O‘ZBEK TILI UCHUN NLPDA PRAGMATIK AKSENTNI ANIQLASH ALGORITMI

Norov Abdisait Muradovich

Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti doktoranti (DSc), abdusaidnorov@gmail.com

Annotatsiya. Pragmatik aksent nutqda muayyan leksema yoki grammatik tuzilmaning kommunikativ ahamiyatlilik darajasini belgilaydigan lingvistik ko‘rsatkich bo‘lib, u onlayn muloqot dasturlari, intellektual ta’lim platformalari, mashinaviy tarjima tizimlari va boshqalarni kompyuterli modellashtirishda asosiy omillardan biri hisoblanadi. Mazkur maqolada o‘zbek tilida pragmatik aksentni aniqlash vazifasiga mo‘ljallangan tabiiy tilni avtomatik qayta ishlash (NLP) algoritmi taklif etilib, uning matemtaik modeli hamda Python dasturlash tilidagi psevdokodi ham keltirilgan.

Kalit so‘zlar: NLP, sintaktik tuzilma, pragmatik aksent, Python, psevdokod, Syntax Score, Accent Score, gibrid baholash.

Аннотация. Прагматический акцент представляет собой лингвистический показатель, определяющий степень коммуникативной значимости определённой лексемы или грамматической конструкции в высказывании. Он является одним из ключевых факторов при компьютерном моделировании онлайн-коммуникационных приложений, интеллектуальных образовательных платформ, систем машинного перевода и других подобных технологий. В данной статье предлагается алгоритм автоматической обработки естественного языка (NLP), предназначенный для идентификации прагматического акцента в узбекском языке, а также приводятся его математическая модель и псевдокод на языке программирования Python.

Ключевые слова: NLP, синтаксическая структура, прагматический акцент, Python, псевдокод, Syntax Score, Accent Score, гибридная оценка.

Abstract. Pragmatic accent is a linguistic indicator that determines the degree of communicative salience of a particular lexeme or grammatical structure within an utterance. It serves as one of the key factors in the computational modeling of online communication applications, intelligent educational platforms, machine translation systems, and similar technologies. This article proposes a natural language processing (NLP) algorithm designed to identify pragmatic accent in the Uzbek language and presents its mathematical model along with pseudocode implemented in the Python programming language.

Keywords: NLP, syntactic structure, pragmatic accent, Python, pseudo code, Syntax Score, Accent Score, hybrid evaluation.

Kirish

Pragmatik aksent deganda, gapda muayyan so‘z yoki ma’noga alohida urg‘u berish, ya’ni nutqdagi kommunikativ maqsadga ko‘ra ma’lum elementni ko‘proq ahamiyatli qilib ko‘rsatish tushuniladi. Bu ikki tushunchani alohida-alohida qaraydigan bo‘lsak, pragmatikani “gap ortidagi maqsad, niyat, kontekst” deb , aksentni esa “urg‘u, diqqatni jamlash” deb tushunish mumkin. Demak, pragmatik aksent gapda e’tibor qaysi joyda bo‘lishiga bog‘liq urg‘udir.

O‘zbek tilida pragmatik aksent bir nechta holatlarda yuzaga keladi, xususan, so‘z tartibini o‘zgartirganda, intonatsiyani o‘zgartirganda yoki leksik markerlarni faol ishlatishda yuzaga keladi. Masalan, qoidaga binoan oddiy tartibda berilgan “Men kitobni ko‘rdim” gapida “kitob” so‘ziga urg‘u berilgan bo‘lib, “nimani ko‘rding” savolisha javob bo‘lmoqda. Agar gap urg‘usini “men”ga qaratish kerak bo‘lsa, u holda “Kitobni men ko‘rdim” singari gap tuzilmasini hosil qilish kerak bo‘ladi. Shunday qilib, pragmatik aksent nutqda gapning qaysi qismi muhim ekanligini ko‘rsatuvchi urg‘u, so‘z tartibi, intonatsiya, maxsus so‘zlar va kontekst bilan ifoda etiladi.

