MASHINALI TARJIMA TIZIMLARI UCHUN O‘ZBEK TILI SO‘ZLARINI LEMMALASH VA GAP STRUKTURALARINI TAHLIL QILISH

Sharipov Maqsud Siddiqovich
Texnika fanlari nomzodi, dotsent
UrDU, Kompyuter ilmlari kafedrasi dotsenti.
E-mail: m.sharipov@urdu.uz

Sobirov Og‘abek Ollayor o‘g‘li
UrDU, Kompyuter ilmlari kafedrasi tayanch doktoranti.
E-mail: sobirov.o@urdu.uz

Sharipova Gavharjon Maqsud qizi
UrDU, Kompyuter ilmlari kafedrasi talabasi.
E-mail: gavharjonsharipova1@gmail.com

Annotatsiya: Tarjima butun dunyodagi odamlar o‘rtasida samarali muloqot qilish imkonini beradi. Mashina tarjimasi - bu tabiiy tilni bir tildan boshqasiga avtomatik ravishda, ma’nosini saqlab qolgan holda o‘girishdir. Ushbu maqolada mashinali tarjima tizimlari uchun o‘zbek tili gap strukturasi va o‘zbek tili so‘zlarini lemmalash, lemmalashning tarjima uchun muhimligi haqida ma’lumot berilgan. Mashina tarjimasi turlari bir nechta hisoblanadi, lemmalash aynan qaysi turlar uchun ishlatilishi ham aytib o‘tilgan va misollar keltirilgan.
Kalit so‘zlar. Mashinali tarjima, so‘z, lemma, lemmalash, gap, gap bo‘laklari, gap turlari.

Annotation: Translation enables effective communication between people all over the world. Machine translation is the automatic translation of natural language from one language to another, preserving the meaning. This article provides information on Uzbek sentence structure and Uzbek word lemmatization for machine translation systems, and the importance of lemmatization for translation. There are several types of machine translation, the use of lemmatization specifically for those types is mentioned and examples are given.
Keywords: Machine translation, word, lemma, lemmatization, sentence, sentence fragments, sentence types.

Аннотация: Перевод обеспечивает эффективное общение между людьми по всему миру. Машинный перевод — это автоматический перевод естественного языка с одного языка на другой с сохранением смысла. В этой статье представлена информация о структуре предложений на узбекском языке и лемматизации узбекских слов для систем машинного перевода, а также о важности лемматизации для перевода. Существует несколько видов машинного перевода, упоминается использование лемматизации именно для этих типов и приводятся примеры.
Ключевые слова: Машинный перевод, слово, лемма, лемматизация, предложение, фрагменты предложений, типы предложений.

Mashina tarjimasi (MT) - bu tabiiy tilni bir tildan boshqasiga avtomatik ravishda, ma'nosini saqlab qolgan holda o'girishdir. Tarjima butun dunyodagi odamlar o'rtasida samarali muloqot qilish imkonini beradi. Ta'lim, iqtisod, meditsina, turizm va san'at sohalarida MT tizmlaridan keng ko'lamli va samarali foydalanib kelinmoqda. Bundan tashqari, MT katta hajmdagi matnlarni tezda tarjima qilish uchun ishlatilishi mumkin, bu esa an'anaviy tarjima usullari yordamida imkonsiz bo'lib qoladi .

MTning asosiy turlari:

- Qoidalarga asoslangan mashina tarjimasi (Rule-based machine translation)
- Lug'atga asoslangan mashina tarjimasi (Dictionary-based machine translation)
- Statistik mashina tarjimasi (Statistical machine translation)
- Misol asosidagi mashina tarjimasi (Example-based machine translation)
- Neyron mashina tarjimasi (Neural machine translation)
- Gibrid mashina tarjimasi (Hybrid machine translation)

