UDC 512.643

O‘ZBEK TILI UCHUN UNIVERSAL BOG‘LIQLIK DARAXTI KORPUSI ASOSIDA CHUQUR BI-AFFIN TOBELIK TAHLILINING NEYRON MODELI

Matlatipov San’atbek G‘ayratovich
Mirzo Ulug‘bek nomidagi O‘zbekiston milliy universiteti, Toshkent, O‘zbekiston
s.matlatipov@nuu.uz

ANNOTATSIYA

Ushbu maqolada o‘zbek tili uchun Universal Dependencies (UD) standartlariga mos yangi daraxtsimon korpusi va uning asosida qurilgan chuqur bi-affin neyron tobelik tahlil modeli taqdim etiladi. Korpus o‘zbek adabiy va ilmiy-ommabop matnlaridan tanlangan 686 ta gapni (taxminan 7 800 ta token) o‘z ichiga oladi va INCEpCTION platformasida tilshunoslar hamda NLP muhandislari tomonidan yuqori annotatorlararo moslik (lemmatizatsiya va UPOS bo‘yicha $> 95\%$) bilan belgilandi. Sintaktik tahlil uchun [3] tomonidan taklif etilgan *chuqur bi-affin neyron diqqat mexanizmi* arxitekturasiga asoslangan model qurilib, BiLSTM enkoder va bosh-tobe so‘z juftliklari uchun bi-affin baholash funksiyasi yordamida tobelik grafigi optimallashtirildi. Stanza kutubxonasiga integratsiyalashgan neyron quvur (tokenizatsiya, POS-tagging, morfologik tahlil va dependency parsing) bo‘yicha olib borilgan tajribalar morfologiya kuchli bo’lgan sharoitida Unlabeled Attachment Score (UAS) 69.21% va Labeled Attachment Score (LAS) 53.21% natijalarini ko‘rsatdi; bu ko‘rsatkichlar o‘zbek tili uchun chuqur neyron tobelik tahlilining birinchi mustahkam bazaviy modeli sifatida taklif etiladi va keyingi matematik hamda amaliy tabiiy tillar jarayoni tadqiqotlari uchun poydevor bo‘lib xizmat qiladi.

Kalit so’zlar: Universal Bog’liqliklar, o‘zbek tili, daraxtsimon korpus, tobelik tahlili, chuqur bi-affin neyron diqqat mexanizminig neyron modeli.

KIRISH

Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasida sintaktik tahlil modellari matnli ma’lumotlarni formal struktura sifatida ifodalash va ular ustida algoritmik amallar bajarish imkonini beruvchi asosiy vositalardan biridir. Tobelik grammatikasi yondashuvida gap strukturasi yo‘naltirilgan daraxt ko‘rinishidagi graf bilan modellashtiriladi: har bir so‘z bitta tugunga, sintaktik munosabatlar esa yoylarga mos keladi. Bunday daraxtlarning kombinator xossalarini, barqarorligini va optimallashtirish usullarini o‘rganish lingvistika bilan bir qatorda ehtimollar nazariyasi, grafiklar nazariyasi va optimallashtirish kabi aniq fan yo‘nalishlari uchun ham dolzarbdir.

O‘zbek tili kabi agglyutinativ, erkin so‘z tartibli va kam resursli tillar uchun yuqori sifatli, qat’iy standartlar asosida annotatsiya qilingan sintaktik ma’lumotlar yetishmasligi zamonaviy neyron modellarning samarali o‘qitilishi va ularning xatoliklarini matematik jihatdan tahlil qilishni cheklab kelmoqda. Universal Dependencies (UD) loyihasi turli tillar uchun yagona formalizm va annotatsiya tamoyillarini taklif etib, ko‘p tilli *universal bog‘liqlik daraxti korpuslari* majmuasini shakllantirgan bo‘lsa-da, o‘zbek tili uchun mavjud resurslar hajm va sifat jihatidan cheklangan bo‘lib, to‘liq ilmiy talabga javob beruvchi *oltin standart* universal bog‘liqlik daraxti korpusi va u asosida qurilgan barqaror neyron sintaktik tahlil modellari yetarli emas.

Ushbu maqolada o‘zbek tili uchun Universal Dependencies standartlariga mos yangi universal bog‘liqlik daraxti korpusi va u asosida qurilgan chuqur bi-affin neyron diqqat mexanizmi yordamida ishlovchi tobelik tahlil modeli taqdim etiladi. Yaratilgan korpus o‘zbek adabiy va ilmiy-ommabop matnlaridan tanlab olingan 686 ta gapni (taxminan 7 800 ta token) o‘z ichiga oladi; annotatsiya jarayoni INCEpCTION[2] platformasida bir nechta annotator tomonidan amalga oshirilgan bo‘lib, lemmatizatsiya va so‘z turkumlari darajasida yuqori

[2]https://inception-project.github.io/

annotatorlararo moslikka erishilgan. Sintaktik tahlilda esa chuqur bi-affin neyron diqqat mexanizmi asosida bosh-tobe so‘z juftliklari uchun baholash funksiyasi qurilib, tobelik grafigining optimalligi maxsus yo‘qotish funksiyasini minimallashtirish orqali ta’minlanadi.

