

**SEMANTIK ANNOTATSIYALANGAN KORPUSSLARNING
MASHINA TARJIMASIDAGI AHAMIYATI (O‘ZBEK TILI
MISOLIDA)**

Munisa Xudoyberganova

TDTU assistent o‘qituvchisi

munisaxudayberganova13@gmail.com

Annotatsiya: Mazkur maqolada mashina tarjimasi tizimlarida semantik jihatdan annotatsiyalangan lingvistik korpuslarning o‘rni va ahamiyati tahlil qilinadi. Unda mashina tarjimasining qoidaga asoslangan, statistik va neyron yondashuvlari hamda ularning lingvistik resurslarga bo‘lgan ehtiyoji yoritiladi. Semantik annotatsiyaning predikat, argument va semantik rollarni formal ifodalash orqali matn mazmunini chuqurroq modellashtirish imkoniyatlari ko‘rib chiqiladi. Shuningdek, PropBank, FrameNet, SemCor va MultiSemCor kabi semantik resurslarning mashina tarjimasi va tabiiy tilni qayta ishlashdagi ahamiyati tahlil qilinadi. Tadqiqotda kam resursli tillar, xususan, o‘zbek tili uchun semantik annotatsiyalangan korpuslarni shakllantirishning dolzarb jihatlari yoritiladi.

Kalit so‘zlar: mashina tarjimasi, neyron mashina tarjimasi, parallel korpus, semantik annotatsiya, semantik rol, o‘zbek tili, kam resursli til, tabiiy tilni qayta ishlash.

Bugungi kunda sun‘iy intellekt va tabiiy tilni qayta ishlash texnologiyalarining jadal rivojlanishi natijasida mashina tarjimasi lingvistik tadqiqotlarning eng faol yo‘nalishlaridan biriga aylandi. Mashina tarjimasi dastlab inson tarjimoniga yordam beruvchi avtomatlashtirilgan vosita sifatida shakllangan bo‘lsa, hozirgi kunda neyron tarmoqlar asosida ishlaydigan tizimlar turli tillar o‘rtasida katta hajmdagi matnlarni avtomatik tarjima qilish imkonini bermoqda. Neyron mashina tarjimasi an‘anaviy statistik

yondashuvdan farqli ravishda tarjima jarayonini yagona neyron model orqali o'rganishga yo'naltirilgan bo'lib, attention mexanizmining rivojlanishi manba gapning tarjima qilinayotgan birlik uchun muhim qismlarini aniqlash imkonini kengaytirdi (Bahdanau et al., 2015).

Keyingi bosqichda Transformer arxitekturasining taklif qilinishi neyron mashina tarjimasi rivojida muhim burilish yasadi. Ushbu arxitektura tarjima jarayonida attention mexanizmiga asoslanib, ketma-ket hisoblashga bog'liqlikni kamaytirishga imkon berdi (Vaswani et al., 2017).

Shu bilan birga, zamonaviy neyron mashina tarjimasi tizimlari qator muammolarga duch kelmoqda. Xususan, o'quv ma'lumotlarining hajmi, kam uchraydigan so'zlar, uzun gaplar, so'zlararo moslashtirish va domenlar o'rtasidagi farqlar NMT tizimlari uchun muhim muammolar sifatida qayd etilgan (Koehn & Knowles, 2017).

Mashina tarjimasi tizimlarining rivojlanishini shartli ravishda qoidaga asoslangan, statistik va neyron yondashuvlar bosqichlariga ajratish mumkin. Dastlabki tizimlarda lingvistik qoidalar, grammatik modellar va lug'atlarga asosiy e'tibor qaratilgan. Keyinchalik statistik mashina tarjimasi katta hajmdagi parallel korpuslardan foydalanib, so'z va birikmalar o'rtasidagi ehtimoliy mosliklarni aniqlashga asoslandi. Hozirgi neyron mashina tarjimasi esa chuqur o'rganish va neyron tarmoqlar yordamida manba va maqsad tillari o'rtasidagi murakkab bog'lanishlarni modellashtirish imkoniyatini kengaytirdi (Bahdanau et al., 2015; Vaswani et al., 2017).

Shu bilan birga, kam resursli tillar uchun mashina tarjimasi tizimlarini yaratishda ma'lumotlar yetishmasligi muhim muammolardan biri bo'lib qolmoqda. O'zbek tili ham katta resursli tillar bilan taqqoslaganda parallel, semantik va yuqori sifatli annotatsiyalangan ma'lumotlarga bo'lgan ehtiyoji yuqori bo'lgan tillardan biridir. Shu sababli o'zbek tili uchun zamonaviy lingvistik korpuslarni yaratish va ularni mashina tarjimasi texnologiyalari bilan bog'lash dolzarb masalalardan hisoblanadi.

