LSTM ASOSIDA MATNLARDAN MUHIM OBYEKTLARNI AJRATIB OLISH

Muhamediyeva D.T., Mamatov A.A.

«Toshkent irrigatsiya va qishloq xo'jaligini mexanizatsiyalash muhandislari instituti» milliy tadqiqot universiteti, Namangan davlat universiteti.

DOI: **https://doi.org/10.5281/zenodo.20214972**

Annotatsiya. *Ushbu maqolada matnlardan muhim obyektlarni ajratib olish va bilimlar bazasini yaratish uchun chuqur o'rganishga asoslangan model taklif etiladi. Nomlangan obyektlarni aniqlash (Named Entity Recognition, NER) vazifasi uchun uzoq qisqa muddatli xotira (Long Short-Term Memory, LSTM) modeli qo'llanilgan. Ma'lumotlar oldindan qayta ishlanib, tokenizatsiya va one-hot kodlash usullari orqali raqamli shaklga o'tkaziladi. Model turli obyekt turlarini (shaxs nomlari, sanalar, joy nomlari) ajratib olish uchun o'qitiladi va baholanadi. Eksperimental natijalar modelning samaradorligini ko'rsatadi va turli parametrlarning ta'siri tahlil qilinadi.*

Kalit so'zlar: *Matnlarni qayta ishlash, nomlangan obyektlarni aniqlash, LSTM, mashinaviy o'rganish, bilimlar bazasi, tokenizatsiya, one-hot kodlash.*

Аннотация. *В данной статье предлагается модель на основе глубокого обучения для извлечения важных объектов из текста и построения базы знаний. Для задачи распознавания именованных сущностей (NER) используется модель долговременной кратковременной памяти (LSTM). Данные предварительно обрабатываются и преобразуются в цифровую форму с помощью токенизации и one-hot кодирования. Модель обучается и оценивается для извлечения различных типов объектов (имена, даты, названия мест). Экспериментальные результаты показывают эффективность модели, а также анализируется влияние различных параметров.*

Ключевые слова: *Обработка текста, распознавание именованных сущностей, LSTM, машинное обучение, база знаний, токенизация, one-hot кодирование.*

Abstract. *This paper proposes a deep learning-based model for extracting important objects from texts and building a knowledge base. A long short-term memory (LSTM) model is used for the task of Named Entity Recognition (NER). The data is preprocessed and converted into digital form using tokenization and one-hot encoding. The model is trained and evaluated to extract different*

object types (personal names, dates, place names). Experimental results show the effectiveness of the model and the effects of various parameters are analyzed.

Keywords: *Text processing, named entity recognition, LSTM, machine learning, knowledge base, tokenization, one-hot encoding.*

1.Kirish

Hozirgi kunda matnlarni tahlil qilish va ulardan bilimlar bazasini yaratish ko‘plab sohalarda muhim ahamiyat kasb etmoqda. Jumladan, tabiiy tilni qayta ishlash (NLP) usullari yordamida matnlardan muhim ma’lumotlarni ajratib olish imkoniyati mavjud. Nomlangan obyektlarni aniqlash (NER) – bu matndagi shaxs ismlari, sanalar, joy nomlari kabi muhim elementlarni aniqlashga qaratilgan muhim yo‘nalishdir. Ushbu tadqiqotda LSTM modeli asosida matnlardan muhim obyektlarni ajratib olish va ularni bilimlar bazasiga joylashtirish usuli taklif etiladi.

LSTM – chuqur o‘rganish usullaridan biri bo‘lib, u ketma-ketliklarni o‘rganish va uzoq muddatli bog‘liqliklarni saqlashda samarali ishlaydi. Ushbu model turli sohalarda, jumladan, matnni qayta ishlash va tabiiy til tahlilida muvaffaqiyatli qo‘llanilmoqda. Tadqiqotimizda LSTM yordamida matnlardan nomlangan obyektlarni ajratib olish jarayoni amalga oshiriladi.

