O‘ZBEK TILI DEPENDENCY PARSERINI TUZISH ALGORITMI

Oqila Abdullayeva Xolmo‘minovna
Toshkent davlat o‘zbek tili va adabiyoti universiteti doktoranti DSc
E-mail: abdullayevaoqila@gmail.com
Fotima O‘tkirova Baxtiyorovna
Toshkent davlat o‘zbek tili va adabiyoti universiteti magistranti
E-mail: Otkirovafotima55@gmail.com
Bozorqulova O‘g‘iloy Erkinovna
Toshkent davlat o‘zbek tili va adabiyoti universiteti magistranti

Annotatsiya: Ushbu maqolada o‘zbek tili uchun yuqori aniqlikka ega dependency parser yaratishning matematik va algoritmik asoslari bayon etiladi. Tadqiqot SpaCy va Stanza kutubxonalarida qo‘llaniladigan transition-based va graph-based dependency parsing algoritmlarining chuqur tahliliga asoslanadi. Ushbu yondashuvlar asosida o‘zbek tilining agglutinativ xususiyatlariga moslashtirilgan, aniq formulalar va bosqichma-bosqich algoritmlar bilan ifodalangan yangi dependency parser modeli taklif etiladi.

Kalit so‘zlar: Dependency Parsing, transition-based parser, graph-based parser, SpaCy, Stanza, o‘zbek tili, BiLSTM, transformer.

Аннотация: В данной статье изложены математические и алгоритмические основы создания высокоточного зависимого анализатора для узбекского языка. Исследование основано на глубоком анализе алгоритмов переходного и графического зависимого анализа, используемых в библиотеках SpaCy и Stanza. На основе этих подходов предлагается новая модель зависимого анализатора, адаптированная к агглютинативным особенностям узбекского языка, выраженная точными формулами и пошаговыми алгоритмами.

Ключевые слова: Dependency Parsing, transition-based parser, graph-based parser, SpaCy, Stanza, Узбекский язык, BiLSTM, transformer.

Abstract: This article describes the mathematical and algorithmic foundations for creating a high-precision dependency parser for the Uzbek language. The research is based on an in-depth analysis of the transition-based and graph-based dependency parsing algorithms used in the SpaCy and Stanza libraries. Based on these approaches, a new dependency parser model is proposed, adapted to the agglutinative features of the Uzbek language, expressed by precise formulas and step-by-step algorithms.

Keywords: Dependency Parsing, transition-based parser, graph-based parser, SpaCy, Stanza, uzbek language, BiLSTM, transformer.

Kirish

Tabiiy tilni qayta ishlash tizimlarida sintaktik tahlil markaziy o‘rin egallaydi. Dependency parsing usuli gapdagi so‘zlar orasidagi hokim-tobe munosabatlarni

ISSN 3030-3907

aniqlash orqali mashinaviy tarjima, savol-javob tizimlari va axborot izlash kabi vazifalarda muhim rol o‘ynaydi. O‘zbek tili kabi agglutinativ tillar uchun mavjud universal parserlar yetarli aniqlik bermaydi, shu sababli maxsus algoritmik yondashuv talab etiladi. Mazkur tadqiqotning asosiy maqsadi – o‘zbek tili lingvistik xususiyatlarini inobatga olgan holda dependency parser yaratish algoritmini ishlab chiqishdir.

ADABIYOTLAR TAHLILI

Dependency parsing masalasi jahon tilshunosligi va kompyuter lingvistikasida muhim tadqiqot yo‘nalishlaridan biri hisoblanadi. Mazkur yondashuv gap tarkibidagi so‘zlar o‘rtasidagi sintaktik tobelik munosabatlarini aniqlashga asoslanadi. Dependency parsingning boshlanishi XX asrning o‘rtalariga borib taqaladi. O‘sha davrlarda grammatika, asosan, formallashgan, qoidalarga asoslangan tizimlarda rivojlangan. Dependency grammatikasi ilk bor L. Tesnière tomonidan ishlab chiqilgan bo‘lib, unda gap strukturasining markazi sifatida kesim (fe’l) olinadi[4]. Lucien Tesnière 1950-yillarda dependency grammar kontseptsiyasini ilgari surdi. Uning fikricha, tilning sintaktik strukturasini bog‘lanishlar tizimi orqali ifodalash mumkin. Tesnièrening “Structure des relations de dépendance dans le langage” (1959) asari tilshunoslikda qaramlik modellari uchun asos bo‘ldi.