Mashina tarjimasi tizimlarining sifatini ta'minlovchi asosiy omillardan biri lingvistik ma'lumotlarning hajmi va sifatidir. Ayniqsa statistik va neyron mashina tarjimasi tizimlari uchun parallel korpuslar muhim ma'lumot manbai hisoblanadi. Parallel korpusda bir tildagi matn va uning boshqa tildagi tarjimasi o'zaro moslashtirilgan holda taqdim etiladi. Parallel korpuslardan foydalanish mashina tarjimasi tizimlariga ikki til birliklari o'rtasidagi mosliklarni o'rganish imkonini beradi. Biroq tarjima jarayonida faqat leksik birliklarning mos kelishi yetarli emas. Tillar o'rtasidagi grammatik, sintaktik va semantik farqlar ham tarjima sifatiga bevosita ta'sir qiladi. Neyron mashina tarjimasi bo'yicha tadqiqotlarda ham ma'lumot hajmi, so'zlararo alignment va uzoq masofali bog'lanishlar kabi masalalar alohida muammo sifatida ko'rsatiladi (Koehn & Knowles, 2017). Masalan, o'zbek va ingliz tillaridan tashkil topgan parallel korpusda o'zbek tilidagi gap va uning ingliz tilidagi tarjimasi bir-biriga moslashtiriladi. Bunday ma'lumotlar mashina tarjimasi modeliga ikki til birliklari o'rtasidagi ehtimoliy mosliklarni o'rganish imkonini beradi. Biroq tillar o'rtasidagi farqlar faqat leksik birliklarning mos kelishi bilan chegaralanmaydi. Har bir til voqea, harakat va ishtirokchilarni turlicha ifodalashi mumkin. Bir tildagi gapda ma'lum semantik rolni ifodalovchi birlik boshqa tilda boshqa grammatik yoki leksik vosita orqali berilishi mumkin. Shuning uchun mashina tarjimasi tizimlari uchun semantik ma'lumotlarning mavjudligi alohida ahamiyatga ega. Semantik annotatsiya matndagi lingvistik birliklarning mazmuniy xususiyatlarini formal shaklda belgilash jarayonidir. Bunda predikat, argument, semantik rol, hodisa va ular o'rtasidagi munosabatlar aniqlanadi. Semantik annotatsiyaning muhim namunalaridan biri PropBank bo'lib, unda sintaktik tuzilmalarga predikat-argument ma'lumotlari, ya'ni semantik rollar qatlami qo'shiladi (Palmer et al., 2005). Mazkur yondashuv orqali gap tarkibidagi birliklarning nafaqat grammatik vazifasi, balki predikat bilan bog'liq semantik funksiyasi ham formal ravishda ifodalanadi. Masalan, "Talaba kitobni kutubxonadan oldi"

ilmiy –amaliy anjuman

gapida “oldi” fe’li asosiy predikatni ifodalaydi. “Talaba” harakatni amalga oshiruvchi ishtirokchi sifatida Agent, “kitobni” harakat ta’sir qiladigan obyekt sifatida Theme, “kutubxonadan” esa harakatning boshlang‘ich manbasini bildiruvchi Source rolda annotatsiya qilinishi mumkin. Ushbu birliklarning semantik vazifalari 1-jadvalda ko‘rsatilgan.

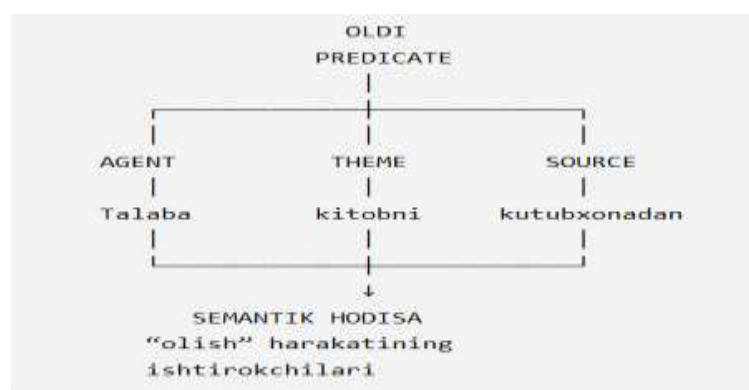
1-jadval. “Talaba kitobni kutubxonadan oldi” gapining semantik annotatsiyasi.