Har bir narsaning ustun va kamchilik tomonlari bo'lganidek, MT tizimlarini qurishda ham o'ziga yarasha kamchilik tomonlari bor. MT tizimlarini qurishda uchraydigan kamchiliklar sifatida quyidagilarni aytish mumkin, *bir nechta ma'noli so'zlar, bir nechta grammatik tuzilishga ega bo'lgan jumlalar* va *grammatikaning boshqa muammolari*. Ushbu maqolada, o'zbek tili uchun MT tizimlari qurishda yuqoridagi muammolardan, aynan *murakkab grammatik tuzilishga ega jumlalar (gaplar)* haqida to'liq ma'lumot beriladi. Bundan tashqari maqolada mashinali tarjima tizimlari uchun o'zbek tili so'zlarini lemmalash, lemmalashning tarjima uchun muhimligi haqida ma'lumot beriladi. Mashina tarjimasi turlari bir nechta hisoblanadi, lemmalash aynan qaysi turlar uchun ishlatilishi ham aytib o'tilgan va misollar keltirilgan.

Lemmalash. Lemma so'zning asosiy yoki lug'at shaklidir. Tilshunoslikda, ayniqsa morfologiya va leksikografiyada lemma so'zning kanonik shaklini ifodalaydi, undan uning barcha flektiv shakllari hosil qilish mumkin. Lemmalash - bu so'zlarni asosiy yoki lug'at shakliga, ya'ni lemmasiga qisqartirish jarayonidir. Lemmalash o'zgaruvchanlikni kamaytirish uchun turli xil tabiiy tillarni qayta ishlash [inglizcha: Natural Language Processing (NLP)] vazifalarida, jumladan, mashina tarjimasi, so'z turkumlarini belgilash va hissiyotlarni tahlil qilishda qo'llaniladi. Misol uchun, *'matn', 'matnni', 'matnga', 'matnlar'* kabi so'z shakllarining lemmasi *'matn'* hisoblanadi.

Gap strukturasini o'rganish orqali MTning bir nechta turlari, xususan, misol asosidagi mashina tarjimasi (example-based machine translation) orqali quriladigan MT tizimlarini yaxshilash mumkin. Misol asosidagi mashina tarjimasi - bu MTning turi bo'lib, u ko'pincha oniy vaqtda parallel matnlardan iborat ikki tilli korpusdan foydalanish bilan tavsiflanadi. Quyidagi misolda MAMT tizimini qurishda gap struktursi va lemmalashning ahamiyatini ko'rib chiqamiz:

I. Juft jumlalarni o'z ichiga olgan ikki parallel korpus olinadi.
*Ana u **ko'k soyobon**ning narxi qancha? - How much is that **blue umbrella**?*
*Ana u **katta kamera**ning narxi qancha? - How much is that **big camera**?*

II. Bu juftliklar o'xshash va farq qiluvchi qismlarga ajratiladi, bunda farqlanuvchi va o'xshash qismlar so'zlari lemmalash orqali topiladi. Keyin o'xshash va farqli so'zlar o'zaro moslanadi, shu tariqa matnlar tarjima qilinadi.
*Ana u **X**ning narxi qancha? – How much is that **X**?*
ko'k soyobon – blue umbrella

katta kamera – big camera

O‘zbek tili ilmiy faoliyatida lemmalash va mashinali tarjima jarayonlari uchun muhim ishlar amalga oshirilib natijalari e’lon qilingan va davom etmoqda. Ushbu maqolani yozishda shunday ilmiy ishlar bilan tanishib qilgan. Ushbu maqolada qoidaga asoslangan lemmalash jarayonining algoritmini qurish haqida aytilgan, bunda so‘zlar va qo‘shimchalar bazasi yig‘ilgan va qoidalar asosida qo‘shimchalarni olib tashlab so‘zlarning lemmasini topish chekli holatlar mashinasi orqali amalga oshirilgan [1]. Ushbu maqolada o‘zbek tilida qo‘llaniladigan uchta yozuv shakli: kirill, hozirda rasmiy lotin va yangi e’lon qilingan yangi lotin alifbolari o‘rtasida mashina transliteratsiyasi (harflarni almashtirish) vositasini taqdim etishgan bo‘lib, bu jarayon qoidalarga asoslangan va sozlash usullari kombinatsiyasidan foydalangan holda yaratilgan [2]. Bu maqolada, avtomatik morfologik tahlil qiluvchi dasturning jarayoni qisqacha ko‘rsatib o‘tilgan bo‘lib. ushbu dastur mashina tarjimasi va o‘zbek tilidagi so‘zlar birliklarining tartibi va qoidalari morfologik jihatdan o‘rganilgan [3]. Bu maqolada agglyutinativ tillar lug‘atida mavjud bo‘lmagan so‘zlar uchun teglash masalasi, lemmalash va o‘zakni topish jarayonlari bilan hamohangligida tushuntirilib misollar bilan keltirib o‘tilgan [4]. Ushbu maqolada teglash metodi bo‘yicha ya’na bir muhim ma’lumotlar keltirib o‘tilgan bo‘lib, bu jarayonning yechimlari turli xil chizmalar va jadvallar asosida keltirilgan, aynan *‘uzun’* so‘zining ko‘p manoligi 1-jadvalda keltirib o‘tilgan [5].

