MATNNI QAYTA ISHLASH VA TAHLIL QILISH USULLARI

Abduraxmonova Umida Rustamovna
uabdurahmanova06@gmail.com
Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti,
Axborot texnologiyalari kafedrasi o‘qituvchisi,
Toshkent axborot texnologiyalari universiteti mustaqil izlanuvchisi

Annotatsiya. Ushbu maqolada tabiiy tilni qayta ishlashning asosiy muammolari muhokama qilinadi. Qayta ishlashning asosiy yo'nalishlari, usullari, hozirda mavjud vositalar va kutubxonalarining asosiy yo'nalishlarini tahlil qiladi. Tabiiy tilda matnlarni qayta ishlash va tahlil qilish bo’yicha ikki-uchta usul ko’rib chiqildi va tahlil qilindi. Matn ma'lumotlarini tahlil qilish zamonaviy dunyo uchun juda muhimdir.

Kalit so'zlar: *mashina tarjmasi, fonetik tamoyil, simantik, sintaktik, morfologik tahlil, grammatik leksik tahlil.*

TEXT PROCESSING AND ANALYSIS METHODS

Annotation. This article discusses the main problems of natural language processing. Analyzes the main directions, methods of processing, the main directions of currently available tools and libraries. Two or three methods for processing and analyzing texts in natural language were considered and analyzed. Analysis of text data is very important for the modern world.

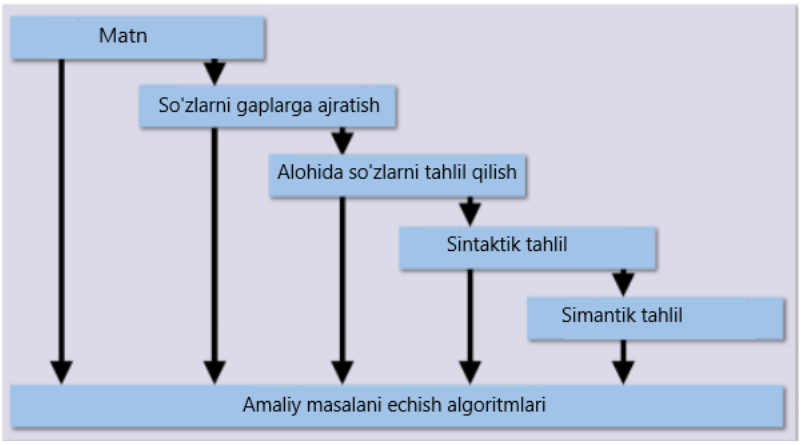
Keywords: *machine translation, phonetic principle, semantic, syntactic, morphological analysis, grammatical lexical analysis.*

O'tgan o'n yil ichida, tabiiy tilda ishlash va so'z sohasiga bo’lgan qiziqish tobora o'sishi kuzatilmoqda, ammo ijtimoiy tarmoqlarda matnlarni qayta ishlashga oid monitoring yangi dasturlari paydo bo'lsa-da, ko'p muammolar hali ham yechimini topmasdan qolmoqda. Bularning barchasi bizni matnni qayta ishlash va tahlil qilish tizimlariga bo’lgan qiziqishimizni va yangidan yangi ihtirolar qilishga undaydi.

Tabiiy tilda matnlarni qayta ishlash muammolari o’tgan asr davomida mutaxassislarni qiziqtirgan [1], usha davr ichida axborot olish muammolarini hal qilish usullari ishlab chiqilgan [2], mashina tarjimasi va hk. Shu bilan asosiy tamoyillar yani axborot olish muammolari kompyuter lingvistikasi bilan bog'liq bo’lgan, lekin ko'plab yangi muammolar (masalan, hujjatlardagi kalit so'zlarni

ajratib ko'rsatish yoki ijtimoiy tarmoqlarni kuzatish) uning tili "kanonik" dan juda farq qilishi mumkinligi va uni qo'llamasdan hal qilish mumkinligi ko’zda tutulgan.

