AKADEMIK MAQOLALARDAN METAMA'LUMOTLARNI EKSTRAKSIYA QILISHDA MAVJUD YECHIMLAR

Rashid Turg'unboyev

Qo'qon davlat universiteti
https://orcid.org/0000-0003-1443-7900
E-mail: atmdsmi@gmail.com

KEYWORDS

katta til modellari, metama'lumotlarni ekstraksiya qilish, akademik maqolalar, GPT-4, BERT, multimodal yondashuv

ABSTRACT

Ushbu maqolada katta til modellarining akademik maqolalardan metama'lumotlarni ekstraksiya qilishdagi imkoniyatlari tahlil qilinadi. GPT-4, BERT va LLaMA kabi modellar an'anaviy usullarga nisbatan yuqori moslashuvchanlik, murakkab ma'lumotlarni ajratish va ko'p tilli ishlash afzalliklariga ega. ChatExtract usuli 90% F1-natijaga erishgan. Biroq resurs talabchanligi, faktik xatolar va maxfiylik muammolari asosiy cheklovlardir. Tor sohaga moslashtirilgan modellar tejamkorligi bilan ajralib turadi.

Kirish

Akademik maqolalardan metama'lumotlarni avtomatik ekstraksiya qilish sohasida so'nggi yillarda sun'iy intellektning rivojlanishi, xususan katta til modellarining paydo bo'lishi tub o'zgarishlarga olib keldi. Ushbu modellar - GPT oilasi, BERT va uning variantlari, LLaMA, Falcon, Vicuna va boshqalar - ulkan hajmdagi matnli korpuslarda oldindan o'qitilgan bo'lib, ular inson tilini chuqur tushunish, kontekstni saqlash va murakkab matnli vazifalarni bajarish qobiliyatiga ega. An'anaviy usullar - qoidalarga asoslangan tizimlar, support vector mashinalari, yashirin Markov modellari va hatto konvolyutsion hamda takrorlanuvchi neyron tarmoqlar - ko'pincha ma'lum bir jurnal yoki nashriyot formatiga moslashtirilgan bo'lib, yangi formatlar paydo bo'lganda qayta sozlashni talab qiladi. Katta til modellari esa matnni nafaqat yuzaki statistik xususiyatlariga ko'ra, balki mazmuni, konteksti va semantik ma'nosiga qarab tahlil qila oladi. Bu xususiyat metama'lumotlarni avtomatik ekstraksiya qilish vazifasiga butunlay yangicha yondashish imkonini beradi va an'anaviy usullar oldida turgan ko'plab muammolarni yengishda katta salohiyatga ega. An'anaviy usullar va katta til modellarining asosiy xususiyatlari 1-jadvalda qiyosiy tahlil qilingan.

1-jadval

Metama'lumotlarni ekstraksiya qilish yondashuvlarini taqqoslash

Xususiyat	Qoidalarga asoslangan	An'anaviy ML (CRF/SVM)	Chuqur o'rganish (CNN/RNN)	Katta til modellari (GPT/BERT/LLaMA)
Qo'lda qoidalar yozish zarurati	Yuqori	Yo'q	Yo'q	Yo'q
Katta o'quv ma'lumotlari talabi	Yo'q	Yuqori	Yuqori	Past
Turli formatlarga moslashuvchanlik	Past	O'rtacha	O'rtacha	Yuqori
Kontekstni tushunish	Yo'q	Cheklangan	O'rtacha	Yuqori
Ko'p tilli ishlash qobiliyati	Past	Past	O'rtacha	Yuqori
Resurs talabchanligi	Past	O'rtacha	Yuqori	Juda yuqori
Tushunarlilik (shaffoflik)	Yuqori	O'rtacha	Past	Past

(cc) BY

Izoh. Metama'lumotlarni ekstraksiya qilish usullarining qiyosiy tahlili Dagdelen va boshq., 2024; Polak va Morgan, 2023 asosida tuzilgan.

Ushbu maqolada katta til modellarining akademik maqolalardan metama'lumotlarni ekstraksiya qilishdagi imkoniyatlari, ularning an'anaviy yondashuvlarga nisbatan afzalliklari, amaliy qo'llanish misollari hamda mavjud cheklovlari tahlil qilinadi.