Mashinali tarjima uchun lemmalashning muhimligini xorijiy ilmiy ishlarda ham ko‘rish mumkin. Ushbu maqolada mualliflar ingliz-malay parallel korpusini olish orqali ingliz-malay statistik mashina tarjimasini amalga oshirishgan, bunda, ular lug‘at va inglizcha lemmalash jarayoni yordamida ingliz-malay statistik mashina tarjimasini yaxshilashga harakat qilishgan va tadqiqotilari shuni ko‘rsatadiki, ikki tilli lug‘at va inglizcha lemmatizatsiyadan foydalanish yondashuvining kombinatsiyasi ingliz tilidan malay tiliga tarjima uchun BLEU ballini 12,90 dan 15,41 gacha oshirdi [6]. Bu maqolada xitoy tilidan ingliz tiliga statistik mashina tarjimasini yaxshilash uchun ingliz lemmatizatsiyasidan foydalanishni taklif qilishgan. Bunda mualliflar CLAW-5 teglar to‘plami uchun ingliz tilidagi so‘z turkumlarini teglash vositasini joriy qilishgan va ingliz tilini lemmatizatsiya qilish uchun ingliz morfologik analizatoridan foydalanishgan, ularning yondashuvi oz miqdordagi o‘quv ma’lumotlari bilan statistik mashina tarjimasini uchun juda samarali ekanligi isbotlangan va ular keng qamrovli tajribalar o‘tkazishgan, xulosalar chiqarish uchun eng zamonaviy baholash ko‘rsatkichlaridan (BLEU, METEOR va TER) foydalanishgan [7]. Ushbu maqolada ma’lumotlarning siyrakligi statistik mashina tarjimasini yomonlashtiradigan omillardan biri ekani aytib o‘tilgan, ularning tegishli ishlar shuni ko’rsatdiki, morfo-sintaktik ma’lumotlardan foydalanish ma’lumotlarning siyrakligini samarali hal qiladi, biroq, inglizcha morfo-sintaktik tahlildan foydalangan holda xitoy-ingliz tilidagi SMT uchun kamroq harakatlar qilingan, tajribalarda ingliz lemmalaridan foydalanish so‘zlarni siyrakligi sifatini sezilarli darajada yaxshilashi mumkinligini aniqlashgan, bu esa tarjima samaradorligini oshirishga olib keladi [8].

Sintaksis. Gaplarning qurilishi hamda ifoda xususiyatlarini o‘rganuvchi tilshunoslik bo‘limi sintaksis deyiladi. Sintaksis so‘zlarning bir-biriga bog‘lanishi hamda gap tuzilishi qonun-qoidalarini o‘rganishiga ko‘ra ikkiga bo‘linadi: *so‘z birikmasi sintaksisi* va *gap sintaksisi*. So‘z birikmasi mustaqil so‘zlarning bog‘lanishidan hosil bo‘ladi, masalan: *do‘stimning kitobi, maktab bog‘i*. Ohang va fikr tugalligiga ega bo‘lib, kesimlik shartlri orqali ifodalanuvchi so‘z va so‘zlar qo‘shilmasiga gap deyiladi, masalan: *Bugun xavo ochiq. Yurtimiz qanday chiroyli!* Ilmiy va rasmiy

uslublarda gapda so‘zlarning odatdagi tartibi, asosan, qat’iy saqlanadi. Lekin so‘zlashuv, badiiy va publistik uslublarda bu tartib o‘zgarishi, ya’ni inversion tartib qo‘llanishi mumkin.

