ILMIY MATNLARNI AVTOMATIK QISQACHA XULOSALASH

Sulaymonova Sug‘diyona Zafarovna
Kompyuter lingvistikasi yo‘nalishi 2-kurs talabasi
sulaymonovasugdiyona7@gmail.com
ToshDO‘TAU

Annotatsiya. Ilmiy-tadqiqot faoliyatida chop etilayotgan maqolalar sonining tobora ortib borishi tadqiqotchi va mutaxassislar oldida ulkan hajmdagi matnli manbalarni tezkor tahlil qilish muammosini keltirib chiqarmoqda. Bitta o‘rtacha ilmiy maqolani chuqur o‘qib chiqish va undan zarur ma’lumotni ajratib olish ancha vaqt talab etadi, ayniqsa bir necha o‘nlab maqola bilan bir vaqtda ishlash lozim bo‘lganda bu jarayon yanada murakkablashadi. Ushbu muammoning yechimi sifatida tabiiy tilni qayta ishlash (NLP) sohasida rivojlanayotgan avtomatik matn xulosalash (summarization) texnologiyalari ko‘rib chiqiladi. Maqolada xulosalashning ekstraktiv, abstraktiv va gibrid turlari, ularning statistik usullardan tortib chuqur o‘rganish, oldindan o‘qitilgan til modellari (BERT, BART, PEGASUS) va zamonaviy yirik til modellari (LLM) davrigacha bo‘lgan rivojlanish bosqichlari tahlil qilinadi. Shuningdek, ROUGE, BERTScore, faktual moslik kabi baholash mezonlari, mavjud korpuslar hamda o‘zbek tili uchun xulosalash tizimlarini yaratishdagi lingvistik qiyinchiliklar (agglyutinativlik, morfologik xilma-xillik) muhokama qilinadi. Tahlil natijasida ilmiy matnlarni avtomatik xulosalash tizimlarini yaratish va takomillashtirish nafaqat xorijiy, balki o‘zbek tili uchun ham dolzarb ilmiy-amaliy vazifa ekanligi asoslanadi.

Kalit so‘zlar: matn xulosalash, summarization, ekstraktiv referatlash, abstraktiv referatlash, tabiiy tilni qayta ishlash, ROUGE, yirik til modellari, o‘zbek tili.

Abstract. The rapid growth in the number of published scientific papers creates a significant challenge for researchers who need to quickly analyze large volumes of textual material. Reading a single average-length scientific article

thoroughly and extracting the necessary information takes considerable time, and this task becomes even more demanding when dozens of articles must be processed simultaneously. Automatic text summarization, a rapidly developing branch of natural language processing (NLP), is considered as a solution to this problem. The article reviews extractive, abstractive and hybrid summarization approaches, tracing their evolution from statistical methods through deep learning and pre-trained language models (BERT, BART, PEGASUS) to the current era of large language models (LLMs). Evaluation metrics such as ROUGE, BERTScore and factual consistency measures, widely used datasets, and the specific linguistic challenges of building summarization systems for the agglutinative and morphologically rich Uzbek language are also discussed. The analysis shows that developing and improving automatic summarization systems for scientific texts is an urgent scientific and practical task both internationally and for the Uzbek language.

Keywords: text summarization, extractive summarization, abstractive summarization, natural language processing, ROUGE, large language models, Uzbek language.

Kirish

Fan-texnika taraqqiyoti bilan bir qatorda ilmiy nashrlar hajmi ham yildan-yilga sezilarli darajada ortib bormoqda. Har kuni turli sohalarda minglab yangi ilmiy maqolalar, konferensiya materiallari va monografiyalar chop etilmoqda. Bunday sharoitda tadqiqotchi, olim yoki mutaxassis o‘z sohasidagi eng so‘nggi ilmiy yutuqlardan xabardor bo‘lish uchun katta hajmdagi matnlarni o‘qib chiqishga majbur bo‘ladi. Amaliyotda tez-tez uchraydigan holat quyidagicha tasavvur qilinishi mumkin: professor yangi ilmiy maqolani qo‘lga kiritadi, biroq u yigirma sahifadan iborat bo‘lib, uni to‘liq o‘qib chiqish uchun bir necha soat vaqt talab etiladi, natijada maqolaning barcha qismini diqqat bilan o‘rganish amaliy jihatdan qiyinlashadi.

