

TABIIY TILNI QAYTA ISHLASH (NLP)

O'ZBEK TILI MATNLARINI MORFOLOGIK TAHLIL QILISH UCHUN LINGVISTIK MANBA YARATISH VA LEMMALASH

Sharipov Maqsud Siddiqovich

Texnika fanlari nomzodi, dotsent,

m.sharipov@urdu.uz,

UrDU, Kompyuter ilmlari kafedrası dotsenti

Samandarova Dilfuza Shonazar qizi

dilfuzasamandarova44@gmail.com,

UrDU, Kompyuter ilmlari kafedrası magistranti

Sobirov Og'abek Ollayor o'g'li

sobirov.o@urdu.uz,

UrDU, Kompyuter ilmlari kafedrası tayanch doktoranti

Annotatsiya. Til resurslari Tabiiy Til Jarayonlari (TTJ) vazifalarida hal qiluvchi rol o'ynaydi va tadqiqotchilarga katta hajmdagi matnlarni samarali tahlil qilish va qayta ishlash imkonini beradi. Ushbu maqolada biz korpus tahlili va lemmalash uchun moslashtirilgan lingvistik manba yaratishga kompleks yondashuvni taqdim etganmiz. Biz korpus lingvistikasida lingvistik manbalarning ahamiyatini muhokama qilganmiz va korpus yaratish jarayonini tasvirlaganmiz. Bundan tashqari, biz lemmalash tushunchasini va uning lingvistik tahlildagi ahamiyatini ko'rib chiqqanmiz. Maqolada korpus tahlili uchun lingvistik manbalarni yaratish bosqichlari ko'rsatilgan va lemmalash usullari haqida tushuncha berilgan.

Annotation. Language resources play a crucial role in Natural Language Processing (NLP) tasks, enabling researchers to analyze and process large volumes of text efficiently. In this paper, we present a comprehensive approach to linguistic resource creation tailored for corpus analysis and lemmatization. We discuss the importance of linguistic resources in corpus linguistics and describe the process of corpus creation. Additionally, we delve into the concept of lemmatization and its significance in linguistic analysis. The paper outlines the steps involved in creating linguistic resources for corpus analysis and provides insights into lemmatization techniques.

Аннотация. Языковые ресурсы играют решающую роль в задачах обработки естественного языка (NLP), позволяя исследователям эффективно анализировать и обрабатывать большие объемы текста. В этой статье мы представляем комплексный подход к созданию лингвистических ресурсов, адаптированный для корпусного анализа и лемматизации. Мы обсуждаем важность лингвистических ресурсов в корпусной лингвистике и описываем процесс создания корпуса. Кроме того, мы углубляемся в понятие лемматизации и ее значение в лингвистическом анализе. В статье

описываются шаги по созданию лингвистических ресурсов для корпусного анализа, а также даются сведения о методах лемматизации.

Kalit so‘zlar. *Tabiiy Til Jarayonlari (TTJ), korpus, korpus tahlili, lingvistik manba, lemma, lemmalash, O‘zbek tili.*

Kirish

Har qanday tildagi matnlarni kompyuterda tahlil qilish, umuman lingvistik dasturlar yaratish uchun, ma’lum lingvistik manba bo‘lishi kerak. Xususan, ushbu maqolada o‘zbek tilidagi matnlarni avtomatik tahlil qilish uchun lingvistik manba yaratish va uni aynan Tabiiy Til Jarayonlari (TTJ) [Natural Language Processing (NLP)] muammolarini hal qilishga qo‘llash va lemmalash masalasi haqida alohida so‘z yuritilgan. Bunda lingvistik manba ma’lumotlar bazasi bo‘lishi mumkin, ushbu ma’lumotlar bazasi esa korpus [corpus] deb ataladi va u yozma nutqning har qanday to‘plami bo‘lishi mumkin. Bunda lingvistik tahlil dasturlari ushbu bazalardan tezkor qidirish va tahlil qilish orqali foydalanadi.