Tadqiqotning asosiy hissasi quyidagilardan iborat:

1. o‘zbek tili uchun UD standartlariga mos, statistik jihatdan barqaror yangi *universal bog‘liqlik daraxti korpusini* yaratish va uning annotatsiya jarayonini formal tavsiflash;
2. o‘zbek tili sintaksisini modellashtirishga moslashtirilgan chuqur bi-affin neyron diqqat mexanizmi asosidagi tobelik tahlil modelini qurish;
3. Stanza kutubxonasi neyron arxitekturasiga integratsiyalashgan holda tokenizatsiya, POS-tagging, morfologik tahlil va tobelik tahlili bo‘yicha eksperimental natijalarni taqdim etish hamda Unlabeled Attachment Score (UAS, belgilarsiz birikish aniqligi) va Labeled Attachment Score (LAS) ko‘rsatkichlari asosida modelning kuchli va zaif tomonlarini matematik jihatdan tahlil qilish.

Natijada, yaratilgan universal bog‘liqlik daraxti korpusi va chuqur bi-affin neyron modeli o‘zbek tili sintaksisini formal modellash, shuningdek, mashina tarjimasi, ma’lumotlarni qidirish va boshqa NLP vazifalarida keyingi nazariy hamda amaliy tadqiqotlar uchun poydevor vazifasini bajaradi.

ADABIYOTLAR TAHLILI

Sintaktik tahlil va universal bog‘liqlik daraxti korpuslarini qurish tabiiy tilni qayta ishlash (NLP) hamda formal til modellari nazariyasining muhim yo‘nalishlaridan biridir. Tobelik grammatikasi yondashuvi gap strukturasini yo‘naltirilgan daraxt ko‘rinishidagi graf bilan ifodalab, sintaktik munosabatlarni tugunlar orasidagi yoylar sifatida modellashtiradi. Ushbu yondashuvning nazariy asosi, shuningdek, ma’lumotlarga asoslangan (data-driven) tobelik tahlil algoritmlari Kubler, McDonald va Nivre tomonidan batafsil yoritilgan bo‘lib, ular turli grafik modellar, ehtimollik yondashuvlari va optimallashtirish usullarini yagona tizimda umumlashtiradi [1].

Ko‘p tilli sintaktik tahlilni yagona formalizm asosida rivojlantirish maqsadida Universal bog‘liqlik daraxti korpusi (Universal Dependencies, UD) loyihasi taklif etilgan bo‘lib, unda tillararo mos keluvchi so‘z turkumlari, morfologik xususiyatlar va tobelik munosabatlari tizimi ishlab chiqilgan [2]. Ushbu loyiha turli tillar uchun universal bog‘liqlik daraxti korpuslarini yaratish va ularni yagona formatda taqdim etish orqali sintaktik tahlil algoritmlarini solishtirish, baholash va ko‘p tilli modellarni o‘qitish imkonini beradi.

Chuqur o‘rganish davrida tobelik tahlili uchun grafga asoslangan neyron modellar keng qo‘llanila boshladi. Xususan, chuqur bi-affin neyron diqqat mexanizmi asosida qurilgan model bosh-tobe so‘z juftliklari uchun bi-affin baholash funksiyasini qo‘llab, tobelik grafigini global optimallashtirish imkonini beradi [3]. Ushbu yondashuvda so‘zlarning kontekstli tasvirlari ko‘p qatlamli BiLSTM enkoder yordamida olinadi, so‘ngra bosh va tobe proyeksiyalar uchun alohida neyron tarmoqlar va bi-affin operatorlar orqali yoy (arc) va munosabat (label) ehtimollari hisoblanadi. Bu model bugungi kunda ko‘plab tillar bo‘yicha universal bog‘liqlik daraxti korpuslarida yetakchi natijalar ko‘rsatgan.

Ko‘p tilli neyron sintaktik tahlilni amaliy dasturiy ta’minot ko‘rinishida taqdim etuvchi Stanza kutubxonasi turli tillar, jumladan, o‘zbek tili uchun ham tayyor neyron quvurlar (tokenizatsiya, POS-tagging, morfologik tahlil va tobelik tahlili)ni taqdim etadi [4]. Stanza arxitekturasi aynan chuqur bi-affin neyron diqqat mexanizmi asosidagi tobelik tahlil modellariga tayanadi va UD formatidagi universal bog‘liqlik daraxti korpuslarida o‘qitishni qo‘llab-quvvatlaydi.

O‘zbek tili bo‘yicha universal bog‘liqlik daraxti korpusini yaratish yo‘nalishida so‘nggi yillarda dastlabki natijalar olingan bo‘lib, ular UD standartlariga mos annotatsiya qilingan korpusni, milliy tilga xos morfologik va sintaktik hodisalarni formal tavsiflashni, shuningdek, ushbu korpus asosida dastlabki sintaktik tahlil modellarini qurishni o‘z ichiga oladi [5]. Ushbu ishlar o‘zbek tili uchun universal bog‘liqlik daraxti korpusining mavjudligini ta’minlab, chuqur neyron modellarni o‘qitish, UAS/LAS metrikalari asosida baholash va boshqa tillar bilan solishtirish imkonini beradi. Mazkur maqola ana shu yo‘nalishni davom ettirib, o‘zbek tili uchun yangilangan universal bog‘liqlik daraxti korpusi va chuqur bi-affin neyron diqqat mexanizmi asosidagi tobelik tahlil modelining matematik va eksperimental tahlilini taqdim etadi.

KORPUS

Ushbu tadqiqotda foydalanilgan universal bog‘liqlik daraxti korpusi o‘zbek tilining zamonaviy adabiy me’yoriga asoslangan bo‘lib, Universal bog‘liqlik daraxti korpusi (Universal Dependencies, UD) standartlariga mos ravishda tuzilgan [2]. Korpusning asosiy maqsadi – o‘zbek tili sintaksisini formal grafik model ko‘rinishida ifodalash, tobelik grammatikasi uchun “oltin standart” ma’lumotlar bazasini shakllantirish va chuqur neyron modellarni o‘qitish hamda baholash uchun yetarlicha barqaror poydevor yaratishdir [1].

