O‘ZBEK TILIDA GAPLARDAGI SEMANTIK ROLLARNI BELGILASH MASALASINI YECHISH

Axmedova Xolisxon Ilxomovna
O‘zbekiston xalqaro islomshunoslik akademiyasi
Orcid: 0000-0002-9828-1650
E-mail: xolisa90292mail.ru

Annotatsiya.Tabiiy tilni qayta ishlash sohasida hanuzgacha to‘liq yechimini topmagan muammolar talaygina bo‘lib, ularning muhimlaridan biri semantik rollarni belgilash (Semantic Role Labeling — SRL) masalasidir. O‘zbek kompyuter tilshunosligida predikatga bog‘liq semantik rollarni aniqlash vazifasi ochiq muammo bo‘lib qolmoqda. Ushbu tadqiqotning asosiy maqsadi mazkur muammoni zamonaviy mashinaviy o‘rganish modellaridan foydalangan holda hal etishga qaratilgan. Mazkur maqolada semantik rollarni belgilash bo‘yicha mavjud ilmiy ishlar tahlil qilinib, ayniqsa, zamonaviy yondashuvlar va ularning xalqaro tajribadagi o‘rni muhokama qilindi. Shuningdek, PropBank va FrameNet kabi lingvistik resurslarda mavjud semantik va tematik rollar tizimi o‘rganildi. Olib borilgan tahlillar asosida o‘zbek tilidagi gaplar uchun moslashtirilgan semantik rollar teglari majmuasi ishlab chiqildi. Yakuniy bosqichda esa ketma-ketlikdan ketma-ketlikka (sequence-to-sequence) yondashuvi asosida kichik hajmdagi ma’lumotlar to‘plamida semantik rollarni belgilash vazifasi amaliy jihatdan sinovdan o‘tkazildi.

Kalit so‘zlar: semantik rollarni belgilash, PropBank, semantik tahlil, argument, agent, ma’lumotlar to‘plami.

Abstract. In the field of Natural Language Processing, there are still many problems that have not yet been fully resolved, one of the most important of which is Semantic Role Labeling (SRL). In Uzbek computational linguistics, the task of identifying predicate-related semantic roles remains an open problem. The main objective of this research is to address this issue using modern machine learning models. In this article, existing scientific studies on semantic role labeling are analyzed, with particular emphasis on modern approaches and their role in international practice. In addition, the systems of semantic and thematic roles available in linguistic resources such as PropBank and FrameNet are examined. Based on the conducted analyses, a set of semantic role tags adapted for Uzbek sentences was developed. In the final stage, the task of semantic role labeling was experimentally tested on a small-scale dataset using a sequence-to-sequence approach.

Keywords: semantic role labeling, PropBank, semantic analysis, argument, agent, dataset.

ISSN 3030-3907

Аннотация. В области обработки естественного языка по-прежнему существует множество нерешённых проблем, одной из наиболее значимых среди которых является задача семантической разметки ролей (Semantic Role Labeling — SRL). В узбекской компьютерной лингвистике задача определения семантических ролей, связанных с предикатом, остаётся открытой проблемой. Основная цель данного исследования заключается в решении этой задачи с использованием современных моделей машинного обучения. В статье проанализированы существующие научные работы по семантической разметке ролей, при этом особое внимание уделено современным подходам и их месту в международной практике. Также были изучены системы семантических и тематических ролей, представленные в лингвистических ресурсах, таких как PropBank и FrameNet. На основе проведённого анализа был разработан набор тегов семантических ролей, адаптированный для предложений на узбекском языке. На заключительном этапе задача семантической разметки ролей была экспериментально апробирована на небольшом наборе данных с использованием подхода «последовательность–последовательность» (sequence-to-sequence).

Ключевые слова: семантическая разметка ролей, PropBank, семантический анализ, аргумент, агент, набор данных.

KIRISH.

So‘nggi yillarda tabiiy tilni tushunish masalalari jadal rivojlanib, mazkur yo‘nalish yangi va murakkab bosqichga kirdi. Avvallari asosan nutqni tanish, so‘z turkumlarini aniqlash hamda sintaktik tahlil bilan cheklangan tabiiy tilni qayta ishlash tizimlarida endilikda mazmun va semantikaga asoslangan vazifalar ustuvor ahamiyat kasb eta boshladi.

Bugungi kunda oddiy nutqqa asoslangan va tijoratda keng qo‘llanilayotgan tabiiy tilni tushunish tizimlari mavjud bo‘lib, ular samolyot reyslarining yetib kelish vaqtlari, yo‘nalishlar haqida ma’lumot berish, shuningdek, bank mijozlariga hisob holatini boshqarishda yordam ko‘rsatish imkonini yaratmoqda. Bundan tashqari, ilmiy maqolalarning qisqacha mazmunini avtomatik yaratish, faktlarga asoslangan savollarga javob berish hamda murakkab semantik va muloqot tuzilmalarini aniqlashga qodir axborot tizimlari ishlab chiqilgan. Shunga qaramay, mazkur sohada hal etilishi lozim bo‘lgan muammolar hanuz saqlanib qolmoqda. Xususan, tabiiy til matnlarining mazmunini to‘liq tushunish, gaplarning sintaktik tuzilishini aniqlash va gap tarkibidagi har bir birlikning semantik vazifasini belgilash hozirgi kundagi eng dolzarb masalalardan biridir. Oldingi davrlarga nisbatan bu vazifalarni avtomatlashtirish bugungi kunda ancha real imkoniyatga ega. Bunga WordNet kabi yirik semantik resurslarning mavjudligi, turli teglarga

ISSN 3030-3907

ega bo‘lgan katta hajmdagi korpuslar hamda mashinaviy o‘rganish sohasidagi so‘nggi yutuqlar mustahkam poydevor bo‘lib xizmat qilmoqda.

