O‘ZBEK TILI MATNLARIDA SAVOLLARNI ANIQLASHNING AVTOMATLASHTIRILGAN TIZIMI UCHUN MASHINANI O‘RGANISH YONDASHUVI

Sharipov Maksud Siddiqovich,
Texnika fanlari nomzodi, dotsent
maqsbek72@gmail.com
Urganch davlat universiteti

Adinayev Xushnudbek Saylboyevich,
hushnudbek.adinaev@gmail.com
TATU Urganch filiali o‘quv-uslubiy boshqarma boshlig‘i

Yusupova Mukaddas Muratovna,
1-kurs magistranti
mukaddas0783@gmail.com
Urganch Ranch texnologiya universiteti

Qadamboyeva Durdona Dilshod qizi,
1-kurs magistrant
durdona06@gmail.com
Urganch davlat universiteti

ANNOTATSIYA. Bu ishda biz o‘zbek tili matnlari undov gaplarini tahlil qilishning qoidaga asoslangan algoritmini ishlab chiqishni ko‘rib chiqamiz. O‘zbek tili kam resursli til hisoblangaligi sababali hozirgacha bunday algoritmlar ishlab chiqilmagan. O‘zbek tili matnlaridagi gaplarning undov belgilari to‘g‘ri yoki noto‘g‘ri qo‘yilganligini aniqlash masalasini yechishning qoidaga asoslangan algoritmi ishlab chiqilgan va qoidalar bazasi yaratilgan. Til - bu muloqot usuli bo‘lib, uning yordamida biz gapirish, o‘qish va yozishimiz mumkin. Misol uchun, biz o‘ylaymiz, biz tabiiy tilda qarorlar, rejalar va boshqalarni qabul qilamiz; aniq, so‘z bilan. Biroq, ushbu AI davrida bizni duch keladigan katta savol shundaki, biz kompyuterlar bilan xuddi shunday tarzda muloqot qila olamizmi? Boshqacha qilib aytadigan bo‘lsak, inson o‘z tabiiy tilida kompyuterlar bilan muloqot qila oladimi? NLP dasturlarini ishlab chiqish biz uchun qiyin, chunki kompyuterlar tuzilgan ma’lumotlarga muhtoj, lekin inson nutqi tuzilmagan va tabiatan ko‘pincha noaniq.

ABSTRACT. In this work, we consider the development of a rule-based algorithm for analyzing exclamatory sentences in Uzbek texts. Since Uzbek is considered a low-resource language, such algorithms have not been developed so far. A rule-based algorithm for solving the problem of determining whether exclamation marks are placed correctly or incorrectly in sentences in Uzbek texts has been developed and a rule base has been created. Language is a means of communication with which we can speak, read, and write. For example, we think, we make decisions, plans, and so on in natural language; specifically, with words.

However, the big question that faces us in this AI era is whether we can communicate with computers in the same way? In other words, can humans communicate with computers in their natural language? Developing NLP programs is difficult for us because computers need structured information, but human speech is unstructured and often ambiguous by nature.

АННОТАЦИЯ. В данной работе рассматривается разработка алгоритма анализа восклицательных предложений в узбекских текстах на основе правил. В связи с тем, что узбекский язык считается языком с ограниченными ресурсами, такие алгоритмы пока не разработаны. Разработан алгоритм на основе правил и создана база правил для решения задачи определения правильности или неправильности расстановки восклицательных знаков в предложениях узбекских текстов. Язык — это способ общения, посредством которого мы можем говорить, читать и писать. Например, мы думаем, принимаем решения, планируем и т. д. на естественном языке; ясно, словами. Однако главный вопрос, стоящий перед нами в эпоху искусственного интеллекта, заключается в том, сможем ли мы общаться с компьютерами таким же образом? Другими словами, могут ли люди общаться с компьютерами на их естественном языке? Нам сложно разрабатывать программы обработки естественного языка, поскольку компьютерам нужны структурированные данные, а человеческая речь по своей природе неструктурирована и часто неоднозначна.

