ТУРКИЙ ТИЛЛАР ГРАММАТИКАСИНИНГ КОМПЬЮТЕРГА ЙЎНАЛТИРИЛГАН ФОРМАЛ МОДЕЛЛАРИ ТАҲЛИЛИ

Норов Абдисаит Мурадович

Муҳамад ал-Хоразмий номидаги ТАТУ докторанти

K E Y W O R D S

Statistical Models, Neural Network Models, Rule Based Grammars, Hybrid NLP

A B S T R A C T

Ушбу мақолада туркий тиллар грамматикасини компьютерли қайта ишлаш учун формал моделлаштириш тамойиллари ёритилади. Морфологик агглютинация, эркин аффиксация каби хусусиятларга эга туркий тиллар учун грамматик қоидаларни алгоритмлаштиришда қўлланиладиган компьютерли моделлар: қоидаларга асосланган грамматикалар (Rule Based Grammars), статистик моделлар (Statistical Models), нейрон тармоқлар (Neural Network Models), гибрид NLP (Hybrid NLP) архитектуралари таҳлил қилинади.

Кириш.

Туркий тиллар морфологик жиҳатдан агглютинатив тиллар гуруҳига мансуб бўлиб, бунда грамматик маъно асосан аффиксация орқали берилади. Компьютер лингвистикасида бундай тиллар учун морфологик ва синтактик моделлар яратиш NLPнинг энг мураккаб йўналишларидан бири ҳисобланади. Чунки битта сўз ўнлаб грамматик қўшимчаларни ўзига олиши мумкин. Масалан, "***Мактабларингиздагиларнинг***" сўзи қуйидаги морфемаларга бўлинади: "***мактаб***" – асос (лексик морфема); "***-лар***" – кўплик қўшимчаси; "***-ингиз***" – иккинчи шахс кўплик/ҳурмат эгалик қўшимчаси; "***-даги***" – сифатдош/аниқловчи аффикс; "***-лар***" – боғланган гуруҳга тегишли кўплик қўшимчаси; "***-нинг***" – қаратқич келишиги. Бундай мураккаб сўз тузилмасини автоматик қайта ишлаш учун қатъий компьютерли моделлар зарур бўлади.

Грамматика аслида икки асосий бўлимдан: морфология (сўзнинг ички қурилиши) ва синтаксис (сўзларнинг гапдаги бирикиш қоидалари) бўлимларидан ташкил топган бўлиб, бироқ амалий жиҳатдан замонавий тилшунослик ва NLPда грамматикага қўшимча равишда морфосинтаксис (морфология ва синтаксис чегарасидаги қоидалар), грамматик семантика (грамматик шакллар орқали маъно ифодаланиши) ва прагматика (прагматик акцентнинг грамматикага алоқадор қисми) ҳам киради.

Компьютерли моделлаштиришда грамматика тушунчаси Н.Чўмский муаллифлигидаги формал грамматика (Computational Grammar) номи билан янада кенгайганлигини кўриш мумкин, хусусий ҳолда улар Context-Free Grammar (CFG), Dependency Grammar, Categorial Grammar (CCG), Lexical Functional Grammar (LFG), Head-driven Phrase Structure Grammar (HPSG) ва Finite State Grammars каби номлар билан аталади.

Умуман олганда, ўзбек ва бошқа туркий тилларни автоматик қайта ишлаш, бу тилларнинг грамматикаси ва фонетикасини NLP методлари асосида компьютерли моделлаштириш учун тўлақонли тадқиқот ишлари олиб борилмаган. Шу боис, туркий тиллар грамматикасининг компьютерга йўгалтирилган формал моделларини таҳлил қилган ҳолда замонавий рақамли технологияларга асосланган моделлар ва методларни ишлаб чиқиш бугунги куннинг долзарб масалаларидан бири саналади.

Адабиётлар таҳлили.

[1] да туркий тилларнинг фонетик, лексик, грамматик, морфологик, синтактик

жиҳатдан қиёсий таҳлили келтирилган. Шунингдек, китобда сўз туркумлари бўйича туркий тилларнинг ўхшаш ва фарқли томонлари кенг очиб берилган.

[2] да ҳар бири кейинги категория учун танланган n та элементдан ташкил топган ҳар қандай табиий тил конструкцияси учун мумкин бўлган сўз тартиблари, уларнинг танланиш чекловлари таклиф қилинади.

[3] да Tree-Adjoining Grammar (TAG) ҳамда Combinatory Categorial Grammar (CCG) моделларининг табиий тилларни автоматик қайта ишлашдаги ўхшашлик ва фарқли функциялари очиб берилган.

[4] да табиий тилдан фойдаланиб маълумотлар базасига осонгина кириш усули таклиф қилинади. Бунда ёндашув қоидаларга асосланган (Rule Based) бўлиб, компьютерда синаб кўриш учун прототип тизими ишлаб чиқилган.