Biroq zamonaviy axborotni ajratib olish va matnni tushunish tizimlarining asosida hanuzgacha soha uchun maxsus ishlab chiqilgan shablonlar va slot-andozalar yotadi. Bunday mazmuniy tavsiflar semantik rollar deb ataladi. Ya’ni, ular hodisa doirasida predikat argumentlarining qanday vazifani bajarishini ifodalaydi. Ushbu turdagi semantik rollar PropBank va FrameNet kabi tabiiy til resurslarida oldindan belgilab qo‘yilgan.

Semantik rollar. Semantik rol tushunchasini yanada aniqroq anglash uchun ikki gapni ko‘rib chiqamiz:

1. Anvar oynani sindirdi.
2. Ra’no eshikni ochdi.

Ushbu misollarda *Anvar* va *Ra’no* mos ravishda oynani sindirish va eshikni ochish harakatlarini amalga oshiruvchi shaxslar bo‘lib, ularning umumiy jihati — har ikkisi ham faol ijrochi hisoblanishidir. Tilshunoslikda bunday umumiyliklar tematik rollar tushunchasi orqali ifodalanadi. Keltirilgan gaplarda fe’l kesimlarning ega bo‘lagi bo‘lgan *Anvar* va *Ra’no* agent sifatida talqin qilinadi.

Boshqacha aytganda, agent (AGENT) — bu ongli ravishda harakatni boshlab beruvchi va hodisaning yuzaga kelishiga ataylab sabab bo‘luvchi tematik roldir. Xuddi shuningdek, har ikkala gapdagi to‘ldiruvchi — ya’ni sindirilayotgan va ochilayotgan obyektlar (*oyna* va *eshik*) odatda jonsiz bo‘lib, bajarilgan harakat natijasida ma’lum bir o‘zgarishga uchraydi. Ingliz tilidagi PropBank tizimida bunday ishtirokchilarga beriladigan semantik rol theme (mavzu/obyekt) deb ataladi.

Demak, ushbu gaplarda:

- Anvar va Ra’no — *agent* roli,
- oyna va eshik — *theme* roli sifatida belgilanadi.

Ingliz tilida tez-tez uchraydigan tematik rollar tizimi keltirilgan [1]. Semantik rollarning hisoblash tizimlarida qo‘llanilishining asosiy sababi shundan iboratki, ular sayoz semantik tasvirlash shakli (shallow semantic representation) sifatida xizmat qiladi. Bunday tasvirlash usuli faqat so‘zlarning ketma-ketligi yoki hatto sof sintaktik daraxt asosida aniqlash mushkul bo‘lgan oddiy mantiqiy xulosalarni chiqarish imkonini beradi.

ADABIYOTLAR SHARHI

Katta hajmdagi ma’lumotlar (Big Data) muhitida eng asosiy muammolardan biri tuzilmasiz (unstructured) matnli axborotning nihoyatda katta miqyosi va turli xilligidir [12]. Bunday murakkabliklar tuzilmasiz ma’lumotlarning bir qator ichki xususiyatlari bilan izohlanadi, xususan:

ISSN 3030-3907

(1) ularning hajmi juda katta bo‘lishi;
(2) internet platformalari, ijtimoiy tarmoqlar va mobil ilovalar kabi turli manbalardan hosil bo‘lishi;
(3) matn, audio, video va vizual axborot kabi ko‘p turdagi shakllarni o‘z ichiga olishi;
(4) yuqori darajadagi shovqin, noaniqlik va semantik jihatdan foydasiz elementlarning mavjudligi tahlil jarayonini murakkablashtirishi [12], [13].

Tuzilmasiz matnli ma’lumotlardan mazmunli axborot ajratib olish murakkab jarayon bo‘lib, shu sababli bunday ma’lumotlarni tuzilmali ko‘rinishga keltirish muhim ahamiyat kasb etadi [1]. Axborotni ajratib olish (Information Extraction — IE) tabiiy tilni qayta ishlash (NLP) sohasining muhim yo‘nalishlaridan biri bo‘lib, u matndan obyektlar, munosabatlar va semantik rollarni aniqlash orqali tuzilmasiz ma’lumotlarni tuzilmali reprezentatsiyaga aylantiradi. Bu esa keyingi tahliliy jarayonlarni sezilarli darajada yengillashtiradi [12], [14].