Korpus tarkibiga o‘zbek adabiyoti va ilmiy-ommabop matnlaridan tanlab olingan 686 ta gap (taxminan 7 800 ta token) kiritilgan. Matnlar turli janr va uslublarni qamrab olishi orqali sintaktik konstruksiyalarning xilma-xilligini ta’minlashga harakat qilindi: bular qatoriga badiiy proza, ilmiy-ommabop maqolalar, ta’limiy matnlar va rasmiy-yozma uslub elementlari kiradi. Korpusni shakllantirishda korpus lingvistikasi tamoyillariga, xususan, namunalar tanlashda reprezentativlik va muvozanatlilik (balanslanganlik) talablariga tayangan holda ishlashga intilindi [6].

Har bir gap UD formatidagi to‘liq satr sifatida ifodalanib, quyidagi annotatsiya qatlamlarini o‘z ichiga oladi: tokenlar (so‘z shakllari), lemmalar, umumiy so‘z turkumlari (UPOS), tilga xos so‘z turkumlari (XPOS), morfologik xususiyatlar (UFeats), shuningdek, har bir token uchun bosh so‘z (HEAD) indeksi va tobelik munosabati turi (DEPREL). Shunday qilib, korpus har bir gapni yo‘naltirilgan daraxt ko‘rinishidagi sintaktik graf sifatida talqin qilish imkonini beradi. O‘zbek tilining agglyutinativ xususiyatlarini inobatga olib, affikslar so‘z tarkibida qolgan holda, lemmatizatsiya va morfologik xususiyatlar orqali grammatik ma’no alohida qatlam sifatida ifodalangan; bu yondashuv bir tomondan UD standartlariga moslikni, ikkinchi tomondan esa morfologik boylikni yo‘qotmaslikni ta’minlaydi [2].

Korpus INCEpCTION platformasi yordamida annotatsiya qilingan bo‘lib, unda 4 nafar tilshunos va 2 nafar NLP muhandisi ishtirok etgan. Annotatsiya jarayoni ikki bosqichli sifat nazorati asosida tashkil etildi: dastlabki belgilashdan so‘ng o‘zaro tekshirish (cross-verification) va yakuniy hakamlik (adjudication) bosqichlari amalga oshirildi. Annotatorlararo moslik lemmatizatsiya va UPOS darajasida 95% dan yuqori, morfologik xususiyatlar darajasida esa qariyb 90% ni tashkil etgani korpusning statistik barqarorligini va keyinchalik neyron modellarni o‘qitishda ishonchli etalon sifatida foydalanish mumkinligini ko‘rsatadi [5].

Universal bog‘liqlik daraxti korpusi sintaktik tahlil modellari uchun nafaqat o‘quv ma’lumotlari, balki baholash (evaluation) standarti sifatida ham xizmat qiladi. Ushbu maqolada aynan shu korpus asosida chuqur bi-affin neyron diqqat mexanizmi yordamida qurilgan tobelik tahlil modeli o‘qitilib, Unlabeled Attachment Score (UAS, belgilarsiz birikish aniqligi) va Labeled Attachment Score (LAS) kabi standart metrikalar orqali sinovdan o‘tkaziladi. Natijada olingan ko‘rsatkichlar o‘zbek tili uchun universal bog‘liqlik daraxti korpusi va chuqur neyron tahlil modellarining bir-birini to‘ldiruvchi, yagona matematik platforma sifatida qaralishi mumkinligini ko‘rsatadi.

Quyidagi 1-rasmda annotatsiya jarayonida foydalanilgan INCEpCTION platformasining ko’rinishi aks ettirilgan.

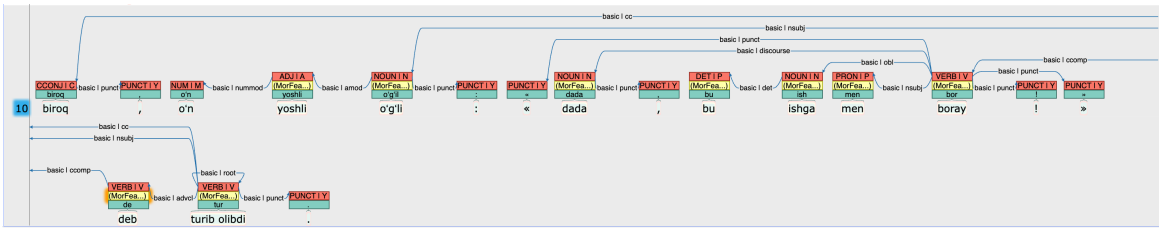

Рис. 8: INCEpCTION platformasida o’zbek tili matnlarini annotatsiya qilish jarayoni.

MASALANING QO‘YILISHI

Ushbu tadqiqotda sintaktik tahlil (tobelik tahlili) masalasi chuqur bi-affin neyron diqqat mexanizmi asosidagi grafik model sifatida qaraladi. Tobelik grammatikasi doirasida gap $S = (w_1, \ldots, w_n)$ ketma-ketlikdagi so‘zlardan tashkil topgan bo‘lib, unga $w_0 = \text{ROOT}$ maxsus ildiz tuguni qo‘shiladi. Sintaktik daraxt esa har bir $i \in \{1, \ldots, n\}$ so‘z uchun bosh so‘z indeksi $h(i) \in \{0, \ldots, n\}$ va tobelik munosabati turini $r(i)$ beruvchi yoylar to‘plami

$$T = \{(h(i), i, r(i)) \mid i = 1, \ldots, n\}$$

ko‘rinishida tasvirlanadi. Daraxt “yaxlit” bo‘lishi uchun bu yoylar to‘plami $n+1$ tugunga ega bo‘lgan, siklsiz va bitta ildizli yo‘naltirilgan daraxt hosil qilishi kerak [1].