Shu ma’noda aytishimiz mumkinki, tabiiy tilni qayta ishlash (NLP) bu kompyuter fanining, ayniqsa sun’iy intellektning (AI) kichik sohasi bo‘lib, u kompyuterlarga inson tilini tushunish va qayta ishlashga imkon beradi. Texnik jihatdan, NLP ning asosiy vazifasi tabiiy tildagi katta hajmdagi ma’lumotlarni tahlil qilish va qayta ishlash uchun kompyuterlarni dasturlashdir.

Kalit so‘zlar. *Tinish belgilari, NLP, so‘roq belgisi, so‘roq gaplar, QA, SVM, LSTM, KNN.*

Kirish

Dunyo bo‘ylab har soniyada turli savollar beriladi. Ushbu savollarga javob topish uchun biz ushbu mavzu bo‘yicha ma’lumotga ega bo‘lishimiz kerak. Bu ma’lumotlar kitoblarda, qog‘ozlarda va eng muhimi internetda. Internet va Internetning takomillashuvi bilan tobora ko‘proq ma’lumotlar mavjud bo‘lib, hamma uchun topish osonroq bo‘ldi. Muayyan savolga javob topishimiz kerak bo‘lganda, biz ushbu savolga tegishli ba’zi ma’lumotlardan javob olamiz. Biz ushbu mavzudagi mavjud ma’lumotlarni qo‘lda o‘qiymiz va javob topishga harakat qilamiz. Qo‘shimcha ma’lumot Internet va Internetda mavjud bo‘lgani uchun, agar

biz ba’zi hujjatlarni o‘qish va ularga javob olish jarayonini avtomatlashtira olsak, ko‘proq savollarga avtomatik javob berishimiz mumkin bo‘ladi. Ushbu maqolada biz bengal tilida avtomatlashtirilgan savollarga javob berish tizimini (QA) yaratish uchun qo‘llanilgan dastlabki qadamlar va qo‘llaniladigan usullarni tasvirlab beramiz. Savollarni aniqlash QA tizimida katta maqsadga xizmat qiladi [1]. Odamlar ko‘pincha tabiiy tilda savollar berishadi, ular grammatik qoidalarga rioya qilmaydilar. Shunday qilib, jumla savolmi yoki yo‘qmi, avvalo aniqlanishi kerak. Odamlar tomonidan berilgan savollarning turlicha shakli aniqlashni qiyinlashtiradi [2]. Berilgan savolga hujjatlardan javob olish uchun tizim ushbu savolga ishlov berishi kerak va uni qayta ishlashning birinchi bosqichi savollarni aniqlashdir, shuning uchun bu savolga javob berish tizimining birinchi bosqichidir [3]. Savolni aniqlash qismidan so‘ng tizim savolni oldindan belgilangan toifalar toifasiga ajratishi kerak. Ushbu tasnif kutilgan javob turini taxmin qilishga va qidiruv domenini qisqartirishga yordam beradi. Demak, tasniflash QA tizimining muhim qismidir.

Savollarni aniqlash - bu savollarga javob berish tizimining dastlabki vazifalari. Savollarni aniqlash, QA tizimining birinchi bosqichi SVM, Logistik regressiya,
K-Neighbor klassifikatori, Ko‘p qatlamli Perceptron va LSTMni qo‘llash orqali amalga oshirildi. Biz chiziqli yadro hiylasi bilan SVM uchun eng yaxshi ishlashga erishdik. 2000 xususiyatli so‘zlar uchun biz 1,4% xatolik darajasiga ega bo‘ldik, bu biz Bengal tili uchun ishlatgan yondashuvlar orasida eng yaxshisidir. LSTM algoritmi uchun biz 3,22% xatolik darajasiga ega bo‘ldik. LSTM katta va yaxshiroq ma’lumotlar to‘plami bilan istiqbolli ko‘rinadi.