Mazkur jarayonning amaliy qo‘llanilishini elektron tijorat sohasidagi chatbotlar misolida ko‘rish mumkin. Chatbotlar bilan foydalanuvchi o‘rtasidagi muloqotlar asosan tuzilmasiz matn shaklida bo‘ladi. Masalan, foydalanuvchi quyidagi so‘rovni bildirsa:
“Men qizil 36-o‘lchamdagi oyoq kiyimni buyurtma qilmoqchiman.”
IE tizimi ushbu gapdan quyidagi semantik rollarni aniqlaydi:
– “Men” — *agent* (harakatni boshlovchi subyekt),
– “buyurtma qilmoqchiman” — *predikat* (ifodalangan harakat),
– “qizil oyoq kiyimni” — *theme/obyekt* (harakat yo‘naltirilgan obyekt),
– “36-o‘lchamdagi” — qo‘shimcha atribut (obyektni aniqlovchi belgi).

Mazkur semantik komponentlarning aniq aniqlanishi chatbotlarga foydalanuvchiga mos, mantiqan izchil va mazmunli javoblarni shakllantirish imkonini beradi hamda inson–kompyuter o‘zaro ta’siri sifatini oshiradi.

Axborotni ajratib olish jarayoni odatda quyidagi asosiy NLP vazifalariga tayangan holda amalga oshiriladi:

(1) Nomlangan obyektlarni aniqlash (NER) — shaxs, tashkilot, joy nomlari kabi oldindan belgilangan kategoriyalarni aniqlash [15], [17], [18];

(2) Munosabatlarni aniqlash (RE) — gapdagi obyektlar o‘rtasidagi semantik bog‘lanishlarni topish [16], [19];

(3) Semantik rollarni belgilash (SRL) — gapda ifodalangan harakat asosida predikat–argument tuzilmalarini aniqlash va rollarni belgilash [13], [20], [21].

NER va RE vazifalari asosan obyektlar va ularning o‘zaro aloqalarini aniqlashga qaratilgan bo‘lsa, SRL gapning *chuqur semantik tuzilmasini* ochib

ISSN 3030-3907

berishni maqsad qiladi. Semantik rollar hodisada ishtirok etuvchi birliklarning harakat, holat yoki tajribadagi funksional rolini ifodalaydi [22].

Predikatlar gap mazmunining markaziy qismini tashkil etib, odatda fe'llar yoki fe'l shakllari orqali ifodalanadi va asosiy semantik yukni o‘zida mujassam etadi. Argumentlar esa "kim?", "nima qildi?", "kimga?", "qayerda?" yoki "qachon?" kabi savollarga javob berib, hodisaning semantik talqinida muhim o‘rin tutadi [20].

Matnni predikat–argument tuzilmalariga o‘tkazish orqali SRL sintaktik va semantik jihatdan ancha aniq reprezentatsiyani ta'minlaydi. Bu tizimga harakat manbasini, ishtirokchilarning semantik vazifalarini aniqlash imkonini beradi va noaniqlikni kamaytiradi [20], [22]. Shuningdek, SRL tomonidan taqdim etiladigan semantik chuqurlik NER yoki RE natijalariga qaraganda kengroq bo‘lib, gap tarkibidagi har bir birlikning funksional rolini baholaydi. Semantik tasvir to‘liq bo‘lmaganda esa axborotni ajratib olishda xatolar yuzaga kelib, muhim faktlarning yo‘qolishiga olib kelishi mumkin [19].

Mazkur maqolada o‘zbek tili uchun *Semantik rollarni belgilash (Semantic Role Labeling — SRL)* muammosi tahlil qilinib, ushbu vazifani yechishga qaratilgan asosiy yondashuvlar muhokama qilinadi.

ASOSIY QISM

SRL masalasini yechishda turli yondashuvlardan foydalanish mumkin. Ularning barchasida fe'l yoki semantik jihatdan o‘zaro yaqin bo‘lgan fe'llar va otlar guruhi uchun maxsus semantik rollarni belgilashga asoslanadi [1]. Quyidagi bo‘limlarda ushbu turdagi semantik rollardan foydalanadigan ikki mashhur leksik resurs ko‘rib chiqiladi:

(1) PropBank — proto-rollar va fe'lga xos semantik rollardan foydalanadigan tizim;

(2) FrameNet — umumiy semantik g‘oyaga, ya'ni *freym* tushunchasiga asoslangan holda freymga xos semantik rollarni qo‘llaydigan resurs.

PropBank. Propozitsiyalar banki (Proposition Bank), odatda PropBank nomi bilan tanilgan, gaplardagi predikat–argument tuzilmalariga semantik rollarni biriktiruvchi lingvistik resurs hisoblanadi. Ingliz tili uchun PropBank Penn TreeBank korpusidagi barcha gaplarni semantik rollar bilan belgilaydi. Xuddi shuningdek, xitoy tili uchun ham PropBank Penn Chinese TreeBank asosida ishlab chiqilgan. Universal tematik rollar to‘plamini aniqlash murakkab bo‘lgani sababli, PropBank tizimida semantik rollar *har bir fe'l ma'nosi* alohida tarzda belgilangan. Ya'ni, har bir fe'lning muayyan ma'nosi uchun o‘ziga xos rollar majmuasi mavjud bo‘lib, ular an'anaviy nomlar bilan emas, balki raqamli belgilash orqali ifodalanadi: *Arg0, Arg1, Arg2* va hokazo.

