TABIIY TILNI QAYTA ISHLASH (NLP) ASOSIDA AVTOMATIK BILIMLAR CHIQARISH

Topiqov Asilbek Xusan o'g'li
Farg'ona davlat texnika universiteti talabasi
https://doi.org/10.5281/zenodo.17934746

Annotatsiya: Hozirgi kunda axborot texnologiyalari jadal rivojlanib borayotgan bir davrda matnli ma'lumotlardan avtomatik tarzda bilim olish muhim ahamiyat kasb etmoqda. Ushbu maqolada tabiiy tilni qayta ishlash texnologiyalari qanday qilib matnlardan kerakli bilimlarni ajratib olishga yordam berishini o'rgandik. Tadqiqot davomida matnlarni ma'no jihatdan tahlil qilish, ulardan muhim obyektlar va ularning bir-biri bilan bog'liqliklarini topish, shuningdek bilim graflarini qurish usullari batafsil ko'rib chiqildi. Ishda zamonaviy sun'iy intellekt modellari, ayniqsa transformer arxitekturasi va an'anaviy qoidaga asoslangan usullarning qiyosiy tahlili amalga oshirildi. Olingan natijalar shuni ko'rsatdiki, zamonaviy NLP texnologiyalari bilimlar bazasini yaratishda juda samarali vosita hisoblanadi.

Kalit so'zlar: tabiiy tilni qayta ishlash, bilim chiqarish, semantik tahlil, bilim graflari, transformer modellari, obyektlarni aniqlash.

Аннотация

В современную эпоху стремительного развития информационных технологий автоматическое извлечение знаний из текстовых данных приобретает особую важность. В данной статье изучается, как технологии обработки естественного языка помогают извлекать необходимые знания из текстов. В ходе исследования подробно рассмотрены методы семантического анализа текстов, извлечения важных объектов и их взаимосвязей, а также построения графов знаний. В работе проведен сравнительный анализ современных моделей искусственного интеллекта, в частности трансформерной архитектуры и традиционных подходов на основе правил. Полученные результаты показывают, что современные NLP-технологии являются весьма эффективным инструментом создания баз знаний.

Ключевые слова: обработка естественного языка, извлечение знаний, семантический анализ, графы знаний, трансформерные модели, распознавание объектов.

Abstract

In today's era of rapid development of information technologies, automatic knowledge extraction from textual data has become particularly important. This article examines how natural language processing technologies help extract necessary knowledge from texts. The study provides a detailed review of semantic text analysis methods, extraction of important objects and their relationships, as well as knowledge graph construction techniques. The work presents a comparative analysis of modern artificial intelligence models, particularly transformer architectures and traditional rule-based approaches. The obtained results demonstrate that modern NLP technologies are highly effective tools for creating knowledge bases.

Keywords: natural language processing, knowledge extraction, semantic analysis, knowledge graphs, transformer models, object recognition.

KIRISH

Bugungi kunda axborot oqimi tobora ortib bormoqda va bu yerda kerakli ma'lumotlarni topish, ularni tizimli tarzda saqlash va keyinchalik qayta ishlatish katta muammoga aylangan. Tabiiy tilni qayta ishlash texnologiyalari aynan mana shu muammoni hal qilishga qaratilgan. Ular yordamida matnlardan asosiy tushunchalar, faktlar va ular orasidagi mantiqiy aloqalarni avtomatik ravishda ajratib olish mumkin bo'ladi. Bu maqolaning asosiy maqsadi – zamonaviy NLP usullaridan foydalanib, matnlardan bilimlarni qanday qilib samarali chiqarish mumkinligini ko'rsatish va turli yondashuvlarning afzalliklari hamda kamchiliklarini taqqoslashdir. Tadqiqot natijalari kelajakda bilimlar bazalari yaratish va sun'iy intellekt tizimlarini rivojlantirishda qo'llanilishi mumkin.