So'zlarni qayta ishlash usullarini qayta ko'rib chiqishni talab qiladigan vazifalar orasida fikrlarni chiqarib olish, matnlarga rang berish, axborot manbalarining asl mazmunini tahlil qilish (masalan, "taniqli blogger" fikri aslida uning xonadoshlarini qiziqtirishi mumkin) noto'g'ri yoki ataylab buzilgan matnlarni qayta ishlashi mumkin. Bu vazifalarning barchasini yani matnlarni qayta ishlashda har doim kompyuter lingvistikasi metodlarisiz hal qilolimasligimizni eslatib o’tamiz. Barcha zamonaviy matnni qayta ishlashda tizim so'zlarni eng oddiy qidirishdan boshlab mashina tarjimasi bilan tugallaydigan, bir necha bosqichlarni ko'zdan kechiradigan va o'zgarmas tabiiy tilni tanlaydigan algoritmlari mavjud. (1-rasm)



1-rasm. So'zlarni klassik ravishda bosqichma-bosqich qayta ishlash.

Tizim kirish joyida belgilar ketma-ketligini oladi va birinchi bosqichda (leksik tahlil) u alohida so'zlar va jumlalarga ajratadi. Shu bilan birga, ba'zi belgilar ketma-ketligi (masalan, rus tilidagi chiziqcha va nuqta) bir ma'noda talqin qilinishi mumkin. Bundan tashqari, leksik tahlil bosqichida deobfuskatsiya vazifasi paydo bo'ladi - ataylab buzilgan (xiralashgan) so'zlarni aniqlab va tuzatib ketadi. Bunday buzilishlarning odatiy misoli bu so'zni almashtirishdir misol tariqasida ingliz tilidagi «drugs» yani (giyohvand moddalar) so’zini spam-jo'natmalarda "d.r.u.g.s" yoki "d-r-u-g-s" ga almashtirib ko’rsatilishidir.

Keyingi bosqichda alohida so'zlarni qayta ishlash amalga oshiriladi, bu ko'pincha morfologik tahlilga to'g'ri keladi yani so'z (gramm) va asosiy so'z shaklining xususiyatlarini aniqlaydi.

Morfologik tahlilni o'tkazishda ikkita yondashuv mavjud. Birinchisi (aniq

morfologiya) har bir so'zning xususiyatlarini o'z ichiga olgan holda bitta katta lug'at qurilishini nazarda tutadi, masalan rus tili uchun bunday lug'at A.A. asosida tuzilgan. Misol uchun rus tilidagi Zaliznyak grammatik lug'atida 8 milliondan ortiq so'zlar mavjud. Ushbu yondashuvni amalga oshirish nisbatan sodda, ammo u ikkita muhim kamchiliklarga ega. Birinchidan, tizim faqat lug'atdagi so'zlarni tahlil qiladi. Ikkinchida, ko'plab tillarda ushbu so'z boyligi juda katta bo'ladi.

So'zlarni tahlil qilishning muqobil yondashuvi (noaniq morfologiya) qoidalar tizimidan foydalanishdir, unga ko'ra ma'lum bir so'z uchun uning xususiyatlari taxmin qilinadi. Ushbu yondashuvning kamchiligi shundaki, u har doim ham natijalarning 100% aniqligini kafolatlay olmaydi.

