishadi. Bularga ma’lumotlar *sifatini ta’minlash, katta hajmdagi ma’lumotlar to‘plamini samarali boshqarish, konfedensial ma’lumotlarni himoyalash* va *ma’lumotlar o‘sishi bilan masshtabni kengaytirish* kiradi [Cassidy, Schmidt, 2017; Bi, 2018; Arista, 2022]. Matnni annotatsiyalashga asoslangan NLP ilovalarining samaradorligini oshirish uchun ushbu muammolarini hal qilish juda muhimdir.

1. Noaniqlik. Bu so‘z, so‘z birikmasi yoki gap bir nechta talqinga (ma`noga) ega bo‘lganda sodir bo‘ladi. Bu esa matnni annotatsiyalash jarayonidagi nomuvofiqliklar va xatolarga olib kelishi mumkin.

For example, the phrase “I saw the man with the telescope” can be interpreted in two ways: either the speaker saw a man who had a telescope, or the speaker used a telescope to see the man.

2. Subyektivlik. Bu matnda shaxsiy fikrlar, his-tuyg‘ular yoki baholashlarning mavjudligini anglatadi. Subyektiv tilni izohlash qiyin bo‘lishi mumkin, chunki turli izohlovchilar bir xil matnni turlicha talqin qilishlari mumkin, bu esa annotatsiyalarda nomuvofiqliklarga olib keladi. Masalan, mijozlar sharhlarida fikr-mulohazalarni izohlash turli sharhlovchilarning matnni tushunishi va idrokiga asoslanib, bir xil sharhni *ijobiy*, *neytral* yoki *salbiy* deb belgilashiga olib kelishi mumkin.

3. Kontekstni tushunish. Matndagi so‘zlar va so‘z birikmalarning ma’nosi ko‘pincha ular ishlatilgan kontekstga bog‘liq. Annotatorlar ma’lumotlarni aniq belgilash va tasniflash uchun matnni va parchaning umumiy ma’nosini hisobga olishlari kerak. For example, the word “bank” can refer to a financial institution or the side of a river, depending on the context. Failing to consider the context may lead to incorrect annotations and negatively impact the performance of machine learning models.

4. Tillarning xilma-xilligi. Tabiiy tillarning xilma-xilligi matn annotatsiyasida qiyinchilik tug‘diradi. Chunki annotatorlar turli tillardagi ma’lumotlarni to‘g‘ri belgilash va turkumlashtirish uchun bir nechta tillarni bilishi kerak. Bu, ayniqsa, kam resursli tillar yoki lahjalar bilan ishlashda qiyin bo‘lishi mumkin, chunki kerakli tajribaga ega izohlovchilarni topish qiyin bo‘lishi mumkin. Shuningdek, tillarning xilma-xilligi annotatsiyalash jarayonida nomuvofiqliklarga olib kelishi mumkin, chunki turli annotatorlar ma’lum bir tilda turli darajadagi malakaga ega bo‘lishi mumkin.

5. Masshtablilik. Katta hajmdagi ma’lumotlarni annotatsiyalash jarayoni ko‘p vaqt va resurs talab qiladi. Bu esa matn annotatsiyasida jiddiy muammo tug‘diradi. Tadqiqotlar shuni ko‘rsatadiki, AI loyihasi vaqtining 80% dan ortig‘i ma’lumotlarni boshqarishga, jumladan, ularni *yig‘ish, umumlashtirish, tozalash* va *teglash* uchun sarflanadi. Ma’lumotlar hajmi oshgani sayin, ma’lumotlar annotatorlariga bo‘lgan talab ham oshib boradi, bu esa kompaniyalar uchun o‘zlarining izohlash ishlarini samarali ravishda kengaytirishni qiyinlashtiradi.

