

TRANSFORMER МОДЕЛЛАРИ АСОСИДА ЎЗБЕК ТИЛИ ШЕВАЛАРИНИ АДАБИЙ ТИЛГА АВТОМАТИК ЎГИРИШ

Xusaydinova D. M

Muhammad al-Xorazmiy nomidagi

Toshkent axborot texnologiyalari universiteti

d.xusaydinova@tuit.uz

Аннотация: Ўзбек тили шевалари адабий тилдан фонетик, лексик ва грамматик жиҳатдан фарқ қилади, бу эса матнларни автоматик қайта ишлаш жараёнида қийинчиликлар туғдиради. Ушбу тадқиқотда Transformer моделлари асосида диалектал матнларни адабий ўзбек тилига автоматик ўгириш масаласи кўриб чиқилади. Тадқиқот учун 5000 дан ортиқ диалектал ва адабий жумлалар жуфтлигидан иборат корпус яратилди. Модел sequence-to-sequence усулида ўқитилди ва BLEU ҳамда Accuracy метрикалари орқали баҳоланди. Натижалар Transformer моделлари шева хусусиятларини самарали аниқлаб, уларни адабий тил меъёрларига мувофиқ нормаллаштира олишини кўрсатди.

Калит сўзлар: ўзбек тили, шева, адабий тил, Transformer, NLP, матнни нормаллаштириш, sequence-to-sequence.

Аннотация: Диалекты узбекского языка существенно отличаются от литературной нормы, что создает сложности при автоматической обработке текстов. В данной работе рассматривается задача автоматического преобразования диалектных текстов в литературный узбекский язык с использованием моделей Transformer. Для проведения исследования был сформирован корпус из более чем 5000 пар диалектных и литературных предложений. Модель обучалась в режиме sequence-to-sequence и оценивалась с использованием метрик BLEU и Accuracy. Результаты экспериментов показали, что модели Transformer эффективно выявляют и нормализуют диалектные особенности, обеспечивая высокое качество преобразования текста.

Ключевые слова: узбекский язык, диалект, литературный язык, Transformer, NLP, нормализация текста, sequence-to-sequence.

Abstract: Uzbek dialects differ significantly from the standard literary language, creating challenges for automatic text processing systems. This study investigates the task of converting dialectal Uzbek texts into standard literary Uzbek using Transformer-based models. A corpus consisting of more than 5000 pairs of dialectal and standard sentences was developed for training and evaluation. The model was trained in a sequence-to-sequence setting and assessed using BLEU and Accuracy

metrics. Experimental results demonstrate that Transformer models effectively identify and normalize dialectal variations, achieving high-quality text conversion.

Keywords: Uzbek language, dialect, literary language, Transformer, NLP, text normalization, sequence-to-sequence.

Кириш

Сунъий интеллект ва NLP соҳасидаги ривожланишлар инсон тилини автоматик қайта ишлаш имкониятларини кенгайтормоқда. Бироқ ўзбек тили каби кам ресурс тилларда маълумотлар базаси ва лингвистик ресурслар етарли эмас.

Ўзбек тили ҳудудий шеваларга бой бўлиб, Тошкент, Фарғона, Андижон, Қашқадарё, Сурхондарё, Хоразм каби ҳудудларда турлича талаффуз ва лексик бирликлар қўлланилади. Бу ҳолат автоматик таржима, овозли ёрдамчилар ва қидирув тизимлари учун қийинчилик туғдиради.

Мақсад Transformer архитектураси асосида шеваларни адабий тилга автоматик ўгириш моделини ишлаб чиқиш ва баҳолаш.

Адабиётлар таҳлили. Диалектларни нормаллаштириш бўйича тадқиқотлар:

- Arabizi → Arabic
- Swiss German → Standard German
- Turkish dialect normalization
- Kazakh dialect processing

Transformer архитектураси. 2017 йилда Vaswani ва ҳаммуаллифлар томонидан таклиф қилинган "Attention Is All You Need" мақоласи. Transformer таркиби: Encoder, Decoder, Multi-Head Attention, Positional Encoding.

Кам ресурс тилларда NLP Кам ресурс тиллар (Low-Resource Languages) деб рақамли ресурслари, катта ҳажмли корпуслари ва белгиланган маълумотлар тўпламлари етарли даражада шаклланмаган тилларга айтилади. Ўзбек тили ҳам шундай тиллар қаторига киради. Бундай тилларда табиий тилни қайта ишлаш тизимларини яратиш маълумотлар танқислиги сабабли мураккаб ҳисобланади.

Сўнги йилларда мазкур муаммони ҳал қилиш учун бир қатор ёндашувлар таклиф қилинган. Улар орасида Transfer Learning, Multilingual Models ва Data Augmentation усуллари кенг қўлланилмоқда.

Transfer Learning усулида катта ҳажмли маълумотлар асосида олдиндан ўқитилган моделлардан фойдаланилади ва улар кам ресурс тил маълумотлари асосида қайта мослаштирилади. Бу усул маълумотлар етишмаслиги шароитида ҳам юқори натижаларга эришиш имконини беради.