To'liq matnli izlashda va matnlarni tasniflash vazifalarida so'zlarni to'liq morfologik tahlilini o'tkazish talab qilinmaydi, faqat ko'rsatilgan ikkita so'z aslida bir so'zning shakllari ekanligini tekshirib chiqadi. Bu asosiy so'z shakliga qisqartirish yoki so'zlarning ba'zi o'zgarmas qismlarini ajratib ko'rsatishdan iborat bo'lgan stemming yordamida amalga oshirish mumkin. Biroq, morfologik tahlil, lemmatizatsiya va stemming har doim ham "xavfsizlik" va "himoya" kabi bir-biriga bog'liq so'zlarni aniqlay olmaydi. Bog'liq so'zlarni aniqlashda maxsus tezaurus lug'atlari yordamida hal etiladi. Ikki so'zning yaqinlik grafasini ikkita mos keladigan birlashtirma eng qisqa yo'l asosida aniqlanadi. Agar so'zlarning kontekstini hisobga olish zarur bo'lsa, unda vazifa ancha murakkablashadi va uni semantik matnni qayta ishlashga kiritish kerak bo’ladi. So'zlarning bog'lanishini aniqlashning avtomatlashtirilgan usullari mavjud bo’lib, birgalikda yuzaga kelish chastotasi yoki ulardan foydalanish sharoitlarining tasodifiylik darajasiga asoslanadi. Tekshirish paytida manba matnidagi so'zlar ketma-ketligini daraxtga o'xshash ierarxiyaga aylantiradi, bunda barglar alohida so'zlarga, tugunlar so'zlar guruhiga, yoylar so'zlar va so'zlar guruhlari o'rtasidagi munosabatlarga mos keladi. Ushbu o'zgartirish tilning ma'lum bir grammatikasi asosida amalga oshiriladi, bu asosan qat'iy qoidalar to'plami hisoblanadi. Grammatikalardan foydalanish aniq qiyinchiliklar bilan bog'liq bo’lib- tabiiy til uchun uni tavsiflovchi qoidalar tizimini ishlab chiqish va qiyinchilik tug’diradigan ayniqsa murakkab morfologik model va o'zboshimchalik bilan so'z tartibiga ega bo'lgani (masalan, rus tili) uchun qiyin ko’rinadi. Bundan tashqari, inson tomonidan yozilgan matnlarning aksariyat qismida xato yoki tipografik xatoliklar mavjud bo’ladi. Shu sababli har qanday grammatika qo'llanilmasligi mumkin yani urinishlar va xatolarning barcha mumkin bo'lgan variantlarini hisobga olishga natija bermaydi.

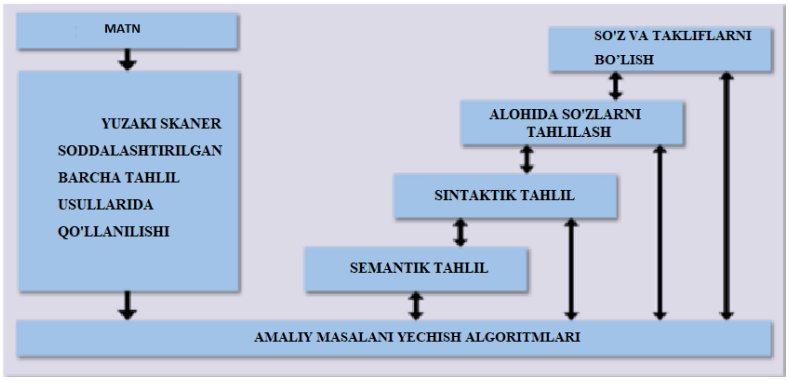
Rus tilidagi matnni tahlil qilish tizimlarining aksariyati turli xil

grammatikalardan foydalanishni o'z ichiga olgan yondashuvlarga asoslanadi. Eng qiziqarli natijalar - Yandex "Tomita-parser" (axborotni chiqarib olish), [3] Abbyy Compreno texnologiyasining ajraluvchisi va ETAP-3 tizimining modullaridir (mashina tarjimasi).

Mahalliy tezaurus har qanday hujjatning mazmunini rasmiy ravishda aniqlashga qodir, ammo bunday taqdimot keyingi ishlov berish uchun juda noqulay bo'lishi mumkin, chunki bir xil fakt turli xil yo'llar bilan ifodalanishi mumkin. Mahalliy tezurus har qanday hujjat tarkibini o'rnatishga qodir, ammo bunday vakil keyingi ishlov berish uchun juda noqulay bo'lishi mumkin, chunki bir xil ko'pgina turli xil usullarni taqdim etadi. Rasmiy ravishda va har qanday ma'noga ega bo'lgan yagona va aniqlik bilan bog'liq bo'lishi kerak. U bilimlarni taqdim etishning har qanday usuliga, shu jumladan semantik tarmoqlardan foydalanishga asoslanishi mumkin. Bunday holda, sizda bunday tarmoqlar turlari va bunday tarmoqlar va ma'lumotlarni mahalliy tezurusdan semantik tarmog'iga aylantirish qoidalarining batafsil tavsifini olishimiz kerak. Ushbu vazifalarni hal qilishga urinishlar allaqachon ilgari qilingan, ammo muvaffaqiyatga erishilmagan, ammo bu vazifalar Abbyy Compreno texnologiyasida hal qilinganligi haqida xabar berib o’tilgan.