Gap bo‘laklari. Gap tarkibida ma’lum so‘roqqa javob bo‘luvchi mustaqil so‘z yoki so‘zlar birikmasi gap bo‘laklari hisoblanadi. Gap bo‘laklari 1) *bosh bo‘laklar* va 2) *ikkinchi darajali gap bo‘laklaridan* iborat bo‘ladi. Bundan tashqari 3) *uyushuq bo‘laklar*, 4) *ajratilgan bo‘laklar* va gap bo‘laklari bilan aloqaga kirishmaydigan bo‘laklar: 5) *undalma* va 6) *kiritmalarga* ham bo‘linadi.

1) Bosh bo‘laklar

- Ega *(kim? nima? qayer?)* - *Bulbul sayradi.*
- Kesim *(nima qildi? nima bo‘ldi?)* - *Qushlar sayraydi (fe’l). Osmonimiz musaffo (ot).*

2) Ikkinchi darajali bo‘laklar

- Aniqlovchi *(qanday? qancha?)* - *Aytar (sifatlovchi) so‘zni ayt. Mehnatning (qaratqich) ziynati ko‘p.*
- To‘ldiruvchi *(kimga?, nimaga?)* - *Inson bilimga (vositali) intiladi. Aqlni (vositasiz) beaqldan o‘rgan.*
- Hol *(qachon? nima uchun?)* - *Aqlni o‘stirmoq uchun hadeb o‘qiyvermasdan ko‘proq fikrlash kerak.*

3) Uyushiq bo‘laklar. Gapda bir xil sintaktik vazifa bajarib, bir xil so‘roqqa javob bo‘luvchi, o‘zaro teng aloqada, boshqa bo‘laklar bilan esa tobe aloqada bo‘lgan bo‘laklar uyushiq bo‘laklar deyiladi. Misol: *Yo‘lda ba’zan katta-katta maydonni qoplab olgan oq, pushti, sariq va qizil gular uchraydi.*

4) Ajratilgan bo‘laklar. Gapning ajratilgan bo‘laklari o‘zaro aloqada bo‘lgan bo‘laklarning ma’nosini izohlab, bo‘rttirib, aniqlashtirib yoki ta’kidlab keladi. Ular boshqa bo‘laklardan maxsus to‘xtam bilan ajralib, gap urg‘usini (mantiqiy urg‘u) oladi. Misol: *Sizga, o‘rta bo‘yli qizga, atlas juda yarashibdi.*

Gap bo‘laklari bilan aloqaga kirishmaydigan bo‘laklar.

5) Undalmalar. Undalma - deb so‘zlovchining nutqi (fikri) qaratilgan shaxs (obyekt) yoki narsalarni bildiruvchi so‘z va so‘zlar birikmasiga aytiladi. Misol: *Bahor, ketma bizning bog‘lardan.*

6) Kiritmalar (Kirish so‘zlar). Kiritmalar so‘zlovchining o‘zi bayon etayotgan fikriga munosabati (*ishonchi, gumoni, tasdig‘i, inkori va h.k.*), fikrning birovga nisbatlanishi (*mansubligi, aloqadorligi*), o‘z fikri tarkibiy qismlarining ahamiyatliligi darajasi (*birinchidan, ikkinchidan va h.k.*), fikri va uning tarkibiy qismlariga doir bayon etadigan qo‘shimcha axboroti yoki izohi kabi rang-barang ma’nolarni ifodalaydi. Misol: *Xayriyat, vaqtida yetib keldik.*

Gapning ifoda maqsadiga ko‘ra turlari. Har bir gapda ma’lum bir maqsad, niyat yoki his-hayajon ifodalanadi. So‘zlovchi biror narsa, voqea-hodisa haqida xabar qiladi, yoki o‘ziga noma’lum bo‘lgan narsa va voqea-hodisalar to‘g‘risida ma’lumot olishni istaydi, tinglovchiga biror ishni bajarish bo‘yicha buyruq-xitobini bildiradi. Gaplar ana shu maqsadlarga ko‘ra turlicha quriladi, o‘ziga xos grammatik xususiyatlarga ega bo‘lib, alohida ohang bilan talaffuz etiladi. Ifoda maqsadiga ko‘ra gaplar: 1) darak gaplar, 2) so‘roq gaplar, 3) buyruq gaplar, 4) istak gaplarga bo‘linadi.