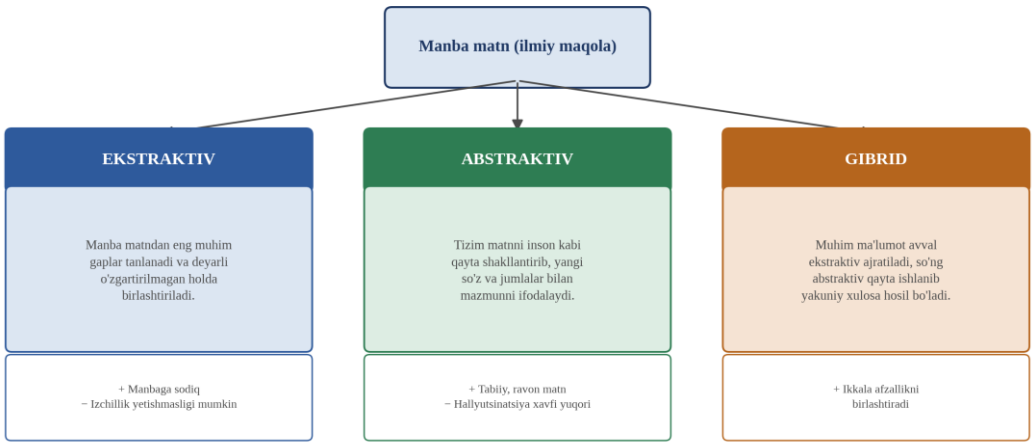
Ana shunday vaziyatda avtomatik matn xulosalash (summarization) tizimlari muhim yechim bo‘lib xizmat qiladi. Bunday tizimlar maqolani avtomatik tarzda o‘rganib chiqib, undan eng muhim mazmunni, ya’ni asosiy g‘oyani, kalit so‘zlarni, foydalanilgan adabiyotlarga oid ma’lumotlarni va boshqa zarur qismlarni ajratib beradi. Natijada foydalanuvchi to‘liq matnni o‘qimasdan turib, uning mohiyati bilan tanishish, ilmiy maqolaning dolzarbligini, metodologiyasini va asosiy xulosalarini qisqa vaqt ichida baholash imkoniyatiga ega bo‘ladi. Bu, o‘z navbatida, ilmiy adabiyotlarni ko‘rib chiqish (literature review) jarayonini tezlashtiradi, tadqiqotchilarning vaqtini tejaydi hamda ilmiy bilimlarning tarqalish sur’atini oshiradi.

Ushbu maqolaning maqsadi ilmiy matnlarni avtomatik xulosalash muammosini kompleks tarzda yoritish, mavjud metodlarning rivojlanish bosqichlarini, ularni baholash mezonlarini hamda mazkur sohadagi ochiq muammolarni tahlil qilishdan iborat. Shuningdek, maqolada o‘zbek tili uchun xulosalash tizimlarini yaratishning o‘ziga xos lingvistik qiyinchiliklari alohida ko‘rib chiqiladi, chunki agglyutinativ til sifatida o‘zbek tili so‘z shakllarining boyligi va sinonimiya kabi xususiyatlari bilan ajralib turadi.

Matn xulosalashning tushunchasi va turlari

Matn xulosalash (text summarization) – bu manba matn (yoki matnlar to‘plami)dan eng muhim axborotni ajratib olib, uni qisqa va izchil shaklda taqdim etish jarayoni sifatida ta’riflanadi. Xulosalash tizimlari kirish manbasining formatiga qarab bir hujjatli (single-document), ko‘p hujjatli (multi-document) va so‘rovga yo‘naltirilgan (query-focused) turlarga bo‘linadi. Bir hujjatli xulosalash bitta maqola yoki hujjat mazmunini qisqartirsa, ko‘p hujjatli xulosalash bir mavzudagi bir nechta manbalarni umumlashtiradi, so‘rovga yo‘naltirilgan xulosalash esa foydalanuvchi bergan aniq so‘rov yoki mavzuga mos javob shakllantiradi.