Matnni qayta ishlash bir necha bosqichlarda sodir bo'ladi, bir bosqichning chiqishi esa keyingi bosqich yoki boshqa bosqichni kiritish uchun mo'ljallangan modullar, masalan, mashina tarjimasi kabi ma'lum bir bosqichni kiritish uchun mo'ljallangan. Har qanday ishining natijalari noaniq - bir xil kirish ma'lumotlari bir nechta mumkin bo'lgan natijalar va aksincha, bir xil natijani to'liq turli xil ma'lumotlardan olish mumkin. Bu bosqichga o'tish paytida qayta ishlangan ma'lumotlar keskin oshishi mumkinligiga olib keladi.

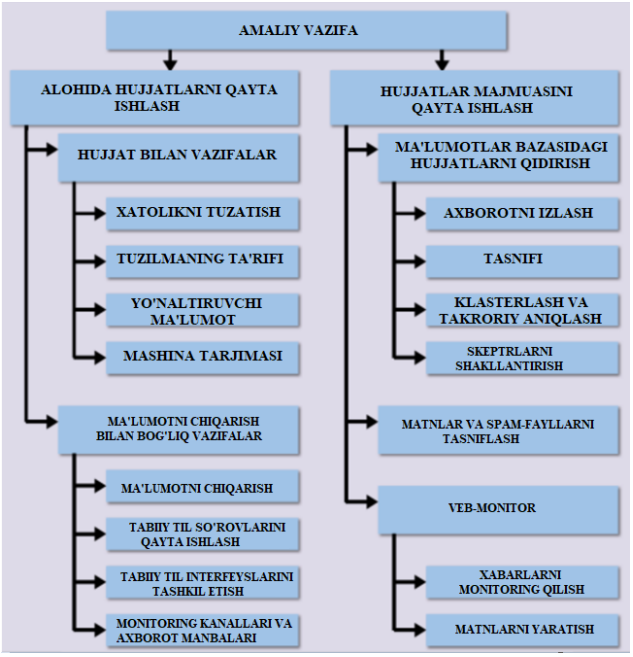
Ko'p bosqichli matnni qayta ishlashning an'anaviy diagrammasi qayta ko'rib chiqishni talab qiladi matnni tahlil qilish ikki bosqichda bo'lishi mumkin (2-rasm): Yuqori modullar kerakli modullar kerakli aniqlik ma'lumotlariga tegishli bo'lgan asosiy modullarga tegishli bo'lgan asosiy modullar ko'rsatilgan holda, yuzaga keladigan asosiy ma’lumotni batafsil tahlil qiladi. Masalan, vazifani yuzaki ko'rib chiqish jarayonida ma'lumot olish vazifasi aniqlanadi.



2-rasm. Matnlarni ikkifazada qayta ishlash.

Yuqorida aytib o'tilganidek, modullar tizimning haqiqiy kirish ma'lumotlarida sinovdan o'tkazilishi kerak va har qanday muammolarni hal qilish uchun mos bo'lgan universal tahlil modullarini tayyorlashga urinishlar bir xil muammoga duch keladi - modullar bitta muammo uchun optimallashtirilgan bo'lib, boshqalar uchun maqbul emas, va agar matn tahlili yomon bajarilgan bo'lsa, unda butun tizimning sifati past bo'ladi. Ikki fazali matnni qayta ishlash bilan tizim qaysi modullardan foydalanish zarurligiga va matnni lingvistik tahlil qilish qanchalik samarali ekanligiga ta'sir o'tkaza oladi.

Amaliy masalalarni echishda ularni ikkita katta guruhga bo'lish muhim (3-rasm): individual hujjatlarni qayta ishlash va ularning massivlarini qayta ishlash.



3-rasm. Amaliy vazifalarni tasniflash.