6. Narx. Bu kompaniyalar uchun katta to‘siqdir, chunki annotatorlarni yollash va o‘qitish, shuningdek, teglash vositalari va texnologiyalariga sarmoya kiritish qimmatga tushishi mumkin. 2030 yilga kelib ma’lumotlarni teglash bo‘yicha global bozor qiymati 13 milliard dollarga yetishi kutilmoqda. Bu esa yuqori sifatli ma’lumotlarni annotatsiyalash uchun zarur bo‘lgan katta sarmoyani talab qiladi. Annotatsiya xizmatlarining narxi bilan aniq va izchil izohlarga bo‘lgan ehtiyojni muvozanatlash sun'iy intellekt va mashinali o‘qitish bo‘yicha yechimlarni joriy qilmoqchi bo‘lgan kompaniyalar uchun muhim muammo hisoblanadi.

Xulosa

Tabiiy tilni kompyuter yordamida tadqiq qilish va qayta ishlashga asoslangan dasturiy ilovalar(axborot tizimlari)ni ishlab chiqish uchun NLP tizimini mavjud ma’lumotlarni o‘rganishini ta’minlash kerak. Ushbu amalni til korpusi orqali amalga oshirish lozim. Til korpusi elektron shaklda taqdim etiladigan katta hajmli va strukturlangan matnlar to‘plami sifatida qaraladi. Til korpusi tabiiy tilni qayta ishlash tizimining asosidir. NLP bilan bog‘liq lingvistik bilimlar leksik, sintaktik, semantik va pragmatik xususiyatlarni o‘z ichiga oladi. Ushbu korpus matnlarini teglash usullari va muammolari haqida keltirildi. Korpus matnlariga ishlov berishdan oldin unga dastlabki ishlov berish matn tozalash tizimiga oid vazifa sanalib, u *nomuhim so‘zlar, maxsus belgi, kulgich, raqam, tinish belgi, ortiqcha bo‘shliq, elektron pochta manzili, HTML teglari hamda URL* kabi elementlarni olib tashlash orqali matn ma’lumotlarining sifatini oshirishga xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Garside, R. G., Leech, G., & Mcenery, A. M. (2020). Introducing corpus annotation. In *Corpus Annotation*. https://doi.org/10.4324/9781315841366-7
2. Hovy, E., & Lavid, J. (2010). Towards a ‘Science’ of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. International Journal of Translation, 22(1).
3. Gries, S. Th., & Berez, A. L. (2017). Linguistic Annotation in/for Corpus Linguistics. In Handbook of Linguistic Annotation. https://doi.org/10.1007/978-94-024-0881-2_15
4. Elov B., Hamroyeva Sh., Abdullayeva O., Uzakova M. O‘zbek tilida POS tegging masalasi: muammo va takliflar / Uzbekistan: Language and Culture. Applied philology. 2022/2(5). – 51-68-b.
5. Elov B., Hamroyeva Sh., Xudayberganov N., Yodgorov U., Yuldashev A. Pos tagging of Uzbek texts using hidden Markov models (HMM) and Viterbi algorithm. O‘zMU xabarlari. Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti ilmiy jurnali. 2023 yil Maxsys son.
6. Elov, B.B., Hamroyeva, Sh.M., Abdullayeva, O.X., Husainova, Z.Y., Xudoyberganov, N.U. 2023. “Agglutinativ tillar uchun pos teglash va stemming masalasi (turk, uyg‘ur, o‘zbek tillari misolida)”. O‘zbekiston: til va madaniyat 2: 6-39.
7. Xusainova Z.Y. O‘zbek tili milliy korpusi qidiruv tizimini optimallashtirishda lemmatizatsiyadan foydalanish / O‘zbekiston: Til va Madaniyat (Kompyuter lingvistikasi). Toshkent, 2023, №2. – B.20-37.
8. Cassidy, S., & Schmidt, T. (2017). Tools for Multimodal Annotation. In Handbook of Linguistic Annotation. https://doi.org/10.1007/978-94-024-0881-2_7
9. Bi, P. (2018). Handbook of Linguistic Annotation. Journal of Quantitative Linguistics. https://doi.org/10.1080/09296174.2018.1424495

10. Arista, J. M. (2022). TOWARD THE MORPHO-SYNTACTIC ANNOTATION OF AN OLD ENGLISH CORPUS WITH UNIVERSAL DEPENDENCIES. Revista de Linguistica y Lenguas Aplicadas, 17. https://doi.org/10.4995/rlyla.2022.16787