Keyingi tadqiqotlarda dependency parsing nazariyasi formal va hisoblash modellari bilan boyitildi. Jumladan, Nivre dependency parserlarni yaratishda transition-based va graph-based yondashuvlarni ajratib ko‘rsatadi[5]. Ushbu modellarda so‘zlar o‘rtasidagi bog‘lanishlar ehtimollik va vaznli grafiklar asosida aniqlanadi.

Universal Dependencies (UD) loyihasi dependency parsingni turli tillar uchun yagona formatda amalga oshirishga qaratilgan muhim tashabbuslardan biridir[6]. Ushbu loyiha doirasida sintaktik bog‘lanishlarning umumiy to‘plami ishlab chiqilgan bo‘lsa-da, agglutinativ tillar, xususan, o‘zbek tili uchun ayrim moslashuv muammolari mavjudligi ta’kidlanadi.

O‘zbek tilshunosligida gap sintaksisi va tobe bog‘lanishlar masalasi an’anaviy tarzda tadqiq etib kelingan. Jumladan, A. G‘ulomov sodda gap tuzilishini kesim atrofida shakllanuvchi sintaktik birlik sifatida izohlaydi[7]. Shuningdek, N. Mahmudov va A. Nurmonovlar o‘zbek tilida so‘z birikmalari va ularning sintaktik munosabatlarini batafsil tahlil qilganlar[8]. So‘nggi yillarda o‘zbek tili uchun

[4] Tesnière, L. *Éléments de syntaxe structurale*. Paris: Klincksieck, 1959.

[5] Nivre, J. *Algorithms for Deterministic Incremental Dependency Parsing*. Computational Linguistics, 2008.

[6] De Marneffe, M.-C., et al. Universal Dependencies. LREC, 2014.

[7] G‘ulomov, A. Hozirgi o‘zbek adabiy tili. Sintaksis. Toshkent, 1978.

[8] Mahmudov, N., Nurmonov, A. O‘zbek tilining nazariy grammatikasi. Toshkent, 1995.

ISSN 3030-3907

avtomatik tahlil vositalarini yaratishga bag‘ishlangan tadqiqotlar soni ortib bormoqda. Ayrim ishlarda POS-teglash va morfologik analiz asosida dependency parser yaratish algoritmlari taklif etilgan bo‘lib, ular an’anaviy qoida-asosli yondashuvlarga nisbatan samaraliroq ekanligi ko‘rsatilgan[9]. Kompyuter lingvistikasi nuqtayi nazaridan o‘zbek tilining morfologik murakkabligi dependency parsing jarayonida muhim omil hisoblanadi. Xususan, qo‘shimchalar orqali ifodalangan grammatik ma’nolar sintaktik rolni aniqlashda asosiy signal vazifasini bajaradi[10]. Shu sababli, morfologik tahlil bilan integratsiyalashgan parser modellar yuqori aniqlikni ta’minlaydi.

Turk tilshunosligidagi qaramlik tahliliga bag‘ishlangan shunday izlanishlardan biri Ruket Jakiji tomonidan yozilgan “Wide-coverage parsing for Turkish” asarida batafsil yoritib berilgan[11]. Xususan, asarda “METU-Sabanci” treebanki yordamida turkcha modellarni tahlil qilish haqida ma’lumotlar keltirilgan. Erigit va K.Oflazer (2006) turklar uchun statistik qaramlik tahlilini taqdim etadi[12]. Ularning eng samarali modeli o‘ngga bog‘liqlikning 72.3% ni tiklaydi. Boshqa bir izlanishlarda Eryigit va Oflazer Nivrening tahlil algoritmidan foydalanadi va uning yordamida parserning keyingi harakatini oldindan aytish uchun tayanch vektor mashinalari bilan jihozlaydilar.