Individual hujjatlarni qayta ishlash bo'yicha topshiriqlar guruhini ikkita kichik

guruhga bo'lish kerak: hujjatlarni to'g'rilash va ma'lumot olish. Birinchisi, kirish va chiqishda matnli hujjat bo'lishini nazarda tutadi (xatolarni tuzatish, matnni tuzatish, uning tuzilishini aniqlash, umumlashtirish, mashinaga tarjima qilish vazifalari). Ikkinchi kichik guruhga rasmiy ravishda taqdim etilgan ma'noni qayta ishlash bilan bog'liq vazifalar kiradi: faktlarni yig'ish, tabiiy tilda so'rovlarni bajarish, tabiiy til interfeyslarini tashkil etish va to'g'ri matnlarni yaratish. Birinchi kichik guruhning barcha vazifalarini amalga oshirish yoki mutaxassislar tomonidan tuzilgan qoidalarga yoki mashinada o'qitish usullarini qo'llash natijasida olingan namunalarga asoslanishi mumkin. Qoidalar tizimidan foydalanish aniqroq va bashorat qilinadigan natijalarni berishi mumkin, ammo uni yaratish uchun yuqori xarajatlarni nazarda tutadi. O'z navbatida, mashinada o'qitish usullarini qo'llash unchalik mashaqqatli emas, balki juda ko'p sonli yuqori sifatli misollarni talab qiladi. Aytaylik, siz boshqa ko'plab Evropa tillari va ingliz tilidagi matnlarini tarjimalarni qilishimiz mumkin, ammo nodir tillarga tarjima namunalarini topish qiyin. Alohida hujjatlarni qayta ishlash bo'yicha vazifalarning ikkinchi kichik guruhiga ma'lumot olish, tabiiy tilda so'rovlarni bajarish, matnlarni yaratish va tabiiy til interfeyslarini tashkil etish kiradi. Ushbu vazifalarning barchasi matnlarni "tushunish" va ko'rsatilgan faktlarni topishni o'z ichiga oladi. Ish ma'lumotlar yig'ish qoidalari tizimi asosida amalga oshiriladi, ularning har biri sintaktik tuzilish shablonini va rasmiylashtirilgan axborot taqdimotining yaratilgan bo'lagi uchun shablonni belgilaydi. Hujjatni qayta ishlashda sintaktik tahlil natijalari ko'rib chiqiladi va parchalari qidirilmoqda, uning tuzilmasi informatsiyadan chiqarib olish qoidalaridan andozalarga mos keladi. Keyingi, "Ishladi" qoidalariga muvofiq so'zlarning bir qismi matndan olinadi va rasmiy tuzilishga aylantiriladi. Matnni "tushunish" vazifasidan asosiy farq shundaki, u belgilangan predmet sohasidagi ma'lumotlar bilan ishlaydi, buning uchun ma'lumotlar kontseptual modeli va ajratib olish qoidalari aniq belgilangan. Tabiiy til interfeyslarini tashkil qilish IBM Watson tizimida birdaniga echilgan uchta vazifaning kombinatsiyasi sifatida qaralishi mumkin: ma'lumot olish, bilim bazasi darajasida javob izlash va matn yaratish. Ushbu texnologiyalar qanday talabga javob berishini aytish hali ham qiyin, ammo tizimdan tibbiy diagnostika uchun foydalanish rejalashtirilgan, ammo tabiiy til interfeysi ekspert tizimlari bilan o'zaro aloqaning boshqa usullaridan ko'ra qulayroq bo'lishi aniq emas.

Foydalanilgan adabiyotlar:

1. Rassel S., Norvig P. Sun'iy intellekt. Zamonaviy yondashuv. M.: Uilyams,

2007. - 1480-yillar.

2. Manning K., Raghavan P., Schütze H. Axborot olish uchun kirish. M.: Uilyams, 2011. - 528 p.

3. Toldova S.Yu. Avtomatik matnni tahlil qilish usullarini baholash 2011–2012: rus tilining sintaktik tahlilchilari // Dialog-2012: konferentsiya referatlari. Moskva, 2012 yil.