QARAQALPAQ TILINDE AVTOMATIKALÍQ SORAW-JUWAP SISTEMASÍN JARATÍWDÍŃ TEORIYALÍQ TIYKARLARÍ

Ótemisov A.Z.
Filologiya ilimleri boyınsha filosofiya doktorı (Phd), docent

Eshanova Mavlyuda Orazniyazovna
Qaraqalpaq filologiyası fakulteti Lingvistika (qaraqalpaq tili) magistrantı
Qaraqalpaq mámleketlik universiteti

DOI: https://10.0.20.161/zenodo.17724033

Annotaciya. *Bul maqalamızda qaraqalpaq tiliniń lingvistikalıq ózgesheliklerin esapqa alǵan halda soraw-juwap algoritimin dúziwdıń áhmiyeti, jasalma intellekt modellerinen (transformer, BERT, T5) il biliminde qollanıw máseleleri sóz etiledi.*

Tayanısh sózler: *Soraw-juwap (SJ) sisteması, ekstraktiv, generativ, gibrid, ashıq-domenli, jabıq-domenli, BERT modeli.*

Аннотация: *В данной статье рассматривается важность построения алгоритма вопроса-ответа с учетом языковых особенностей каракалпакского языка, а также проблемы использования моделей искусственного интеллекта (трансформер, BRT, T5) в обучении языку.*

Ключевые слова: *Система вопроса-ответа (ВС), экстрактивная, генеративная, гибридная, открыто-доменная, закрыто-доменная, модель БЕРТ.*

Abstract: *This article discusses the importance of building an answer-question algorithm taking into account the linguistic features of the Karakalpak language, and the problems of using artificial intelligence models (transformer, BRT, T5) in language learning.*

Keywords: *Answer-question system (AQ), extractive, generative, hybrid, open-domain, closed-domain, BERT model.*

Zamanagóy informaciya-texnologiyaları rawajlanıwı nátiyjesinde jasalma intellektke tiykarlanǵan soraw-juwap sistemaları bilimlendiriw, medicina, huqıq hám basqa da tarawlarda keń qollanılmaqta. Inglis, rus, qıtay tillerinde bunday sistemalar álleqashan rawajlanǵan bolsa da, ózbek hám qaraqalpaq tilinde avtomatikalıq soraw-juwap sistemasın jaratıw elege shekem aktual ilmiy hám ámeliy problema bolıp kelmekte.

Soraw-juwap (SJ) sistemaları kompyuter lingvistikası hám jasalma intellekt tarawlarınıń áhmiyetli jónelislerinen biri. Bul sistemalar tábiyiy tildi qayta islew (NLP) hám informaciya izlew (Information Retrieval) ushın baǵdarlanǵan bolıp, olar paydalanıwshı tárepinen tábiyiy tilde berilgen sorawlarǵa tap sol tilde (máselen, qaraqalpaq tilinde) juwap qaytarıw ushın jaratılǵan [1]. Bul sistemalar sorawdıń mazmunın anıqlaydı, tiykarǵı bazaǵa kirgizilgen úlken kólemli maǵlıwmatlar ishinen tiyisli maǵlıwmatlardı ajratıp aladı hám berilgen sorawǵa qolaylı túsinerli formada anıq juwap beredi [2: 13].

Ápiwayı dástúriy xabar izlew sistemaları ádette berilgen sorawǵa juwap beriwshı kitaplardıń, hújjetlerdiń yaki silkalardıń dizimin kórsetip beriwge baǵdarlanǵan [3]. Al, soraw-juwap (SJ) sistemasınıń tiykarǵı maqseti bolsa, paydalanıwshıǵa durıs, anıq hám

kerekli juwaptı usınadı. Bul avtomatikalıq sistema “kerekli maǵlıwmattı” jetkerip beriwge qaratılǵan bolıp, dástúriy usıllarǵa qaraǵanda birqansha jaqsıraq [4: 34].