1) Darak gap. Darak, xabar ma’nosini bildirib, kesimi aniqlik maylidagi fe’l orqali ifodalanadi. Misol: *Filiologiya fakulteti shahar markazidan xiyla chetda joylshgan.*

2) So‘roq gap. Narsalar, voqea-hodisalar haqida so‘rash ma’nosini ifodalaydigan gaplar. Misollar: *1. Bugungi darsga kim kelmadi?(So‘roq olmoshlari yordamida: kim? nima? qayer?*

qachon? ...). 2. Kyingi darsga kelasanmi? (So'roq yuklamalari yordamida: -mi, -chi, -a, -ya). 3. Soat ikki bo'ldi. Ikki bo'ldi?

3) Buyruq gap. Davat qilish ma'nosini bildirib, kesimi buyruq shaklidagi fe'l orqali ifodalanadi. Misol: *Avval o'yla, keyin so'yla.*

4) Istak gap. Istak ma'nosini bildirib, kesimi shart (*-sa*) shaklidagi fe'l orqali ifodalanadi. Misol: *Millatning dardiga darmon bo'lsangiz.*

Gapning tuzilishiga ko'ra turlari. Grammatik asoslarning miqdoriga ko'ra gaplar ikkiga bo'linadi: 1) soda gaplar va 2) qo'shma gaplar.

1) Sodda gap (kesim=1). Grammatik asosi bitta bo'lib, ma'lum bir fikr ifodalovchi gaplar sodda gaplar deyiladi.

2) Qo'shma gap (kesim≥2). Ikki va undan ortiq sodda gaplarning birikuvidan hosil bo'lgan gap qo'shma gap hisoblanadi.

Foydalanilgan adabiyotlar:

1. M. Sharipov and O. Sobirov, 'Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language', in *CEUR Workshop Proceedings*, V. J. and K. B., Eds., CEUR-WS, 2022, pp. 154 – 159. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146112590&partnerID=40&md5=e1080c39d101c0e351cfed1a8228d391
2. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, 'A Machine Transliteration Tool Between Uzbek Alphabets', *CEUR Workshop Proc*, vol. 3315, pp. 42 – 50, 2022, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146119140&partnerID=40&md5=be670d829670d883b2f8326559ce954a
3. N. Abdurakhmonova and U. Tuliyev, 'Morphological analysis by finite state transducer for Uzbek-English machine translation/Foreign Philology: Language', *Literature, Education*, vol. 3, p. 68, 2018.
4. E. B. Boltayevich, E. Adali, K. S. Mirdjonovna, A. O. Xolmo'minovna, X. Z. Yuldashevna, and X. Nizomaddin Uktamboy O'G'li, 'The Problem of Pos Tagging and Stemming for Agglutinative Languages (Turkish, Uyghur, Uzbek Languages)', in *UBMK 2023 - Proceedings: 8th International Conference on Computer Science and Engineering*, 2023, pp. 57 – 62. http://doi.org/10.1109/UBMK59864.2023.10286792 .
5. A. M. Abdurashetona and I. O. Ismailovich, 'Methods of Tagging Part of Speech of Uzbek Language', in *Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021*, 2021, pp. 82 – 85. https://10.1109/UBMK52708.2021.9558900
6. Y.-L. Yeong, T.-P. Tan, and S. K. Mohammad, 'Using Dictionary and Lemmatizer to Improve Low Resource English-Malay Statistical Machine Translation System', *Procedia Comput Sci*, vol. 81, pp. 243–249, 2016, doi: https://doi.org/10.1016/j.procs.2016.04.056
7. R. Zhang, H. Yamamoto, and E. Sumita, 'On the Use of Lemmatization for Statistical Machine Translation'.
8. R. Zhang and E. Sumita, 'Boosting statistical machine translation by lemmatization and linear interpolation', in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 2007, pp. 181–184.