[5] да табиий тилни таҳлил қилиш учун учта статистик модель таклиф этилган. Ушбу моделлар бир қатор лингвистик ҳодисаларни кузатувчи параметрларни ўз ичига олади: биграм лексик боғлиқликлар, субкатегоризация рамкалари, слаш категорияларининг тарқалиши ва бошқалар. Моделлар генератив моделлар бўлиб, уларда таҳлил дарахтлари дарахтнинг юқоридан пастга қараб ҳосил бўлишида бир қатор босқичларга бўлинади ва ҳосил бўлишдаги қарорлар шартли эҳтимолликлар сифатида моделлаштирилади.

Усуллар

Туркий тилларнинг компьютерли моделлаштиришдаги грамматик хусусиятларини қуйидаги бир нечта омилларда кўриш мумкин:

1. Агглютинация ва морфема чегараси. Туркий тиллардаги сўзнинг қўшимчалари бу сўз таркибида изчил тартибда келиш қонуниятига эга. Бу қонуниятларни формал грамматика (Finite-State Transducer, FST) асосида моделлаштириш анча қулай. FST формал тиллар назариясидаги чекли автомат (Finite-State Automaton, FSA)нинг кенгайтирилган бир кўриниши бўлиб, у кириш символини чиқиш символига мослайдиган, яъни трансформация амалларини бажарадиган математик моделдир. FST фонология, морфология, орфография, графема-фонема (Grapheme-to-Phoneme, G2P) трансформацияси ва TTS/STT жараёнларини компьютерли моделлаштиришда кўп қўлланилади.

2. Сингармонизм ва фонологик қоидалар. Сўз таркибидаги унлилар ва уларнинг аллофонларини компьютерли моделлаштириш учун фонологик қоидалар тўплами талаб қилинади. Масалан, биргина унлиларга хос бўлган редукция, ассимиляция, диссимиляция, элизия ва эпентеза каби фонетик ҳодисалар туркий тиллардаги бош фонологик қоидалардир.

3. Сўзнинг гапдаги эркин позицияси. Туркий тилларда гапнинг умумий тузилмаси учун SOV (Subject, Object, Verb) ya’ni (эга гуруҳи, иккинчи даражали бўлаклар, кесим гуруҳи) асосий тартиб бўлса-да, аммо прагматик акцентга қараб бу тартиб, яъни сўзнинг гапдаги позицияси ўзгариши мумкин, масалан:

- *Мен китобни эртага олиб бораман (қачон? – эртага).*
- *Мен эртага китобни олиб бораман (ким? мен).*
- *Эртага китобни мен олиб бораман (нимани? китобни).*

Бундай синтактик тузилмалар учун махсус Dependency Grammar (боғланишли грамматика) моделлари талаб қилинади.

Қўйилган масаланинг турига қараб туркий тиллар учун компьютерга йўналтирилган грамматик моделлар қуйидагилардан иборат бўлиши мумкин:

1. Қоидаларга асосланган моделлар (Rule-based NLP). Туркий тиллар учун энг самарали классик моделлардан бири саналади, бу моделлардан морфологик таҳлил дастурларини яратишда самарали фойдаланиш мумкин. Бунда асосий элементлар аффикслар луғати ва фонологик

трансформация қоидаларидан иборат бўлиб, амалдаги алгоритмлар сифатида Finite-State Transducer (FST) ва Two-level morphology алгоритмларидан фойдаланилади.

2. Статистик моделлар (HMM, CRF). Маълумки, туркий тилларда агглютинация сабабли морфологик бўлинишлар жуда кўп, шу боис HMM (Hidden Markov Model) ва CRF (Conditional Random Field) каби статистик усуллардан теглаш, POS-теггинг ва аффикс сегментациясида самарали фойдаланиш мумкин.

3. Нейрон тармоқларга асосланган моделлар (Neural NLP). Кейинги йилларда туркий тиллар бўйича BiLSTM-based morphological tagger, BERT, mBERT, XLM-R каби трансформер-базавий моделлар, Seq2Seq морфологик генераторлар ва гибрид моделлар энг самарали деб топилмоқда.

BiLSTM-based Morphological Tagger сўзларга морфологик тегларни автоматик равишда белгилайдиган нейрон модел бўлиб, у икки томонлама LSTM (BiLSTM) нейрон тармоғига асосланган. Бу модел морфологик таҳлил учун NLPда энг кўп қўлланиладиган усуллардан биридир. Моделнинг асосий вазифаси жумла ичидаги сўзларнинг тўлиқ морфологик тегларини башорат қилишдан иборат. Бошқа тилларга қараганда туркий тиллар учун ушбу модель самарали натижалар бериши тадқиқотларда исботланган.

