

IV SHO‘BA. MASHINA TARJIMASI

GRAMMATIK TEGGLASHDA MASHINALI O‘QITISH YONDASHUVI

Hamroyeva Shahlo Mirdjanovna,
Filologiya fanlari doktori, professor v.b.
shaxlo.xamrayeva@navoiy-uni.uz
ToshDO‘TAU

Erkinjonova Go‘zal Baxtiyor qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
uzbekistan-24@mail.ru
ToshDO‘TAU

Annotatsiya. Grammatik teglash tabiiy tilni qayta ishlashning (NLP) muhim sohasi hisoblanadi. So‘zlarning gapdagi sintaktik va morfologik rolini aniqlash uchun ishlatiladi. Mashinali o‘qitish yondashuvlari grammatik teglashda an‘anaviy qoidalarga asoslangan usullarga nisbatan sezilarli yaxshilanishlarni ta‘minlamoqda. Ushbu maqolada mashinali o‘qitish asosida grammatik teglash usullari, ularning ishlash tamoyillari va o‘zbek tilida qo‘llanilishi tahlil qilinadi.

Annotation: Grammar tagging is an important field in natural language processing (NLP). It is used to determine the syntactic and morphological roles of words in a sentence. Machine learning approaches provide significant improvements over traditional rule-based methods in grammar tagging. This article analyzes machine learning-based grammar tagging methods, their working principles, and their application in the Uzbek language.

Аннотация: Грамматическая разметка является важной областью обработки естественного языка (NLP). Она используется для определения синтаксических и морфологических ролей слов в предложении. Подходы машинного обучения обеспечивают значительные улучшения по сравнению с традиционными методами, основанными на правилах, в грамматической разметке. В этой статье анализируются методы грамматической разметки на основе машинного обучения, их принципы работы и применение в узбекском языке.

Kalit so‘zlar: *Mashinali o‘qitish, chuqur o‘rganish, grammatik teglash, qoidalarga asoslangan yondashuv, NLP, o‘zbek tili, POS tagging.*

KIRISH

Mashinali o‘qitish (Machine Learning, ML) – bu sun‘iy intellektning (SI) bir bo‘limi hisoblanadi. Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) va til modellari uchun mashinali o‘qitish juda muhim, chunki bu texnologiya tilni tushunish, tahlil qilish va avtomatik tarzda ishlov berish uchun juda zarur bo‘ladi. Mashinali o‘qitish tilni avtomatik tarzda tushunish va qayta ishlashni

ta'minlaydigan turli usullarni rivojlantirishga yordam beradi: Mashinali o'qitishning teglashdagi o'zni shundaki, model o'rgatish jarayonida ma'lum bir til va uning tarkibiy qismlarini taniy oladi. Bu jarayon odatda katta hajmdagi annotatsiyalangan ma'lumotlar bilan amalga oshiriladi. Mashinali o'qitish metodlari, xususan, supervised o'qitish (supervised learning), bu turdagi tahlillarni avtomatlashtirish uchun ishlatiladi.

Jahon kompyuter lingvistikasida grammatik teglash bo'yicha ko'plab tadqiqotlar olib borilgan. Masalan, Penn Treebank korpusi asosida yaratilgan ingliz tili uchun taggerlar, hamda Universal Dependencies loyihasi doirasida turli tillar bo'yicha teglash tizimlari ishlab chiqilgan. So'nggi yillarda BERT, GPT va XLM-R kabi transformer modellaridan foydalangan holda yuqori aniqlikdagi grammatik teglash tizimlari ishlab chiqilmoqda, Bojanowski va boshqalar (2017)[1] tomonidan ishlab chiqilgan fastText modeli turli tillardagi grammatik teglashga qo'llanilishi mumkinligini ko'rsatgan.