Chiqish natijasi shakliga ko‘ra xulosalash usullari ekstraktiv, abstraktiv va gibrid turlarga ajratiladi. Ekstraktiv usulda tizim manba matndan eng muhim gaplarni tanlab, ularni deyarli o‘zgartirmagan holda birlashtiradi; bunday xulosalar mazmunan asl matnga sodiq bo‘ladi, biroq ba’zan izchillik va uslubiy yaxlitlik yetishmasligi mumkin. Abstraktiv usulda esa tizim inson kabi matnni qayta shakllantirib, yangi so‘z va jumlalar bilan mazmunni ifodalaydi; bu yondashuv tabiiyroq matn hosil qilsa-da, faktik xatolar (hallyutsinatsiya) xavfi ko‘proq uchraydi. Gibrid usullar dastlab muhim ma’lumotni ekstraktiv tarzda ajratib, so‘ngra uni abstraktiv qayta ishlash orqali yakuniy xulosani shakllantiradi, bu esa ikkala yondashuvning afzalliklarini birlashtirishga xizmat qiladi. Ushbu uch yondashuvning umumlashtirilgan sxemasi 1-rasmda keltirilgan.



1-rasm. Matn xulosalashning asosiy turlari: ekstraktiv, abstraktiv va gibrid yondashuvlar

Xulosalash metodlarining rivojlanish bosqichlari

So‘nggi tadqiqotlarda matn xulosalash sohasining rivojlanishi to‘rtta asosiy bosqichga bo‘lib tahlil qilinadi: statistik bosqich, chuqur o‘rganish bosqichi, oldindan o‘qitilgan til modellari (PLM) bosqichi va zamonaviy yirik til modellari (LLM) bosqichi. Har bir bosqich o‘ziga xos texnologik yondashuv va imkoniyatlar bilan tavsiflanadi.

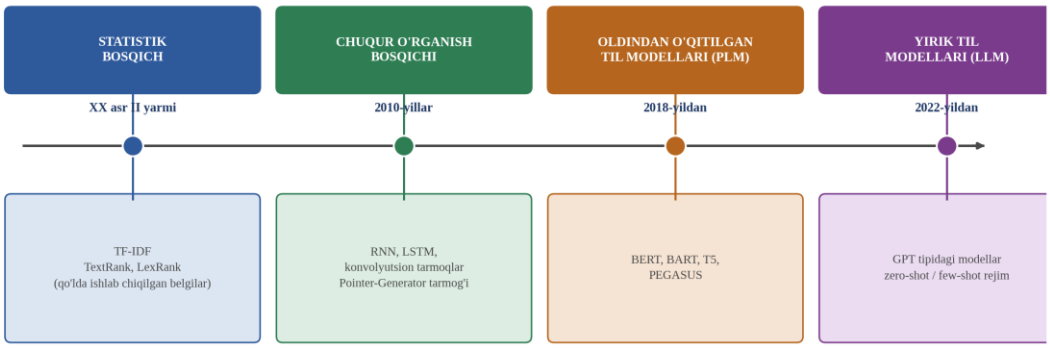
Statistik bosqich XX asrning ikkinchi yarmida boshlangan bo‘lib, unda gaplarning muhimligi so‘z chastotasi, gap uzunligi, matndagi joylashuvi kabi qo‘lda ishlab chiqilgan (hand-crafted) belgilar asosida baholangan. Ushbu davrning eng mashhur usullaridan biri TF-IDF (Term Frequency – Inverse Document Frequency) ko‘rsatkichi bo‘lib, u hujjat ichida ko‘p uchraydigan, ammo boshqa hujjatlarda kamdan-kam qayd etiladigan so‘zlarga yuqori vazn beradi. Grafik asosidagi TextRank va LexRank algoritmlari esa gaplar orasidagi o‘xshashlikni graf ko‘rinishida ifodalab, eng markaziy gaplarni ajratib olishga asoslangan.

2010-yillarda chuqur neyron tarmoqlar (rekurrent tarmoqlar, LSTM, konvolyutsion tarmoqlar) yordamida nazorat ostidagi (supervised) o‘qitish usullari keng qo‘llanila boshlandi. Bu bosqichda matn-xulosa juftliklaridan iborat katta hajmdagi ma’lumotlar to‘plamlaridan foydalanib, tizimlar avtomatik ravishda muhim gaplarni tanlash yoki yangi matn generatsiya qilishni o‘rganishga qodir bo‘ldi. Ayniqsa pointer-generator tarmog‘i kabi arxitekturalar manba matndan so‘zlarni to‘g‘ridan-to‘g‘ri “nusxalash” imkoniyatini berib, kamdan-kam uchraydigan atamalarni to‘g‘ri ifodalash muammosini qisman hal etdi.

