JAHON TILLARIDA TINISH BELGILARINI TIKLASHGA OID TADQIQOTLAR TAHLILI

Elov Botir Boltayevich,
Texnika fanlari doktori (DSc), dotsent
elov@navoiy-uni.uz
ToshDO‘TAU
Sobirova Zarnigor G‘anijon qizi,
Tayanch doktorant (PhD)
sobirovazarnigor1996@gmail.com
ToshDO‘TAU

Annotatsiya. Mazkur maqolada jahon tillarida tinish belgilarini avtomatik tiklashga oid ilmiy yondashuvlarning rivojlanish tendensiyalari tahlil qilinadi. Tadqiqot davomida punktuatsiya tiklash masalasi faqat grammatik belgilarni joylashtirish jarayoni emas, balki matnning yashirin semantik va pragmatik chegaralarini qayta tiklovchi intellektual mexanizm sifatida talqin qilinadi. Maqolada qoidaviy tizimlardan boshlab transformer va multimodal sun’iy intellekt modellarigacha bo‘lgan evolyutsion bosqichlar o‘rganilib, ularning lingvistik imkoniyatlari hamda cheklovlari qiyosiy tahlil qilinadi. Shuningdek, o‘zbek tilining agglyutinativ tabiati, erkin so‘z tartibi va morfologik murakkabligi punktuatsiya tiklash tizimlari uchun alohida ilmiy muammo ekani asoslanadi. Tadqiqotda punktuatsiyani “semantik navigatsiya vositasi” sifatida talqin qiluvchi yangi konseptual qarash ilgari surilib, kelajakdagi NLP tizimlarida prosodik signal, diskurs tahlili va adaptiv sun’iy intellekt integratsiyasining ahamiyati yoritiladi.

Kalit so‘zlar: *tinish belgilarini tiklash, tabiiy tilni qayta ishlash, transformer modellar, avtomatik nutqni tanish, multimodal sun’iy intellekt, semantik rekonstruksiya, diskurs tahlili, o‘zbek tilini qayta ishlash.*

Abstract. This article analyzes the developmental trends of scientific approaches to automatic punctuation restoration across world languages. The study interprets punctuation restoration not merely as a process of inserting grammatical symbols, but as an intelligent mechanism for reconstructing the hidden semantic and

pragmatic boundaries of a text. The paper examines the evolutionary stages of punctuation restoration systems, ranging from rule-based methods to transformer and multimodal artificial intelligence models, and comparatively evaluates their linguistic capabilities and limitations. In addition, the study demonstrates that the agglutinative nature, flexible word order, and morphological complexity of the Uzbek language constitute a distinct scientific challenge for punctuation restoration systems. The article also proposes a new conceptual perspective that interprets punctuation as a “semantic navigation tool” and highlights the importance of integrating prosodic signals, discourse analysis, and adaptive artificial intelligence into future NLP systems.

Keywords: *punctuation restoration, natural language processing, transformer models, automatic speech recognition, multimodal artificial intelligence, semantic reconstruction, discourse analysis, Uzbek language processing.*

Kirish

Raqamli axborot almashinuvi kengayib borayotgan hozirgi davrda inson nutqini avtomatik qayta ishlash, og‘zaki ma’lumotlarni matnga aylantirish, real vaqt rejimida subtitr yaratish, ovozli yordamchilar bilan muloqot qilish va katta hajmdagi matnlarni tahlil qilish tabiiy tilni qayta ishlash sohasining asosiy yo‘nalishlaridan biriga aylandi. Bunday tizimlarda matnning faqat so‘zlardan iborat bo‘lishi yetarli emas, chunki inson tafakkuri matnni ma’lum sintaktik, semantik va ohangiy bo‘laklarga ajratgan holda qabul qiladi. Ana shu bo‘laklarning yozma nutqdagi asosiy ko‘rsatkichlaridan biri tinish belgilaridir.

Tinish belgilarini avtomatik tiklash masalasi, ayniqsa, avtomatik nutqni tanish tizimlari uchun muhim hisoblanadi. Chunki ko‘plab ASR tizimlari og‘zaki nutqni matnga aylantirganda so‘zlar ketma-ketligini beradi, biroq gap chegaralari, vergul, nuqta, so‘roq yoki undov belgilarini har doim ham to‘g‘ri tiklay olmaydi. Natijada

hosil bo‘lgan matn grammatik jihatdan to‘liq ko‘rinmaydi, o‘qilishi qiyinlashadi va mazmunni noto‘g‘ri talqin qilish ehtimoli ortadi. Masalan, bir xil so‘zlar ketma-ketligi tinish belgilariga qarab buyruq, so‘roq, izoh yoki xabar mazmunini berishi mumkin.

