

КАМ РЕСУРС ТИЛЛАР УЧУН НУТҚНИ МАТНГА АЙЛАНТИРИШ ВА ДИАЛЕКТЛАРНИ НОРМАЛЛАШТИРИШ: ЎЗБЕК ТИЛИ МИСОЛИДА

Xusaydinova D.

Muhammad al-Xorazmiy nomidagi

Toshkent axborot texnologiyalari universiteti

d.xusaydinova@tuit.uz

Аннотация: Табиий тилни қайта ишлаш (Natural Language Processing, NLP) соҳасида сезиларли ютуқларга эришилган бўлса-да, аксарият ривожланишлар юқори ресурс тилларга қаратилган. Ўзбек тили учун эса белгиланган маълумотлар тўпламлари ва лингвистик ресурсларнинг етишмаслиги NLP тизимлари самарадорлигини чеклаб келмоқда. Ушбу тадқиқот ўзбек тили шеваларини автоматик нормаллаштириш вазифасига бағишланган. Тадқиқот доирасида 5000 дан ортиқ диалектал жумлалардан иборат маълумотлар тўплами яратилди ва Transformer архитектурасига асосланган sequence-to-sequence модели ўқитилди. Модел сифати BLEU ва Assurasy метрикалари орқали баҳоланди. Эксперимент натижалари тақлиф этилган ёндашув ўзбек тилидаги диалектал матнларни қайта ишлашда самарали эканлигини кўрсатди.

Калит сўзлар: NLP, ўзбек тили, диалект, Transformer, нутқни матнга айлантириш (Speech-to-Text), матнни нормаллаштириш, кам ресурс тиллар.

Аннотация: Несмотря на значительный прогресс в области обработки естественного языка (NLP), развитие технологий в основном сосредоточено на высокоресурсных языках. Для узбекского языка по-прежнему существует недостаток размеченных данных и языковых ресурсов, что снижает эффективность NLP-систем. В данной работе рассматривается задача автоматической нормализации диалектов узбекского языка. Для исследования был создан корпус, содержащий более 5000 диалектных предложений. Нормализация выполнялась с использованием модели sequence-to-sequence на основе архитектуры Transformer. Качество модели оценивалось с помощью метрик BLEU и Assurasy. Полученные результаты подтвердили эффективность предложенного подхода для обработки диалектных текстов узбекского языка.

Ключевые слова: NLP, узбекский язык, диалект, Transformer, Speech-to-Text, нормализация текста, низкоресурсные языки.

Abstract: Although significant progress has been achieved in Natural Language Processing (NLP), most advancements have focused on high-resource languages. For the Uzbek language, the lack of annotated datasets and linguistic resources continues to limit the performance of NLP systems. This study addresses the task of

automatic dialect normalization for Uzbek. A dataset containing more than 5,000 dialectal sentences was created and used to train a Transformer-based sequence-to-sequence model. The model was evaluated using BLEU and Accuracy metrics. Experimental results demonstrate the effectiveness of the proposed approach for processing Uzbek dialectal texts.

Keywords: NLP, Uzbek language, dialect, Transformer, Speech-to-Text, text normalization, low-resource languages.

Кириш

Сунъий интеллект ва рақамли технологияларнинг жадал ривожланиши инсон тилини автоматик қайта ишлаш масалаларини долзарб қилиб қўймоқда. Табиий тилни қайта ишлаш соҳаси машиналарга инсон тилини тушуниш, таҳлил қилиш ва қайта ишлаш имконини беради. Бугунги кунда овозли ёрдамчилар, автоматик таржима тизимлари, чат-ботлар ва ахборот қидирув тизимлари кенг қўлланилмоқда.

Шунга қарамай, кўпчилик тадқиқотлар юқори ресурс тилларга қаратилган. Ўзбек тили каби кам ресурс тиллар учун маълумотлар базасининг чекланганлиги NLP соҳасида жиддий муаммо бўлиб қолмоқда. Айниқса, нутқни матнга айлантириш тизимларида шева ва диалектларнинг кўплиги таниб олиш аниқлигига салбий таъсир кўрсатади.

Ўзбекистоннинг турли ҳудудларида Тошкент, Фарғона, Қашқадарё, Хоразм, Бухоро ва Сурхондарё шевалари кенг тарқалган. Бу шевалар фонетик, морфологик ва лексик жиҳатдан адабий тилдан маълум даражада фарқ қилади. Натижада автоматик тил технологиялари бундай матнларни қайта ишлашда хатоларга йўл қўяди.

Ушбу тадқиқотнинг мақсади ўзбек тили шеваларини адабий тилга автоматик нормаллаштириш моделини ишлаб чиқиш ва унинг самарадорлигини баҳолашдан иборат.

Адабиётлар таҳлили. Сўнгги йилларда Transformer архитектураси NLP соҳасида катта ўзгаришларга сабаб бўлди. Vaswani ва ҳаммуаллифлар томонидан таклиф қилинган Transformer модели рекуррент нейрон тармоқларга нисбатан юқори самарадорликни намоён қилди.

BERT, GPT, T5 ва mBART каби моделлар кўплаб тилларда юқори натижалар кўрсатган бўлса-да, уларнинг аксарияти юқори ресурс тиллар маълумотлари асосида ўқитилган.