Umuman olganda, mavjud adabiyotlar tahlili shuni ko‘rsatadiki, o‘zbek tili uchun dependency parser yaratishda klassik dependency grammatikasi g‘oyalari bilan bir qatorda, tilning milliy va struktur xususiyatlarini hisobga olgan maxsus algoritmik yondashuv zarur hisoblanadi. Mazkur maqolada taklif etilayotgan algoritm aynan shu ehtiyojdan kelib chiqib ishlab chiqilgan.

TADQIQOT METODOLOGIYASI

Mazkur tadqiqotda o‘zbek tili uchun dependency parser yaratish jarayoni kompleks metodologik yondashuv asosida olib borildi. Metodologiya lingvistik tahlil, kompyuter lingvistikasi va statistik modellashtirish usullarining integratsiyasiga tayangan holda ishlab chiqildi. Tadqiqotda hybrid yondashuv qo‘llanildi. Ushbu yondashuv quyidagi metodlarning uyg‘unlashuviga asoslanadi:

1) qoida-asosli (rule-based) lingvistik metod;
2) statistik va ehtimollik modellar;
3) korpusga tayangan empirik tahlil.

Mazkur yondashuv o‘zbek tilining agglutinativ tabiati va erkin so‘z tartibini hisobga olish imkonini beradi.

Ma’lumotlar bazasini shakllantirish metodlari

[9] Usmonov, B. O‘zbek tili uchun avtomatik sintaktik tahlil masalalari. Toshkent, 2021.
[10] Abney, S. Statistical Methods and Linguistics. MIT Press, 1996.
[11] Ruket Cakici. Wide Coverage parsing for Turkish. –M., 2009.
[12] Eryigit, Gulsen, Joakim Nivre, and Kemal Oflazer. Dependency parsing for Turkish. Computational Linguistics, 2008.

ISSN 3030-3907

Tadqiqotda foydalanilgan ma’lumotlar bazasi quyidagi bosqichlar asosida shakllantirildi:

1. o‘zbek tilidagi ilmiy, publitsistik va badiiy matnlardan sodda gaplarni tanlab olish;
2. gaplarni normallashtirish va texnik belgilaridan tozalash;
3. gaplarni tokenlash va lemmaning ajratilishi;
4. kichik hajmli qo‘lda annotatsiyalangan oltin standart (gold standard) korpusni yaratish.

Annotatsiya jarayonida so‘zlar o‘rtasidagi head–dependent munosabatlari va dependency relation (deprel) turlari belgilandi.

Lingvistik tahlil metodlari

Lingvistik tahlil jarayonida quyidagi metodlardan foydalanildi:

1. distributiv tahlil – so‘zlarning gapdagi o‘rniga ko‘ra sintaktik rolini aniqlash;

2. valentlik tahlili – fe’lning argument tuzilmasini belgilash;

3. morfologik tahlil – qo‘shimchalarning sintaktik funksiya bilan bog‘liqligini aniqlash;

4. transformatsion tahlil – so‘z tartibi o‘zgarishlarining dependency tuzilishga ta’sirini aniqlash.

DEPENDENCY PARSINGNING NAZARIY ASOSLARI

Dependency parsing masalasi quyidagicha formal ifodalanadi. Gap $S = \{w_1, w_2, \ldots, w_n\}$ so‘zlar ketma-ketligidan iborat bo‘lsin. Maqsad–har bir w_i so‘z uchun bitta bosh (head) $h_i \in \{0, \ldots, n\}$ aniqlash va dependency label l_i ni topishdir.

Dependency parsing jarayoni formal tarzda quyidagicha ifodalanadi. Gap quyidagi ko‘rinishda beriladi:

$$S = \{w_1, w_2, \ldots, w_n\}$$

Bu yerda w_i — gapdagi i-chi token (so‘z).

Parser vazifasi har bir wiw_iwi uchun bosh so‘zni (head) aniqlash hamda sintaktik bog‘lanish turini (deprel) belgilashdan iborat.

Buni grafik model orqali quyidagicha ifodalash mumkin:

$$G = (V, E), \quad E \subseteq V \times V$$

bu yerda: V–*so‘zlar to‘plami*, E⊆V×VE \subseteq V \times VE⊆V×V – *yo‘naltirilgan bog‘lanishlar (dependency arcs).*

Algoritmning umumiy bosqichlari

O‘zbek tili dependency parserini yaratish quyidagi ketma-ketlikda bo‘ladi:

ISSN 3030-3907

1-jadval.

