Automatic Information Extraction Technologies From Textual Data

Boimurodova Gulnoza
Shahrisabz State Pedagogical Institute,
Preschool Education, 1st year master's student

Annotation
This article provides a comprehensive review of technologies for automatic information extraction from textual data. Methods based on natural language processing (NLP), machine learning and deep learning approaches are analyzed in detail. Basic techniques such as Named Entity Recognition (NER), relationship extraction and data description are studied from the point of view of their effectiveness and areas of application. The article presents a comparative analysis of methods based on rules, statistics and neural networks. The results obtained serve to expand automation in modern information systems and improve the quality of education.
Keywords: text analytics, information technology, NLP, NER, machine learning, deep learning, natural language processing, BERT, transformer, digital technologies, educational systems.
Annotatsiya
Ushbu maqolada matnli ma'lumotlardan avtomatik axborot ajratib olish texnologiyalari keng ko'lamda ko'rib chiqilgan. Tabiiy tilni qayta ishlash (NLP), mashina o'qitish va chuqur o'rganish yondashuvlariga asoslangan usullar batafsil tahlil qilingan. Named Entity Recognition (NER), munosabatlarni ajratib olish va ma'lumotlarni tavsiflash kabi asosiy texnikalar samaradorligi va qo'llanilish sohalari nuqtai nazaridan o'rganilgan. Maqolada qoida asosidagi, statistik va neyron tarmoqlarga asoslangan usullarning qiyosiy tahlili amalga oshirilgan. Olingan natijalar zamonaviy axborot tizimlarida avtomatlashtirishni kengaytirishga va ta'lim sifatini oshirishga xizmat qiladi.
Kalit so'zlar: matn tahlili, axborot ajratib olish, NLP, NER, mashina o'qitish, chuqur o'rganish, tabiiy tilni qayta ishlash, BERT, transformer, raqamli texnologiyalar, ta'lim tizimlari.

KIRISH

Zamonaviy axborot jamiyatida har kuni millionlab matnli hujjatlar yaratilmoqda: ilmiy maqolalar, yangiliklar, ijtimoiy tarmoq xabarlari, tibbiy hisobotlar, huquqiy hujjatlar va boshqa ko'plab manbalardagi ma'lumotlar. IDC (International Data Corporation) ma'lumotlariga ko'ra, dunyo bo'yicha har yili yaratilayotgan ma'lumotlar hajmi eksponent darajada o'sib bormoqda va 2025-yilga kelib 175 zettabaytga yetishi kutilmoqda. Bunday katta hajmdagi tuzilmagan matnlardan foydali aniq axborotni qo'lda ajratib olish inson uchun deyarli mumkin emas. Shu sababli, avtomatik axborot ajratib olish (Information Extraction — IE) texnologiyalari zamonaviy sun'iy intellekt va ma'lumotlar fanining eng muhim yo'nalishlaridan biriga aylandi [1, 2]. IE tizimlari tuzilmagan matndan foydali ma'lumotlarni avtomatik ravishda ajratib, ularni tizimli ko'rinishga — jadval, grafik yoki bilimlar bazasiga — aylantirish imkonini beradi. Natijada, katta hajmdagi hujjatlarni tez va aniq tahlil qilish, qaror qabul qilishni qo'llab-quvvatlash va inson mehnatini tejash mumkin bo'ladi.

Axborot ajratib olish — bu tuzilmagan yoki yarim tuzilmagan matnli ma'lumotlardan aniq faktlar, munosabatlar, hodisalar va boshqa semantik birliklarni avtomatik tarzda topish va tizimlashtirish jarayonidir. Raqamli texnologiyalarning jadal rivojlanishi va tarmoq

texnologiyalarining takomillashishi [4] bu sohadagi imkoniyatlarni yanada kengaytirmoqda. Xususan, LiFi va boshqa yuqori tezlikli uzatish texnologiyalari IE tizimlarini real vaqt rejimida ishlashini ta'minlaydi.