2018-yilda BERT modelining paydo bo‘lishi bilan oldindan o‘qitilgan til modellari (PLM) davri boshlandi. BERT kabi kodlovchi (encoder) modellar ekstraktiv xulosalashda, BART va T5 kabi kodlovchi-dekodlovchi (encoder-decoder) arxitekturalar esa abstraktiv xulosalashda samarali qo‘llanildi. Xususan, xulosalash uchun maxsus mo‘ljallangan PEGASUS modeli, matndagi muhim gaplarni niqoblab, ularni qayta generatsiya qilishga o‘rgatilgan bo‘lib, ko‘plab benchmark ma’lumotlar to‘plamida yuqori natijalarni namoyish etdi.

2022-yildan boshlab GPT tipidagi yirik til modellari (LLM) sohaga sezilarli o‘zgarish olib keldi. Endilikda xulosalash vazifasi ko‘pincha maxsus o‘qitishni talab qilmaydigan nol-shot (zero-shot) yoki bir necha misolli (few-shot) rejimda amalga oshirilmoqda. Tadqiqotlar shuni ko‘rsatadiki, LLM yordamida yaratilgan xulosalar

ko‘pincha inson tomonidan yozilgan xulosalar bilan solishtirilganda yuqori baholanadi, biroq bunday tizimlar hisoblash resurslari nuqtai nazaridan ancha qimmat bo‘lib, faktik nomuvofiqlik (hallyutsinatsiya) muammosi hamon dolzarbligicha qolmoqda. Xulosalash metodlarining to‘rt bosqichli evolyutsiyasi 2-rasmda sxematik tarzda aks ettirilgan.



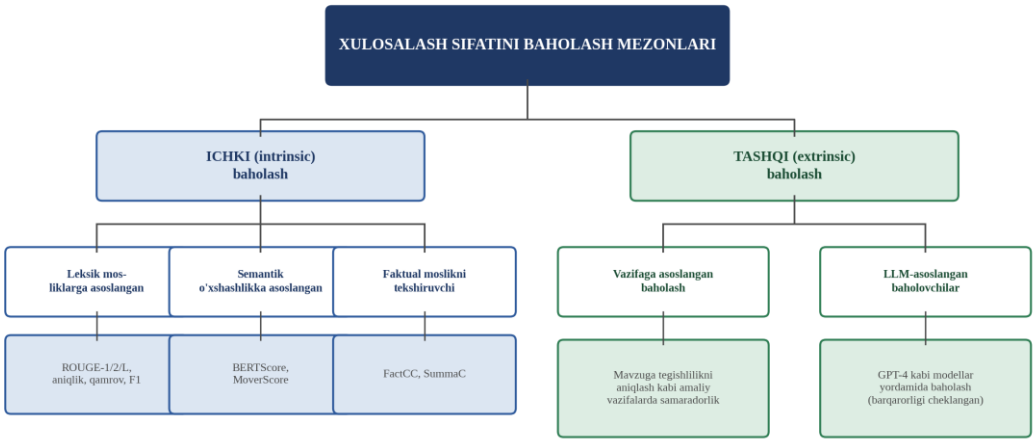
2-rasm. Matn xulosalash metodlarining rivojlanish bosqichlari va ularga xos usullar

Baholash mezonlari va ma’lumotlar to‘plamlari

Avtomatik xulosalash tizimlarining sifatini baholash sohaning eng murakkab masalalaridan biri hisoblanadi, chunki bitta matn uchun bir nechta yaxshi xulosa varianti mavjud bo‘lishi mumkin. Baholash usullari odatda ichki (intrinsic) va tashqi (extrinsic) turlarga bo‘linadi. Ichki baholashda xulosaning o‘zi – mazmuni, izchilligi va informativligi – tahlil qilinadi, tashqi baholashda esa xulosaning qandaydir amaliy vazifani (masalan, hujjat mavzusiga tegishliligini aniqlash) bajarishga qanchalik yordam berishi o‘lchanadi.