Metodlar

Ushbu tadqiqot adabiyotlar tahlili asosida katta til modellarining metama'lumotlarni ekstraksiya qilish sohasidagi qo'llanilishini o'rganishga qaratilgan. Tahlil uchun so'nggi besh yil ichida nufuzli ilmiy jurnallar va konferensiyalarda chop etilgan maqolalar, shuningdek ochiq kodli tizimlarning texnik hujjatlari tanlab olindi. Tadqiqot quyidagi bosqichlardan iborat bo'ldi. Birinchi bosqichda katta til modellarining umumiy arxitekturasi va ularning matnli ma'lumotlarni qayta ishlashdagi asosiy mexanizmlari o'rganildi. Xususan, transformer arxitekturasi asosidagi modellarning diqqat mexanizmi va kontekstli ifodalarni yaratish qobiliyatiga e'tibor qaratildi. Ikkinchi bosqichda ushbu modellarning metama'lumotlarni ekstraksiya qilish vazifasiga qo'llanilgan empirik tadqiqotlari tahlil qilindi. Bular qatoriga Polak va Morgan tomonidan ishlab chiqilgan ChatExtract usuli [1], Dagdelen va boshqalarning Nature Communications jurnalida chop etilgan ishi [2], Krause va boshqalarning kirill alifbosidagi nashrlar bo'yicha tadqiqoti [3], shuningdek Choi va boshqalarning ko'p tilli korpus asosidagi ishi kiritilgan [4]. Uchinchi bosqichda multimodal yondashuvlar - matn va vizual ma'lumotlarni birgalikda qo'llaydigan tizimlar - tahlil qilindi. Buning uchun Boukhers va Yangning qiyosiy tadqiqoti hamda Kohút va boshqalar tomonidan yaratilgan BiblioPage ma'lumotlar to'plamida vizual katta til modellarini baholash natijalari o'rganildi. To'rtinchi bosqichda ixtisoslashtirilgan tor sohalarda katta til modellarining samaradorligi, jumladan Saier va boshqalarning giperparametr ma'lumotlarini ajratish bo'yicha ishi va Usta tomonidan turk tilidagi sarlavha sahifalaridan MARC yozuvlarini yaratish bo'yicha tadqiqoti tahlil qilindi. Beshinchi bosqichda katta til modellarini qo'llashning iqtisodiy samaradorligi va resurs talabchanligi masalalari o'rganildi. Natijalarni baholashda asosan F1-ko'rsatkich, aniqlik va to'liqlik kabi standart metrikalardan foydalanilgan. F1-ko'rsatkich quyidagi formula asosida xisoblanadi:

$$F1 = 2 \times \frac{P \times R}{P + R}$$

bunda,

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

TP - to'g'ri musbat
FP - noto'g'ri musbat
FN - noto'g'ri manfiy

Natijalar

Adabiyotlar tahlili va empirik tadqiqotlar natijalari katta til modellarining metama'lumotlarni ekstraksiya qilish sohasida an'anaviy usullarga nisbatan bir qator muhim afzalliklarga ega ekanligini ko'rsatdi. Eng muhim natijalardan biri shundaki, katta til modellari turli xil formatdagi va tuzilishdagi ilmiy hujjatlar bilan ishlashda yuqori moslashuvchanlikni namoyish etadi. Polak va Morgan tomonidan taklif qilingan ChatExtract usuli GPT-4 modeli yordamida materialshunoslik sohasidagi maqolalarda aniqlik va to'liqlik ko'rsatkichlarini 90 foizga yaqin darajada olgan. ChatExtract usulining asosiy g'oyasi modelning ma'lumotni saqlab qolish xususiyatidan foydalanib, bir nechta ketma-ket so'rovlar orqali ekstraksiya qilingan ma'lumotlarni tekshirish va tasdiqlashdan iborat. Bu yondashuv katta til modellarining ba'zan faktik jihatdan noto'g'ri javoblar berishi muammosini sezilarli darajada kamaytiradi. Turli tadqiqotlarda katta til modellari yordamida erishilgan asosiy natijalar 2-jadvalda jamlangan.