Ta'lim sohasida ham ushbu texnologiyalar tobora muhim ahamiyat kasb etmoqda. Masofaviy ta'limning rivojlanishi [6] va ta'lim tizimlarini boshqarishning zamonaviy yondashuvlari [5] IE texnologiyalariga bo'lgan talabni oshirmoqda. Maktabgacha ta'lim muassasalarida pedagogik adabiyotlardan kalit tushunchalarni avtomatik ajratib olish, o'quv materiallarini tizimlashtirish va bolalar rivojlanishini kuzatish uchun bu texnologiyalardan samarali foydalanish mumkin. Maqolaning maqsadi — matnli ma'lumotlardan avtomatik axborot ajratib olishning asosiy texnologiyalari, usullari va algoritmlarini ilmiy jihatdan tahlil qilish, ularning afzalliklari hamda kamchiliklarini aniqlash va amaliy qo'llanilish sohalarini ko'rsatishdir. Maqolaning vazifalari: (1) IE texnologiyalarining tarixiy rivojlanish bosqichlarini o'rganish; (2) asosiy usullarni qiyosiy tahlil qilish; (3) O'zbek tili uchun mavjud yechimlarni baholash; (4) ta'lim sohasida qo'llanilish imkoniyatlarini aniqlash.

ADABIYOTLAR SHARHI

Axborot ajratib olish sohasidagi tadqiqotlar 1970-yillarda MUC (Message Understanding Conference) konferensiyalari doirasida boshlangan. Dastlabki tizimlar qoida asosidagi yondashuvlarga tayanib, belgilangan shablonlar orqali matndan ma'lumotlarni ajratib olgan. MUC-6 konferensiyasida (Grishman & Sundheim, 1996) shaxslar, tashkilotlar va joylarni tanish vazifalari bo'yicha birinchi standartlashtirilgan baholash metodologiyasi taklif etildi. Manning va Schutze (1999) [3] o'zlarining fundamental asarida statistik tilshunoslik va mashina o'qitishning matn tahlilida qo'llanilishi imkoniyatlarini batafsil bayon etishgan. Ushbu asar NLP sohasida asosiy manba sifatida bugungi kunda ham keng foydalaniladi. Mikolov va boshqalar (2013) tomonidan taklif etilgan Word2Vec modeli so'zlarni vektor fazosida ifodalash imkonini berdi va IE tizimlarining sifatini sezilarli oshirdi.

Lafferty va boshqalar (2001) [9] tomonidan ishlab chiqilgan Shartli Tasodifiy Maydonlar (CRF) modeli NER va matn teglovchi tizimlar uchun keng tarqalgan standart usulga aylandi. CRF modelining asosiy afzalligi — u nafaqat joriy so'zning xususiyatlarini, balki atrofidagi so'zlar bilan bog'liq kontekstual ma'lumotlarni ham hisobga ola

olishidir. Bu xususiyat NER vazifalarida F1 ko'rsatkichini sezilarli darajada oshirdi. 2018-yilda Google tomonidan taqdim etilgan BERT modeli [1] axborot ajratib olish sohasida haqiqiy inqilob yasadi. Bidirectional Encoder Representations from Transformers (BERT) o'ng va chap kontekstni bir vaqtda hisobga olishi tufayli matnning chuqur semantik ma'nosini tushunish imkonini berdi. Kompyuter grafikasi va raqamli tasvir qayta ishlash sohasidagi yutuqlar [7] ham matnli ma'lumotlarni vizual tahlil qilishda yangi imkoniyatlar ochmoqda. Vaswani va boshqalar (2017) [10] tomonidan taklif etilgan Attention mexanizmi zamonaviy barcha transformer modellarining poydevori hisoblanadi. So'nggi yillarda GPT-3, GPT-4, RoBERTa, XLNet, T5 kabi katta til modellari (Large Language Models — LLM) ham IE sohasida yuqori natijalar bermoqda. Virtual reallik texnologiyalari [8] esa IE natijalarini interaktiv va vizual tarzda taqdim etishda yangi imkoniyatlar yaratmoqda.