Eng keng tarqalgan avtomatik ko‘rsatkich – ROUGE (Recall-Oriented Understudy for Gisting Evaluation) bo‘lib, u tizim yaratgan xulosa bilan inson yozgan namunaviy xulosa orasidagi n-gramma darajasidagi mosliklarni hisoblaydi. ROUGE-1 va ROUGE-2 mos ravishda bir va ikki so‘zli birikmalar mosligini, ROUGE-L esa eng uzun umumiy ketma-ketlikni o‘lchaydi. Bundan tashqari, aniqlik (precision), qamrov (recall) va F1 ko‘rsatkichlari ham keng qo‘llaniladi. Biroq ROUGE faqat leksik mosliklarni hisobga olgani sababli, sinonim so‘zlar yoki qayta

ifodalangan jumlalarni to‘g‘ri baholay olmaydi. Shu sababli so‘nggi yillarda BERTScore, MoverScore singari kontekstual joylashtirishlarga (embeddings) asoslangan semantik o‘xshashlik ko‘rsatkichlari, shuningdek FactCC va SummaC kabi faktual moslikni tekshiruvchi metrikalar ishlab chiqilgan. Yirik til modellari davrida esa xulosalarni baholashda GPT-4 kabi modellardan foydalanilayotgani, ammo bunday baholovchilarning ham izchil va barqaror natija berishi hali to‘liq ta’minlanmaganligi qayd etilmoqda. Baholash mezonlarining ichki va tashqi turlarga bo‘linishi hamda ularning kichik toifalari 3-rasmda umumlashtirilgan.



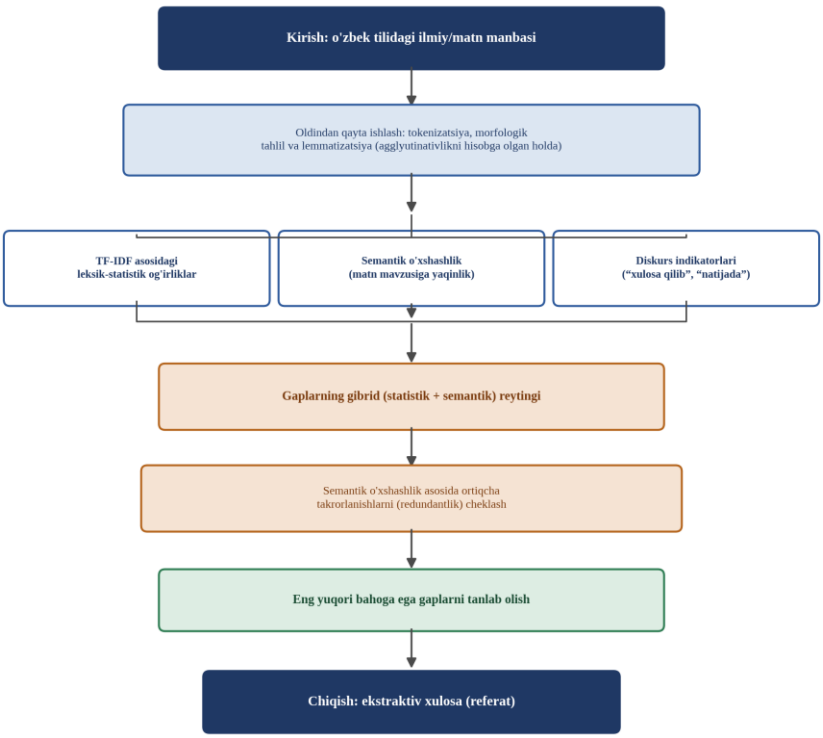
3-rasm. Xulosalash sifatini baholash mezonlarining taksonomiyasi

Xulosalash tizimlarini o‘qitish va sinovdan o‘tkazish uchun ko‘plab ochiq ma’lumotlar to‘plamlari yaratilgan. Yangiliklar sohasida CNN/DailyMail va XSum eng ko‘p qo‘llaniladigan manbalar hisoblanadi, uzun ilmiy matnlar uchun esa arXiv, PubMed va BigPatent kabi to‘plamlar mavjud bo‘lib, ularda maqola matni kirish, mos abstrakt esa maqsadli xulosa sifatida ishlatiladi. Ko‘p hujjatli xulosalash uchun MultiNews va DUC to‘plamlaridan, so‘rovga yo‘naltirilgan xulosalash uchun esa QMSum kabi maxsus to‘plamlardan foydalaniladi.