ISSN 3030-3907

Umumiy holda PropBank rollari quyidagi tendensiyalarga ega:

(1) **Arg0** odatda PROTO-AGENTni, ya'ni harakatni yuzaga keltiruvchi sababchi subyektni bildiradi;

(2) **Arg1** ko'pincha PROTO-PATIENT, ya'ni harakat ta'siriga uchraydigan shaxs yoki obyektni ifodalaydi.

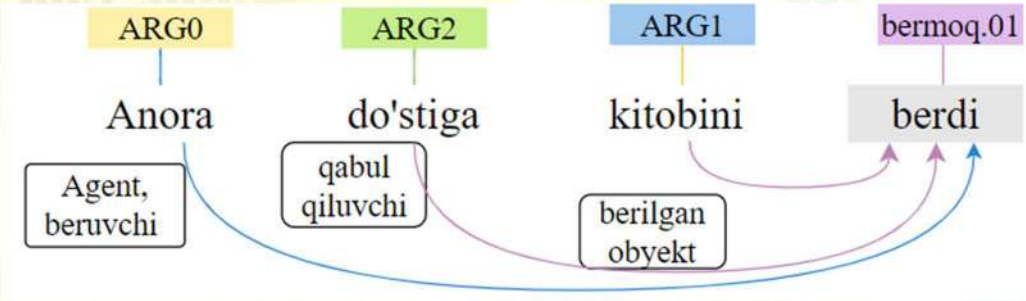
Qolgan argumentlarning ma'nosi esa kamroq barqaror bo'lib, ko'pincha har bir fe'l uchun alohida tarzda aniqlanadi. Shunga qaramay, ayrim umumiy yo'nalishlarni ko'rsatish mumkin:

(1) **Arg2** ko'pincha foyda oluvchi, vosita, xususiyat yoki yakuniy holatni bildiradi;

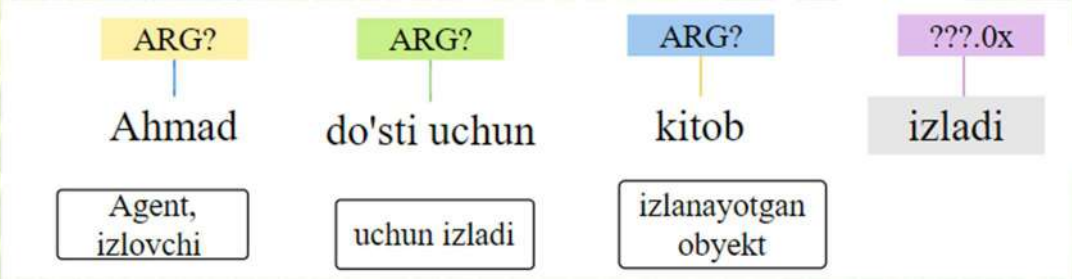
(2) **Arg3** boshlang'ich nuqta, foyda oluvchi, vosita yoki xususiyat rolida kelishi mumkin;

(3) **Arg4** odatda harakatning yakuniy nuqtasini ifodalaydi.

PropBank resursining vazifasi va ishlash prinsipi quyida keltirilgan rasm orqali yanada aniqroq ko'rsatiladi.



1-rasm: ProbBankdagi Semantik rollar teglari



2-rasm: SRL axborot tizimining vazifasi

Quyida *rozi bo'lmoq* va *pasaymoq* fe'llarining har biridan bittadan ma'noga oid soddalashtirilgan PropBank yozuvlari keltiriladi. Bunday PropBank yozuvlari freym fayllar deb ataladi. Ta'kidlash joizki, freym fayllarida har bir semantik rol uchun berilgan tavsiflar (masalan, *"boshqa rozi bo'luvchi shaxs"*, *"pasayish miqdori yoki darajasi"*) rasmiy ta'rif emas, balki inson tomonidan o'qish va tushunishni osonlashtirishga qaratilgan izohiy tavsiflardir.

Rozi bo'lmoq fe'li uchun semantik rollar

— Arg0 — rozi bo'luvchi shaxs

— Arg1 — taklif yoki mazmun (propozitsiya)

ISSN 3030-3907

— Arg2 — boshqa rozi bo‘luvchi shaxs yoki narsa

Namuna:

1. [Guruh]Arg0 [taklif bermaslikka]Arg1 rozi bo‘ldi.
2. [Odatda]ArgM-TMP [Anvar]Arg0 [Ziyoda bilan]Arg2 [hamma narsaga]Arg1 rozi bo‘ladi.

Birinchi gapda Arg0 predikatni bajaruvchi ishtirokchini bildiradi va o‘zbek tilshunosligida bu gapning egasi sifatida talqin qilinadi. Arg1 esa predikat ta’siriga uchrayotgan mazmunni ifodalaydi va u o‘zbek grammatikasida to‘ldiruvchi vazifasida keladi. Ko‘rinib turibdiki, bunday to‘ldiruvchi ayrim hollarda bir nechta tokenlardan tashkil topishi mumkin.