METODOLOGIYA

Tadqiqotda quyidagi ilmiy metodlardan foydalanildi: adabiyotlarni tizimli tahlil qilish, mavjud texnologiyalarni qiyosiy baholash, algoritmlarga asoslangan tasnif va tizimlashtirish, shuningdek eksperimental natijalarni taqqoslash. Tadqiqot predmeti sifatida 2000-2025-yillar oralig'ida nashr etilgan 50 dan ortiq ilmiy maqola, texnik hisobot va loyihalar materiallari ko'rib chiqildi.

Axborot ajratib olish usullari uchta asosiy kategoriyaga ajratildi va har birining samaradorligi standart baholash o'lchovlari — Precision (aniqlik), Recall (qamrov) va F1-score (ularning

garmonik o'rtachasi) — asosida taqqoslandi. Virtual reallik va kengaytirilgan reallik texnologiyalari [8] ham axborot ajratib olish tizimlarining vizualizatsiyasida yangi imkoniyatlar yaratmoqda.

3.1. Qoida asosidagi usullar (Rule-based Methods)

Dastlabki IE tizimlari qo'lda yozilgan leksik, morfologik va sintaktik qoidalarga asoslangan. Masalan, 'Janob', 'professor', 'doktor' so'zlari oldidan kelgan so'z shaxs ismi sifatida belgilanishi mumkin. Yoki 'kompaniya', 'korporatsiya', 'OAJ' so'zlari oldidan kelgan so'z tashkilot nomi sifatida tasniflanadi. Regulyar ifodalar (Regular Expressions) va lug'at asosidagi usullar ham bu kategoriyaga kiradi.

Qoida asosidagi yondashuvning asosiy afzalligi — yuqori aniqlik (precision) ko'rsatkichi va tizimning ishlash prinsipi to'liq tushunarli bo'lishi (interpretability). Kamchiliklari: yangi domenga moslashish uchun mutaxassis tomonidan qo'lda qoidalar yozilishi kerak, bu katta vaqt va mehnat talab etadi; shuningdek, qamrov (recall) ko'rsatkichi odatda past bo'ladi, chunki barcha mumkin bo'lgan variantlarni qoidalarda qamrab olish qiyin.

3.2. Statistik va mashina o'qitish usullariMashina o'qitishga asoslangan yondashuvlarda model belgilangan (annotated) korpuslardan o'qitiladi. Asosiy algoritmlar orasida quyidagilarni ajratib ko'rsatish mumkin: Yashirin Markov Modeli (HMM) — ketma-ket belgilash uchun ehtimollik asosidagi model; Maksimal Entropiya (MaxEnt) — turli xususiyatlarni hisobga oladigan klassifikator; Qo'llab-quvvatlash vektor mashinalari (SVM) — yuqori o'lchamli xususiyat fazosida optimal ajratuvchi tekislikni topadigan model; Shartli Tasodifiy Maydonlar (CRF) [9] — kontekstli xususiyatlarni hisobga oladigan kuchli model.

CRF modeli NER vazifalarida ayniqsa samarali ishlaydi, chunki u nafaqat alohida so'zning belgilarini, balki qo'shni so'zlar bilan kontekstual aloqalarni ham hisobga oladi. Masalan, 'New York' birikmasida 'New' va 'York' so'zlari alohida tahlil qilinganda oddiy sifat va ism bo'lib ko'rinishi mumkin, lekin CRF ularning birgalikdagi kontekstini hisobga olib, joy nomini to'g'ri aniqlaydi. Bu usul LiFi kabi zamonaviy tarmoq infratuzilmalari [4] orqali real vaqt rejimida ham samarali ishlashi mumkin.