Soraw-juwap sistemaları rawajlanıw tariyxı 1960-jıllarǵa barıp taqaladı. Dáslepki programmalardan biri BASEBALL bolıp, ol Beysbol ligası haqqındağı sorawlarǵa juwap beriwge baǵdarlanǵan edi. Jáne bir mısal LUNAR sisteması bolıp, ol Apollon missiyası tárepinen alıp kelingen ay bólekshelerin geoloiyalıq analizi haqqında sorawlarǵa juwap bergen. Bul sistemalar óz tarawlarında joqarı jetiskenlikke erisken bolsa da, olardıń bilimi júda tar sheńberde shegaralanǵan edi. SHRDLU siyaqlı eń dáslepki sistemalar jasalma intellekt hám kompyuter lingvistikasınıń teoriyalıq tiykarların rawajlandırıwǵa járdem berdi [5: 23].

Soraw-Juwap sistemalarınıń túrleri hám klassifikaciyaları. Soraw-juwap sistemaların bir neshe kriteriyalar boyınsha klassifikaciyalaw múmkin. Bul klassifikaciyalar sistemanıń juwaptı qanday qáliplesiwi, bilim sheńberi hám xabarǵa kirisiw usılı sıyaqlı táreplerge tiykarlanadı [2: 32].

Juwaptı qáliplestiriw usılına kóre klassifikaciya:

• **Ekstraktiv (Extractive) sistemalar:** Bul sistemalar juwaptı aldınan berilgen tekst yamasa maǵlıwmatlar bazadan tuwrıdan-tuwrı tawıp hám ajıratıp aladı [2:34]. Olar pútkil tekstti oqıp, sorawǵa sáykes keletuǵın tekst aralıǵın yamasa segmentin anıqlaw ushın sol atamadağı obyektlerdi anıqlaw (Named Entity Recognition) hám tekst aralıǵın boljaw (span predicition) sıyaqlı texnikalardan paydalanadı [2:65]. Mısalı, BERT sıyaqlı transformator modellerine tiykarlanǵan sistemalar tap usı jantasıw menen isleydi.

• **Generativ (Generative) sistemalar:** Ekstraktiv sistemalardan ayrıqsha bolıp esaplanıw, generativ sistemalar juwaptı ózleriniń ishki bilimleri tiykarında sintez etedi hám tolıq jańa tekst jaratadı [2:3]. Olar juwaptı sózbe-sóz ajıratıp alıw menen sheklenbeydi, balki dóretiwshilik hám tereń analitikalıq juwaplar jaratıwǵa baǵdarlanǵan. Bul sistemalar, ádetde, ChatGPT yamasa GPT-3 sıyaqlı úlken til modellerine (LLM) tayanadı [2:45].

• **Gibrid (Hybrid) sistemalar:** Bul sistemalar bir neshe kózqaraslar hám qatnaslardı, atap aytqanda, ekstraktiv hám generativ metodlardı birlestiredi [3:21]. Gibrid qatnas jasawdıń payda bolıwı birden-bir túrdegi sistemalardıń sheklewlerine juwap retinde júzege keldi. Dástúriy generativ modeller geyde faktlerdi nadurıs bayanlawı (gallucinaciya) yamasa tek úyretilgen waqıttağı bilimge súyenip qalıwı múmkin [6:26]. Basqa tárepinen, ekstraktiv sistemalar tek maǵlıwmat dáreginde bar bolǵan juwaplardı taba aladı, biraq jańa tekst jarata almaydı. RAG (Retrieval-Augmented Generation) sıyaqlı gibrid arxitekturalar bul mashqalalardı saplastırıwǵa xızmet etedi. Olar informaciya izlewdiń anıqlıǵın generativ modellerdiń dóretiwshilik penen biriktirip, isenimli hám aktual maǵlıwmatlarǵa tiykarlanǵan juwaplar jaratıwǵa múmkinshilik beredi.

Bilim sheńberine qaray klassifikaciya:

Ashıq-domenli (Open-domain) sistemalar: Bul sistemalar derlik hárqanday tema boyınsha sorawlarǵa juwap beriw ushın arnalǵan. Olar pútkil internet yamasa Wikipedia

sıyaqlı úlken ulıwma bilim bazalarına súyenedi. Bunday sistemalar virtual járdemshiler yamasa izlew sistemaları ushın júdá sáykes keledi [2:36].