Ma’lumotlar to‘plamini tayyorlash va tahlil qilish

Biz jami 8000 qatorli o‘zbek tili matnidan iborat ma’lumotlar to‘plamini tayyorladik. Bizning ma’lumotlar to‘plamimizda 3100 ta savol va 4900 ta savol bo‘lmagan. Ushbu ma’lumotlar turli veb-saytlardan to‘plangan. O‘bekistonning mashhur shoirlari haqidagi ba’zi maqolalar Vikipediyadan to‘plangan. Shuningdek, biz matnlardan savollarga javoblar juftligini to‘pladik. Ushbu ma’lumotlar qo‘lda tayyorlangan. Keyin biz ushbu ma’lumotlarni savolni aniqlash maqsadida belgilashimiz kerak edi. Biz ma’lumotlarni ikki toifaga ajratdik: savol va savolsiz. Biz ishimiz uchun butun ma’lumotlar to‘plamini qo‘lda tozaladik. Keyin biz keyingi ishlov berish uchun jumlalarni tokenizatsiya qildik. Keyin biz olgan jumlalarda so‘zlarni leksema qildik. Keyin countvectorizer yordamida so‘zlarni vec ga aylantirdik. Shundan so‘ng, TF/IDF yordamida biz so‘zlarni to‘pladik va eng qimmatli so‘zlarni oldik,

Xususiyatlarni chiqarish

Jumlalarni tokenizatsiya qilish: Birinchidan, biz har bir jumlani keyingi qayta ishlash uchun to‘plangan parchadan ajratdik. Jumlalarni tokenizatsiya qilishda yaxshi natijaga erishish uchun ko‘p narsalar ko‘rib chiqildi.

So‘zlarni tokenizatsiya qilish: jumlalar tokenizatsiyasidan so‘ng biz ushbu jumlalarda so‘z tokenizatsiyasini qo‘lladik. So‘zlarni tokenizatsiya qilish orqali biz jumlalarning har bir so‘zini ajratdik va ularni keyingi qayta ishlash uchun saqladik.

Wh-so‘z: So‘z tokenizatsiyasidan keyin vazifamiz biz uchun osonlashdi. Bengal tilida xuddi ingliz tilidagi wh so‘zi kabi ba’zi so‘zlar mavjud. Bu so‘zlar biz uchun juda foydali bo‘ldi. Ular asosan savollarda uchraydi. Shunday qilib, bu Wh so‘zlari so‘z tokenizatsiyasidan keyin eng ko‘p uchraydigan so‘zlar ekanligi aniq.

Biz bu so‘zlarga boshqa so‘zlarga qaraganda ustunlik berdik. Wh-so‘zga ko‘proq ustunlik berilgandan so‘ng, bu so‘zlar sodir bo‘lsa, gapning savol bo‘lish ehtimoli ortadi. Jadval II Bengal Wh-so‘zlarning ba’zi misollarini ko‘rsatadi.

Trening va tan olish

Trening maqsadida biz SVM, MLP, KNN, Logistic Regression va LSTM algoritmlaridan foydalandik. Ushbu algoritmlar 8000 Bengal matnidan iborat katta ma’lumotlar to‘plamimizda ishga tushirildi. Dastlab biz SVM algoritmini ishga tushirdik va SVM ning 3 xil yadrosidan foydalandik. Yadrolar Lineer, Polynomial va RBF. Biz MLP dan foydalandik, bu erda yashirin qatlamlar o‘lchami 12 va hal qiluvchi lbfg bo‘lgan. KNN biz ishlatgan yana bir mashinani o‘rganish algoritmidir. Barg o‘lchami 30 ta bo‘lib, algoritm avtomatik hisoblanadi. KNN metrikasi minkowski, ishlatilgan og‘irliklar esa masofa edi. Maqsadimiz bo‘yicha matn ma’lumotlar to‘plamini tasniflash uchun biz mashinani o‘rganish algoritmini LR (Logistik regressiya) ishlatdik. LSTM matn ma’lumotlarini tasniflash uchun eng ko‘p ishlatiladigan algoritmlardan biridir. Biz foydalangan oxirgi algoritm LSTM. LSTM modelini amalga oshirishda xususiyatli so‘zlar soni 20000 tani tashkil etdi. Ketma-ket model qo‘llanildi va kirish uzunligi 42. Bizning modelimizda foydalanilgan jami qatlamlar 21 ta, faollashtirish funksiyasi sigmasimon. SVM uchun eng yaxshi natijaga erishiladi. Agar ma’lumotlar to‘plami yaxshilansa, LSTM bilan ishlash yaxshilanadi.