Ikkinchi gapda ArgM-TMP raqamlanmagan semantik argument bo‘lib, predikatning amalga oshish vaqtini bildiradi. Garchi Arg0 ko‘pincha harakatni bajaruvchini anglatsa-da, u har doim ham grammatik ega bilan to‘liq mos kelavermaydi. Ushbu holat keyingi tahlil bo‘limlarida batafsil ko‘rib chiqiladi. Arg2 roli esa predikat kim bilan yoki nima bilan bog‘liq holda bajarilayotganini bildiradi va misolda ko‘rinib turganidek, bu rol ham bir nechta tokenlar orqali ifodalanishi mumkin.

pasaymoq fe’li uchun semantik rollar

Quyidagi gaplarda *pasaymoq* fe’liga mos semantik rollar quyidagicha belgilanadi:

— Arg1 — mantiqiy ega, ya’ni harakat ta’siriga uchrayotgan obyekt (nima pasaymoqda);

— Arg2 — pasayish miqdori yoki darajasi;

— Arg3 — boshlang‘ich nuqta (qayerdan pasaydi);

— Arg4 — yakuniy nuqta yoki natijaviy holat.

Namuna:

3.[Sotuvlar]Arg1 [27 milliondan]Arg3 [25 millionga]Arg4 pasaydi.
4.[O‘rtacha past sifatli obligatsiya]Arg1 [4,2% ga]Arg2 pasaydi.

Bunday gaplarda harakatning sodir bo‘lganligi va uning miqdori yoki chegarasi aniq ifodalanadi. O‘zbek tilshunosligida mazkur turdagi gaplar ko‘pincha egasiz gaplar sifatida talqin qilinadi, chunki bu yerda agent ochiq ifodalanmagan bo‘ladi.

oshirmoq fe’li misolida umumlashtirish:

PropBank tizimidagi semantik rollar fe’l argumentlariga oid sayoz semantik axborotni aniqlashda samarali hisoblanadi. Masalan, *oshirmoq* fe’li uchun quyidagi rollar ajratiladi:

— Arg0 — oshirishga sabab bo‘luvchi subyekt;

— Arg1 — oshirilayotgan obyekt;

ISSN 3030-3907

— Arg2 — oshirish miqdori yoki usuli;
— Arg3 — boshlang‘ich nuqta;
— Arg4 — yakuniy nuqta.

PropBank tizimida, shuningdek, ArgM deb ataluvchi raqamlanmagan argumentlar majmuasi ham mavjud bo‘lib, ular qo‘shimcha semantik ma’lumotlarni ifodalaydi (masalan, ArgM-TMP — vaqt, ArgM-LOC — joy va boshqalar). Ushbu argumentlar ko‘plab predikatlar uchun deyarli o‘zgarmas bo‘lgani sababli, har bir freym faylida alohida sanab o‘tilmaydi. ArgM orqali belgilangan ma’lumotlar predikat bilan bog‘liq vaqt, joy yoki yo‘nalish kabi modifikatsiyalarni aniqlashda muhim ahamiyatga ega bo‘lib, bunday xususiyatlar modelni o‘qitish jarayonida katta foyda keltiradi. ArgM turlarining to‘liq ro‘yxati quyidagi jadvalda keltirilgan [1].

1-Jadval

ArgM sematik rollar oilasida mansub teglar

	ArgM	So‘rog‘i	Izohi	Misol
.	TMP	Qachon?	Vaqtni bildiradi	*Kecha, bugun, ertaga*
.	LOC	Qayerda?	Joyni bildiradi	*Kinoda, tog‘da*
.	DIR	Qayerga/ Qayerdan?	Yo‘nalishni bildiradi	*Pastga, Turkiyadan*
.	PRP/CAU	Nega?	Sabab yoki maqsadni bildiradi	*Hukmga javoban, chunki ...*
.	REC	-	O‘zaro harakatni bildiradi	*O‘zi uchun, har biri*
.	ADV	-	Qo‘shimcha ma’lumot beradi (sifatdosh, ravishdosh)	*odatda*
.	PRD	-	Ikkinchi darajali holat yoki holatdagi subyekt	*Ovqatni qaynoq holda yemoq*

SRL yondashuvlariga kelsak, tanlangan lingvistik resurslardan qat’i nazar, vazifaning murakkabligi sababli dastlabki tadqiqotlarda odatda SRL jarayonining har bir bosqichi uchun alohida-alohida tizimlar qo‘llanilgan [5–6]. Biroq so‘nggi yillarda neyron arxitekturalarning jadal rivojlanishi natijasida to‘liq integratsiyalashgan (end-to-end) yechimlarni yaratish imkoniyati paydo bo‘ldi [7–8]. Shundan so‘ng, SRL vazifasida ketma-ketlikdan ketma-ketlikka (sequence-to-sequence) modellar yangi yechim sifatida ommalasha boshladi, chunki ular predikat–argument tuzilmalarini to‘g‘ridan-to‘g‘ri chiziqli ketma-ketlik

ISSN 3030-3907

ko‘rinishida generatsiya qilish imkonini beradi [9]. Mazkur yondashuvning muhim jihati shundaki, u predikat ma’nosini aniqlash (predicate sense disambiguation), argumentlarni aniqlash va semantik rollarni belgilash jarayonlarini birgalikda, sintaktik tahlildan mustaqil holda bajarishga imkon beradi. Bundan tashqari, u ham span-based, ham dependency-based yondashuvlarni qo‘llab-quvvatlay oladi.