Mazkur muammo faqat matnni chiroyli ko‘rsatish bilan cheklanmaydi. Punktuatsiyasiz yoki noto‘g‘ri punktuatsiyalangan matn keyingi avtomatik tahlil jarayonlariga ham bevosita ta’sir qiladi. Mashina tarjimasi, matnni qisqartirish, savol-javob tizimlari, hissiyot tahlili, nomlangan obyektlarni aniqlash va sintaktik tahlil kabi vazifalarda gap chegaralarining noto‘g‘ri belgilanilishi umumiy natija sifatini pasaytiradi. Shu sababli tinish belgilarini tiklash bugungi kunda tabiiy tilni qayta ishlashning yordamchi bosqichi emas, balki matnning semantik yaxlitligini ta’minlovchi muhim intellektual komponent sifatida qaralmoqda. So‘nggi yigirma yil ichida tinish belgilarini tiklash bo‘yicha tadqiqotlar sezilarli darajada rivojlandi. Dastlab qoidaviy va statistik modellar yetakchi bo‘lgan bo‘lsa, keyinchalik recurrent neural network, LSTM, BiLSTM, CRF, transformer, ko‘p tilli va end-to-end arxitekturalar paydo bo‘ldi. Hozirda bu yo‘nalish faqat ASR post-processing muammosi sifatida emas, balki til tizimlarining strukturaviy sifatini yaxshilovchi alohida ilmiy va amaliy yo‘nalish sifatida qaralmoqda. [5:1-2]

Ushbu maqolaning maqsadi jahon tillarida tinish belgilarini tiklash bo‘yicha olib borilgan tadqiqotlarni bosqichma-bosqich tahlil qilish, asosiy metodlarni solishtirish, tillar kesimidagi tajribalarni umumlashtirish va istiqboldagi ilmiy yo‘nalishlarni ko‘rsatib berishdan iborat. [6:1-3]

1. Tinish belgilarini tiklash masalasining nazariy mohiyati

Tinish belgilarini tiklash masalasi tabiiy tilni qayta ishlashda matnning yo‘qolgan yoki noto‘g‘ri qo‘yilgan punktuatsion belgilarini avtomatik tarzda aniqlash va ularni tegishli joylarga joylashtirish jarayoni hisoblanadi. Ushbu vazifa tashqi tomondan oddiy ko‘ringani bilan, aslida u matnning sintaktik tuzilishi,

semantik mazmuni, nutq ohangi va kommunikativ maqsadini chuqur tahlil qilishni talab qiladi. Chunki tinish belgisi faqat yozuvdagi yordamchi belgi emas, balki matnning ichki mantiqiy tuzilishini ko‘rsatuvchi muhim lingvistik signal vazifasini bajaradi. Nazariy jihatdan tinish belgilarini tiklash matndagi tokenlar ketma-ketligini tahlil qilish orqali har bir so‘zdan keyin qanday belgi qo‘yilishi kerakligini aniqlashga asoslanadi. Masalan, model har bir so‘zdan keyin “belgi yo‘q”, “vergul”, “nuqta”, “so‘roq belgisi” yoki “undov belgisi” kabi sinflardan birini tanlaydi. [7:2-4]

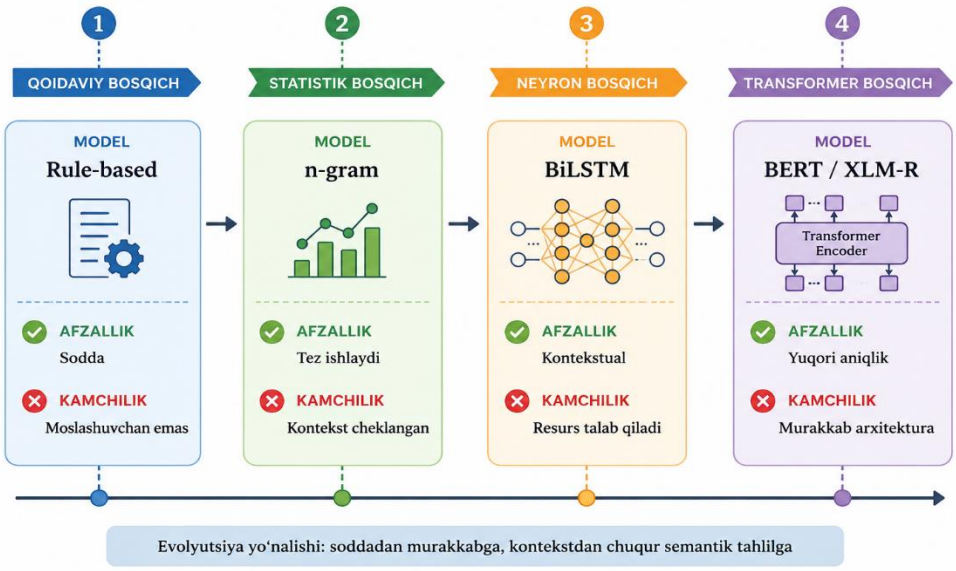
Shu sababli bu masala ko‘pincha ketma-ketlikni belgilash, ya’ni sequence labeling vazifasi sifatida qaraladi. Bunda model faqat alohida so‘zga emas, balki undan oldingi va keyingi kontekstga ham tayanadi. Tinish belgilarining asosiy nazariy vazifasi gap va fikr chegaralarini belgilashdan iborat. Nuqta odatda tugallangan fikrni bildiradi, vergul gap ichidagi semantik va sintaktik bo‘laklarni ajratadi, so‘roq belgisi savol mazmunini ko‘rsatadi, undov belgisi esa emotsional yoki ta’sirchan nutqni ifodalaydi. Demak, punktuatsiyani tiklashda model faqat grammatik shaklni emas, balki matndagi ma’no oqimini ham tushunishi kerak.