Metodologiyasi

O‘zbek tilida yozilgan matnlarda savol gaplarni avtomatik tarzda aniqlash ko‘plab tizimlar uchun muhim: chatbotlar, matn tahlili, avtomatik javob berish tizimlari, savol-javob platformalari va boshqalar. Biroq, O‘zbek tili uchun NLP resurslar kamligi bu vazifani murakkablashtiradi.

Metodologiya bosqichlari

1. Ma’lumotlar to‘plami tayyorlash

- O‘zbek tilidagi matnlar (forumlar, maqolalar, chatlar) to‘planadi.
- Har bir gap **"savol"** yoki **"savol emas"** deb belgilanadi.
- Agar mavjud bo‘lsa, mavjud datasetlardan (masalan, O‘zbekcha OpenCorpora yoki O‘zNLP) foydalaniladi.

2. Matnni oldindan qayta ishlash

- Gaplarni ajratish (tokenizatsiya).
- Belgilarni tozalash (punktuatsiya, emoji, html taglar).
- Harf o‘lchamlarini bir xil qilish (masalan, kichik harflarga o‘tkazish).
- Lemmatizatsiya yoki stemming (ixtiyoriy).

3. Belgilarni (Feature) ajratish

Matndan quyidagi belgilar chiqariladi:

- Savol so‘zlari (kim, nima, qachon, qayerda, nega, nima uchun, qanday)
- Gap oxirida ? belgisi mavjudligi.
- -mi, -chi kabi savol qo‘shimchalari.
- Gap uzunligi.
- Gapdagi fe’l yoki yordamchi fe’llar.

4. Model tanlash va qurish

Variant A: Qoida asosidagi yondashuv (Rule-based)

- Regex (muntazam ifodalar) orqali savolni aniqlash.
- Foydalanuvchiga tezkor natija berish uchun soddalashtirilgan algoritm.

Variant B: Mashina o‘rganish (ML) asosida

- Klassifikatsiya modellari:
  - Naive Bayes
  - Logistic Regression
  - Random Forest
- Belgilar asosida model o‘rgatiladi va test qilinadi.

Variant C: Transformer modeli (Deep Learning)

- O‘zbek tilida o‘qitilgan BERT model (masalan, uzbekBERT, XLM-R, yoki mBERT) fine-tuning qilinadi.
- Klassifikatsiya uchun BERTning so‘nggi qatlamiga softmax layer qo‘shiladi.

5. Modelni baholash

Modellar quyidagi metrikalar bo‘yicha baholanadi:

- **Aniqlik (Accuracy)**
- **Aniqlovchanlik (Precision)**
- **Qamrov (Recall)**
- **F1-ko‘rsatkich**

Kross-valikatsiya usuli bilan turli matn turlari (chat, maqola, ijtimoiy tarmoq)da sinovdan o‘tkaziladi.

Natija

Ma’lumotlar to‘plamimizni muvaffaqiyatli o‘qitishdan so‘ng biz yuqori aniqlik va kamroq xatolikka erishdik, bu juda qoniqarli edi. LSTM dan tashqari har bir algoritmda bizning ma’lumotlar poezdi testi bo‘linishi 80/20 ni tashkil etdi, bu umumiy ma’lumotlarning 80% ta’lim uchun, 20% esa sinov uchun ishlatilgan. LSTM algoritmi uchun bizning o‘quv va sinov ma’lumotlarimiz mos ravishda 66,67% va 33,33% edi.