Definition Modeling vazifasi dastlab berilgan so‘z embeddingi asosida uning tabiiy tildagi ta’rifini generatsiya qilishni maqsad qilgan edi [10]. Keyinchalik bu yondashuv ko‘pma’nolilik (polisemiya) hodisasini hisobga olgan holda umumlashtirildi, chunki bitta so‘z turli kontekstlarda turli ma’nolarni ifodalashi mumkin. Garchi mazkur yondashuv bir necha yil avval taklif etilgan bo‘lsa-da, so‘nggi paytlarda unga bo‘lgan ilmiy qiziqish yana kuchaydi [11].

Definition Modeling yondashuvini chuqur o‘rganish asosida biz predikat ma’nosini aniqlash, argumentlarni topish va argumentlarni tasniflash vazifalarini yagona end-to-end jarayon sifatida amalga oshirishni taklif qilamiz. Ushbu vazifaning maqsadi — gap s tarkibidagi predikat p ning argument tuzilmasini tabiiy til orqali tavsiflashdir. Hosil qilinadigan tavsif t^p nafaqat predikat p ning ma’nosini aniqlab berishi, balki p bilan bog‘liq $a_1, a_2, \ldots, a_n$ argumentlar o‘rtasidagi semantik munosabatlarni ham aniq ifodalashi lozim. Bu yerda n — predikatga bog‘langan argumentlar sonini bildiradi.

Masalan, agar p = “bermoq”, va gap s = “Anora kitobni do‘stiga berdi” bo‘lsa, predikat va uning argument tuzilmasining tabiiy tildagi tavsifi quyidagicha ifodalanadi:

t^p =*“bermoq: transfer. [Anora]{beruvchi} berdi [kitobni]{berilgan narsa} [do‘stiga]{berilgan obyekt}”*

Mazkur ketma-ketlik quyidagi elementlarni o‘z ichiga oladi:

i) predikatning ma’nosini aniqlashtiruvchi tavsif,

ii) gap tarkibidagi barcha argumentlarning qavslar ichida aniqlanishi,

iii) har bir argumentga mos semantik rolning biriktirilishi.

Taklif etilayotgan SRL vazifasini hal etish uchun oddiy, ammo samarali end-to-end avtoregressiv yondashuv ishlab chiqiladi. Ushbu usul berilgan gap s va predikat p asosida argument tuzilmasining tabiiy til shaklidagi tavsifi t^p ni generatsiya qiladi.

Xususan, biz sequence-to-sequence modelini ishlab chiqamiz, unda kiruvchi ketma-ketlik s_p quyidagicha aniqlanadi:

(keyingi qism model arxitekturasi va kirish–chiqish formatlari bilan davom ettiriladi).

$$s^p = w_1 \ldots w_i \ldots$$

ISSN 3030-3907

$$< p > \ p_1 \dots p_k < /p > \dots w_n$$

Bu yerda w_i gap s tarkibidagi i-tartibli so‘zni bildiradi, va esa predikat p ning boshlanishi va tugashini ko‘rsatuvchi maxsus belgilardir. Agar predikat p bir nechta tokenlardan iborat bo‘lsa ($k > 1$), ushbu belgilar predikatning butun oraliq qismini qamrab oladi. Shunga muvofiq, model berilgan kirish ma’lumotlari asosida semantik jihatdan boyitilgan t_p gapini generatsiya qilish vazifasini bajaradi.

Hosil qilinadigan t_p gapida:

- predikat p ning ma’nosiga oid ta’rif gap boshida joylashtiriladi;
- p ga tegishli argumentlar kvadrat qavslar ([...]) ichida ko‘rsatiladi;
- har bir argumentdan so‘ng darhol uning semantik roli jingalak qavslar ({...}) orqali belgilanadi.

$$t^p = p_1 \dots p_k : d_1^p \dots d_{k'}^p . w_1 \dots \left[w_1^{\alpha_1} \dots w_{m_1}^{\alpha_1}\right]\left\{d_1^{\alpha_1} \dots d_{m_1'}^{\alpha_1}\right\} \dots$$
$$\left[w_1^{\alpha_2} \dots w_{m_2}^{\alpha_2}\right]\left\{d_1^{\alpha_2} \dots d_{m_2'}^{\alpha_2}\right\}$$
$$\vdots$$
$$\dots\left[w_1^{\alpha_j} \dots w_{m_j}^{\alpha_j}\right]\left\{d_1^{\alpha_j} \dots d_{m_j'}^{\alpha_j}\right\} \dots w_n$$

Bu yerda p_i predikat p ning i-tartibli so‘zini bildiradi;
d_i^p — predikat p ma’nosining ta’rifida ishtirok etuvchi i-tartibli so‘zni anglatadi;
$w_i^{\alpha_j}$ — predikat p ning j-argumentiga tegishli bo‘lgan i-tartibli so‘zni ifodalaydi;
$d_i^{\alpha_j}$ — j-argumentga biriktirilgan semantik rol tavsifidagi i-tartibli so‘zni bildiradi.

Bundan tashqari, k' predikat ta’rifining uzunligini, m_j esa a_j argumentining uzunligini ifodalaydi. Shuningdek, m_j' a_j argumentiga mos keluvchi semantik rol tavsifining uzunligini anglatadi.