O‘zbek tili uchun avtomatik xulosalashning o‘ziga xos jihatlari

O‘zbek tilida matnlarni avtomatik xulosalash bo‘yicha tadqiqotlar hozircha ingliz tilidagi ishlanmalarga nisbatan kamroq rivojlangan bo‘lib, bu bir qator lingvistik va texnik omillar bilan izohlanadi. Yaqinda o‘tkazilgan tadqiqotlarda

o‘zbek tilidagi matnlar uchun gibrid – statistik, semantik va strukturaviy belgilarni birlashtirgan – ekstraktiv referatlash algoritmi taklif etilgan bo‘lib, unda TF-IDF asosidagi leksik-statistik og‘irliklar, matn mavzusiga yaqinlikni ifodalovchi semantik o‘xshashlik hamda diskurs indikatorlari (masalan, “xulosa qilib”, “natijada” kabi belgilovchi so‘zlar) birgalikda qo‘llaniladi. Ushbu tadqiqot natijalari o‘zbek tilining agglyutinativ xususiyati – ya’ni bitta so‘z ildizidan ko‘plab affiksal shakllar hosil bo‘lishi – chastotaviy ko‘rsatkichlarning tarqoq bo‘lib ketishiga olib kelishini, shu sababli lemmatizatsiya yoki hech bo‘lmaganda soddalashtirilgan normalizatsiya bosqichi zarurligini ko‘rsatgan. Bundan tashqari, semantik o‘xshashlik asosida ortiqcha takrorlanishlarni (redundantlik) cheklash mexanizmi o‘zbek matnlarida ham xulosaning qamrovini oshirishga yordam berishi aniqlangan, chunki sinonimiya va perifraz holatlari leksik jihatdan farqli, ammo mazmunan yaqin gaplar shaklida namoyon bo‘lishi mumkin. Ushbu gibrid algoritmning umumiy ish jarayoni 4-rasmda tasvirlangan.



4-rasm. O‘zbek tili uchun taklif etilgan gibrid ekstraktiv referatlash algoritmining arxitekturasi

Boshqa tadqiqotlarda ta’kidlanishicha, o‘zbek tili hozircha resurslari cheklangan (low-resource) tillar toifasiga kiradi, ya’ni yirik va sifatli matn korpuslari, belgilangan (annotatsiyalangan) ma’lumotlar bazalari hamda o‘zbek tiliga moslashtirilgan yirik til modellari soni hali yetarli emas. Shunga qaramay, so‘nggi yillarda UzBERT va BERTbek kabi o‘zbek tilida oldindan o‘qitilgan Transformer arxitekturali modellarning yaratilishi, shuningdek yangiliklar saytlaridan yig‘ilgan matn korpuslarining shakllantirilishi ushbu yo‘nalishda ijobiy siljish sifatida baholanadi. Bunday resurslarning ortib borishi kelajakda o‘zbek tili uchun ham sifatli abstraktiv xulosalash tizimlarini yaratish imkonini beradi.

Muhokama: ochiq muammolar va rivojlanish istiqbollari

Sohadagi sezilarli yutuqlarga qaramay, avtomatik xulosalashda bir qator ochiq muammolar saqlanib qolmoqda. Birinchidan, hallyutsinatsiya, ya’ni tizim tomonidan manba matnda mavjud bo‘lmagan yoki unga zid bo‘lgan ma’lumotning generatsiya qilinishi, ayniqsa abstraktiv usullarda jiddiy muammo hisoblanadi va tibbiyot yoki huquq kabi jiddiy sohalarda xavfli oqibatlarga olib kelishi mumkin. Ikkinchidan, hisoblash resurslarining yetishmasligi yirik til modellarini amaliyotda keng qo‘llashni qiyinlashtiradi, shu sababli parametrlarni tejamli sozlash (masalan, LoRA usuli) va bilim distillyatsiyasi kabi yechimlar faol o‘rganilmoqda. Uchinchidan, xulosalarning shaxsiylashtirilishi – foydalanuvchining qiziqishlari va bilim darajasiga moslashtirilgan xulosalarni shakllantirish – kelgusi tadqiqotlarning muhim yo‘nalishlaridan biri sifatida ko‘rsatilmoqda. Nihoyat, ilmiy matnlar uchun xulosalash tizimlarida faktlarning aniqligi va manbaga havolalarning to‘g‘riligini ta’minlash alohida ahamiyat kasb etadi, chunki bunday matnlarda noto‘g‘ri xulosa ilmiy xulosalarning noto‘g‘ri talqin qilinishiga olib kelishi mumkin.