Avtomatik nutqni tanish tizimlarida bu masala yanada muhim ahamiyat kasb etadi. Chunki og‘zaki nutq matnga aylantirilganda ko‘pincha tinish belgilarisiz, bosh harflarsiz va gap chegaralarisiz transkript hosil bo‘ladi. Bunday matnni o‘qish qiyin bo‘ladi, ayrim hollarda esa mazmun butunlay noto‘g‘ri talqin qilinishi mumkin. Masalan, “keling kutamiz” va “keling, kutamiz” jumlalari shaklan yaqin bo‘lsa-da, vergul ularning kommunikativ tuzilishini aniqroq ko‘rsatadi. Punktuatsiya tiklashning nazariy mohiyatida uchta asosiy qatlam muhim hisoblanadi. Birinchisi – sintaktik qatlam. Bu qatlam gap bo‘laklari, qo‘shma gaplar, uyushiq bo‘laklar va ergash gaplar chegarasini aniqlash bilan bog‘liq. Ikkinchisi – semantik qatlam. Bunda matndagi fikr tugallanishi, ma’no bo‘laklari va mazmuniy aloqalar

aniqlanadi. Uchinchisi – pragmatik qatlam. Bu qatlam muallifning kommunikativ niyati, savol, buyruq, ta’kid yoki emotsional munosabatini ifodalashga xizmat qiladi.

O‘zbek tili uchun tinish belgilarini tiklashning nazariy mohiyati yanada murakkabroqdir. Chunki o‘zbek tili agglyutinativ til bo‘lib, bitta so‘z tarkibida bir nechta grammatik ma’no ifodalanishi mumkin. Masalan, “kelmaganligingizni” kabi birlikda inkor, egalik, otlashish va kelishik ma’nolari birlashgan. Bu holat modeldan faqat so‘z ketma-ketligini emas, balki morfologik tuzilmani ham tahlil qilishni talab qiladi. Shuningdek, o‘zbek tilida so‘z tartibining nisbatan erkinligi ham punktuatsiya tiklashni murakkablashtiradi. Bir fikr turli shakllarda ifodalanishi mumkin, lekin tinish belgilarining joylashuvi mazmunga qarab o‘zgaradi. Umuman olganda, tinish belgilarini tiklash masalasining nazariy mohiyati matndagi yo‘qolgan belgilarni mexanik tarzda qo‘yishdan iborat emas. Bu jarayon matnning yashirin sintaktik, semantik va pragmatik chegaralarini aniqlash, fikrlar orasidagi munosabatni tiklash hamda yozma nutqni inson tomonidan oson qabul qilinadigan shaklga keltirishdan iborat murakkab intellektual vazifadir.

1-jadval. Punktuatsiya tizimlari evolyutsiyasi



2. Til kesimidagi tadqiqotlar

Ingliz tili punktuatsiyani tiklash bo‘yicha eng ko‘p o‘rganilgan tillardan biridir. Bunga katta annotatsiyalangan korpuslar, kuchli ASR platformalari va boy ilmiy ekotizim sabab bo‘lgan. Ingliz tili ko‘pincha yangi arxitekturalarni sinash uchun benchmark sifatida ishlatiladi. [9:1-2]

Portugal tilida dastlab broadcast news korpuslari asosida statistik va klassifikatsion usullar qo‘llanilgan bo‘lsa, keyinchalik transformer va explainable AI yondashuvlari paydo bo‘ldi. Bu esa punktuatsiya tiklash nafaqat natija, balki model qarorlarini izohlash nuqtayi nazaridan ham tadqiq qilinayotganini ko‘rsatadi.