NLP metodlaridan foydalangan holda o‘zbek tili uchun bunday pragmatik aksentni kompyuterli modellashtirishda dastlab fokus so‘zlarni aniqlash, intonatsiya markazini topish, yangi deb qaralgan axborotni belgilab olish hamda model uchun muhim sanalgan so‘zni o‘rganishga to‘g‘ri keladi.

Bu o‘zbek tilidagi gaplarning sintaktik tuzilmalarini kompyuterili modellashtirishdagi dastlabki bosqichlardan biri bo‘lib, hozirgacha bu boradagi muammolar ilmiy-amaliy jihatdan oxirigacha o‘rganilmagan.

Adabiyotlar tahlili

[1] da prosodik fokusni avtomatik ravishda aniqlashga oid amaliy tadqiqot ishlari berilgan bo‘lib, unda intonatsiyadan aksentni chiqarish uchun akustik-statistik xususiyatlar va mashinaviy o‘rganishni qo‘llash haqida aniq metodlar bayon etilgan.

[2] da amaliy jihatdan pragmatikaga bag‘ishlangan baholash to‘plami, ya’ni pragmatik fenomenlarni o‘rganish va modellarni tekshirish uchun resurslar tuzish borasida muhim fikr-mulohazalar yuritilgan bo‘lib, NLPda pragmatika va fokus tasnifiga yordam beradigan metodlar yoritilgan.

[3] da spontan nutqdan pragmatik va semantik farqni ajratib olish uchun ikkita lingvistik tushuncha: kontrast va fokusni hisoblash usuli ko‘rib chiqilgan. Kontrast va fokus zamonaviy tilshunoslikda intonatsiya va axborot/diskurs tuzilishini bog‘laydigan kategoriyalar sifatida keng o‘rganilgan.

[4] da mahalliy va mahalliy bo‘lmagan tillardagi gap urg‘usi bu ikki tilga mansub bo‘lgan prosodik elementlarga bog‘liqligi, shuningdek, tillararo farqlarda mavju bo‘lgan fon shovqinining mavjudligi gap urg‘usini aniqlashga turlicha ta’sir ko‘rsatishi tadqiq etilgan.

ISSN 3030-3907

[5] da prosodik xususiyatlar orasida qaysi biri pragmatik funksiyalarni yetkazishda muhim ekanligi va bu modelda qaysi akustik xususiyatlarga e'tibor qaratish kerakligi empirik ravishda tahlil qilingan.

Usullar.

O'zbek tili uchun NLPda pragmatik aksentni aniqlashga doir algoritmini qurish va dasturlash bir nechta bosqichlarda amalga oshiriladi (Python misolida).

1. Foydalanuvchi tomonidan kiritilgan matnni normalizatsiya qilish bosqichi (dastlabki bosqich). Bunda matn tarkibida bir nechta amallar bajariladi, xususan, Unicode tizimidagi " ‘ " (kodi: u8216), "ҳ" (kodi: u1203), "қ" (kodi: u1179) va shunga o'xshash simvollarni ASCII formatiga o'tkazish uchun orfografik normalizatsiya (o' → o', g' → g') o'tkaziladi. So'ng bosh harf bilan boshlangan barcha so'zlar mos kichik harflar bilan almashtiriladi. Matndagi tinish belgilari olib tashlanadi.

Psevdokod:

text_norm = Normalize(text)

2. Tokenizatsiya va POS tagging bosqichi. Bunda matn so'zlar bo'yicha, so'zlar esa harflar bo'yicha tokenizatsiya qilinadi. Tokenizatsiya qilingan har bir so'zning qaysi so'z turkumiga xosligi aniqlanadi va tahlil qilinadi.

Psevdokod:

tokens, pos_tags = POS_Tag(text_norm)

3. Sintaktik tuzilmani aniqlash (Dependency Parsing) bosqichi. Gapda pragmatik aksent ko'pincha kesimdan oldin joylashgan hol (rema (new info)) pozitsiyasida keladi. Kompyuterli modellashtirishda biz izlagan obyekt aynan shuni ifodalaydi.