Matndagi birlik	Grammatik vazifasi	Semantik roli	Predikat bilan munosabati
Talaba	Ega	Agent	Harakatni amalga oshiruvchi
kitobni	To‘ldiruvchi	Theme	Harakat ta’sir qiladigan obyekt
kutubxonadan	Hol	Source	Harakatning boshlang‘ich manbai
oldi	Kesim	Predicate	Asosiy harakatni ifodalovchi birlik

Jadvaldan ko‘rinadiki, semantik annotatsiyada gap tarkibidagi birliklar alohida so‘zlar sifatida emas, balki predikat va uning argumentlari o‘rtasidagi munosabatlar tizimi sifatida ham ifodalanadi. Bunday yondashuv matn mazmunini formal modellashtirish, semantik munosabatlarni avtomatik aniqlash va turli tillardagi strukturaviy farqlarni qiyosiy tahlil qilish imkonini beradi. Ayniqsa, mashina tarjimasida bunday semantik tuzilmani hisobga olish muhim ahamiyatga ega. Chunki mashina faqat “talaba → student”,

“kitob → book”, “oldi → took” kabi alohida leksik birliklarning mosligini aniqlash bilan cheklanmay, ularning gapdagi semantik vazifalarini ham hisobga olishi mumkin. Bunda **Agent + Theme + Source → Predicate** ko‘rinishidagi predikat-argument tuzilmasi tarjima jarayonida mazmuniy munosabatlarni saqlashga xizmat qiladi. Shu jihatdan, semantik annotatsiyalangan korpuslar, xususan PropBank kabi resurslar, tabiiy tilni qayta ishlash, mashina tarjimasi va semantik tahlil tizimlari uchun muhim lingvistik asos vazifasini bajaradi. Ushbu tamoyillarni o‘zbek tilidagi matnlarga tatbiq etish esa o‘zbek tilining semantik xususiyatlarini formal tavsiflash va kelgusida UzSemCor kabi semantik korpuslarni shakllantirish uchun nazariy va amaliy asos yaratadi.

Semantik annotatsiyaning yana bir muhim modeli FrameNet loyihasi bilan bog‘liq. FrameNet leksik birliklarni ma’lum semantik freymlar va freym elementlari bilan bog‘lashga asoslanadi. Shu sababli u til birliklarining faqat shakliy emas, balki vaziyat va ma’no bilan bog‘liq xususiyatlarini ham modellashtirish imkonini beradi (Baker et al., 1998). Bu yondashuv mashina tarjimasida bir tildagi ma’no tuzilmasini boshqa tilda qanday ifodalash mumkinligini tadqiq qilish uchun nazariy asos yaratadi.



1-sxema. **Predikat-argument munosabatlarning semantik modeli.**

O'zbek tili agglutinatív til bo'lib, grammatik ma'nolarning katta qismi affikslar yordamida ifodalanadi. Shu sababli o'zbek tilini mashina tarjimasí tizimlarida qayta ishlashda morfologik segmentatsiya, lemma aniqlash, grammatik belgilarni ajratish va sintaktik tuzilmani modellashtirish muhim ahamiyatga ega. O'zbek tilidagi gaplarning semantik tuzilishini modellashtirishda kelishik shakllari ham muhim ma'lumot manbai bo'lishi mumkin. Masalan, -ni, -ga, -dan, -da kabi qo'shimchalar otning gapdagi grammatik munosabatlarini aniqlashda yordam beradi. Biroq ularni to'g'ridan-to'g'ri semantik rol bilan tenglashtirish mumkin emas. Semantik rolni aniqlashda predikatning xususiyati, kontekst va sintaktik munosabatlar ham hisobga olinishi kerak.

Kam resursli tillarda yana bir muammo sifatli parallel korpuslarning cheklanganligidir. Korpus hajmining kichikligi neyron mashina tarjimasí modellarining umumlashtirish qobiliyatiga ta'sir qilishi mumkin. Koehn va Knowles NMT uchun trening ma'lumotlari miqdorini, kam uchraydigan so'zlarni, uzun gaplarni va alignment masalasini asosiy muammolar qatorida ko'rsatadi (Koehn & Knowles, 2017). Shu sababli mavjud korpuslarni nafaqat hajm jihatidan kengaytirish, balki ularga qo'shimcha lingvistik ma'lumotlar, jumladan, morfologik, sintaktik va semantik annotatsiya qatlamlarini kiritish muhimdir.