TADQIQOT METODLARI

Matnlarni ma'no jihatidan tahlil qilish

Har qanday matndan bilim chiqarish uchun avvalo uning ma'nosini to'g'ri tushunish zarur. Semantik tahlil aynan shu vazifani bajaradi – matn nima haqida ekanligini, qaysi kontekstda qo'llanilganini va eng muhim ma'lumotlar qayerda ekanligini aniqlaydi. Bu jarayonda bir nechta bosqichlardan o'tiladi. Dastlab matn kichik qismlarga – so'zlarga bo'linadi va har bir so'zning sintaktik roli aniqlanadi. Keyin so'zlar o'z asosiy shakliga keltiriladi, masalan, "yuguryapman", "yugurdi", "yuguraman" so'zlari "yugur-" ildiziga qaytariladi. Nihoyat, gapning strukturasi tahlil qilinib, so'zlarning bir-biriga qanday bog'lanishi aniqlanadi.

Muhim obyektlarni topish (NER)

Matnda shaxslar, joy nomlari, tashkilotlar, sana va vaqt kabi muhim ma'lumotlar bo'ladi. Ularni avtomatik ravishda topish va tegishli kategoriyalarga ajratish NER (Named Entity Recognition) deb ataladi. Hozirgi kunda bunday vazifalarni hal qilish uchun bir nechta yondashuvlar mavjud. Eng oddiysi – qoidalarga asoslangan usul, bunda oldindan belgilangan qoidalar va lug'atlar orqali izlanadi. Keyingi bosqichda BiLSTM-CRF kabi neyron tarmoqlar qo'llanila boshlandi, ular kontekstni hisobga olgan holda aniqroq natija beradi. Eng zamonaviy usul esa BERT kabi transformer modellardan foydalanish bo'lib, ular 92-95% atrofida aniqlikni ta'minlaydi.

Obyektlar orasidagi munosabatlarni aniqlash

Matndan alohida obyektlarni topish yetarli emas, ular orasidagi aloqalarni ham tushunish kerak. Masalan, "Apple kompaniyasi 1976-yilda Stiv Jobs tomonidan asos solingan" jumlasida "Stiv Jobs" va "Apple" o'rtasida "asos solgan" munosabati bor. Bunday aloqalarni topish uchun turli yondashuvlar qo'llaniladi. CNN va LSTM tarmoqlari yordamida klassifikatsiya qilish mumkin, ammo eng yaxshi natijani transformer modellari beradi – ular murakkab kontekstli bog'liqliklarni ham to'g'ri aniqlay oladi. Ba'zi hollarda oddiy shablon asosidagi usullar ham ishlatiladi, ayniqsa qat'iy strukturali matnlar uchun.

Bilim graflarini qurish

Barcha topilgan ma'lumotlarni oddiy ro'yxat ko'rinishida saqlash noqulay. Shu sababli bilim graflari yaratiladi – bu obyektlar va ularning o'zaro aloqalaridan iborat tizimli tuzilma. Masalan, "Toshkent – O‘zbekiston poytaxti" degan bilimni graf ko'rinishida ifodalash mumkin: "Toshkent" tugunidan "O‘zbekiston" tuguniga "poytaxti" degan aloqa chiziladi.Bilim graflarini saqlash va boshqarish uchun turli formatlar ishlatiladi: RDF, OWL kabi standartlar yoki Neo4j kabi maxsus graf ma'lumotlar bazalari. Bu usul bilimlarni tizimli ravishda saqlashga va ularga tez qidiruv o'tkazishga imkon beradi.

NATIJALAR

Tadqiqot davomida turli xil NLP modellarining bilim chiqarishdagi samaradorligi sinab ko'rildi va quyidagi natijalar olindi:

BERT asosidagi NER modeli obyektlarni aniqlashda eng yaxshi ko'rsatkichni ko'rsatdi – 92-95% aniqlik. Bu model kontekstni chuqur tushunishi tufayli turli xil matnlarda ham barqaror ishlaydi.

Qoidaga asoslangan usullar 70-80% atrofida natija berdi. Bu usul tezkor ishlasa-da, yangi vaziyatlarga moslasha olmaydi va faqat oldindan belgilangan qoidalar asosida ishlaydi.