(cc) BY

2-jadval

Metama’lumotlarni ekstraksiya qilishda kata til modellarining asosiy empirik natijalari

Tadqiqot	Qo‘llanilgan model	Vazifa	Erishilgan natija (F1)
Polak & Morgan (ChatExtract)	GPT-4	Materialshunoslik maqolalaridan ma’lumot ajratish	≈90%
Dagdelen et al.	GPT-3, Llama-2	Materiallar kimyosi (3 vazifa)	Yuqori aniqlik
Krause et al.	GROBID (qayta o‘qitilgan)	Kirill alifbosidagi nashrlar	Yaxshilanish (aniq raqam keltirilmagan)
Choi et al.	BERT (multilingual)	Ingliz/koreys adabiyotlaridan metama’lumot	Yuqori samaradorlik
Kohút et al.	GPT-4o (vision)	Skanerlangan sarlavha sahifalari	67 F1
Saier et al. (HyperPIE)	BERT (sozlangan)	Giperparametrlarni ajratish	+29% (mavjud usulga nisbatan)
Usta	BERT (sozlangan)	Turk tilidagi sarlavha sahifalari	0.666 F1

Izoh. Katta til modellari yordamida metama’lumotlarni ekstraksiya qilish bo‘yicha empirik tadqiqotlar natijalari Brinkmann va boshq., 2025; Jumashev va boshq., 2025; BMC Medical Informatics, 2025 asosida tuzilgan

Katta til modellarining yana bir muhim natijasi ularning murakkab va ierarxik tuzilgan ma’lumotlarni ekstraksiya qilish imkoniyatidir. Dagdelen va boshqalar tomonidan Nature Communications jurnalida chop etilgan tadqiqotda GPT-3 va Llama-2 kabi modellar materiallar kimyosi sohasida uchta reprezentativ vazifani - dopantlar va ularning tashuvchi materiallari o‘rtasidagi bog‘liqlikni aniqlash, metall-organik ramka strukturalarini kataloglash va umumiy tarkib, faza, morfologiya hamda qo‘llanish ma’lumotlarini ajratish - muvaffaqiyatli amalga oshirgan. Tadqiqot shuni ko‘rsatdiki, katta til modellari bir jumla yoki butun paragrafdan ma’lumotlarni ajratib olishda yuqori samaradorlikka erisha oladi va natijalarni oddiy ingliz jumlalari yoki JSON ob’ektlari kabi tuzilgan formatlarda qaytarishi mumkin.

Ko‘p tilli kontekstda ham katta til modellari an’anaviy tizimlarga nisbatan sezilarli darajada yaxshi natijalar ko‘rsatdi. Krause va boshqalar kirill alifbosidagi nashrlardan metama’lumotlarni ekstraksiya qilishda GROBIDni qayta o‘qitish orqali sarlavha va muallif ma’lumotlarini ajratishda yaxshilanishlarga erishgan bo‘lsa-da, katta til modellari oldindan o‘qitish jarayonida ko‘p tilli ma’lumotlar bilan ishlaganligi sababli, ular turli tillardagi matnlarni tushunish va qayta ishlash qobiliyatiga ega. Choi va boshqalar tomonidan yaratilgan ko‘p tilli ingliz va koreys jurnal maqolalari adabiyotlaridan metama’lumotlarni avtomatik ekstraksiya qilish uchun mo‘ljallangan anotatsiyalangan korpus BERT asosidagi transfer o‘rganish modelini o‘qitishda qo‘llanilgan. Ushbu tadqiqotda BERT asosidagi model nafaqat ingliz tilidagi, balki koreys tilidagi adabiyotlardan ham metama’lumotlarni ekstraksiya qilishda yuqori samaradorlik ko‘rsatgan.