Mazkur kodlash asosida ketma-ketlikdan ketma-ketlikka (sequence-to-sequence) model o‘qitiladi va u hosil qilinadigan tabiiy til tavsifiga nisbatan kross-entropiya yo‘qotish funksiyasini minimallashtirishni maqsad qiladi. Bu jarayonda ehtimollik funksiyasi

$$p(t^p \mid s^p)$$

quyidagi ko‘rinishda faktorizatsiya qilinadi:

$$p(t^p|s^p) = p(t_1^p|s^p) \prod_{j=2}^{|t^p|} p(t_j^p|t_{1:j-1}^p, s^p)$$

Agar tabiiy til konvensiyalariga asoslangan holda predikat ma’nolari va semantik rollar bilan belgilangan semantik annotatsiyalangan ma’lumotlar to‘plami mavjud bo‘lsa, uni har doim yuqorida taklif etilgan formulaga mos

ISSN 3030-3907

ko‘rinishga keltirish mumkin. Bundan tashqari, asosiy maqsad gap darajasidagi semantikani ifodalovchi chiqish ketma-ketligini generatsiya qilish bo‘lsa-da, ayrim holatlarda predikat ma’nolari va semantik rollar uchun an’anaviy diskret teglar bilan ishlash ham foydali bo‘lishi mumkin.

Masalan, 2-rasmda ko‘rsatilganidek, “izlamoq” fe’li PropBank ma’lumotlar to‘plamida mavjud emas.

Agar predikat tanlangan predikatlar inventarida mavjud bo‘lsa, u holda tabiiy til tavsifi $d^{\bar{y}}$ predikat p yoki uning a_i argumentlariga biriktirilgan ta’rifga eng yaqin bo‘lgan $\bar{y}$ tegni tanlash imkoniyati paydo bo‘ladi. Mazkur $\bar{y}$ teg quyidagi tarzda aniqlanadi:

$$\bar{y} = argmax_{y \in Y}\, \sigma(f(d^{y}), f(d^{\cdot}))$$

Bu yerda

σ(·) o‘xshashlik funksiyasini (masalan, kosinus o‘xshashligi) ifodalaydi,

f(·) — ta’rifning vektor ko‘rinishini hosil qiluvchi funksiya hisoblanadi,

Y — mavjud teglar to‘plamini anglatadi,

d^{y} — tanlangan ma’lumotlar to‘plamida berilgan y tegqa mos keluvchi tabiiy til ta’rifi hisoblanadi.

Shuni ta’kidlash joizki, soddalik nuqtayi nazaridan generatsiya qilingan chiqish natijalarining to‘g‘riligini tekshirish uchun hech qanday post-processing (keyingi qayta ishlash) usullari qo‘llanilmadi, murakkabroq va aniqligi yuqori bo‘lgan metodlar esa kelgusidagi tadqiqotlar doirasida ko‘rib chiqilishi rejalashtirilgan.

NATIJALAR

Ushbu tadqiqotning muhim bosqichlaridan biri o‘zbek tilidagi gaplarda semantik rollarni aniqlashga mo‘ljallangan axborot tizimi uchun ma’lumotlar to‘plamini tuzish va annotatsiyalashdan iboratdir. Datasetni yaratishda CoNLL formati asos qilib olindi. Quyida annotatsiyalangan ma’lumotlar to‘plamidan namunaviy parcha keltiriladi.

2-Jadval

SRL qilingan CoNNL formatdagi dataset

Sentence ID	Token ID	Token	BIO	Sementic role
1	1	Nodira	B	Arg0
1	2	ko‘zlarini	B	Arg1
1	3	katta	B	ArgM-MNR
1	4	ochdi	B	P
2	1	Otam	B	Arg0
2	2	meni	B	Arg1
2	3	jonidan	B	ArgM-MNR

ISSN 3030-3907

2	4	ortiq	I	ArgM-MNR
2	5	ko‘rar	B	P
2	6	edi	I	P
2	7	.		
3	1	Shunday	B	ArgM-TMP
3	2	do‘ppi	I	ArgM-TMP
3	3	tor	I	ArgM-TMP
3	4	kelgan	I	ArgM-TMP
3	5	paytda	O	ArgM-TMP
3	6	Azim	B	Arg0
3	7	boyvachcha	B	Arg2
3	8	Xolxo‘janing	I	Arg2
3	9	joniga	B	P
3	10	ora	I	P
3	11	kirdi	O	P
3	12	.		

4-jadvalda CoNLL formatining 3–9 ustunlariga oid ma’lumotlar keltirilmagan, chunki ushbu maqola aynan *semantik rollarni belgilash (SRL)*ga qaratilgan. Ushbu format asosida jami 5000 ta gap annotatsiyalangan. Tadqiqotning dastlabki bosqichi sifatida ushbu ma’lumotlar to‘plamiga tayangan holda model o‘qitilgan. Datasetning kichik hajmi yana bir bor uning aniqlikka salbiy ta’sir ko‘rsatganini namoyon etdi. Kelgusidagi ishlarimizda semantik rollar bilan annotatsiyalangan gaplar sonini oshirish va modelni qayta o‘qitish orqali aniqlik darajasini yaxshilashni rejalashtirmoqdamiz.