2.Usullar va yondashuvlar

Matnlardan nomlangan obyektlarni ajratib olish uchun bir necha bosqich bajariladi. Dastlab, matnlarni raqamli shaklga o‘tkazish uchun tokenizatsiya jarayoni amalga oshiriladi. Bu jarayonda matn alohida so‘zlarga ajratilib, har bir so‘zga indeks tayinlanadi. Keyinchalik, matnlarning uzunligi bir xil darajaga keltirish uchun padding usuli qo‘llaniladi. Kichikroq uzunlikdagi matnlarga maxsus to‘ldirish belgilarini qo‘shish orqali barcha matnlarning uzunligi tenglashtiriladi. Shundan so‘ng, obyektlarni modelga moslashtirish maqsadida one-hot kodlash amalga oshiriladi. Bu bosqichda obyektlarga mos raqamli ifodalar tayinlanadi va ular modelga kiritish uchun tayyorlanadi.

Nomlangan obyektlarni aniqlash jarayonida har bir so‘zga mos obyekt turi belgilanadi. Bu jarayonda model matndagi shaxs nomlari, sanalar va joy nomlarini ajratib oladi. Ajratilgan obyektlar turli kategoriyalar bo‘yicha tasniflanadi va natijalar model tomonidan qayta ishlanadi. Bu bosqichda shaxs nomlari ("Alice", "Bob", "Charlie"), sanalar ("January 5th, 2023", "March 15th, 2023") va joy nomlari ("New York") kabi obyektlarni aniqlash maqsad qilib qo‘yiladi.

Nomlangan obyektlarni aniqlash uchun chuqur o‘rganish usuli – LSTM modelidan foydalaniladi. Model bir necha asosiy qatlamlardan tashkil topgan bo‘lib, birinchi qatlam so‘zlarni

raqamli vektorlarga o‘tkazuvchi embedding qatlami hisoblanadi. Ushbu qavatdan keyin LSTM qatlami kelib, u vaqt bo‘ylab so‘zlarning o‘zaro bog‘liqligini saqlaydi va tahlil qiladi. Modelning haddan tashqari moslashuvini oldini olish uchun dropout qatlami qo‘shiladi. Ushbu qatlam neyronlarning bir qismini vaqtincha o‘chirib, modelni umumlashuvchan qilishga yordam beradi. Modelning yakuniy qatlami esa softmax funksiyasiga asoslangan bo‘lib, har bir so‘zning tegishli kategoriya ehtimolini hisoblab chiqadi.

Modelni o‘qitish jarayonida categorical_crossentropy yo‘qotish funksiyasi qo‘llaniladi. Bu funksiya modelning nomlangan obyektlarni to‘g‘ri tasniflash darajasini oshirishga xizmat qiladi. Model vaznlari Adam optimizer yordamida yangilanadi va model 10 epoch davomida o‘qitiladi. Har bir o‘qitish qadamida mini-batch usuli qo‘llanilib, har bir qadamda 2 ta namunadan iborat ma’lumotlar ishlatiladi.

Model samaradorligini baholash uchun aniqlik (accuracy), F1-score va xatolik matritsasi hisoblab chiqiladi. Aniqlik modelning umumiy to‘g‘ri tasniflangan obyektlar ulushini o‘lchashga yordam beradi. F1-score esa aniqlik va to‘g‘rilik muvozanatini hisoblash imkonini beradi. Xatolik matritsasi orqali modelning noto‘g‘ri va to‘g‘ri tasniflash darajasi tahlil qilinadi. Bu natijalar asosida modelning samaradorligi aniqlanadi va uning ishlash sifati baholanadi.

3. Natigalar.

Modelni o‘qitish va sinov jarayonlaridan so‘ng, uning samaradorligi bir nechta mezonlar asosida baholandi. Aniqlik (Accuracy), F1-score va xatolik matritsasi tahlil qilindi. Modelni o‘qitish davomida aniqlik oshib bordi va yo‘qotish funksiyasi qiymati kamaydi. 10 epoch davomida o‘qitish natijalariga ko‘ra, trening aniqligi 92.4% ni, test aniqligi esa 89.7% ni tashkil etdi. Shu bilan birga, trening yo‘qotish funksiyasi qiymati 0.19, test yo‘qotish funksiyasi esa 0.23 ga teng bo‘ldi. Bu natijalar shuni ko‘rsatadiki, model trening to‘plami bo‘yicha yaxshi o‘rgangan, test to‘plamida esa biroz pastroq natija qayd etgan. Biroq, farq katta emas va bu modelning ortiqcha moslashib qolmaganligini (overfitting yo‘qligini) bildiradi.