Chuqur bi-affin neyron diqqat mexanizmi g‘oyasi shundan iboratki, model gapdagi har bir bosh-tobe so‘z juftligi (j,i) uchun raqamli baho $s_{j,i}^{(\mathrm{arc})}$ hisoblaydi va bu baholar asosida eng ehtimolli tobelik daraxtini tanlaydi [3]. Boshqa tomondan qaralganda, $s_{j,i}^{(\mathrm{arc})}$ lar shunday tanlanadiki, ular har bir so‘z i uchun barcha mumkin bo‘lgan boshlar $\{0,\dots,n\}$ ustida taqsimot hosil qiladi:

$$P(h(i)=j \mid S)=\frac{\exp\left(s_{j,i}^{(\mathrm{arc})}\right)}{\sum\limits_{k=0}^{n}\exp\left(s_{k,i}^{(\mathrm{arc})}\right)}.$$

Shunga o‘xshash tarzda, to‘g‘ri bosh $h(i)$ tanlanganidan so‘ng, model munosabat turi $r(i)$ uchun ham alohida ehtimollik taqsimoti $P(r(i) \mid h(i), S)$ ni o‘rganadi. Amaliyotda bu ikki qism (yoy mavjudligi va munosabat yorlig‘i) uchun alohida neyron chiqish qatlamlari va yo‘qotish funksiyalari ishlatiladi.

O‘qitishning maqsadi – universal bog‘liqlik daraxti korpusi asosida berilgan “oltin standart” daraxt T^* ni maksimal ehtimollik bilan tiklay oladigan parametrlar θ ni topishdir. Boshqacha aytganda, model

$$P(T^* \mid S;\theta)$$

qiymatini maksimal qilishga intiladi. Buni amalda yoylar va munosabatlar mustaqil deb faraz qilinadigan faktorizatsiyalangan ko‘rinishda yozish mumkin:

$$\log P(T^* \mid S;\theta)=\sum_{i=1}^{n}\log P\big(h^*(i) \mid S;\theta\big)+\sum_{i=1}^{n}\log P\big(r^*(i) \mid h^*(i),S;\theta\big),$$

bu yerda $h^*(i)$ va $r^*(i)$ – korpusda berilgan bosh va munosabatning “oltin” qiymatlari. Neyron modelni o‘qitishda ushbu ifodaning manfiy qiymati Cross-Entropy ko‘rinishidagi yo‘qotish funksiyasi sifatida minimallashtiriladi:

$$\mathcal{L}(\theta)=\mathcal{L}_{\mathrm{arc}}(\theta)+\mathcal{L}_{\mathrm{rel}}(\theta)=-\sum_{i=1}^{n}\log P\big(h^*(i) \mid S;\theta\big)-\sum_{i=1}^{n}\log P\big(r^*(i) \mid h^*(i),S;\theta\big).$$

Bu maqsad funksiyasini intuitiv misol orqali tushuntirish mumkin. Masalan, *“O‘quvchi kitobni o‘qidi”* gapida to‘g‘ri daraxt quyidagi yoylardan iborat bo‘ladi: ROOT → o‘qidi (predikat), o‘qidi → O‘quvchi (ega), o‘qidi → kitobni (tiniqli to‘ldiruvchi). Agar model dastlab *“O‘quvchi”* uchun noto‘g‘ri bosh tugun sifatida *“kitobni”* so‘zini yuqori ehtimollik bilan tanlasa, $\log P(h^*(i) \mid S)$ qiymati kichik bo‘ladi va yo‘qotish $\mathcal{L}_{\mathrm{arc}}$ ortadi. O‘qitish jarayonida gradientlar yordamida parametrlar θ shunday yangilanadiki, to‘g‘ri yoylarning ehtimolligi oshadi, noto‘g‘ri yoylar esa penalizatsiya qilinadi. Xuddi shunday, *nsubj* (ega) va *obj* (to‘ldiruvchi) kabi munosabat turlarini farqlashda xato belgilash ham $\mathcal{L}_{\mathrm{rel}}$ ni oshiradi va model keyingi iteratsiyalarda bu xatoni kamaytirishga intiladi.

1. Enkoder (BiLSTM)

Modelning birinchi komponenti – ketma-ketlikni kodlovchi neyron tarmoq (enkoder) bo‘lib, u gapdagi so‘zlarni kontekstga bog‘liq vektor tasvirlarga o‘tkazadi. Har bir so‘z w_i uchun so‘z shakli, lemma va so‘z turkumi t_i ga asoslangan boshlang‘ich embeddinglar $\mathbf{e}_i$ hosil qilinadi; amaliyotda ular so‘z embeddingi va UPOS/XPOS embeddinglarining birlashtirilgan ko‘rinishi sifatida yozish mumkin:

$$\mathbf{x}_i=[\mathbf{e}_i^{(\mathrm{word})};\,\mathbf{e}_i^{(\mathrm{pos})}].$$

Hosil bo‘lgan $\mathbf{x}_1,\dots,\mathbf{x}_n$ ketma-ketligi kontekstual ma’lumotlarni o‘zlashtirish uchun ko‘p qavatli Ikki Tomonlama Uzun Qisqa Muddatli Xotira (BiLSTM) tarmog‘iga uzatiladi:

$$\mathbf{r}_i=\mathrm{BiLSTM}(\mathbf{x}_1,\dots,\mathbf{x}_n)_i,\qquad i=1,\dots,n. \tag{1}$$