Кам ресурс тиллар учун олиб борилган тадқиқотларда маълумотларни кўпайтириш (Data Augmentation), трансфер ўқитиш (Transfer Learning) ва кўп тилли моделлардан фойдаланиш кенг тарқалган. Турк, қozoқ, қирғиз ва уйғур тиллари учун ҳам диалектларни нормаллаштириш бўйича бир қатор ишлар амалга оширилган.

Ўзбек тили бўйича эса морфологик таҳлил, автоматик таржима ва матн таснифлаш соҳаларида айрим ишлар мавжуд бўлса-да, шеваларни нормаллаштириш бўйича тадқиқотлар чекланган ҳисобланади.

Тадқиқот методологияси. Маълумотлар базасини шакллантириш

Тадқиқот учун 5000 дан ортиқ диалектал жумлалардан иборат корпус шакллантирилди. Маълумотлар турли ҳудудларга мансуб шева намуналаридан йиғилди. Ҳар бир диалектал жумла адабий ўзбек тилидаги мос варианты билан жуфтланди. Мисол:

- Диалект: «Ман борувдим бозорга»;
- Адабий тил: «Мен бозорга борган эдим»;
- Диалект: «Келяпману»;
- Адабий тил: «Келяпман»;

Маълумотларга дастлабки ишлов бериш. Маълумотларга қуйидаги ишлов бериш босқичлари қўлланилди:

- ортиқча белгиларни ўчириш;
- регистрларни бир хиллаштириш;
- токенизация;
- нормаллаштириш;
- маълумотларни ўқитиш ва тест қисмларига ажратиш.

Transformer модели. Модел Encoder-Decoder архитектураси асосида қурилди. Encoder қисми кирувчи диалектал матнни яширин векторларга айлантиради. Decoder қисми эса ушбу векторлар асосида адабий матнни генерация қилади.

Attention механизми сўзлар ўртасидаги боғлиқликларни аниқлаш имконини беради ва узоқ контекстларни яхшироқ ҳисобга олади.

Ўқитиш параметрлари. Модел қуйидаги параметрлар билан ўқитилди:

- Epochs: 20
- Batch Size: 32
- Learning Rate: 0.0001
- Optimizer: Adam
- Loss Function: Cross-Entropy Loss

Эксперимент натижалари. Моделнинг сифати BLEU ва Accuracy метрикалари орқали баҳоланди.

Метрика	Қиймат
BLEU	78.4
Accuracy	84.7%

1-Жадвал. Эксперимент натижалари

Натижалар моделнинг диалектал матнларни адабий тилга муваффақиятли ўгиришини кўрсатди. Мисоллар:

- Кириш: «Кетвомман ишга»
- Натижа: «Ишга кетяпман»
- Кириш: «Борувдим кеча»
- Натижа: «Кеча борган эдим»

Муҳокама. Олинган натижалар Transformer архитектурасининг ўзбек тили шеваларини қайта ишлашда самарали эканлигини кўрсатади. Айниқса фонетик ўзгаришлар ва қисқартирилган шаклларни нормаллаштиришда яхши натижалар қайд этилди.

Бироқ маълумотлар ҳажмининг нисбатан кичиклиги баъзи ҳолларда нотўғри прогнозларга сабаб бўлди. Келгусида корпус ҳажмини 50–100 минг жуфтликкача кенгайтириш мақсадга мувофиқ ҳисобланади.

Хулоса

Ушбу тадқиқотда ўзбек тили учун диалектларни автоматик нормаллаштириш масаласи кўриб чиқилди. 5000 дан ортиқ жумлалардан ташкил топган корпус яратилди ва Transformer модели асосида ўқитилди.

Эксперимент натижалари тақлиф қилинган ёндашувнинг юқори самарадорликка эга эканлигини кўрсатди. Тадқиқот натижалари келажакда ўзбек тили учун Speech-to-Text тизимлари, овозли ёрдамчилар ва бошқа NLP ечимларини ривожлантиришда муҳим аҳамият касб этади.

Фойдаланилган адабиётлар:

1. Vaswani A. et al. Attention Is All You Need. NIPS, 2017.
2. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers. NAACL, 2019.
3. Brown T. et al. Language Models are Few-Shot Learners. NeurIPS, 2020.
4. Raffel C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR, 2020.
5. Conneau A. et al. Unsupervised Cross-lingual Representation Learning. ACL, 2020.
6. Jurafsky D., Martin J. Speech and Language Processing. Pearson, 2023.
7. Koehn P. Neural Machine Translation. Cambridge University Press, 2020.
8. Sennrich R. Neural Machine Translation of Rare Words. ACL, 2016.
9. Xue L. et al. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. 2021.
10. Wolf T. et al. Transformers: State-of-the-Art NLP. EMNLP, 2020.
11. Abdurakhmonov S. Uzbek NLP Resources and Challenges. 2022.
12. Tursunov B. Uzbek Language Technologies: Current State and Future Directions. 2023.
13. Goodfellow I. Deep Learning. MIT Press, 2016.

14. Goldberg Y. Neural Network Methods for NLP. Morgan & Claypool, 2017.
15. Manning C., Schütze H. Foundations of Statistical NLP. MIT Press, 1999.