Psevdokod:

dep_tree = DependencyParse(tokens)

4. Tematik-rematik tuzilmani (Theme-Rheme) aniqlash bosqichi. Bunda o'zbek tilidagi maxsus qoidalar asosida jumla boshidagi muhim so'z (aksent), kesimdan oldin kelgan qo'shimcha ma'lumot (Rema) hamda "hatto", "aynan", "hatto", "faqat", "esa" kabi markerlar aksent sifatida aniqlanadi.

Psevdokod:

if token.dep == "obj" and token.position == 1:
token.is_rheme = True

5. Pragmatik markerlarni aniqlab olish bosqichi. O'zbek tilida shunday so'zlar borki, ularni gap boshida, gap o'rtasida yoki gapning oxirida ishlatish eshituvchining diqqatini tortadigan jumlaning paydo bo'lishiga olib keladi. Bunday so'zlar jumlasiga "aynan", "faqat", "faqatgina", "biroq", "lekin", "ammo", "esa", "xuddi" va shunga o'xshash ko'plab markerlarni kiritish mumkin. Bu singari markerlarning aksariyati jumlaning ma'lum shartlar bilan berilganligini ifoda etadi.

ISSN 3030-3907

Psevdokod:

```
for token in tokens:
    if token.lemma in pragmatic_markers:
        token.accent_score += 0.4
```

6. Attention-based Neural Weighting bosqichi. Transformersda attention mexanizmi pragmatik aksentni tabiiy ravishda “chizadi”.

Psevdokod:

```
attn = BERT_Attention(text_norm)
```

Bunda aksent uchun quyidagi formula o‘rinli bo‘ladi:

$$\text{Aksent} = AW + SW + PM,$$

bu yerda:

AW – attention weight;
SW – syntactic weight;
PM – pragmatic markers.

7. Accent Score qiymatini hisoblash bosqichi. Bu uch komponentli gibrid baholashni amalga oshirish uchun psevdokod quyidagicha yoziladi:

```
accent_score =
    0.45 * SyntaxFocus(token)
  + 0.35 * AttentionWeight(token)
  + 0.20 * PragmaticMarkers(token)
```

Ushbu bosqichda NLPda pragmatik aksentni gibrid (syntax + pragmatic markers + attention) model orqali aniqlash uchun quyidagicha hisoblashlar amalga oshiriladi:

1. Syntax Score (SYN) ning qiymatini topish uchun quyidagi formuladan foydalanamiz:

$$SYN(t) = w_s \cdot f_{front}(t) + w_r \cdot f_{role}(t).$$

2. Marker Score (MKR) ni hisoblash uchun quyidagi formulaga murojaat qilamiz:

$$MKR(t) = \sum_{i=1}^{n} m_i(t)$$

3. BERT/Transformer modelining tokenga ajratgan diqqat vazni uchun **Attention Score (ATT)** ning qiymati quyidagicha hisoblanadi:

$$ATT(t) = \frac{1}{H}\sum_{h=1}^{H}\left(\frac{1}{L}\sum_{l=1}^{L} att_{h,l}(t)\right)$$

4. Yakuniy Accent Score (FINAL) protsedurasi asosida Syntax + Marker + Attention birlashtirilib, yakuniy pragmatik aksent qiymati hisoblanadi:

$$FINAL(t) = \alpha \cdot SYN(t) + \beta \cdot MKR(t) + \gamma \cdot ATT(t)$$

ISSN 3030-3907

8. Aksent eng yuqori bo‘lgan so‘z yoki so‘zlarni aniqlab beradigan yakuniy bosqich. Bu bosqichda berilgan gap (jumla)dagi urg‘u qaysi so‘zga aynan qaratilganligi aniqlanadi va natija ekranga chiqariladi.

Psevdokod:

accent_token = argmax(accent_score)

return accent_token

Natija va muhokama.