Jabıq-domenli (Closed-domain) sistemalar: Bul sistemalar medicina, huqıq, texnika yamasa korporativ maǵlıwmatlar sıyaqlı arnawlı bir tarawlarǵa qánigelesken [2:38]. Olar óz tarawlarına tán bilim dereklerinen paydalanǵan halda anıq hám tolıq juwaplar beredi [3:86]. Mısalı, lingvistika tarawındaǵı jabıq-domenli SJ sisteması tek tilge baylanıslı sorawlarǵa juwap beredi.

Qaraqalpaq tilinde avtomatikalıq soraw-juwap sistemasın jaratıwda joqarıda keltirilgen ekstraktiv jabıq-domenli soraw-juwap sistemasın jaratıwdı aldımızǵa maqset etkenbiz. Sebebi, ekstraktiv sistemada juwaptı tiykarǵı bazaǵa kirgizilgen maǵlıwmatlar tekstinen ajıratıp aladı. Onıń abzal tárepi, berilgen sorawǵa juwap anıq hám faktke tiykarlanǵan boladı. Máselen, “Antonim degenimiz ne?”, juwap “Mánilik jaqtan bir-birine qarama-qarsı sózler”, [Házirgi qaraqalpaq tili. Leksikologiyası Nókis-1994-jıl, 51-bet] dep anıq faktler, maǵlıwmatlar menen keltirilip dálilenip beriledi.

Biz avtomatikalıq soraw-juwap sistemasın BERT siyaqlı transformator modeline tiykarlanıp jaratamız. Transformatorlardı eki tárepleme kodlawshı kórinisler (Bidirectional encoder representations from transformers (BERT)) 2018-jılı oktyabr ayında Google izertlewshileri tárepinen usınıs etilgen til modeli[7:56]. Ol ózin-ózi qadaǵalawshı, úyreniw járdeminde teksti vektorlar izbe-izliginde kórinislerdi payda etedi. Ol tek kodlawshı transformator arxitekturasında paydalanıladı. BERT úlken til modelleri ushın zamanagóylikti sezilerli dárejede jaqsılaydı. 2020-jıl statistikalıq maǵlıwmatlarǵa qaraǵanda, BERT tábiyiy tildi qayda islew (NLP) tájriybelerinde barlıq jerde bar bolǵan tiykarǵı baǵdar esaplanǵan [8:68]. Sol ushın da, biz keleshekte usı modelden paydalanıp qaraqalpaq tilinde jasalma intellektti payda etiwde maqset etkenbiz.

Juwmaqlap aytqanda, “Qaraqalpaq tilindegi avtomatikalıq soraw-juwap sisteması” til biliminde bar bolǵan teoriyalar tiykarında paydalanıwshıǵa anıq faktlerge tiykarlanǵan juwaplardı jaratıw. Bul óz gezeginde waqıttı únemlep, qaraqalpaq tili boyınsha bilmegen sorawlarımızǵa anıq, durıs juwap alıwǵa kómeklesedi.

Paydalanǵan ádebiyatlar:

1. https://ibm.com
2. https://www.ibm.com/think/topics/question-answering
3. https://spotintelligence.com/2023/01/20/question-answering-qa-system-nlp/
4. https://start.csail.mit.edu/
5. https://wisconsin.pressbooks.pub/naturallanguage/chapter/information-retrieval/
6. https://huggingface.co/tasks/question-answering
7. "Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing". Google AI Blog. November 2, 2018. Retrieved November 27, 2019.
8. Rogers, Anna; Kovaleva, Olga; Rumshisky, Anna (2020). "A Primer in BERTology: What We Know About How BERT Works". Transactions of the Association for Computational Linguistics. 8: 842–866. arXiv:2002.12327. doi:10.1162/tacl_a_00349. S2CID 211532403.