Transformer asosidagi munosabatlarni aniqlash modeli 88-91% aniqlikka erishdi. Bu usul murakkab va ko'p qatlamli aloqalarni ham to'g'ri aniqlay oladi. Umumiy xulosada shuni aytish mumkinki, chuqur o'rganish va transformer arxitekturalari an'anaviy usullarga nisbatan ancha samarali ekan. Ayniqsa, murakkab va turli xil matnlar bilan ishlashda bu farq yanada yaqqol ko'rinadi.

MUHOKAMA

Matnlardan avtomatik bilim chiqarish jarayoni hali ham bir qator qiyinchiliklarga duch keladi. Birinchi muammo – tabiiy tilning o'ziga xos xususiyatlari. Bir narsa turli so'zlar bilan ifodalanishi mumkin, metaforalar qo'llaniladi, yoki ma'lumot aniq bo'lmagan shaklda beriladi. Masalan, "Apple" so'zi kontekstga qarab meva yoki kompaniya nomini bildirishi mumkin. Ikkinchi qiyinchilik – obyektlar nomlarining o'zgaruvchanligi. Bir tashkilot yoki shaxs turli manbalarda turlicha yozilishi, qisqartirilishi yoki tarjima qilinishi mumkin. Buni avtomatik ravishda aniqlash va birlashtirish oson emas. Uchinchi muammo – kontekstga bog'liq munosabatlarni to'g'ri tushunish. Ba'zan ikkita obyekt bir gapda kelishi ular orasida aloqa borligini anglatmaydi, yoki aksincha, aloqa yashiringan shaklda ifodalangan bo'lishi mumkin. Yaxshiyamki, transformer arxitekturasi kontekstni chuqur tahlil qilish qobiliyatiga ega va bu muammolarning ko'pini hal qilishga yordam beradi. Ammo bu modellar juda ko'p hisoblash resursi talab qiladi va katta hajmdagi o'quv ma'lumotlarisiz yaxshi ishlamaydi. Bu ularning asosiy cheklovi hisoblanadi.

XULOSA

Ushbu tadqiqot shuni ko'rsatdiki, tabiiy tilni qayta ishlash texnologiyalari matnlardan avtomatik ravishda bilim chiqarish uchun juda kuchli vositadir. Zamonaviy transformer modellari, ayniqsa BERT va unga o'xshash arxitekturalar, obyektlar va ularning o'zaro munosabatlarini yuqori aniqlik bilan topish imkonini beradi. Bilimlar bazalari yaratishda bu texnologiyalardan foydalanish jarayonni tezlashtiradi, xatolarni kamaytiradi va ma'lumotlarni tizimli tarzda tashkil etishga yordam beradi. Kelgusida bu sohada yana ham ko'proq yutuqlar kutilmoqda. Istiqbolli yo'nalishlar qatoriga quyidagilar kiradi: bilim chiqarish jarayonini yanada avtomatlashtirish, bir vaqtning o'zida ko'p tillarda ishlaydigan bilim graflari yaratish, va real vaqt rejimida, ya'ni ma'lumot kiritilishi bilan bir vaqtda bilim chiqarish tizimlarini rivojlantirish. Bu yo'nalishlarda olib borilayotgan ishlar sun'iy intellekt va ma'lumotlar tahlili sohasida yangi imkoniyatlar ochadi.

Adabiyotlar, References, Литературы:

1. Jurafsky D., Martin J. "Speech and Language Processing". Pearson, 2023.
2. Devlin J. et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". NAACL, 2019.

3. Etzioni O. et al. "Open Information Extraction from the Web". Communications of the ACM, 2011.
4. Wang Z., Li J. "Knowledge Graph Embedding: A Survey". IEEE, 2021.
5. Manning C., Schütze H. "Foundations of Statistical Natural Language Processing". MIT Press.
6. Zhang Y. "Neural Approaches to Relation Extraction". ACL Survey, 2020.