Bu yerda $\mathbf{r}_i$ – i-so‘zning chap va o‘ng kontekstini birgalikda hisobga olgan, yuqori o‘lchamli yashirin holat vektori bo‘lib, keyingi bi-affin baholash qatlamlari uchun kirish nuqtasini tashkil etadi. Shu tarzda butun gap

$$S=(w_1,\dots,w_n)\quad\longmapsto\quad R=(\mathbf{r}_1,\dots,\mathbf{r}_n)$$

ko‘rinishidagi vektorli fazoga ko‘chiriladi va sintaktik bog‘lanishlar aynan shu fazoda o‘rganiladi.

2. Bi-affin baholash (Scoring)

Chuqur bi-affin neyron diqqat mexanizmining asosiy g‘oyasi – har bir bosh-tobe so‘z juftligi uchun alohida vektor proyeksiyalarni hisoblab, ular ustida bi-affin (ikki chiziqli) operator yordamida baho olishdir. BiLSTM dan chiqqan har bir $\mathbf{r}_i$ vektori to‘rt xil ko‘p qatlamli perseptron (MLP) orqali ikki guruhga bo‘linadi:

$$\mathbf{h}_i^{(\text{arc-dep})} = \text{MLP}_{\text{arc-dep}}(\mathbf{r}_i), \qquad \mathbf{h}_j^{(\text{arc-head})} = \text{MLP}_{\text{arc-head}}(\mathbf{r}_j),$$
$$\mathbf{h}_i^{(\text{rel-dep})} = \text{MLP}_{\text{rel-dep}}(\mathbf{r}_i), \qquad \mathbf{h}_j^{(\text{rel-head})} = \text{MLP}_{\text{rel-head}}(\mathbf{r}_j).$$

Bu yerda **Arc-Head** va **Arc-Dep** proyeksiyalar bosh-tobe juftligi o‘rtasida umuman bog‘lanish mavjud yoki yo‘qligini (yoy mavjudligi) baholash uchun, **Rel-Head** va **Rel-Dep** esa bog‘lanish turini (masalan, *nsubj*, *obj*, *obl* kabi yorliqlarni) aniqlash uchun xizmat qiladi.

Ikki so‘z – i (tobe) va j (bosh) – o‘rtasidagi bog‘lanish uchun skalyar baho $s_{j,i}^{(\text{arc})}$ quyidagi bi-affin formula orqali hisoblanadi:

$$s_{j,i}^{(\text{arc})} = \mathbf{h}_i^{(\text{arc-dep})\top} \mathbf{U}^{(\text{arc})}\, \mathbf{h}_j^{(\text{arc-head})} + \mathbf{w}^{(\text{arc})\top} \begin{bmatrix} \mathbf{h}_i^{(\text{arc-dep})} \\ \mathbf{h}_j^{(\text{arc-head})} \end{bmatrix} + b^{(\text{arc})}, \tag{2}$$

bu yerda $\mathbf{U}^{(\text{arc})}$ – bosh va tobe vektorlarning o‘zaro ta’sirini modellashtiruvchi o‘rganiluvchi matritsa, $\mathbf{w}^{(\text{arc})}$ va $b^{(\text{arc})}$ esa chiziqli va skalyar siljitish parametrlaridir. Amaliyotda $s_{j,i}^{(\text{arc})}$ qiymatlar har bir tobe so‘z i uchun barcha mumkin bo‘lgan boshlar $\{0, \ldots, n\}$ ustida softmax funktsiyasi orqali ehtimollik taqsimotiga aylantiriladi:

$$P(h(i) = j \mid S) = \frac{\exp\left(s_{j,i}^{(\text{arc})}\right)}{\sum_{k=0}^{n} \exp\left(s_{k,i}^{(\text{arc})}\right)}.$$

Xuddi shunday, munosabat turi bo‘yicha baholash uchun $\mathbf{h}_i^{(\text{rel-dep})}$ va $\mathbf{h}_j^{(\text{rel-head})}$ vektorlari ko‘p kanalli bi-affin transformatsiyaga yuborilib, har bir mumkin bo‘lgan yorliq r uchun $s_{j,i,r}^{(\text{rel})}$ bahosi olinadi va ular ustida softmax qo‘llanadi. Natijada model har bir juftlik (j, i) uchun “bosh kim?” va “munosabat turi nima?” savollariga ehtimollik nuqtai nazaridan javob beradi.

3. Model Arxitekturasi Vizualizatsiyasi

Quyidagi 9 o’zbek tilidagi *"O’quvchi kitobni o’qidi"* gapi misolida arxitekturaning ishlash prinsipi tasvirlangan.

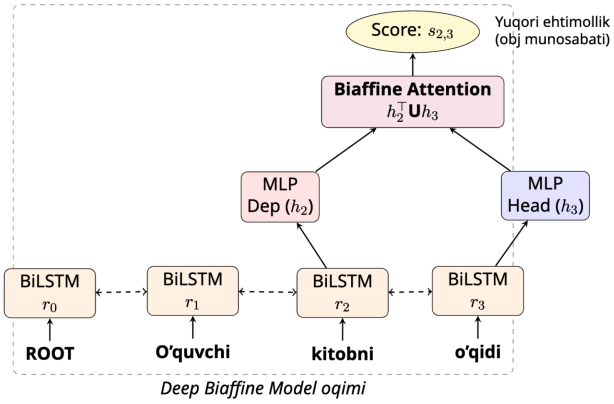

Рис. 9: O’zbek tili uchun Deep Biaffine Attention arxitekturasi sxemasi. *"Kitobni"* (tobe) va *"o’qidi"* (bosh) so’zlari o’rtasidagi bog’liqlikni hisoblash jarayoni.