Multimodal yondashuvlar, ya’ni matn va vizual ma’lumotlarni birgalikda ishlatish, ayniqsa turli xil shablonlar qo‘llaniladigan hujjatlarda an’anaviy unimodal usullarga nisbatan sezilarli darajada yuqori natijalar beradi. Boukhers va Yang tomonidan olib borilgan tadqiqotda nemis ijtimoiy fanlari nashrlarida multimodal yondashuv an’anaviy NLP usullaridan ustun kelgan. [5] Kohút va boshqalar tomonidan yaratilgan BiblioPage ma’lumotlar to‘plamida skanerlangan sarlavha sahifalaridan bibliografik metama’lumotlarni ekstraksiya qilish uchun turli vizual katta til modellari baholangan. [6] Tadqiqot natijalariga ko‘ra, GPT-4o modeli 67 F1-natijaga erishgan bo‘lsa, YOLO va DETR

(cc) BY

kabi ob’ekt aniqlash modellari 52 mAP va 59 F1-natija ko‘rsatgan. Bu esa vizual katta til modellarining an’anaviy kompyuter ko‘rish modellariga nisbatan murakkabroq vizual-mantiqiy vazifalarni bajarishdagi ustunligini ko‘rsatadi.

Ixtisoslashtirilgan tor sohalarda katta til modellarining qo‘llanilishi ham o‘z natijalarini berdi. Saier va boshqalar tomonidan giperparametr ma’lumotlarini ajratish vazifasida sozlangan BERT modellari mavjud eng yaxshi texnikaga nisbatan 29 foizlik F1 yaxshilanishiga erishgan. [7] Katta til modellari uchun esa YAML formatida tuzilgan chiqish JSON formatiga nisbatan ob’ektlarni aniqlashda o‘rtacha 5.5 foiz yaxshilanish ko‘rsatgan. Alyafeai va boshqalar tomonidan ishlab chiqilgan MOLE tizimi arab tilidan tashqari boshqa tillardagi maqolalardan metama’lumotlarni ekstraksiya qilishda sxema asosidagi metodologiyaga asoslanadi va natijalarning izchilligini ta’minlash uchun mustahkam tekshirish mexanizmlarini o‘z ichiga oladi. [8]

Katta til modellarining iqtisodiy samaradorligi va resurs talabchanligi bo‘yicha Usta tomonidan olib borilgan tadqiqot muhim natijalarni keltirib chiqardi. [9] Turkiy tilidagi sarlavha sahifalaridan MARC yozuvlarini avtomatik yaratishda sozlangan BERT modeli (F1-natija 0.666) vizual katta til modellaridan (Gemini 2.5 Flash F1-natija 0.533) sezilarli darajada yuqori ko‘rsatkichga erishgan. Bundan tashqari, resurs iste’moli tahlili shuni ko‘rsatdiki, BERT asosidagi model har bir namuna uchun xarajat va uglerod izi bo‘yicha vizual katta til modellariga nisbatan ancha tejamkor. Bu natija muhim xulosani beradi: ixtisoslashtirilgan, tor sohaga moslashtirilgan modellar umumiy maqsadli katta modellarga nisbatan ko‘p hollarda yuqori samaradorlik, tejamkorlik va barqarorlikka ega bo‘lishi mumkin.

Muhokama

Katta til modellarining metama’lumotlarni ekstraksiya qilish sohasiga qo‘llanilishi an’anaviy usullar oldida turgan ko‘plab muammolarni hal qilish imkoniyatini beradi. Qoidalarga asoslangan tizimlarning moslashuvchanligi pastligi va doimiy yangilab borish zarurati, an’anaviy mashinaviy o‘rganish modellarining o‘quv ma’lumotlari sifati va hajmiga bog‘liqligi, shuningdek ularning matn mazmunini to‘liq tushuna olmasligi kabi cheklovlar katta til modellari yordamida sezilarli darajada yengillashtiriladi. Katta til modellarining kontekstni chuqur tushunish qobiliyati ularga murakkab va nostandart hollarda ham barqaror natijalar berish imkonini beradi. Ayniqsa, ChatExtract kabi takroriy so‘rovlar asosidagi tekshirish mexanizmlari bilan birgalikda qo‘llanilganda, bu modellar inson darajasidagi aniqlikka yaqin natijalarni ko‘rsatmoqda. Katta til modellarini qo‘llashdagi asosiy cheklovlar va ularni kamaytirishning mumkin bo‘lgan usullari 3-jadvalda tizimlashtirilgan.