O'zbek tilida grammatik teglash bo'yicha tadqiqotlar so'nggi yillarda boshlangan bo'lishiga qaramay, izlanishlar olib borilmoqda. O'zbek tilining aglutinativ tabiati tufayli, klassik HMM va qoidalarga asoslangan yondashuvlar samaradorligi cheklangan. Shuningdek, chuqur o'rganish modellaridan foydalanish bo'yicha ham ayrim eksperimental ishlanmalar mavjud. Xususan, morfologik tahlil va sintaktik teglash bo'yicha ilg'or tadqiqotlar olib borilmoqda. Xususan Sh. Hamrayeva “O'zbek tili morfologik analizatorining lingvistik ta'minoti”[2] dissertatsiyasida grammatik teglash haqida batafsil ma'lumot bergan.

Mashinali o'qitish usullari grammatik teglashni avtomatlashtirishda muhim hisoblanadi. Turli korpuslar matnni tahlil qilish, so'z turkumlarini teglash va sintaktik tahlilni aniqlash imkonini beradi. Ayniqsa, nazoratli o'rganish va chuqur o'rganish usullari bu sohada samarali natijalar bermoqda. Grammatik teglashda statistik modellar, neyron tarmoqlar va qoidalarga asoslangan tizimlar ishlatiladi.

Nazoratli o'rganish (supervised learning) modellari tegli korpuslardan foydalanadi. Ular oldindan belgilangan so'z turkumlari asosida o'qitiladi va yangi matnlarda teg qo'yish uchun ishlatiladi. Bunday modellar orasida Naive Bayes, SVM, CRF va LSTM asosida qurilgan modellar ajralib turadi.

Nazoratsiz o'rganish (unsupervised learning) modellari esa avvaldan teglanmagan matnlar asosida o'z-o'zidan tuzilmalarni aniqlashga harakat qiladi. Bu yondashuvning asosiy kamchiligi u nazoratli modellar darajasida yuqori aniqlikni ta'minlay olmaydi. Ammo, bu modellar yangi til resurslari uchun moslashuvchan yechim bo'lishi mumkin. Chuqur o'rganish (supervised learning) grammatik teglashda katta yutuqlarga erishgan usullardan biridir. Ayniqsa, RNN, LSTM va Transformer modellarining ishlatilishi natijasida yuqori aniqlikga erishish mumkin bo'ldi. Ushbu modellar kontekstni hisobga olish orqali murakkab tuzilmalarda ham

yaxshi natijalar ko'rsatadi. Chuqur o'rganish modellarining samaradorligi katta hajmdagi ma'lumotlar to'plamidan foydalanganda oshadi.

Qoidalarga asoslangan yondashuv sintaktik va morfologik qoidalarga asoslanadi. Ushbu tizimlar lingvist mutaxassislar tomonidan ishlab chiqilgan qoidalarga tayangan holda teglash jarayonini amalga oshiradi. Grammatik qoidalarga asoslangan teglash tizimlari an'anaviy yondashuv bo'lib, turli tillarda ishlatiladi. Ular tilda mavjud grammatik xususiyatlarga asoslanib, so'zlarning rolini aniqlaydi. Masalan, o'zbek tilida ot, fe'l, sifat singari so'z turkumlarini aniq teglash uchun morfologik xususiyatlarga tayanish talab etiladi.

Neyron tarmoqlar sathma-sath o'rganish printsipiga asoslanadi. Har bir qatlam oldingi qatlamdan olingan ma'lumotlarni qayta ishlaydi va chuqurroq xususiyatlarni ajratib oladi. Ayniqsa, konvolyutsion neyron tarmoqlar (CNN) va rekurrent neyron tarmoqlar (RNN) tabiiy tilni qayta ishlash jarayonlarida samarali natijalar beradi. Transformer arxitekturasida esa uzoq masofali bog'liqliklarni o'rganish va tahlil qilishda yuqori samaradorlikni ko'rsatadi.