3.3. Chuqur o'qitish usullari (Deep Learning)

Neyron tarmoqlarga asoslangan usullar, xususan LSTM (Long Short-Term Memory), BiLSTM-CRF va Transformer arxitekturalari (BERT [1], RoBERTa, XLNet) hozirgi kunda eng yuqori natijalarni bermoqda. LSTM modeli matnning uzoq muddatli bog'liqliklarini saqlab qolish qobiliyati tufayli NLP sohasida keng qo'llanildi. BiLSTM-CRF modeli esa ikki yo'nalishli LSTM va CRFni birlashtirgan gibrid arxitektura bo'lib, NER sohasida eng ko'p ishlatiladigan modellardan biri hisoblanadi.

BERT modeli [1] pre-training (oldindan o'qitish) va fine-tuning (sozlash) strategiyasiga asoslanadi. Dastlab model katta miqdordagi belgilanmagan matnda (masalan, Wikipedia va BookCorpus) o'qitiladi. So'ngra kichik miqdordagi belgilangan ma'lumotlar yordamida aniq vazifaga moslashtiriladi. Bu yondashuv kamroq belgilangan ma'lumotlar bilan ham yuqori aniqlikka erishish imkonini beradi. Attention mexanizmi [10] esa modelga matnning qaysi qismlariga ko'proq e'tibor qaratish kerakligini o'rganish imkonini beradi.

3.4. Asosiy axborot ajratib olish vazifalari

Tadqiqotda quyidagi asosiy IE vazifalari o'rganildi va tahlil qilindi:

(1) Nomlangan birliklarni tanish (Named Entity Recognition — NER) — matndagi shaxslar, tashkilotlar, joylar, sanalar, pul miqdorlari va boshqa muhim birliklarni avtomatik aniqlash. Bu IE sohalarining eng ko'p o'rganilgan va amalda qo'llaniladigan vazifasi hisoblanadi.

(2) Munosabatlarni ajratib olish (Relation Extraction — RE) — matnda aniqlangan birliklar o'rtasidagi semantik aloqalarni topish. Masalan, 'Elon Musk Tesla kompaniyasining asoschisi' jumlasidan 'Elon Musk' va 'Tesla' o'rtasidagi 'asoschisi' munosabatini ajratib olish.

(3) Voqealarni ajratib olish (Event Extraction — EE) — matndagi hodisalarni, ularning ishtirokchilarini, vaqti va joyini aniqlash. Masalan, harbiy mojarolar, iqtisodiy hodisalar yoki tabiiy ofatlar haqidagi matnlardan faktlarni ajratib olish.
(4) Ma'lumotni to'ldirish (Slot Filling) — belgilangan shablonlarga mos axborot izlash. Masalan, kompaniya haqida ma'lumot to'plashda: nomi, asoschisi, joylashuvi, moliyaviy ko'rsatkichlari kabi maydonlarni avtomatik to'ldirish.