Semantik annotatsiyalangan korpus mashina tarjimasining bir necha bosqichida foydali bo'lishi mumkin. Birinchidan, u gapning asosiy predikati va uning argumentlarini aniqlash imkonini beradi. Ikkinchidan, turli tillarda bir xil semantik vazifani bajaruvchi birliklarni qiyoslash imkonini yaratadi. Uchinchidan, semantik ma'lumot tarjima modeliga kontekstni chuqurroq modellashtirish uchun qo'shimcha lingvistik signal sifatida xizmat qilishi mumkin. PropBank tajribasi predikat-argument munosabatlarini korpusda tizimli ravishda belgilash mumkinligini ko'rsatadi (Palmer et al., 2005).

Shuningdek, AnCora korpusi morfologik, sintaktik va semantik qatlamlarni birgalikda tashkil etish imkoniyatini namoyish etadi. Unda argument tuzilmalari, tematik rollar, semantik fe'l sinflari va boshqa ma'lumotlar morfologik hamda sintaktik annotatsiya bilan birgalikda berilgan (Taulé et al., 2008).

Bunday ko'p qatlamli model o'zbek tili uchun ham muhim metodologik tajriba bo'la oladi. Chunki mashina tarjimasida so'zning shakli bilan birga uning gapdagi sintaktik va semantik vazifasi ham hisobga olinishi mumkin. Masalan, bir fe'l turli kontekstlarda turlicha ma'no va semantik tuzilishga ega bo'lishi mumkin. Faqat so'z shakliga asoslangan tarjima tizimi bunday holatlarda noaniqlikka duch kelishi mumkin. Semantik annotatsiya esa fe'lning muayyan kontekstdagi ma'nosi va uning argumentlarini aniqlashga yordam beradi. Semantik annotatsiyalangan parallel korpus yaratishda mavjud annotatsiyalangan resurslardan foydalanish ham mumkin. MultiSemCor tadqiqotida manba tildagi semantik annotatsiyalarni parallel matn va so'zlararo alignment orqali boshqa tilga ko'chirish yondashuvi ko'rib chiqilgan. Mualliflar bunday yondashuvni resurslari cheklangan tillar uchun annotatsiyalangan korpuslarni yaratishda mehnatni kamaytiruvchi istiqbolli usul sifatida ko'rsatadilar (Bentivogli & Pianta, 2005).

Mashina tarjima tizimlarining zamonaviy rivojlanish bosqichida katta hajmdagi parallel korpuslar bilan bir qatorda chuqur lingvistik annotatsiyaga ega resurslarga bo'lgan ehtiyoj ham ortib bormoqda. Neyron mashina tarjima rivojida attention va Transformer kabi yondashuvlar muhim rol o'ynagan bo'lsa-da, modelning samaradorligi uchun ma'lumotlarning hajmi va sifati ham muhim bo'lib qolmoqda.

Semantik annotatsiya tarjima jarayonida gapning mazmuniy tuzilishini, predikat va argumentlar o'rtasidagi munosabatlarni hamda semantik rollarni formal ko'rinishda ifodalash imkonini beradi. **PropBank** va **FrameNet** kabi resurslar bunday annotatsiyaning turli modellarini namoyish etadi.

O'zbek tili kabi resurslari nisbatan cheklangan tillar uchun semantik annotatsiyalangan korpuslarning yaratilishi mashina tarjimasi texnologiyalarining lingvistik asoslarini mustahkamlashga xizmat qilishi mumkin. Bunda morfologik, sintaktik va semantik qatlamlarning o'zaro bog'langan holda tashkil etilishi muhim ahamiyat kasb etadi. **AnCora** tajribasi bunday ko'p qatlamli annotatsiya modelining amaliy imkoniyatlarini ko'rsatadi. Shunday qilib, o'zbek tilining semantik korpusini yaratish faqat korpus lingvistikasi doirasidagi vazifa emas, balki o'zbek tilida sun'iy intellekt va mashina tarjimasi texnologiyalarini rivojlantirish uchun zarur bo'lgan lingvistik infratuzilmaning muhim tarkibiy qismi sifatida qaralishi mumkin.

FOYDALANILGAN ADABIYOTLAR

1. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. Proceedings of ICLR.
2. Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. Proceedings of COLING-ACL, 86–90. <https://doi.org/10.3115/980845.980860>
3. Bentivogli, L., & Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. Natural Language Engineering, 11(3), 247–261. <https://doi.org/10.1017/S1351324905003839>
4. Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. Proceedings of the First Workshop on Neural Machine Translation, 28–39. <https://doi.org/10.18653/v1/W17-3204>

5. Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. Computational Linguistics, 31(1), 71–106. <https://doi.org/10.1162/0891201053630264>
6. Taulé, M., Martí, M. A., & Recasens, M. (2008). AnCorra: Multilevel annotated corpora for Catalan and Spanish. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.