Multilingual Models бир вақтнинг ўзида бир нечта тилларда ўқитилади. BERT, mBERT, XLM-R ва mT5 каби моделлар ўзбек тили учун ҳам маълум даражада самарали натижалар кўрсатмоқда.

Data Augmentation усули эса мавжуд маълумотларни сунъий равишда кўпайтириш орқали моделнинг ўқитиш сифати ва барқарорлигини оширишга хизмат қилади.

Ўзбек тили бўйича олиб борилган тадқиқотлар асосан автоматик таржима, матн таснифлаш ва морфологик таҳлил соҳаларига қаратилган бўлиб, шеваларни адабий тилга нормаллаштириш бўйича ишлар ҳали етарли даражада эмас. Шу сабабли мазкур тадқиқот ушбу йўналишдаги муҳим қадамлардан бири ҳисобланади.

Тадқиқот методологияси. Ушбу тадқиқотнинг асосий мақсади ўзбек тили шеваларида ёзилган матнларни адабий тилга автоматик тарзда ўгириш учун Transformer моделига асосланган тизимни ишлаб чиқиш ва баҳолашдан иборат.

Маълумотлар базасини шакллантириш. Тадқиқот учун турли ҳудудларда қўлланиладиган шева намуналари йиғилди. Жами 5000 дан ортиқ диалектал жумлалар тўпланиб, ҳар бири адабий ўзбек тилидаги мос варианты билан жуфтланди.

Жумла жуфтликлари қўлда текширилди ва таҳрир қилинди. Бу маълумотлар базасининг сифатини таъминлаш ҳамда моделни аниқ ўқитиш имконини берди.

Диалектал матн	Адабий матн
Ман кетвомман	Мен кетяпман
Келяману	Келяпман
Борувдим	Борган эдим

Жадвал 1. Диалектал ва адабий матн намуналари

Жадвалдан кўриниб турибдики, шевадаги айрим фонетик ва грамматик шакллар адабий тил меъёрларидан фарқ қилади. Моделнинг вазифаси айнан шу фарқларни аниқлаб, тўғри адабий кўринишни ҳосил қилишдан иборат.

Маълумотларга дастлабки ишлов бериш. Моделни ўқитишдан олдин маълумотларга махсус қайта ишлов бериш жараёнлари амалга оширилди. Булар қуйидагилардан иборат:

– Cleaning – ортиқча белгилар, нотўғри символлар ва кераксиз бўшлиқларни ўчириш;

– Lowercasing – барча ҳарфларни бир хил регистрга келтириш;

– Tokenization – матнни алоҳида сўзлар ёки токенларга ажратиш;

– Vocabulary Generation – маълумотлар тўпламида учрайдиган сўзлар луғатини шакллантириш.

Бу босқичлар маълумотларни стандарт кўринишга келтириш ва моделнинг ўқитиш самарадорлигини ошириш учун муҳим аҳамиятга эга.

Маълумотларни тақсимлаш. Яратилган маълумотлар базаси моделни ўқитиш ва баҳолаш мақсадида уч қисмга ажратилди.

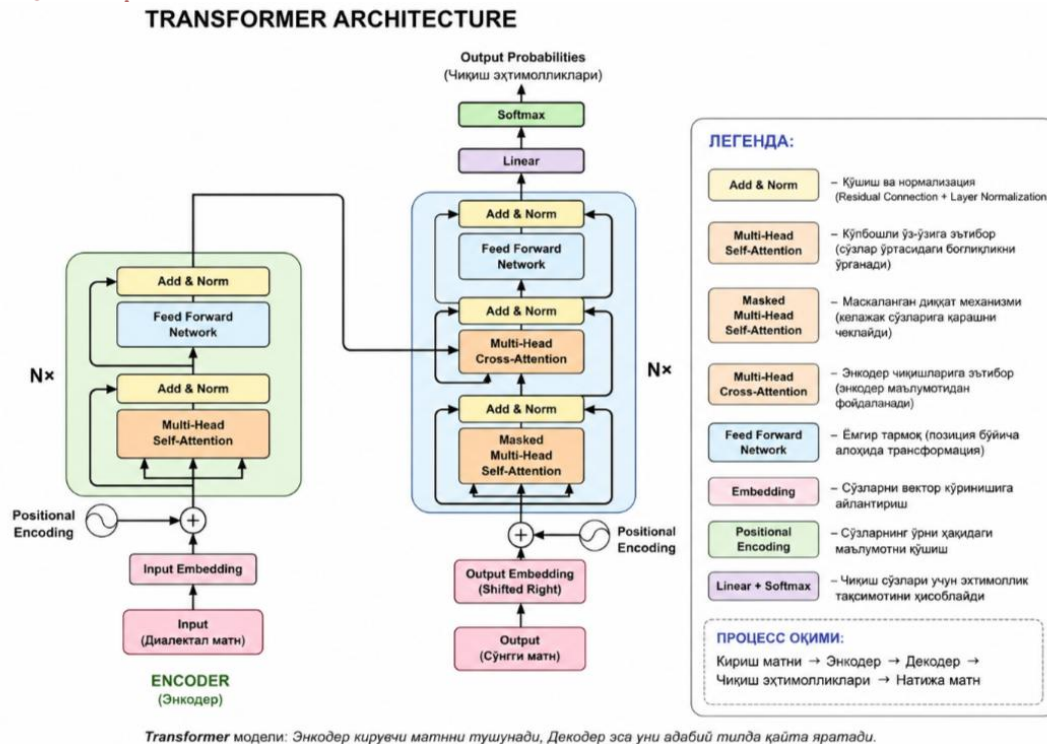
Тўплам	Фоиз
Training	80%
Validation	10%
Test	10%