Model matndagi shaxs nomlari, sanalar va joy nomlarini ajratib olishda quyidagi natijalarga erishdi: shaxs nomlarini aniqlash aniqligi 94.1%, sanalarni aniqlash aniqligi 90.3%, joy nomlarini aniqlash aniqligi esa 87.5% ni tashkil etdi. Shaxs nomlarini ajratib olishdagi aniqlik eng yuqori bo‘ldi. Bu shuni anglatadiki, model shaxs nomlarini yaxshi farqlaydi. Sanalar va joy nomlarining aniqlik darajasi biroz pastroq bo‘lsa ham, qoniqarli natija qayd etildi.

Xatolik matritsasiga ko‘ra, model shaxs nomlarini deyarli mukammal aniqlagan, biroq sanalarni va joy nomlarini noto‘g‘ri tasniflash holatlari mavjud. Ayniqsa, ba’zi sanalar joy nomlari sifatida noto‘g‘ri belgilangan. Masalan, "March 15th, 2023" iborasi ayrim holatlarda joy nomi sifatida tasniflangan. Bu xatoliklar modelning o‘quv ma’lumotlari hajmi va ba’zi matnlardagi murakkab ifodalarga bog‘liq bo‘lishi mumkin. Ushbu muammoni bartaraf etish uchun o‘quv to‘plamini kengaytirish yoki sanalarni yanada aniq aniqlash uchun maxsus qoidalarni (rule-based features) qo‘shish tavsiya etiladi.

Modelning umumiy tasniflash sifatini baholash uchun F1-score quyidagicha bo‘ldi: shaxs nomlari uchun 93.8%, sanalar uchun 88.7%, joy nomlari uchun esa 85.9%. Bu natijalarga asoslanib, model shaxs nomlarini mukammal tasniflashga yaqin natijaga ega, sanalar va joy nomlari bo‘yicha esa yana takomillashtirish talab qilinishi mumkin.

Quyidagi jadval modelning shaxs nomlari, sanalar va joy nomlarini ajratib olishdagi samaradorligini ifodalaydi:

Kategoriyalar	Aniqlik (%)	F1-score (%)	Xatolik (%)
Shaxs nomlari	94.1	93.8	5.9
Sanalar	90.3	88.7	9.7
Joy nomlari	87.5	85.9	12.5

Jadvaldan ko‘rinib turibdiki, shaxs nomlarini aniqlashning aniqlik va F1-score ko‘rsatkichlari eng yuqori bo‘ldi. Sanalar va joy nomlarini aniqlash natijalari nisbatan pastroq bo‘lib, bu ba’zi xatoliklar bilan bog‘liq.

4.Xulosa

Ushbu tadqiqotda LSTM asosida matnlardan shaxs nomlari, sanalar va joy nomlarini ajratib olish modeli yaratildi va uning samaradorligi tahlil qilindi. Tahlillar shuni ko‘rsatadiki, model o‘quv jarayonida matnning kontekstual ma’nosini yaxshi o‘zlashtirgan bo‘lsa-da, ba’zi sanalar va joy nomlarini ajratishda xatoliklarga yo‘l qo‘ygan. Bu xatoliklar asosan ma’lumotlar to‘plamidagi o‘ziga xosliklar yoki modelning ba’zi murakkab kontekstlarni tushunishdagi cheklovlari bilan bog‘liq. Ushbu ish natijalari matnlarni avtomatik qayta ishlash sohasida amaliy ahamiyatga ega bo‘lib, turli sohalarda, jumladan, hujjatlar tahlili, huquqiy hujjatlar va ilmiy maqolalar bo‘yicha ma’lumotlarni saralash uchun foydali bo‘lishi mumkin.

Foydalanilgan adabiyotlar ro‘yxati

1. Duppati, S. K. ., & Babu, A. (2023). Named Entity Recognition for English Language Using Deep Learning Based Bi Directional LSTM-RNN. International Journal on Recent and Innovation Trends in Computing and Communication, 11(5), 330–337. https://doi.org/10.17762/ijritcc.v11i5.6621

2. A. Mohamed and N. Jaitly, 2013, “Hybrid speech recognition with deep bidirectional LSTM,”, IEEE, 2013, pp. 273–278.