4. Xatolik funksiyasi (Loss)

Modelni o‘qitish jarayonida maqsad – universal bog‘liqlik daraxti korpusida berilgan “oltin” boshlar $h^*(i)$ va munosabat turlari $r^*(i)$ ni yuqori ehtimollik bilan qayta tiklashdir. Buning uchun standart **Cross-Entropy** tipidagi yo‘qotish funksiyalaridan foydalaniladi. Umumiy yo‘qotish

$$\mathcal{L} = \mathcal{L}_{\text{arc}} + \mathcal{L}_{\text{rel}} \tag{3}$$

ko‘rinishida faktorizatsiya qilinadi, bu yerda

$$\mathcal{L}_{\text{arc}} = -\sum_{i=1}^{n} \log P\big(h^*(i) \mid S\big), \qquad \mathcal{L}_{\text{rel}} = -\sum_{i=1}^{n} \log P\big(r^*(i) \mid h^*(i), S\big).$$

Intuitiv tarzda, agar model *“O‘quvchi kitobni o‘qidi”* gapida *“O‘quvchi”* uchun to‘g‘ri bosh sifatida *“o‘qidi”* so‘zini emas, balki *“kitobni”* so‘zini yuqori ehtimollik bilan tanlasa, $\log P(h^*(i) \mid S)$ qiymati kamayadi va $\mathcal{L}_{\text{arc}}$ ortadi; shuningdek, *nsubj* o‘rniga xato ravishda *obj* yorlig‘ini berish $\mathcal{L}_{\text{rel}}$ ni oshiradi. Gradientlar yordamida parametrlar θ ni yangilash natijasida model to‘g‘ri bosh-tobe juftliklari va ularning yorliqlariga yuqori ehtimollik berishni “o‘rganadi”, noto‘g‘ri variantlar esa penalizatsiya qilinadi. Shu tariqa yo‘qotish funksiyasini minimallashtirish o‘zbek tilining sintaktik qonuniyatlarini chuqur bi-affin neyron diqqat mexanizmi orqali o‘zlashtirishga olib keladi.

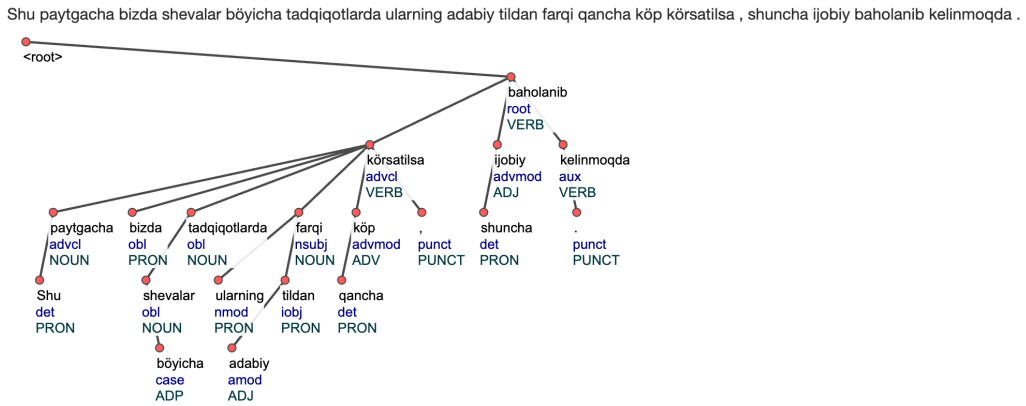

Рис. 10: Model tomonidan yaratilgan o’zbek tili gapining sintaktik tobelik daraxti.

TAJRIBA VA NATIJALAR

Ushbu bo‘limda taklif etilgan chuqur bi-affin neyron diqqat mexanizmi asosidagi tobelik tahlil modelining amaliy samaradorligi tahlil qilinadi. Tajribalar UzUDT universal bog‘liqlik daraxti korpusi asosida, Stanza kutubxonasi neyron quvuri (tokenizatsiya, so‘z turkumlarini aniqlash, morfologik tahlil va tobelik tahlili) bilan integratsiyalashgan holda olib borildi [4, 5]. Baholashda universal bog‘liqlik daraxti korpusi hamjamiyatida qabul qilingan standart metrikalar – so‘z turkumlari aniqligi (UPOS/XPOS), morfologik xususiyatlar aniqligi (UFeats), shuningdek, Unlabeled Attachment Score (UAS, belgilarsiz birikish aniqligi), Labeled Attachment Score (LAS, belgilangan birikish aniqligi) va CLAS ko‘rsatkichlari ishlatildi [2, 3].

1. So‘z turkumlari va morfologik tahlil

Birinchi bosqichda so‘z turkumlari va morfologik xususiyatlarni aniqlash vazifasi (tagging) ko‘rib chiqildi. UzUDT universal bog‘liqlik daraxti korpusi o‘quv to‘plamida `xlm-roberta-base` modeli mayin sozlandi (fine-tuning) va u o‘zbek tili uchun maxsus o‘qitilgan FastText vektorlari (`cc.uz.300.vec`) bilan birgalikda ishlatildi. Stanza quvuri doirasida so‘z shakli, lemma va POS teglaridan tashkil topgan kirish vektorlari neyron tagging moduli orqali qayta ishlanib, har bir token uchun UPOS, XPOS va UFeats qatlamlari qayta baholandi.