Adabiyotlar tahlili. O‘zbek kompyuter lingvistikasida [computational linguistics] lingvistik manbalardan foydalangan holda amalga oshirilgan ishlar bilan tanishib chiqar ekanmiz, quyidagi ilmiy ishlar va ularning natijalarini aytib o‘tish joiz. So‘zlar [Words] va Qo‘shimchalar [Affixes] bazalari shakllantirilgan holda qoidaga asoslangan lemmalash algoritmi ishlab chiqilgan, algoritmini qurishda Chekli Holatlar Mashinasidan (ChHM) [Finite State Machine (FSM)] foydalanilgan [Sharipov, Sobirov, 2022]. So‘z asosini topish uchun qo‘shimchalarning XML fayli yaratilib, ular o‘n beshta sinfga ajratilib, har bir sinf uchun alohida ChHM qurilgan va ular birlashtirilib asosiy yagona ChHM yaratilib u asosida stemming jarayoni amalga oshirilgan [Sharipov, Yuldashov, 2022]. Matnlarda Ot va Olmosh so‘z turkumlarini aniqlash uchun ularning tup shakldagi bazasi va ularning qo‘shimchalar bazasi shakllantirilga [Sharipov, Mavlonova, Sobirov, 2023]. O‘zbek tili fe’l so‘z turkumi uchun stemming algoritmini ishlab chiqish ChHM yordamida amalga oshirilgan [Sharipov, Salayev, Matlatipov, 2021]. Stemmingdan foydalangan holda matnlarni belgilash jarayoni amalga oshirilgan, turli sohalarga tegishli bo‘lgan ~23 000 atrofida so‘zlar testlab ko‘rilgan va 89.78% aniqlikka erishilgan [Sharipov, Kuriyozov, Yuldashov, Sobirov, 2023]. Lug‘atda mavjud va lug‘atda mavjud bo‘lmagan o‘zbek, turk va uyg‘ur tillari uchun so‘z turkumlarini belgilash va stemming bo‘yicha nazariy va amaliy ishlar amalga oshirilib, ularning tahliliy natijalari turli xil jadvallar va grafiklar bilan chop qilingan [Elov, Adali, Hamroyeva, Abdullayeva, Xusainova, 2023]. Kompyuterda TTJ muammolarini hal qilish uchun lingvistik manba tarkibida mavjud bo‘lishi kerak bo‘lgan lingvistik resurslar haqida so‘z boradi [Abjalova, Rashidova, Abdumuminov, 2022]. TF-IDF yordamida matnlarni xulosalash, bunda semantik tahlil muhim bo‘lib, n-gram metodi yordamidan foydalanilgan va yig‘ilgan “Maktab korpusi”dan foydalanilgan [Madatov, Bekchanov, Vichich, 2023].

Lingvistik manba yaratish. Morfologik tahlilning asosiy maqsadi tilni tushunish va uni qo'llashga yordam berishdir. O'zbek tilidagi morfologik tahlil, so'zlar va ularning qo'shimcha shakllarini aniqlash, funksiyasini belgilash va tarkibiy qurilishini tushuntirishga oid bo'ladi. Morfologik tahlil, tilni tushunish, o'rganish va foydalanishda muhim bir qadamdir. Bu tilning qoidalarini va tuzilmalarini tushunishda va unga oid matnlar va yozuvlardan foydalanishda yordam beradi. Lingvistik manba, tilning tadqiqotlarini, ma'lumotlarini va ilmiy o'rganishlarini taqdim etuvchi muhim manba yoki manbalardir. Bu manbalar tilshunoslikda, yozuvchilarga, tarjimanlarga va boshqa lingvistik soha bilan shug'ullanuvchi o'quvchilar, olimlar va tadqiqotchilar uchun amaliy yordamchi bo'ladi. Lingvistik manba turli xil shaklda bo'lishi mumkin: tarjima lug'atlar, izohli lug'atlar, imlo lug'atlari va hattoki qo'llanmalar ham. O'zbek tilini morfologik tahlil qilish uchun lingvistik manba sifatida biz so'zlarning tegishli belgi bilan belgilangan bazasini yaratdik. Bunda so'zlarni darsliklar, maqolalar, adabiyotlar va boshqa manbalardan oldik [Sapayev, 2009; Mirzayev, Begmatov, Madvaliyev, 2008]. Ushbu manba TTJ vazifalarida hal qiluvchi rol o'ynaydi va tadqiqotchilarga katta hajmdagi matnlarni samarali tahlil qilish va qayta ishlash imkonini beradi. Bu esa barcha so'z turkumlari doirasida, xususan ot, fe'l, sifat, son, ravish, olmosh, oraliq va yordamchi so'z turkumlaridagi so'zlarni qamrab oladi. Manba quyidagicha tuzilishga ega:

- Kitob/xon (dastlab: ot, keyin: ot)
- Qora/ymoq (dastlab: sifat, keyin: fe'l)
- Iz/chil (dastlab: ot, keyin sifat)
- O'q/dek (dastlab: ot, keyin: ravish)
- Bog'/bon (dastlab: ot, keyin: ot)
- Ikki/lanmoq (dastlab: son, keyin: fe'l)