4. NATIJALAR

Tahlil natijasida matnli ma'lumotlardan avtomatik axborot ajratib olishning uchta asosiy generatsiyasi aniqlandi va ularning samaradorligi standart baholash o'lchovlari asosida qiyosiy tahlil qilindi. Har bir usulning kuchli va zaif tomonlari aniqlanib, qo'llanilish sohalari belgilandi. Birinchi generatsiya (qoida asosidagi tizimlar) ingliz tili uchun NER vazifasida F1 ko'rsatkichini 70-75% gacha ko'tara olgan, ammo yangi sohaga moslashish juda qiyin va mehnat talab qilgan. Ikkinchi generatsiya (statistik usullar, xususan CRF [9]) F1 ni 85-88% gacha oshirgan va yangi domenga moslashish osonlashgan. Uchinchi generatsiya (chuqur o'qitish, BERT [1] va undan keyingi modellar) hozirda ingliz tilida CoNLL-2003 testida 93%+ F1 ko'rsatkichiga erishmoqda.
Munosabatlarni ajratib olish (RE) sohasida chuqur o'qitish modellari NYT10 va DocRED kabi standart to'plamlarda 65-75% F1 ko'rsatkichiga erishgan. Bu vazifa NERga nisbatan murakkab, chunki birliklar o'rtasidagi munosabatlar juda xilma-xil bo'lishi mumkin. Zero-shot va few-shot o'qitish metodlari esa GPT-4 kabi katta til modellari yordamida ushbu ko'rsatkichlarni yanada oshirmoqda.
O'zbek tili uchun vaziyat murakkabroq: annotatsiyalangan korpuslarning kamligi, agglutinativ morfologiya va standartlashtirilgan NLP vositalarining cheklanganligi tufayli O'zbek tili uchun NER tizimlari hali rivojlanish bosqichida — 75-82% F1 darajasida ishlaydi. Tarmoq texnologiyalari [4] yordamida bu tizimlarni tezkor va ishonchli ishlashini ta'minlash mumkin. Ta'limni boshqarish tizimlari tasnifi [5] sohasida IE texnologiyalarini qo'llash o'quv materiallarini avtomatik tavsiflash, kalit tushunchalarni ajratib olish va bilimlar bazasini shakllantirish imkonini beradi. Masofaviy ta'lim sharoitida [6] axborot ajratib olish
texnologiyalari o'quv materiallarini avtomatik tizimlashtirish va shaxsiylashtirilgan ta'lim yo'llarini shakllantirish imkonini beradi.
Kompyuter grafikasi va raqamli tasvir qayta ishlash [7] sohasidagi yutuqlar IE natijalarini vizual tarzda taqdim etishda muhim rol o'ynaydi. Virtual reallik [8] texnologiyalari bilan birgalikda IE tizimlari ta'lim jarayonini yanada interaktiv va samarali qilish imkonini beradi.

5. MUHOKAMA

Tadqiqot natijalari shuni ko'rsatadiki, zamonaviy chuqur o'qitish texnologiyalari matnli axborot ajratib olishda insonning o'rtacha ko'rsatkichlariga yaqinlashib qolmoqda. Bir qator muhim masalalar tahlil qilindi.
Birinchi muhim masala — domenlararo ko'chirish (domain transfer). Model bir sohada, masalan tibbiyotda o'qitilsa, boshqa sohada — huquq yoki moliyada — samarasi keskin pasayadi. Bu muammoni hal qilish uchun domain adaptation va transfer learning metodlari qo'llanilmoqda. Xususan, sohaga oid ma'lumotlar bilan BERT modelini fine-tuning qilish yaxshi natijalar bermoqda.
Ikkinchi muhim masala — past resursli tillar muammosi. O'zbek, qozoq, tojik kabi tillar uchun katta annotatsiyalangan korpuslar mavjud emas, bu esa modellarni o'qitishni qiyinlashtiradi. Bu muammoni hal qilish uchun cross-lingual transfer learning — ingliz tilida o'qitilgan modelni boshqa tillarga moslash — texnikasi qo'llanilmoqda. Masalan, mBERT va XLM-RoBERTa modellari 100 dan ortiq tilda ishlash imkoniyatiga ega.
Uchinchi muhim jihat — modellarning interpretatsiyasi (explainability). Chuqur neyron

tarmoqlar 'qora quti' (black box) sifatida ishlaydi va ularning qanday qaror qabul qilishi aniq emas. Bu tibbiyot va huquq kabi mas'uliyatli sohalarda ishonchlilik muammosini keltirib chiqaradi. Kompyuter grafikasi va tasvir qayta ishlash [7] sohasidagi texnologiyalar modellarning vizual tushuntirish qobiliyatini oshirishga yordam beradi. Shu sababli, ilmiy hamjamiyat XAI (Explainable AI) yo'nalishida faol ish olib bormoqda. Shahrisabz davlat pedagogika institutida maktabgacha ta'lim sohasida IE texnologiyalarini qo'llash katta istiqbolga ega. Pedagogik adabiyotlardan kalit tushunchalarni ajratib olish, o'quv rejalarini tahlil qilish, bolalar rivojlanishiga oid ilmiy ma'lumotlarni tizimlashtirish va pedagogik tavsiyalar berish uchun IE texnologiyalaridan samarali foydalanish mumkin. Virtual reallik [8] va masofaviy ta'lim [6] bilan birgalikda bu texnologiyalar ta'lim jarayonini yanada boyitishi mumkin.