Jadval 12: UzUDT universal bog‘liqlik daraxti korpusida so‘z turkumlari va morfologik tahlil natijalari.

To’plam	UPOS aniqligi	XPOS aniqligi	UFeats aniqligi
Dev	82.92%	83.85%	67.39%
Test	**86.10%**	**83.96%**	**70.06%**

Rivojlantirish (Dev) va Test to‘plamlari bo‘yicha olingan natijalar 1-jadvalda keltirilgan.

Jadvaldan ko‘rinib turibdiki, model Test to‘plamida UPOS bo‘yicha 86.10%, XPOS bo‘yicha 83.96% va morfologik xususiyatlar (UFeats) bo‘yicha 70.06% aniqlikka erishgan. Dev va Test ko‘rsatkichlari o‘rtasidagi tafovutlarning katta emasligi modelning umumlashma qobiliyati (generalization) yetarli ekanini ko‘rsatadi. Shu bilan birga, UFeats aniqligi UPOS/XPOS ko‘rsatkichlariga nisbatan ancha pastroq bo‘lib qolmoqda; bu o‘zbek tilidagi kelishik, shaxs-son, zamon, egalik kabi murakkab morfologik xususiyatlarning neyron model uchun qiyinroq ekanligidan dalolat beradi. Bunday natija agglyutinativ tillar uchun tabiiy bo‘lib, morfologik analizator bilan chuqur neyron modelni birlashtirish (‘morphology-aware’ arxitekturalar) zaruriyligini ko‘rsatadi [1, 5].

2. Sintaktik tobelik tahlili

Sintaktik tobelik tahlili bosqichida chuqur bi-affin neyron diqqat mexanizmi asosidagi modelning sof sintaktik qobiliyatini baholash maqsadida *oltin morfologiya* (*gold morphology*) sharoiti tanlandi. Ya’ni, kirish sifatida modelga universal bog‘liqlik daraxti korpusidan olingan to‘liq to‘g‘ri lemmalar, UPOS va XPOS teglar uzatildi; shu orqali morfologik xatolarning sintaksisga tarqalishi (error propagation) istisno qilinib, faqat tobelik tahlil komponentining xatti-harakati o‘lchandi [3].

Test to‘plami bo‘yicha olingan UAS, LAS va CLAS ko‘rsatkichlari 2-jadvalda keltirilgan.

Jadval 13: Oltin morfologiya sharoitida sintaktik tobelik tahlilining baholash natijalari.

Metrika	Natija	Ta’rif
UAS	**69.21%**	Unlabeled Attachment Score (belgilarsiz birikish aniqligi)
LAS	**53.21%**	Labeled Attachment Score (bosh va munosabat turi to’g’ri)
CLAS	46.32%	Content LAS (mazmunli so’zlar uchun belgilangan birikish aniqligi)

UAS va LAS ko‘rsatkichlari orasidagi taxminan 16 foizlik farq shuni ko‘rsatadiki, model gapning umumiy ierarxik tuzilmasini, ya’ni qaysi so‘z qaysi so‘zga bog‘langanini nisbatan yaxshi o‘rganadi (UAS $\approx$ 69%), biroq bog‘lanishning aniq sintaktik turini (masalan, *nsubj* – ega, *obj* – to‘ldiruvchi, *obl* – hol kabi yorliqlarni) ajratishda qiyinchiliklar mavjud (LAS $\approx$ 53%). CLAS ning LAS dan pastroq bo‘lishi esa aynan mazmuniy so‘zlar (fe’l, ot, sifat) uchun to‘g‘ri belgilash vazifasining yanada murakkabligini aks ettiradi.

Bu natijalar o‘zbek tilining erkin so‘z tartibi, boy affiksatsiya tizimi va mavjud universal bog‘liqlik daraxti korpusining hajmi bilan izohlanishi mumkin. Kichik o‘quv to‘plami sharoitida chuqur bi-affin neyron diqqat mexanizmi sintaktik daraxtning asosiy skeletini (bosh-tobe bog‘lanishlar) yetarli darajada tiklay olsa-da, ingichka farqlanuvchi sintaktik rollarni ajratishda yanada ko‘proq ma’lumot va qo‘shimcha lingvistik priorlar talab etiladi [1, 3]. Shu nuqtai nazardan, mazkur natijalar o‘zbek tili uchun chuqur neyron tobelik tahlili bo‘yicha birinchi mustahkam bazaviy chiziq (baseline) sifatida qaralishi mumkin; kelgusida korpus hajmini oshirish, morfologik xususiyatlarni yanada chuqur integratsiya qilish va ko‘p tilli transfer yondashuvlarini qo‘llash orqali UAS/LAS ko‘rsatkichlarini sezilarli yaxshilash imkoniyati mavjud.

Umuman olganda, so‘z turkumlari va morfologik tahlil natijalari chuqur modelning o‘zbek tili grammatik tizimini o‘zlashtira boshlaganini, sintaktik tobelik tahlili natijalari esa chuqur bi-affin neyron diqqat mexanizmi asosida qurilgan modelning erkin so‘z tartibli, agglyutinativ til sharoitida ham barqaror ishlashini ko‘rsatadi. Ushbu tajribalar universal bog‘liqlik daraxti korpusi va chuqur neyron tobelik tahlili o‘rtasida yagona matematik platforma shakllanayotganini tasdiqlaydi hamda keyingi tadqiqotlar uchun aniq raqamli benchmark vazifasini bajaradi.