Lemmalash. Matnlar ustida lingvistik amallar bajarishda dastlabki muhim ishlov berish usullaridan biri lemmalashdir. Ya'ni TTJ muammolarini hal qilish uchun dastlab matnlar ustida dastlabki ishlov berish usullari belgilash (teglash), tokenlash, lemmalash va stemming kabi ishlarni amalga oshirish kerak bo'ladi. Lemma tilshunoslik sohasida foydalaniladigan muhim konseptdir. Lemma - so'zning o'zgaruvchan qismi, odatda asosiy yoki standart shakli deb ataladi va lemmalar to'g'ri tarjima qilinadi. Lemmalash – so'zni o'z asos shakliga qaytarish amaliyotidir. Tilning morfologik tahlilini amalga oshirib, so'zni undagi asosiy shakliga aylantirishdir. Lemmalash natijasida olingan asosiy shakl lemma deb ataladi. Masalan, “*kitoblar*”, “*kitobni*”, “*kitoblarga*” kabi so'zlar “*kitob*” lemmasiga (asosiy shakliga) aylantirish jarayonidir.

Natija va tahlil. Ushbu lingvistik manba asosida adabiyot sohasiga tegishli matn tahlil qilib ko'rildi. Natija quidagi jadvalda berilgan [1-jadval].

№	Soʻz	Matnda uchrashlar soni	Matnda uchrash foizi	Lemmasi
1	daraxt	14	1.4%	daraxt
...				
8	uning	7	0.7%	u
...				
999	qoʻydi	1	0.1%	qoʻymoq

1-jadval. Lingvistik manba asosida adabiyot sohasiga tegishli matn tahlili
Matn quyidagicha boshlangan “*Hovlimizdan oqib oʻtadigan ariq labidagi bu daraxt qishlogʻimizda yagona hisoblanardi. ...*”. Ushbu matnni biz yaratgan lingvistik manba yordamida tahlil qilish uchun dastur ishlab chiqib, aynan soʻzlar lemmasini topish uchun qoʻlladik. Lingvistik manbadan foydalangan holda TTJlarining boshqa muammolarini yechishda ham foydalanish mumkin.

Xulosa

Xulosa oʻrnida shuni aytish joizki, lingvistik muammolarni hal qilishda, TTJ vazifalarini bajarishda, tarjima dasturlarini ishlab chiqishda lingvistik manbaning oʻrni katta. Ushbu maqolada aynan lingvistik manba yaratishning nazariy va amaliy maʼlumotlari berib oʻtildi va ushbu maʼlumotlar asosida lingvistik manba yaratilib, u asosida matn tahlil qilib koʻrildi, ushbu tahlil natijalari jadval asosida eʼlon qilingan. Kelajakda ushbu manbani mukammallashtirib, uni turli xil janrdagi matnlarni atroflicha tahlil qilish uchun asosiy manbaga aylantirmoqchimiz.

Foydalanilgan adabiyotlar:

1. Sharipov M, Sobirov O. Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language. CEUR Workshop Proceedings, CEUR-WS 2022, 154 – 159.
2. Sharipov M, Yuldashov O. UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language. CEUR Workshop Proceedings, CEUR-WS 2022, 137 – 144.
3. Sharipov M, Mavlonova M, Sobirov O. Oʻzbek tilidagi matnlarda ot va olmosh soʻz turkumlariga tegishli soʻzlarni aniqlovchi algoritmi yaratish. Computer Linguistics: Problems, Solutions, Prospects 2023; 1.
4. Sharipov M, Salaev U, Matlatipov G. Implemented stemming algorithms based on finite state machine for uzbek verbs| computer linguistics: problems. solutions, prospects 2021;
5. Sharipov M, Kuriyozov E, Yuldashev O, Sobirov O. UzbekTagger: The rule-based POS tagger for Uzbek language. arXiv preprint arXiv:230112711 2023;
6. Boltayevich EB, Adali E, Mirdjonovna KS, Xolmoʻminovna AO, Yuldashevna XZ, Nizomaddin Uktamboy OʻGʻli X. The Problem of Pos Tagging



and Stemming for Agglutinative Languages (Turkish, Uyghur, Uzbek Languages). UBMK 2023 - Proceedings: 8th International Conference on Computer Science and Engineering, 2023, 57 – 62.

7. Abjalova M, Rashidova U, Abdumuminov B. Matnlarni qayta ishlash manbalari. computer linguistics: problems, solutions, prospects 2022; 1.

8. Madatov K, Bekchanov S, Vičič J. Uzbek text summarization based on TF-IDF. arXiv preprint arXiv:230300461 2023;

9. Sapayev Q. Hozirgi O'zbek tili (morfemika, so'z yasalishi va morfologiya). Toshkent: Nizomiy nomidagi TDPU, 2009.

10. Mirzayev T, Begmatov E, Madvaliyev A *et al.* O'zbek tilining izohli lug'ati. Toshkent: 'O'zbekiston milliy ensiklopediyasi' Davlat ilmiy nashriyoti, 2008.