Ispan tilida customer support transkriptlari, live transcription va ta’lim bilan bog‘liq sohalarda punktuatsiya tiklash bo‘yicha maxsus tadqiqotlar mavjud. Perez-Enriquez va hammualliflarning live transcription uchun ishlab chiqilgan modeli bu tilning real vaqtli amaliy tizimlarida ham alohida o‘rni borligini ko‘rsatdi. Italyan tilida transformer modellarni baholashga qaratilgan tajribalar BERT oilasining yuqori samaradorligini ko‘rsatdi. Cross-domain baholashlar esa modelning bir domen korpusida o‘qitilib, boshqasida sinovdan o‘tkazilganda sifat qanday o‘zgarishini tahlil qilish imkonini berdi.

Xulosalar

Jahon tajribasi o‘zbek tili uchun bir nechta muhim metodologik xulosalarni beradi. Birinchidan, samarali punktuatsiya tiklash tizimi yaratish uchun sifatli annotatsiyalangan korpus zarur. Bu korpus yozma matn, ASR transkriptlari va turli domenlarga oid namunalarni o‘z ichiga olishi kerak.

Ikkinchidan, punktuatsiya tiklashni bosh harflarni tiklash, gap segmentatsiyasi va hattoki POS tagging kabi vazifalar bilan birgalikda o‘rganish maqsadga muvofiq. Chunki jahon tajribasida punctuation va truecasing ko‘pincha birgalikda o‘rganiladi.

Uchinchidan, o‘zbek tilida punktuatsiya tiklash tizimini faqat yakuniy matn ko‘rinishi uchun emas, balki machine translation, speech processing, summarization

va linguistic analysis tizimlari sifatini oshiruvchi strukturaviy modul sifatida baholash kerak. Bu kelajakdagi amaliy va ilmiy tadqiqotlar uchun asos bo‘la oladi.

Foydalanilgan adabiyotlar ro‘yxati

1. Matusov E., Mauser A., Ney H. Automatic sentence segmentation and punctuation prediction for spoken language translation. IWSLT, 2006. URL: https://aclanthology.org/2006.iwslt-papers.1.pdf

2. Gravano A., Jansche M., Bacchiani M. Restoring punctuation and capitalization in transcribed speech. ICASSP, 2009. URL: https://www.cs.columbia.edu/nlp/papers/2009/gravano_al_09.pdf

3. Batista F., Mamede N., Trancoso I., Nunes M. Automatic recovery of punctuation marks and capitalization in spoken broadcast news. SLTECH, 2009. URL: https://www.isca-archive.org/sltech_2009/batista09_sltech.pdf

4. Courtland M., et al. Efficient automatic punctuation restoration using transformers. IWSLT, 2020. URL: https://aclanthology.org/2020.iwslt-1.33/

5. Peitz S., et al. Modeling punctuation prediction as machine translation. IWSLT, 2011. URL: https://aclanthology.org/2011.iwslt-papers.7.pdf [5:1–2]

6. Tilk O., Alumäe T. LSTM for punctuation restoration in speech transcripts. Interspeech,2015.URL:https://www.iscaarchive.org/interspeech_2015/tilk15_interspeech.pdf [6:1–3]

7. Alam T., Khan A. I., Khan F. Punctuation restoration using transformer models for high-and low-resource languages. W-NUT, 2020. URL: https://aclanthology.org/2020.wnut-1.18/ [7:2–4]

8. Courtland M., et al. Efficient automatic punctuation restoration using transformers. IWSLT, 2020. URL: https://aclanthology.org/2020.iwslt-1.33/

9. Chordia V. PunKtuator: A multilingual punctuation restoration system for spoken and written text. EACL Demos, 2021. URL: https://aclanthology.org/2021.eacl-demos.37/ [9:1–2]

10. Păiș V., Tufiș D. Capitalization and punctuation restoration: A survey. 2021. URL: https://arxiv.org/pdf/2111.10746

11. Miaschi A., et al. Evaluating transformer models for punctuation restoration in Italian. CEUR Workshop Proceedings, 2021. URL: https://ceur-ws.org/Vol-3015/paper156.pdf

12. Lai V. D., et al. A punctuation restoration dataset for livestreaming video transcript text. Findings of NAACL, 2022. URL: https://aclanthology.org/2022.findings-naacl.149/