O'zbek tilidagi grammatik teglash jarayonida Deep Learning modellarining natijalari yuqori bo'lishi mumkin. Bu tilning morfologik xususiyatlari, jumladan, affiksatsiya tizimi va sintaktik strukturalari tahlil qilinib, teg qo'yish jarayoni amalga oshiriladi.

Grammatik teglashning keyingi rivojlanish bosqichlari, ayniqsa, Deep Learning texnologiyalaridan foydalanish bilan bog'liq bo'ladi. Bugungi kunda mashinali o'qitish modellarining samaradorligi faqatgina turli tillarga tegishli murakkab korpuslar orqali yuqori aniqlikni ta'minlash bilan cheklanmaydi, balki o'zbek tilidagi kabi morfologik jihatdan boy tillarda ham yuqori natijalarga erishishga imkon yaratadi. Shu bilan birga, yangi modellar va usullar, masalan, *Transformer* va *BERT* kabi avvalgi metodlardan yanada samaraliroq natijalarga olib kelmoqda.

Transformer arxitekturasida grammatik teglashda yuqori samaradorlikni ko'rsatgan eng ilg'or texnologiyalardan biridir. Ushbu modelda so'zlarning ma'nosi va ularning o'zaro bog'lanishlari davomiy va murakkab kontekstda aniqlanadi. Transformer modelining asosiy afzalliklaridan biri bu uning *uzoq masofali bog'liqliklarni* o'rganish qobiliyatidir. Natijada, grammatik strukturalarni tahlil qilishda samaradorlik oshadi. BERT (Bidirectional Encoder Representations from Transformers) modelining alohida ahamiyati shundaki, u so'zlarni ikki yo'nalishda, ya'ni oldingi va keyingi kontekst asosida o'rganadi. Bu orqali so'zlar orasidagi bog'lanishni yanada aniqroq tahlil qilish imkonini beradi. BERT asosidagi modellar o'zbek tilida grammatik teglashni yanada aniqlik bilan amalga oshirishi mumkin. BERT ing pre-trening qobiliyatlari yordamida model tilning barcha sintaktik va morfologik xususiyatlarini o'rganadi va yangi matnlar uchun moslashadi. Ayniqsa,

o'zbek tilining aglutinativ tabiati tufayli, so'zlarning turli affiksatsiya shakllari va ularning o'zgarishlarini tahlil qilishda BERT modellaridan foydalanish samarali bo'ladi.

Deep Learning ,xususan, *Recurrent Neural Networks (RNN)*, *Long Short-Term Memory (LSTM)* va *Convolutional Neural Networks (CNN)*, grammatik teglashda yana bir qadam oldinga siljish imkonini beradi. Bu modellar, ayniqsa, tilning sintaktik va morfologik strukturalarini chuqurroq o'rganish imkonini beradi, chunki ular uzun muddatli bog'liqliklarni aniqlash va yaxshiroq kontekstni tushunish uchun yaratilgan. *LSTM* va *RNN* kabi modellar so'zlar o'rtasidagi uzluksiz bog'liqliklarni ta'minlashda juda foydali, chunki ular vaqt ketma-ketligini tahlil qilishda kuchli hisoblanadi. Bu modellar, ayniqsa, sintaktik bog'lanishlarni aniqlashda samarali ishlaydi. O'zbek tilidagi so'zlar o'rtasidagi morfologik o'zgarishlarni, affikslar va ularning o'zaro bog'lanishini aniq tasavvur qilishga imkon yaratadi. *CNN* modellarining esa so'zlarning lokal xususiyatlarini aniqlashda samarali ekanligi ma'lum. Bu modellarning yordamida so'zlarning strukturaviy jihatlarini va ularning turkumlarini tez va samarali tarzda aniqlash mumkin.