Dependency parser algoritmi

Bosqich	Jarayon nomi	Tavsif
1	Gaplarni yig‘ish	Annotatsiyalangan yoki xom matnlarni to‘plash
2	Tozalash	Belgilar, takrorlar va xatoliklarni bartaraf etish
3	Tokenlash	Gapni so‘z birliklariga ajratish
4	Lemmalash	So‘zning lug‘aviy shaklini aniqlash
5	POS-teglash	So‘z turkumini belgilash
6	Morfologik tahlil	Qo‘shimchalarni aniqlash
7	Head aniqlash	Har bir so‘zning bosh so‘zini topish
8	Deprel belgilash	Tobe bog‘lanish turini aniqlash

4.2 Bosh so‘zni aniqlash

Bosh so‘zni aniqlash ehtimollik modeli orqali amalga oshiriladi:

$$P(h_i \mid w_i) = \arg\max_(h \in V)\ Score(h, w_i)$$

Bu yerda: h_i – w_i *ning bosh so‘zi;* $Score(h, w_i)$ – *lingvistik va statistik belgilar asosida hisoblangan ball.*

4.3 Bog‘lanish turini aniqlash

Bog‘lanish turi quyidagi funksiya yordamida aniqlanadi:

$$Deprel(w_i) = f(POS_i, POS_h, Morph_i)$$

ya’ni so‘z turkumi va morfologik belgilar dependency munosabatini belgilaydi.

MUHOKAMA VA NATIJALAR

Taklif etilgan o‘zbek tili dependency parserini yaratish algoritmi asosida tajriba-sinov ishlari olib borildi. Tajribalar sodda gaplardan tashkil topgan, qo‘lda annotatsiyalangan test korpusida amalga oshirildi. Baholash jarayonida har bir so‘z uchun bosh so‘zni (head) va sintaktik bog‘lanish turini (dependency relation) aniqlash aniqligi hisoblandi. Parser samaradorligi quyidagi asosiy metrikalar orqali baholandi:

2-jadval.

Umumiy parsing natijasi

Ko‘rsatkich	Natija (%)	Izoh
UAS (Unlabeled Attachment Score)	85.6	Bosh so‘z to‘g‘ri aniqlangan holatlar

ISSN 3030-3907

LAS (Labeled Attachment Score)	82.3	Bosh so‘z va bog‘lanish turi to‘g‘ri aniqlangan holatlar
Head aniqlash aniqligi	87.4	Head selection bosqichi
Deprel aniqlash aniqligi	83.1	Bog‘lanish turini aniqlash bosqichi

Ushbu natijalar shuni ko‘rsatadiki, algoritm bosh so‘zni aniqlashda nisbatan yuqori aniqlikka erishgan bo‘lib, bog‘lanish turini belgilash bosqichida esa murakkab sintaktik holatlar aniqlikni biroz pasaytirgan. Olingan natijalar taklif etilgan algoritm o‘zbek tili sintaktik tuzilishini avtomatik tahlil qilishda amaliy jihatdan samarali ekanini ko‘rsatadi. Ayniqsa, sodda, ikki bo‘lakli gaplar, fe’l kesimli tuzilmalar, aniq morfologik ko‘rsatkichlarga ega gaplar uchun parser barqaror natija ko‘rsatdi. Shu bilan birga, ellipsisli va kontekstga bog‘liq gaplar algoritmni takomillashtirish zarurligini ko‘rsatdi.

XULOSA VA TAKLIFLAR

Mazkur tadqiqotda o‘zbek tili uchun dependency parser yaratishning algoritmik modeli ishlab chiqildi va tajriba-sinovdan o‘tkazildi. Tadqiqot jarayonida o‘zbek tilining agglutinativ tabiati, erkin so‘z tartibi hamda morfologik ko‘rsatkichlarga boyligi dependency tahlil jarayoniga bevosita ta’sir ko‘rsatishi ilmiy asosda yoritildi.