Xulosa

Biz bengal tili uchun QA tizimining dastlabki ishlarini bajardik. To‘liq QA tizimi uchun keyingi qadamlarni bajarish kerak. Buning uchun ko‘p kuch va etarlicha yaxshi ma’lumotlar to‘plami kerak. Agar ma’lumotlar to‘plami ko‘paytirilsa, biz LSTM yordamida yanada yaxshi ishlashga umid qilamiz. Yaxshiroq natijaga erishish uchun boshqa usullarni ham sinab ko‘rish mumkin. Savollarni tasniflash, ma’lumot olish va javoblarni olish QA tizimining qismlari bo‘lib, Bengal tabiiy tilini qayta ishlash bo‘yicha katta tadqiqot sohasi bo‘lishi mumkin. Savollarni aniqlash - bu savollarga javob berish tizimining kichik qismi va boshqa ko‘plab muammolar. U boshqa ko‘plab tadqiqot sohalari uchun eshikni ochishi mumkin. Ta’riflangan savollarni aniqlash tizimi yaxshiroq ma’lumotlar to‘plami bilan yaxshi ishlash imkonini beradi.

Foyadalanilgan adabiyotlar:

1. M. A. , M. N. , A. D. Mahmudov N., O*‘zbek tili meyorlari (Punktuatsiya)*, vol. 1. Tashkent: Zamin nashr, 2021.
2. J. Pokrywka, “Punctuation Prediction for Polish Texts using Transformers,” Feb. 2024. doi: 10.48550/arXiv.2410.04621.
3. P. Żelasko, P. Szymański, J. Mizgajski, A. Szymczak, Y. Carmiel, and N. Dehak, “Punctuation Prediction Model for Conversational Speech,” Feb. 2018, pp. 2633–2637. doi: 10.21437/Interspeech.2018-1096.
4. M. S. Sharipov, H. S. Adinaev, and E. R. Kuriyozov, “Rule-Based Punctuation Algorithm for the Uzbek Language,” in *International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices, EDM*, 2024, pp. 2410 – 2414. doi: 10.1109/EDM61683.2024.10615061.
5. M. Bayraktar, B. Say, and V. Akman, “An analysis of English punctuation: the special case of comma,” International Journal of Corpus Linguistics, vol. 3, Jul. 1998, doi: 10.1075/ijcl.3.1.03bay.
6. D. Hardt, “Comma checking in Danish,” 2001.
7. U. Salaev, E. Kuriyozov, and C. Gomez-Rodríguez, “SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language,” in 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings, 2022, pp. 199 –

206. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138700420&partnerID=40&md5=bf476cd74317f06577dd0548c5c600d6

8. A. M. Abdurashetona and I. O. Ismailovich, “Methods of Tagging Part of Speech of Uzbek Language,” in Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021, 2021, pp. 82 – 85. doi: 10.1109/UBMK52708.2021.9558900.

9. A. M. Abdurashetona and U. Mokhiyakon, “Software Features and Linguistic Features of Uzbek Synonymizer,” in Proceedings - 7th International Conference on Computer Science and Engineering, UBMK 2022, 2022, pp. 171 – 175. doi: 10.1109/UBMK55850.2022.9919447.

10. B. Mengliyev, S. Shahabitdinova, S. Khamroeva, S. Gulyamova, and A. Botirova, “The morphological analysis and synthesis of word forms in the linguistic analyzer,” Journal of Language and Linguistic Studies, vol. 17, no. 1, pp. 558 – 564, 2021, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103797274&partnerID=40&md5=8a4052419f721c3c734f8fe1c48984ec

11. K. Madatov, S. Bekchanov, and J. Vicic, “Dataset of Karakalpak language stop words,” Data Brief, vol. 48, 2023, doi: 10.1016/j.dib.2023.109111.

12. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, ‘Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language’, CEUR Workshop Proceedings, vol. 3315, pp. 93–98, 2022.

13. D. Mengliev, E. Akhmedov, V. Barakhnin, Z. Hakimov, and O. Alloyorov, “Utilizing Lexicographic Resources for Sentiment Classification in Uzbek Language,” Jan. 2023, pp. 1720–1724. doi: 10.1109/APEIE59731.2023.10347765.

14. K. Madatov, S. Bekchanov, and J. Vicic, “Dataset of stopwords extracted from Uzbek texts”, Data in Brief, vol. 43, 2022.