O‘zbek tili nuqtai nazaridan, kelgusi tadqiqotlar quyidagi yo‘nalishlarda olib borilishi maqsadga muvofiq: birinchidan, ilmiy va ta’limiy matnlardan iborat sifatli, belgilangan (annotatsiyalangan) korpuslarni yaratish; ikkinchidan, morfologik tahlil

va lemmatizatsiya vositalarini takomillashtirish orqali leksik-statistik usullarning aniqligini oshirish; uchinchidan, mavjud ko‘p tilli va o‘zbek tiliga moslashtirilgan til modellaridan foydalangan holda gibrid (statistik + semantik) yondashuvlarni rivojlantirish; to‘rtinchidan, xulosalash sifatini baholash uchun o‘zbek tiliga mos standartlashtirilgan baholash protokollarini ishlab chiqish.

XULOSA

Ushbu maqolada ilmiy matnlarni avtomatik qisqacha xulosalash muammosi kompleks tarzda ko‘rib chiqildi. Xulosalash tizimlari statistik usullardan boshlab chuqur o‘rganish, oldindan o‘qitilgan til modellari va zamonaviy yirik til modellari davrigacha sezilarli evolyutsiyani bosib o‘tgani ko‘rsatildi. Har bir bosqich o‘ziga xos afzallik va cheklovlarga ega bo‘lib, hozirgi kunda yirik til modellari yuqori sifatli, tabiiy xulosalar yaratish imkonini bersa-da, faktual aniqlik va hisoblash xarajatlari borasida hali yechilmagan muammolar mavjud. Baholash borasida ROUGE kabi an’anaviy ko‘rsatkichlar bilan bir qatorda semantik va faktual moslikka asoslangan zamonaviy metrikalarning qo‘llanilishi zaruriy hisoblanadi. O‘zbek tili uchun bu sohadagi tadqiqotlar hali boshlang‘ich bosqichda bo‘lib, agglyutinativ morfologiya va resurslarning cheklanganligi asosiy qiyinchiliklar sifatida namoyon bo‘lmoqda; shu bilan birga, so‘nggi yillarda taklif etilgan gibrid algoritmlar va o‘zbek tilida oldindan o‘qitilgan modellar bu boradagi istiqbolli qadamlar ekanligini ko‘rsatmoqda. Yakuniy xulosa sifatida aytish mumkinki, ilmiy maqolalarni avtomatik xulosalash tizimlarini yaratish va takomillashtirish nafaqat xalqaro miqyosda, balki milliy ta’lim va ilmiy tadqiqot tizimida ham dolzarb va istiqbolli yo‘nalish bo‘lib qolmoqda.

Foydalanilgan adabiyotlar ro‘yxati

1. Zhang H., Yu P.S., Zhang J. A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models // ACM Computing Surveys. – 2025. – Vol. 57, No. 11, Article 277. – 41 p.

2. Uppalapati P.J., Dabbiru M., Rao K.V. A Comprehensive Survey on Summarization Techniques // SN Computer Science. – 2023. – 4:560.

3. Dang H.T., Owczarzak K. Overview of the TAC 2008 Update Summarization Task // Proceedings of the Text Analysis Conference (TAC 2008). – Gaithersburg: NIST, 2008.

4. Mani I., Klein G., House D., Hirschman L., Firmin T., Sundheim B. SUMMAC: a text summarization evaluation // Natural Language Engineering. – 2002. – Vol. 8, No. 1. – P. 43–68.

5. Qodirov F.E., Solixova B.S. Matnli ma’lumotlarni avtomatik referatlash va muhim axborotni aniqlash algoritmi // Xalqaro ilmiy-amaliy konferensiya materiallari. – 2025. – B. 293–303.

6. Sobirova Z.G‘. Til modellari va chuqur o‘rganish (Deep Learning) usullari: umumiy tavsif // «Kompyuter lingvistikasi: muammolar, yechim, istiqbollar» V Xalqaro ilmiy-amaliy konferensiya materiallari. – 2025. – Jild 1, №01. – B. 548–556.

7. Luhn H.P. The automatic creation of literature abstracts // IBM Journal of Research and Development. – 1958. – Vol. 2, No. 2. – P. 159–165.

8. Lin C.Y. ROUGE: A package for automatic evaluation of summaries // Proceedings of the ACL-04 Workshop: Text Summarization Branches Out. – Barcelona, 2004. – P. 74–81.

9. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT 2019. – P. 4171–4186.

10. Lewis M., Liu Y., Goyal N. et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension // Proceedings of ACL 2020. – P. 7871–7880.