Grammatik teglash uchun mashinali o'qitish metodlari korpuslarning sifatiga va hajmiga katta bog'liq. *Korpuslar* – bu tilni tahlil qilishda asosiy resursdir, chunki ularning yordamida modellarga tilning sintaktik va morfologik strukturalari o'rgatiladi. O'zbek tilini grammatik teglashini avtomatlashtirishda, tilning barcha affiksatsiya tizimlarini, so'z turkumlarini va ularning o'zaro bog'lanishlarini aniqlash uchun yirik va diversifikatsiyalangan annotatsiyalangan korpuslar kerak bo'ladi.

Fe'l so'z turkumini zamon va shaxs-son qo'shimchalarini aniqlash uchun 15 mingta matnni fe'l so'z turkumi qo'lda grammatik teglandi, bu esa zamon va shaxs-son kategoriyalarning avtomatik aniqlash uchun asos yaratildi. Olingan natijalar mashina tarjimasini, matnni avtomatik tahlil qilish va tabiiy tilni qayta ishlash (NLP) tizimlarida qo'llanilishi mumkin. Ayniqsa, fe'l shakllarining zamon, shaxs-son va nominal shakllardagi taqsimoti avtomatik tarjima algoritmlarida kontekstga mos tarjima variantlarini tanlashda muhim rol o'ynaydi. Keyingi tadqiqotlarda ushbu teglangan matnlar asosida matnlarda ularning chastotasi, bu natijalardan qanday model tuzish mumkinligi ko'rib chiqiladi.

Xulosa

Grammatik teglash tabiiy tilni qayta ishlash (NLP) sohasida muhim yo'nalish bo'lib, mashinali o'qitish usullari bu jarayonni samarali va avtomatik tarzda amalga oshirishga yordam bermoqda. Deep learning va supervised learning usullari grammatik teglashda katta yutuqlarga erishish imkonini yaratdi. Transformer modellarining va boshqa chuqur o'rganish texnologiyalarining qo'llanilishi

grammatik teglashning samaradorligini oshirishda muhim rol o'ynaydi. O'zbek tilida grammatik teglashni yaxshilash uchun zamonaviy modellarni qo'llash, tilning morfologik xususiyatlarini chuqur tahlil qilish va yirik annotatsiyalangan korpuslardan foydalanish zarur. Kelajakda, bu sohadagi tadqiqotlar davom etib, yanada aniqroq va samarali grammatik teglash tizimlarini ishlab chiqish imkonini beradi.

Foydalanilgan adabiyotlar:

1. **Bojanowski, P., Grave, E., Mikolov, T., & Joulin, A.** (2017). Enriching Word Vectors with Subword Information. arXiv preprint arXiv:1607.04606.
2. Хамроева Ш. Ўзбек тили морфологик анализаторининг лингвистик таъминоти: Филол. фан. док. ...дис. – Тошкент, 2021. – Б. 150.
3. **Collobert, R., & Weston, J.** (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (*ICML '08*). ACM.
4. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I.** (2017). Attention is all you need. In Advances in neural information processing systems (NeurIPS 2017).
5. **Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
6. Elova D. Fe'l grammatik shakllarining uslubiy xoslanishiga ko'ra lingvistik annotatsiyalash masalasi xususida.
7. Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Бухоро, 2018. – Б. 3.
8. **Abdulkarimov, S., & Tashkent, A.** (2017). O'zbek tilida morfologik tahlil va sintaktik teglash. O'zbekiston tilshunoslik va informatika jurnali, 5(1), 120-134.
9. **Shodmonov, M.** (2020). O'zbek tili va uning grammatik tizimi: Mashinali o'qitish va tillarni qayta ishlash. Toshkent, O'zbekiston.
10. **O'rinov, S., & Saidov, T.** (2021). Chuqur o'rganish va grammatik teglash: O'zbek tilidagi yondashuvlar. *Informatika va texnologiyalar jurnali*, 7(3), 202-217.
11. https://www.researchgate.net/publication/380823282_O'ZBEK_TILI_KO_RPUSINI_SINTAKTIK_TEGSLASH_MASALASI
12. https://www.researchgate.net/publication/375912394_Korpus_birliklarini_teglash_va_annotatsiyalash