6. XULOSA

Matnli ma'lumotlardan avtomatik axborot ajratib olish texnologiyalari sun'iy intellektning eng jadal rivojlanayotgan yo'nalishlaridan biri hisoblanadi. Ushbu tadqiqot doirasida IE texnologiyalarining rivojlanish tarixi, asosiy usullari, samaradorligi va qo'llanilish sohalari keng ko'lamda o'rganildi. Tadqiqot natijalariga asoslanib, quyidagi asosiy xulosalar chiqarish mumkin:

1. Chuqur o'qitishga asoslangan modellar (BERT [1], RoBERTa, Attention [10]) qoida asosidagi va statistik usullardan sezilarli darajada ustun turadi hamda zamonaviy IE tizimlarining asosini tashkil etadi. F1 ko'rsatkichi bo'yicha chuqur o'qitish usullari 8-15 foiz ustunlikni ta'minlaydi.
2. O'zbek tili uchun axborot ajratib olish tizimlari hali rivojlanish bosqichida bo'lib, milliy annotatsiyalangan korpuslar yaratish va NLP vositalarini ishlab chiqish dolzarb masaladir. Cross-lingual transfer learning bu muammoni qisman hal etsa-da, mahalliy tilga moslashtirilgan modellar yaratish zarur.
3. Ta'lim [5,6], tibbiyot va huquq sohalarida IE texnologiyalarini joriy etish axborot izlash va qaror qabul qilish jarayonlarini sezilarli darajada optimallashtiradi. Maktabgacha ta'lim sohasida IE texnologiyalari pedagogik jarayonni avtomatlashtirishga muhim hissa qo'sha oladi.
4. Virtual reallik [8] va kompyuter grafika [7] sohasidagi yutuqlar bilan birgalikda IE texnologiyalari ta'lim jarayonini yanada interaktiv va samarali qilish imkonini beradi.
5. Kelajakdagi tadqiqotlar ko'p tilli modellar, past resursli tillar uchun transfer learning, interpretatsiyalanadigan IE tizimlari (XAI) yaratish va real vaqt rejimida ishlovchi tizimlarni takomillashtirishga yo'naltirilishi maqsadga muvofiqdir.

FOYDALANILGAN ADABIYOTLAR

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

Manning, C.D., & Schutze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

Kodirov, F.E., Axmatova, S.Z. (2019). LiFi-NEW NETWORK TECHNOLOGIES. Nauka i innovatsii v XXI veke: aktual'nye voprosy, otkrytiya i dostizheniya.

Qodirov, F., Allanazarova, A. (2025). Ta'limni boshqarish tizimlari tasnifi. Central Asian Journal of Multidisciplinary Research and Management Studies, 2(11), 113-117.

Qodirov, F. (2020). Masofaviy ta'limda o'qishning qulayliklari va kamchiliklari. Muhammad al-Xorazmiy nomidagi TATU Qarshi filiali.

Qodirov, F.E., Akbarova, D.A., Shokirov, S.H. (2021). Software for working with computer graphics and their tasks. Application of digital image processing fields, 57-58

Qodirov, F., Sa'dullayeva, M. (2025). Virtual reallik (VR) va kengaytirilgan reallik (AR). Molodye uchenye, 3(8), 139-144.

Lafferty, J., McCallum, A., & Pereira, F.C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML.

Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.