Seq2Seq морфологик генератор – бу киритилган лемма (асосий сўз шакли) ва грамматик теглар асосида шу сўзнинг флексияланган, ўзгарган ёки боғланган морфологик шаклини генерация қиладиган нейрон моделдир. У одатда Encoder-Decoder архитектурасига асосланади ва NLPда морфологияни автоматлаштириш учун ишлатилади.

Компьютер лингвистикаси, NLP ва туркий тиллар морфологиясида энг оммалашган моделлар гибрид моделлар бўлиб, бу моделлар икки ёки бир нечта турли типдаги моделларнинг кучли томонларидан фойдаланган ҳолда мураккаб масалаларни юқори аниқлик билан еча оладиган яхлит модель архитектурасидир.

Агар аниқроқ талқинда айтадиган бўлсак, гибрид моделни қоидаларга асосланган (rule-based), статистик ва нейрон моделларни бир-бирига интеграция қилган яхлит система деб қараш мумкин. NLP да гибрид моделларни қўллаш соҳаларини қуйидаги жадвалдан кўриш мумкин:

Соҳа	Гибрид моделлар
Морфологик таҳлил	FST + BiLSTM
G2P (графема-фонема)	Rule-based + Transformer
POS/NER	CRF + BERT
Синтактик таҳлил	Dependency rules + Neural parser
Машина таржимаси	Neural MT + Rule-based post-editing
Акустик модел (ASR)	HMM + Neural encoder
TTS	DSP rules + Tacotron2

Жадвалдан ҳам кўриниб турибдики, фақатгина Rule Based нинг ўзи ёки фақатгина нейрон моделнинг ўзи барча соҳани қамраб оолмайди ёки барча ҳолатни олий даражада ўргана олмайди, бундай ҳолатда гибрид моделгина идеал ечим бўлиши мумкин. Чунки табиий тилларни қайта ишлашда ҳар бир парадигманинг маълум чекланишлари бор, масалан, Rule-Based модель янги сўзлар, ўзлашма сўзлар ва истисно сўзлар учун жуда заиф бўлса-да, бироқ синтаксис ва морфология қоидаларини тўлиқ ўзига қамраб олган ҳолда аниқ ишлайди. Шунингдек, Machine Learning ёки нейрон моделлар лингвистика соҳасидаги қатъий қоидаларни тўла-тўкис ҳисобга олмаслиги мумкин, бироқ моделлаширишда жуда қулай, контекстни чуқур тушуниши маълум. Гибрид модель иккаласининг ҳам кучли томонларини бирлаштиради, лекин унинг асосий камчилиги мураккаб архитектурага эга, компонентлар ўртасида мувофиқлаштиришни талаб қилади.

Қуйида “CCG + Dependency Parser + Transformer” моделларнинг ўзаро уйғунлигига асосланган гибрид синтактик таҳлил (парсинг)нинг ўзбек тили учун татбиқини амалий мисолда қараймиз.

Ўзбек тилидаги гапларда сўзнинг позиция бўйича эркин жойлашуви ва прагматик акцент хусусиятига кўра уни маъно жиҳатидан фарқлашда битта парсер камлик қилади. Шу сабабли битта гап бир нечта парсерлар орқали текширувдан ўтказилади.

Масалан, "машинани мен ҳайдадим" гапида урғу "мен"га тушганда бу гап "машинани ким ҳайдади?" саволига жавоб тариқасида бўлади, агар урғу "машинани" сўзига тушса, "қайси техникани ҳайдадинг?" саволига жавоб берилгандек бўлади.

Демак дастлаб, биринчи қадамда мазкур жумлани Transformer – базавий морфо-контент таҳлилдан ўтказиб оламиз, бунда транформер ҳар бир токен учун контекстуал боғланишни ярартади:

Token	POS	Morphology	Context Vector
Mashinani	NOUN	ACC	v_1
men	PRON	1SG	v_2
haydadim	VERB	PAST.1SG	v_3

Иккинчи қадамда Dependency Parser модели орқали гап структурасини аниқлашга киришамиз:

Head	Dependent	Relation
haydadim	men	nsubj
haydadim	Mashinani	obj

Учинчи қадамда қоидага асосланган категориялар билан ишловчи CCG (Combinatory Categorial Grammar) парсердан ўтказамиз (лекин CCG парсер ҳам Dependency Parser сингари прагматик акцентни ҳисобга олмайди):

Token	CCG Category
haydadim	(S\NP)/NP
men	NP
Mashinani	NP

Навбатдаги босқич гибрид босқичдан иборат бўлиб, бунда биз Transformer-based Pragmatic Accent Detector орқали "fronted object"ни аниқлаб оламиз:

Token	Position	Syntax Role	Attention Score	Accent Score
Mashinani	sentence-initial	obj	0.82	**0.91**
men	medial	subj	0.34	0.45
haydadim	final	verb	0.21	0.21

Бу ерда Attention Score қийматини аниқлаш учун қуйидаги формула ишлатилди:

$$Attention(Q,K,V) = soft\,max\left(\frac{QK^T)}{\sqrt{d_k}}\right)V$$

Accent Score синтактик, прагматик ва нейрон белгиларнинг бирлашган кўрсаткичи бўлиб, унинг қиймати учун қуйидаги формулани ёзамиз:

$$AccentScore(t_i) = \lambda_1 \cdot SyntaxFocus(t_i) +$$

$$+\lambda_2 \cdot AttentionWeight(t_i) + \lambda_3 \cdot Pr\,agmaticMar\,ker(t_i)$$

,

$$\lambda_1 + \lambda_2 + \lambda_3 = 1.$$

Шундай қилиб, юқоридаги жадвалда Accent Score нинг қиймати максимал 0,91 бўлгани боис Accent = "машинани" бўлади. Бу эса "айнан машинани мен ҳайдадим (бошқа нарсани эмас)" деган маънони беради.

Натижа ва муҳокама.

Агглютинатив типдаги туркий тилларнинг морфологик жиҳатдан мураккаблиги, уларда қўшимчаларнинг юксак продуктивлиги, фонологик қоидаларнинг кўплиги ва истисно ҳолатларнинг мавжудлиги туфайли классик қоидага асосланган моделлар бу тиллар учун етарли даражада самара бермайди. Шунингдек, фақат нейрон тармоқларга асосланган моделлар ҳам баъзан морфологик қоидаларни етарли даражада ҳисобга олмаслиги мумкин. Шу сабабли NLP соҳасида гибрид моделлар охирги ўн йилликда тобора кенг қўлланилмоқда.

Хулоса.

Юқорида кўрганимиздек, туркий тиллар грамматикасини компьютерда формал қайта ишлаш учун универсал бир модел йўқ.

Масалан, морфологик қоидаларни FST орқали, контекстли маъно ва синтактик боғланишларни эса нейрон тармоқлар орқали компьютерли моделлаштириш мумкин. Шу боис энг самарали ёндашув – гибрид моделлаштириш бўлиб, бу йўналиш техник фанларда NLP илмий изланишлари учун долзарб ҳисобланади.

Фойдаланилган адабиётлар рўйхати:

1. Абдурасулов Ё. Туркий тилларнинг қиёсий-тарихий грамматикаси. – Т.: “Фан”, 2009. – 260 б.
2. Steedman, Mark. "A formal universal of natural language grammar." Language, vol. 96 no. 3, 2020, p. 618-660
3. Marco Kuhlmann, Alexander Koller, Giorgio Satta; Lexicalization and Generative Power in CCG. Computational Linguistics 2015; 41 (2): 215–247.
4. Mahmud, Tanzim & Hasan, K. M. & Ahmed, Mahtab & Chak, Thwoi. (2015). A rule based approach for NLP based query processing. 78-82. 10.1109/EICT.2015.7391926.
5. Collins, Michael. (2003). Head-Driven Statistical Models for Natural Language Parsing. Computational Linguistics. 29. 589-637. 10.1162/089120103322753356.
6. Norov A. M. & Tog‘ayev I. B. (2025). FONETIK TRANSKRIPSIYALI NUTQ SINTEZI UCHUN LINGVOTEXNOLOGIK YONDASHUV. Development Of Science, 5(3), pp. 358-369. https://doi.org/0.
7. Norov A. M. (2025). O‘ZBEK TILI UCHUN NLPDA PRAGMATIK AKSENTNI ANIQLASH ALGORITMI. Development Of Science, 12(4), pp. 182-187. https://doi.org/0.
8. Abdisait M. Norov, Ilxom B. Tog'ayev, “THE PROBLEM OF COMPUTER MODELING OF ORTHOGRAPHIC TRANSLITERATION”. Innovative: International Multidisciplinary Journal of Applied Technology (2995-486X), vol. 3, no. 5, June 2025, pp. 74-81, https://multijournals.org/index.php/innovative/article/view/3392.
9. Norov A. M., Tog'ayev I. B., Jorabekov T. K., Murodov S. A., Научно-методический журнал “Открытое и дистанционное образование”. № 1 (85), 2025. ISSN 1609-5944. – С. 7-18 https://journals.tsu.ru/ou/&journal_page=archive&id=2475 doi: 10.17223/16095944/85/1.
10. Norov Abdisait Muradovich, Tog‘ayev Ilxom Baxtiyorovich, Abstract book of the Scientific Conference “Electro-physics and information technology applications-2025”. Toshkent – 2025. – B. 43-46.