3-jadval

Katta til modellarining cheklovlari va bartaraf etish strategiyalari

Cheklov	Tavsif	Bartaraf etish usullari
Resurs talabchanligi	GPU, elektr energiyasi, vaqt talab qiladi	Tor sohaga moslashtirilgan kichik modellar (BERT)
Faktik xatolar (hallucination)	Noto‘g‘ri yoki uydirma ma’lumotlar	ChatExtract usuli (takroriy so‘rovlar)
Maxfiylik xavfi	Bulutli modellar ma’lumotlarni ko‘rishi mumkin	Lokal modellar (LLaMA, Falcon)
Qora quti muammosi	Qarorlarni tushuntirish qiyin	Gibrid yondashuvlar
Cheklangan kontekst uzunligi	Juda uzun maqolalarni to‘liq qayta ishlay olmaslik	Bo‘lib-bo‘lib ishlash (chunking)

Izoh. Katta til modellarining cheklovlari va ularni bartaraf etish usullari Usta, 2026; Saier va boshq., 2024 asosida tuzilgan.

Biroq katta til modellarini qo‘llashning o‘ziga xos cheklovlari ham mavjud. Birinchidan, ularning resurs talabchanligi katta muammo bo‘lib qolmoqda. GPT-4 kabi yirik modellarni ishga tushirish va ular bilan ishlash katta hisoblash resurslarini va elektr energiyasini talab

(cc) BY

qiladi, bu esa atrof-muhitga salbiy ta’sir ko‘rsatadi. Usta tadqiqotida ko‘rsatilganidek, tor sohaga moslashtirilgan BERT modeli resurs iste’moli va uglerod izi bo‘yicha umumiy maqsadli yirik modellardan ancha tejamkor. Shu sababli, amaliyotda qaysi modelni tanlash vazifaning murakkabligi, mavjud resurslar va talab qilinadigan aniqlik darajasiga qarab belgilanishi kerak.

Ikkinchidan, katta til modellarining ba’zan faktik jihatdan noto‘g‘ri javoblar berishi, ya’ni “hallucination” muammosi jiddiy cheklov hisoblanadi. Bu muammo ayniqsa metama’lumotlarni ekstraksiya qilish kabi aniqlik muhim bo‘lgan vazifalarda xavfli bo‘lishi mumkin. Ekstraksiya qilingan ma’lumotlarning ishonchlilik koeffitsiyenti quyidagi formula asosida hisoblanadi:

$$C_{final} = \frac{1}{n}\sum_{i=1}^{n} LLM_i(so'rov, nomzod)$$

bunda, n - takroriy so‘rovlar soni

LLM_i - i sonli so‘rovdan olingan ikkilik yoki ehtimollok kelishuv

ChatExtract usulidagi takroriy so‘rovlar va tekshirish mexanizmlari bu muammoni kamaytirishga yordam bersa-da, uni to‘liq bartaraf eta olmaydi. Uchinchidan, maxfiylik va ma’lumotlar xavfsizligi masalalari ham muhimdir, chunki ko‘plab yirik katta til modellari bulutli xizmatlar orqali taqdim etiladi va tijorat maqsadlarida foydalaniladi.

Multimodal yondashuvlar, xususan vizual katta til modellari, skanerlangan va murakkab formatdagi hujjatlar bilan ishlashda katta imkoniyatlarni ko‘rsatmoqda. Kohút va boshqalar tadqiqotida GPT-4o an’anaviy kompyuter ko‘rish modellaridan sezilarli darajada yuqori natijalarga erishgan. Bu holat vizual va matnli ma’lumotlarni birgalikda qayta ishlash nafaqat matnli kontekstni, balki sahifadagi elementlarning vizual joylashuvini ham tushunish muhim bo‘lgan hollarda muhim afzallik beradi. Biroq bu modellarning resurs talabchanligi an’anaviy matnli modellarga nisbatan ham yuqori.