XULOSA

Xulosa qilib aytganda, semantik rollarni belgilash tabiiy tilni qayta ishlash sohasidagi murakkab va dolzarb vazifalardan biri hisoblanadi. An’anaviy so‘z darajasidagi semantik tahlildan farqli o‘laroq, ushbu muammoni hal qilish matnlardan avtomatik ravishda turli ma’lumotlarni — masalan, semantik tuzilma, maqsad, yo‘nalish va obyektlar — aniqlash imkonini beradi. O‘zbek kompyutatsion tilshunosligida bunday vazifalarni avtomatlashtirish milliy til korpusi imkoniyatlarini sezilarli darajada kengaytiradi. Kelgusidagi ishlarimizda o‘zbek tili uchun semantik rollarni belgilashda yuqori aniqlik natijalariga erishishni maqsad qilganmiz.

ISSN 3030-3907

FOYDALANILGAN ADABIYOTLAR

1. Daniel Jurafsky & James H. Martin. Speech and Language Processing. Copyright © 2024. All rights reserved. Draft of January 12, 2025.21-chapter, 1-21-pp.
2. Levin, B. 1993. English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press.
3. Levin, B. and M. Rappaport Hovav. 2005. Argument Realization. Cambridge University Press.
4. Kipper, K., H. T. Dang, and M. Palmer. 2000. Class-based construction of a verb lexicon. AAAI.
5. Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1192–1202, Berlin, Germany. Association for Computational Linguistics.
6. Diego Marcheggiani and Ivan Titov. 2020. Graph convolutions over constituent trees for syntax-aware semantic role labeling. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3915–3928, Online. Association for Computational Linguistics
7. Jiaxun Cai, Shexia He, Zuchao Li, and Hai Zhao. 2018. A full end-to-end semantic role labeler, syntacticagnostic over syntactic-aware? In Proceedings of the 27th International Conference on Computational Linguistics, pages 2753–2765, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
8. Shexia He, Zuchao Li, and Hai Zhao. 2019. Syntaxaware multilingual semantic role labeling. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5350–5359, Hong Kong, China. Association for Computational Linguistics.
9. Rexhina Blloshmi, Simone Conia, Rocco Tripodi, and Roberto Navigli. 2021. Generating senses and roles: An end-to-end model for dependency- and spanbased semantic role labeling. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pages 3786–3793. International Joint Conferences on Artificial Intelligence Organization.
10. Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, pages 3259–3266. AAAI Press.

ISSN 3030-3907

11. Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.
12. K. Adnan and R. Akbar, ‘‘Limitations of information extraction methods and techniques for heterogeneous unstructured big data,’’ Int. J. Eng. Bus. Manage., vol. 11, Jan. 2019, Art. no. 184797901989077, doi: 10.1177/1847979019890771.
13. A. D. P. Ariyanto, C. Fatichah, and D. Purwitasari, ‘‘Semantic role labeling for information extraction on Indonesian texts: A literature review,’’ in Proc. Int. Seminar Intell. Technol. Appl. (ISITIA), Jul. 2023, pp. 119–124, doi: 10.1109/ISITIA59021.2023.10221008.
14. Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo, and Z. Wang, ‘‘A survey of information extraction based on deep learning,’’ Appl. Sci., vol. 12, no. 19, p. 9691, Sep. 2022, doi: 10.3390/app12199691.
15. A. Goyal, V. Gupta, and M. Kumar, ‘‘Recent named entity recognition and classification techniques: A systematic review,’’ Comput. Sci. Rev., vol. 29, pp. 21–43, Aug. 2018, doi: 10.1016/j.cosrev.2018.06.001.
16. K. Detroja, C. K. Bhensdadia, and B. S. Bhatt, ‘‘A survey on relation extraction,’’ Intell. Syst. Appl., vol. 19, Sep. 2023, Art. no. 200244, doi: 10.1016/j.iswa.2023.200244.
17. J. Li, A. Sun, J. Han, and C. Li, ‘‘A survey on deep learning for named entity recognition,’’ IEEE Trans. Knowl. Data Eng., vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.
18. K. Adnan and R. Akbar, ‘‘An analytical study of information extraction from unstructured and multidimensional big data,’’ J. Big Data, vol. 6, no. 1, p. 91, Dec. 2019, doi: 10.1186/s40537-019-0254-8
19. G. Frisoni, G. Moro, and A. Carbonaro, ‘‘A survey on event extraction for natural language understanding: Riding the biomedical literature wave,’’ IEEE Access, vol. 9, pp. 160721–160757, 2021, doi: 10.1109/ACCESS.2021.3130956.
20. D. Larionov, A. Shelmanov, E. Chistova, and I. Smirnov, ‘‘Semantic role labeling with pretrained language models for known and unknown predicates,’’ in Proc. Natural Lang. Process. a Deep Learn. World, Oct. 2019, pp. 619–628, doi: 10.26615/978-954-452-056-4_073.
21. H. Fei, S. Wu, Y. Ren, and D. Ji, ‘‘Second-order semantic role labeling with global structural refinement,’’ IEEE/ACM Trans. Audio, Speech, Lang., Process., vol. 29, pp. 1966–1976, 2021, doi: 10.1109/TASLP.2021.3082706.

ISSN 3030-3907

22. J. Akanya and C. G. Omachonu, ‘‘Meaning and semantic roles of words in context,’’ Int. J. English Lang. Linguistics Res., vol. 7, no. 2, pp. 1–9, 2019. [Online]. Available: https://www.eajournals.org

ISSN 3030-3907