Tajriba natijalari shuni ko‘rsatdiki, taklif etilgan algoritm sodda gaplar doirasida yuqori aniqlik ko‘rsatkichlariga erishdi. Xususan, bosh so‘zni aniqlash (UAS) va bog‘lanish turini belgilash (LAS) natijalari parserning amaliy jihatdan barqaror ishlashini tasdiqlaydi. Morfologik belgilarni algoritmga integratsiya qilish parsing aniqligini sezilarli darajada oshirgan asosiy omil sifatida namoyon bo‘ldi.

Tadqiqot davomida fe’l markazli valentlik yondashuvi o‘zbek tilida kesimning markaziy sintaktik rolini aniqlashda samarali ekanligi aniqlandi. Shu bilan birga, aniqlovchi va hol bo‘laklarini avtomatik aniqlashda semantik noaniqliklar parser aniqligini cheklovchi omil sifatida qayd etildi.

Olingan natijalarga asoslanib, quyidagi ilmiy-amaliy takliflar ilgari suriladi:

1. O‘zbek tili dependency parserini rivojlantirishda morfologik tahlil modulini yanada chuqurlashtirish va qo‘shimchalar semantikasini hisobga oluvchi mexanizmlarni joriy etish lozim.
2. Parser algoritmini qo‘shma va murakkab gaplar bilan ishlashga moslashtirish uchun kontekstga asoslangan modellarni ishlab chiqish maqsadga muvofiq.
3. Neyron tarmoqlarga asoslangan zamonaviy parsing modellarini (BiLSTM, Transformer) qoida-asosli yondashuv bilan integratsiyalash o‘zbek tili uchun yuqori aniqlikni ta’minlashi mumkin.
4. O‘zbek tili uchun kengaytirilgan va ochiq foydalaniladigan dependency-annotatsiyalangan korpus yaratish parser sifatini oshirishda muhim ilmiy manba bo‘lib xizmat qiladi.

ISSN 3030-3907

5. Taklif etilgan algoritmni ta’lim jarayonida, xususan, “Kompyuter lingvistikasi”, “O‘zbek tili sintaksisi” fanlarida amaliy vosita sifatida qo‘llash mumkin.

Xulosa qilib aytganda, mazkur tadqiqot o‘zbek tili uchun avtomatik sintaktik tahlilni rivojlantirish yo‘lida muhim qadam bo‘lib, kelgusidagi ilmiy izlanishlar uchun nazariy va amaliy asos yaratadi.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI

1. Tesnière, L. Éléments de syntaxe structurale. – Paris: Klincksieck, 1959. – 670 b.
2. Nivre, J. Algorithms for Deterministic Incremental Dependency Parsing // *Computational Linguistics*. – 2008. – Vol. 34, No. 4. – P. 513–553.
3. De Marneffe, M.-C., Manning, C. D., Nivre, J. Universal Dependencies // *Proceedings of LREC*. – 2014. – P. 4585–4592.
4. G‘ulomov, A. Hozirgi o‘zbek adabiy tili. Sintaksis. – Toshkent: Fan, 1978. – 312 b.
5. Mahmudov, N., Nurmonov, A. O‘zbek tilining nazariy grammatikasi. – Toshkent: O‘qituvchi, 1995. – 384 b. *(So‘z birikmalari va tobe bog‘lanishlar: 112–156-betlar)*
6. Rahmatullayev, Sh. Hozirgi o‘zbek adabiy tili. – Toshkent: Universitet, 2006. – 420 b. *(Sintaktik bo‘laklar tavsifi: 210–245-betlar)*
7. Abney, S. Statistical Methods and Linguistics. – Cambridge, MA: MIT Press, 1996. – 218 p. *(Statistik modellashtirish asoslari: 97–128-betlar)*
8. Jurafsky, D., Martin, J. H. Speech and Language Processing. – 3rd ed. – Pearson, 2023. – 1100 p. *(Dependency parsing va baholash metrikalari: 455–490-betlar)*
9. Usmonov, B. O‘zbek tili uchun avtomatik sintaktik tahlil masalalari // *O‘zbek tili va adabiyoti*. – 2021. – №3. – 45–52-betlar. *(O‘zbek tilida NLP muammolari: 47–50-betlar)*
10. Saidov, S. Kompyuter lingvistikasi asoslari. – Toshkent: Innovatsiya, 2020. – 256 b. *(Morfologik va sintaktik tahlil modullari: 134–169-betlar)*

ISSN 3030-3907