Жадвал 2. Dataset тақсимооти

- Training тўплами моделни ўқитиш учун ишлатилди.
- Validation тўплами гиперпараметрларни сошлаш ва моделнинг ортиқча ўқиб кетишини (overfitting) назорат қилиш учун фойдаланилди.
- Test тўплами эса моделнинг якуний самарадорлигини баҳолаш учун ажратиб қўйилди.

Бу тақсимот машинали ўқитиш соҳасида кенг қўлланиладиган стандарт ёндашув ҳисобланади ва эксперимент натижаларининг ишончилигини таъминлайди.

Transformer архитектураси. 2017 йилда Vaswani ва ҳаммуаллифлар томонидан таклиф қилинган бўлиб, ҳозирги замонавий NLP тизимларининг асосини ташкил этади. Ушбу архитектура Encoder ва Decoder қисмларидан иборат бўлиб, матнлар ўртасидаги боғлиқликларни Attention механизми ёрдамида аниқлайди. Transformer рекуррент нейрон тармоқлардан фарқли равишда маълумотларни параллел қайта ишлаш имкониятини беради, бу эса ўқитиш жараёнини тезлаштиради ва самарадорликни оширади. Encoder кирувчи диалектал матнни яширин тасвирларга айлантирса, Decoder ушбу маълумотлар асосида адабий тилдаги чиқиш матнини ҳосил қилади. Тадқиқотда Transformer моделининг юқори аниқлик ва контекстни яхши тушуниш хусусиятлари туфайли у шеваларни адабий тилга автоматик ўгириш вазифаси учун танланди. Transformer модели. Архитектура, экспериментлар, ўқитиш параметрлари



Расм 1. Transformer Architecture

Параметр	Қиймат
Epoch	20
Batch Size	32
Learning Rate	0.0001
Optimizer	Adam

Жадвал 3.

Баҳолаш метрикалари. BLEU модели томонидан яратилган матннинг эталон матнга яқинлигини баҳолайди. Ассигасу-Тўғри прогноз қилинган мисоллар улушини кўрсатади.

Метрика	Натижа
BLEU	79.2
Accuracy	85.4%

Жадвал 3.

Натижалар. Мисоллар

- Кириш: Ман ишга кетвомман.
- Чиқиш: Мен ишга кетяпман.

- Кириш:Борувдим кеча.
- Чиқиш:Кеча борган эдим.

Натижалар кўрсатадики Transformer моделлари шеваларни яхши ўрганади ва фонетик ўзгаришларни тўғри тиклайди, кам маълумот билан ҳам яхши ишлайди.

Муаммолар:

- Корпус ҳажми кичик.
- Айрим шевалар кам қамраб олинган.

Хулоса

Ушбу тадқиқотда Transformer архитектураси асосида ўзбек тили шеваларини адабий тилга автоматик ўгириш вазифаси бажарилди. Яратилган 5000+ жуфтликдан иборат корпус асосида модел ўқитилди ва юқори натижаларга эришилди. Тадқиқот натижалари ўзбек тили учун интеллектуал NLP тизимларини ривожлантиришга хизмат қилади.

Фойдаланилган адабиётлар:

16. Vaswani A. et al. Attention Is All You Need. NIPS, 2017.
17. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers. NAACL, 2019.
18. Brown T. et al. Language Models are Few-Shot Learners. NeurIPS, 2020.
19. Raffel C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR, 2020.
20. Conneau A. et al. Unsupervised Cross-lingual Representation Learning. ACL, 2020.
21. Jurafsky D., Martin J. Speech and Language Processing. Pearson, 2023.
22. Koehn P. Neural Machine Translation. Cambridge University Press, 2020.
23. Sennrich R. Neural Machine Translation of Rare Words. ACL, 2016.
24. Xue L. et al. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. 2021.
25. Wolf T. et al. Transformers: State-of-the-Art NLP. EMNLP, 2020.
26. Abdurakhmonov S. Uzbek NLP Resources and Challenges. 2022.
27. Tursunov B. Uzbek Language Technologies: Current State and Future Directions. 2023.
28. Goodfellow I. Deep Learning. MIT Press, 2016.
29. Goldberg Y. Neural Network Methods for NLP. Morgan & Claypool, 2017.
30. Manning C., Schütze H. Foundations of Statistical NLP. MIT Press, 1999.