XULOSA

Ushbu maqolada o‘zbek tili uchun universal bog‘liqlik daraxti korpusi asosida sintaktik tahlil masalasi chuqur bi-affin neyron diqqat mexanizmi nuqtai nazaridan o‘rganildi. Avvalo, zamonaviy adabiy va ilmiy-ommabop matnlardan tanlab olingan 686 ta gap (taxminan 7 800 ta token) uchun universal bog‘liqlik daraxti korpusi standartlariga mos annotatsiya amalga oshirildi; lemmatizatsiya, so‘z turkumlari va morfologik xususiyatlar bo‘yicha yuqori annotatorlararo moslikka erishilgani korpusning “oltin standart” sifatida qo‘llanishi mumkinligini ko‘rsatdi. Bu korpus o‘zbek tili sintaksisini formal graf modeli sifatida tavsiflaydigan birinchi barqaror ma’lumotlar bazasi bo‘lib, keyingi nazariy va amaliy tadqiqotlar uchun asos yaratadi.

Taklif etilgan chuqur bi-affin neyron diqqat mexanizmi asosidagi tobelik tahlil modeli BiLSTM enkoder, bi-affin baholash qatlamlari va Cross-Entropy yo‘qotish funksiyasi orqali bosh-tobe so‘z juftliklarini ehtimollik nuqtai nazaridan baholash va optimal sintaktik daraxtni tiklashga qaratildi. Oltin morfologiya sharoitida olingan natijalar – so‘z turkumlari uchun UPOS aniqligi 86.10%, morfologik xususiyatlar uchun UFeats aniqligi 70.06%, sintaktik tahlilda esa UAS=69.21% va LAS=53.21% – o‘zbek tili uchun chuqur neyron tobelik tahlilining birinchi aniq raqamli bazaviy ko‘rsatkichlarini belgilaydi. UAS va LAS o‘rtasidagi farq erkin so‘z tartibli, agglyutinativ til sharoitida sintaktik rollarni nozik farqlashning murakkabligini aks ettirsa-da, model gapning umumiy ierarxik strukturasi va bosh-tobe bog‘lanishlarini yetarli darajada o‘rganayotganini tasdiqlaydi.

Kelgusidagi tadqiqotlar uchun universal bog‘liqlik daraxti korpusining hajmini kengaytirish, morfologik analizator bilan chuqur neyron modelni yanada chambarchas integratsiya qilish, ko‘p tilli va transfer-o‘rganish yondashuvlarini qo‘llash hamda qo‘shimcha lingvistik priorlar kiritish orqali UAS/LAS ko‘rsatkichlarini sezilarli yaxshilash imkoniyati mavjud. Shunday qilib, yaratilgan universal bog‘liqlik daraxti korpusi va chuqur bi-affin neyron tahlil modeli o‘zbek tili sintaksisining matematik va hisoblash nuqtai nazaridan tadqiqi uchun poydevor va kelajakdagi NLP tizimlari uchun ishonchli komponent sifatida qaralishi mumkin.

ADABIYOTLAR RO‘YXATI:

1. John Carroll. 2010. Book Review: Dependency Parsing by Sandra Kubler, Ryan McDonald, and Joakim Nivre. Computational Linguistics, 36(1).
2. Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 4034-4043, Marseille, France. European Language Resources Association.
3. Dozat, T., & Manning, C. D. (2017). Deep Biaffine Attention for Neural Dependency Parsing. *ICLR 2017*.
4. Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 101-108, Online. Association for Computational Linguistics.
5. Matlatipov, S. G., et al. (2024). *UzUDT: Universal Dependencies Treebank for Uzbek*. National University of Uzbekistan.
6. McEnery T, Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge University Press; 2011.

Resume

This article introduces a new Universal Dependencies (UD) treebank for the Uzbek language and a dependency parser based on a deep biaffine neural attention mechanism. The corpus contains 686 sentences (7,800 tokens) from literary and popular-science texts, manually annotated with lemmas, POS tags, morphological features and dependency relations, achieving inter-annotator agreement above 95% for lemmatization and UPOS. On top of this gold-standard resource, we train and evaluate a BiLSTM-based deep biaffine dependency parser implemented in the Stanza pipeline, obtaining 86.10% UPOS accuracy, 70.06% UFeats accuracy and, under gold morphology, 69.21% UAS and 53.21% LAS on the test set. The treebank and model define the first strong neural baseline for dependency parsing in Uzbek and provide a mathematically grounded platform for further NLP research on the language.

Key words: Universal Dependencies, Uzbek language, dependency parsing, deep biaffine neural attention, treebank, NLP.

Резюме

В статье представлен новый трибанк Universal Dependencies (UD) для узбекского языка и построенный на его основе парсер зависимостей с глубокой биаффинной нейронной моделью внимания. Корпус включает 686 предложений (7 800 токенов) из художественных и научно-популярных текстов, вручную аннотированных леммами, частями речи, морфологическими признаками и отношениями зависимостей; межаннотаторское согласие для лемматизации и UPOS превышает 95%. На базе этого «золотого стандарта» обучается и оценивается BiLSTM-биаффинный парсер, реализованный в конвейере Stanza, который при золотой морфологии достигает 86,10% точности UPOS, 70,06% точности UFeats и 69,21% UAS / 53,21% LAS на тестовой выборке. Полученный трибанк и модель задают первую сильную нейронную базу для анализа зависимостей в узбекском языке и создают математически обоснованную платформу для дальнейших исследований в области НЛП.

Ключевые слова: Universal Dependencies, узбекский язык, анализ зависимостей, глубое биаффинное нейронное внимание, трибанк, НЛП.