Xulosa

Ushbu maqolada katta til modellarining akademik maqolalardan metama’lumotlarni avtomatik ekstraksiya qilish sohasidagi imkoniyatlari va cheklovlari tahlil qilindi. Tadqiqot shuni ko‘rsatdiki, zamonaviy katta til modellari - GPT-4, BERT, LLaMA, Gemini va boshqalar - an’anaviy usullarga nisbatan bir qator muhim afzalliklarga ega. Ularning turli format va tillarga moslasha olishi, murakkab ierarxik ma’lumotlarni ajratib olish qobiliyati, shuningdek multimodal (matn va vizual) yondashuvlarni qo‘llash imkoniyati ularni istiqbolli vositalarga aylantiradi. ChatExtract kabi takroriy so‘rovlar asosidagi tekshirish mexanizmlari faktik xatoliklar muammosini kamaytirishga yordam beradi. Ko‘p tilli tadqiqotlar ingliz tilidan tashqari boshqa tillardagi ilmiy nashrlar uchun ham katta til modellarining samarali qo‘llanilishi mumkinligini ko‘rsatdi.

Biroq katta til modellarining resurs talabchanligi, faktik xatolar muammosi, maxfiylik va ma’lumotlar xavfsizligi kabi cheklovlari mavjud. Ixtisoslashtirilgan, tor sohaga moslashtirilgan modellar (masalan, sozlangan BERT) umumiy maqsadli yirik modellarga nisbatan ko‘p hollarda yuqori samaradorlik, tejamkorlik va barqarorlikka ega bo‘lishi mumkin. Shu sababli, amaliyotda umumiy maqsadli katta modellar va ixtisoslashtirilgan modellarni uyg‘unlashtirgan gibrid yondashuvlar istiqbolli hisoblanadi. Kelgusidagi tadqiqotlar o‘zbek tilidagi ilmiy maqolalar uchun katta til modellarini moslashtirish, resurs cheklangan muhitda samarali ishlaydigan yengil modellarni ishlab chiqish hamda faktik aniqlikni oshirishning yangi usullarini yaratishga qaratilishi kerak.

Foydalanilgan adabiyotlar

1. Polak, M.P., Morgan, D.: Extracting accurate materials data from research papers with conversational language models and prompt engineering - Example of ChatGPT (Mar 2023). https://doi.org/10.48550/arXiv.2303.05352
2. J. Dagdelen et al., “Structured information extraction from scientific text with large language models,” Nature Communications, vol. 15, no. 1, 2024, doi: 10.1038/s41467–024-45563-x.

(cc) BY

3. Johan Krause, Igor Shapiro, Tarek Saier, and Michael Färber. 2021. Bootstrapping Multilingual Metadata Extraction: A Showcase in Cyrillic. In Proceedings of the Second Workshop on Scholarly Document Processing, pages 66–72, Online. Association for Computational Linguistics.
4. Choi, W., Yoon, H. M., Hyun, M. H., Lee, H. J., Seol, J. W., Lee, K. D., ... & Kong, H. (2023). Building an annotated corpus for automatic metadata extraction from multilingual journal article references. PloS one, 18(1), e0280637.
5. Boukhers, Z., & Yang, C. (2025). Comparison of Feature Learning Methods for Metadata Extraction from PDF Scholarly Documents. arXiv preprint arXiv:2501.05082.
6. Kohút, J., Dočekal, M., Hradiš, M., Vaško, M. (2026). BiblioPage: A Dataset of Scanned Title Pages for Bibliographic Metadata Extraction. In: Yin, XC., Karatzas, D., Lopresti, D. (eds) Document Analysis and Recognition – ICDAR 2025. ICDAR 2025. Lecture Notes in Computer Science, vol 16025. Springer, Cham. https://doi.org/10.1007/978-3-032-04624-6_17
7. T. Saier, M. Ohta, T. Asakura, and M. Färber, “HyperPIE: Hyperparameter Information Extraction from Scientific Publications,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 14609 LNCS, pp. 254–269, 2024, doi: 10.1007/978-3-031-560606_17.
8. Alyafeai, Z., Al-Shaibani, M. S., & Ghanem, B. (2025). MOLE: Metadata extraction and validation in scientific papers using LLMs. arXiv preprint arXiv:2505.19800.
9. Usta, G. (2026). Automating MARC Records from Turkish Title Pages: A Performance and Efficiency Comparison of Fine-Tuned BERT vs. VLMs. Journal of Library Metadata, 1–19. https://doi.org/10.1080/19386389.2026.2627732

(cc) BY