Shunday qilib, “**Men mashinani haydadim**” jumlasini aksentlash orqali qayta hosil qiladigan bo‘lsak, hosilaviy jumlalardan aynan biri “**Mashinani men haydadim**” ko‘rinishidagi tuzilmaga ega bo‘ladi. Bu jumladagi pragmatik aksentni topish uchun yuqoridagi formulular asosida hisoblangan quyidagi jadvalga e’tibor qaratamiz:

Token	Syntax	Marker	Attention	Calculation	Final Score
Mashinani	1.0	0	0.30	0.45×1.0 + 0.35×0.30 + 0.20×0	0.555
men	0.1	0	0.40	0.45×0.1 + 0.35×0.40 + 0.20×0	0.185
haydadim	0.0	0	0.20	0.45×0.0 + 0.35×0.20 + 0.20×0	0.07

Jadvaldagi hisob-kitoblardan ko‘rinib turibdiki, accent_score ning maksimal qiymati 0.555 ga teng bo‘lib, u “Mashinani” so‘ziga (tokeniga) mos kelyapti. Demak, pragmatik aksent “**Mashinani**” tokenining ustida joylashgan.

Xulosa.

Hozirgi vaqtda o‘zbek tilini semantik, morfologik va fonetik jihatdan kompyuterli modellashtirish uchun kam bo‘lsa-da resurslar mavjud, biroq pragmatik aksentni aniqlovchi korpus deyarli yo‘q. Bu esa NLP metodlarini to‘la tatbiq etishda, intellektual ta’lim hamda mashinaviy tarjima tizimlarini yaratishda jiddiy to‘siq bo‘lib hisoblanadi. Chunki maqolada keltirilgan modellar xorijiy tillarga nisbatan o‘tkazilgan tadqiqotlar mahsuli bo‘lib, o‘zbek tili uchun pragmatik aksentga mo‘ljallangan katta korpuslar hali mavjud emas. Shu bois o‘zbek tiliga oid pragmatik aksentni aniqlovchi algoritmlarni qurish va kompyuterli modellarni yaratish uchun pragmatik korpuslarga kuchli ehtiyoj seziladi.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI

1. Sunghye Cho, Mark Liberman, and Yong-cheol Lee, Automatic Detection of Prosodic Focus in American English. INTERSPEECH 2019, September 15–19, 2019, Graz, Austria, pp. 3470-3474.

2. Damien Sileo, Tim Van de Cruys, Camille Pradel, Philippe Muller, A Pragmatics-Centered Evaluation Framework for Natural Language Understanding.

ISSN 3030-3907

Proceyedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), 20-25 June 2022, pp. 2382-2394.
3. Tong Zhang, Mark Hasegawa-Johnson, Stephen E. Levinson, Extraction of pragmatic and semantic salience from spontaneous spoken English, Speech Communication, Volume 48, Issues 3–4, 2006, Pages 437-462.
4. Scharenborg O, Kakouros S, Post B, Meunier F. Cross-linguistic Influences on Sentence Accent Detection in Background Noise. Lang Speech. 2020 Mar;63(1):3-30. doi: 10.1177/0023830918819573. Epub 2019 Jan 4. PMID: 30606083; PMCID: PMC7066482.
5. Nigel G. Ward, Divette Marco, Olac Fuentes, WHICH PROSODIC FEATURES MATTER MOST FOR PRAGMATICS? arXiv:2408.13240v1 [cs.CL] 23 Aug 2024. – 5 p.
6. Norov A.M. The initial practical matters of creation formal grammar of Uzbek language // East European Scientific Journal, #6 (46), 2019. P. 27-30.
7. Norov A.M., Aripov M., Ismailov K.A. The algorithmic bases of development the semantic-syntactic parser belong to numerical information in the textual form // “Science and Education in Karakalpakstan”. – Nukus, 2019. – №3(11). – P. 75-80.
8. Арипов М., Норов А. Языковое моделирование разговорной узбекской речи / Материалы республиканской научной конференции «Актуальные проблемы и применения анализа». КарГУ. Карши, 4-5 октябрь, 2019. С. 290-292.
9. Арипов М., Норов А. Разработка онтологической модели синтаксических правил узбекского языка / Материалы международной конференции «Актуальные проблемы прикладной математики и информационных технологий». НУУз. Ташкент, 14-15 ноябрь, 2019. С. 275-276.
10. Norov A.M., Jorabekov T.K., Murodov Sh.A. The Relationship of linguistic models and Lingua-mathematical models of Turkic languages // Science and Education in Karakalpakstan. №3(18). 2021. P. 70-72.

ISSN 3030-3907