

AVTOMATIK LINGVISTIK VA EKSTRALINGVISTIK TEGlash VOSITALARI UMUMIY TAVSIFI

UDK:81:33

Yuldashev Aziz Uyg'un o'g'li
yuldashevaziz@navoiy-uni.uz
ToshDO'TAU o'qituvchisi

Annotatsiya. Ushbu maqolada avtomatik lingvistik va ekstralingvistik teglash vositalari umumiy tavsifi keltiriladi. Mystem, pymorphy2 kabi avtomatik morfologik teglash vositalari; Stanford CoreNLP, spaCy avtomatik sintaktik teglash vositalari; WordNet, ruWordNet kabi avtomatik semantik teglash vositalari tavsiflanadi. Korpus tuzish, interfeys va qidiruv tizimidan foydalanish uchun ham tayyor dasturiy vositalar bo'lib, Sketch Engine, AntConc kabi vositalarning umumiy xususiyatlari, imkoniyatlari, afzalliklari tahlil qilinadi.

Kalit so'zlar: avtomatik lingvistik teglash, avtomatik ekstralingvistik teglash, Mystem, pymorphy2, Stanford CoreNLP, spaCy, WordNet, ruWordNet, Sketch Engine, AntConc.

Автоматические лингвистические и экстралингвистические средства тегирования: общий обзор

Абстракт. В данной статье дается общее описание инструментов автоматической лингвистической и экстралингвистической маркировки. Инструменты автоматического морфологического тегирования, такие как Mystem, pymorphy2; Stanford CoreNLP, инструменты автоматической синтаксической разметки spaCy; Описаны инструменты автоматической семантической разметки, такие как WordNet и ruWordNet. Представлены готовые программные инструменты для создания корпусов, интерфейс и использование поисковых систем, а также проанализированы общие характеристики, возможности и преимущества таких инструментов, как Sketch Engine и AntConc.

Ключевые слова: автоматическая лингвистическая разметка, автоматическая экстралингвистическая разметка, Mystem, pymorphy2, Stanford CoreNLP, spaCy, WordNet, ruWordNet, Sketch Engine, AntConc.

Automatic Linguistic and Extralinguistic Tagging Tools: A General Overview

Abstract. This article provides a general description of automatic linguistic and extralinguistic tagging tools. Automatic morphological tagging tools such as Mystem, pymorphy2; Stanford CoreNLP, automatic syntactic markup tools spaCy;

Automatic semantic markup tools such as WordNet and ruWordNet are described. Ready-made software tools for creating corpora, interface and use of search engines are presented, and general characteristics, capabilities and advantages of such tools as Sketch Engine and AntConc are analyzed.

Keywords: automatic linguistic markup, automatic extralinguistic markup, Mystem, pymorphy2, Stanford CoreNLP, spaCy, WordNet, ruWordNet, Sketch Engine, AntConc.

KIRISH

Til korpuslarini teglashda Mystem, pymorphy2 kabi avtomatik morfologik teglash vositalari; Stanford CoreNLP, spaCy avtomatik sintaktik teglash vositalari; WordNet, ruWordNet kabi avtomatik semantik teglash vositalari ahamiyatli va zaruriy vositalar sanaladi.

Shuningdek, korpus tuzish, interfeys va qidiruv tizimidan foydalanish uchun ham tayyor dasturiy vositalar bo'lib, Sketch Engine, AntConc kabi vositalarning umumiy xususiyatlari, imkoniyatlari, afzalliklari tahlil qilinadi.

ASOSIY QISM

Lingvistik teglash tabiiy tilni qayta ishlashda muhim bosqich bo'lib, bu jarayonda ko'plab yordamchi vositalar mavjud. Quyida shunday dasturiy vositalarning tavsifi keltiriladi.

1. Avtomatik morfologik teglash vositalari:

Mystem: Yandeks tomonidan ishlab chiqilgan, Ingliz va rus tilining morfologik tahlilini taqdim etadi. Tez va aniq ishlaydi.

pymorphy2: morfologik tahlil uchun Python kutubxonasi. Rus, ukrain va boshqa tillarini qo'llab-quvvatlaydi. Moslashuvchan va foydalanish oson.

2. Avtomatik sintaktik teglash vositalari:

Stanford CoreNLP: Stanford universitetining tabiiy tilni qayta ishlash uchun vositalar to'plami. Sintaktik tahlil (parsing) dasturi bo'lib, nomlangan obyektlarni aniqlash (Named Entity Recognition) va boshqa vositalarni o'z ichiga oladi, ko'p tillarni qo'llab-quvvatlaydi.

spaCy: tabiiy tilni qayta ishlash uchun ilg'or Python kutubxonasi. Tez va samarali ishlaydi, sintaktik tahlil (parsing), nomlangan obyektlarni aniqlash (Named Entity Recognition) va boshqa funksiyalarni taqdim etadi.

3. Avtomatik semantik teglash vositalari:

WordNet: Ingliz tilining leksik ma'lumotlar bazasi. So'zlarni ularning ma'nolari va munosabatlari bo'yicha guruhlaydi. Semantik tahlil va matnni qayta ishlashda ishlatiladi.

ruWordNet: WordNet ma'lumotlar bazasining ruscha versiyasi.

Mazku ma'lumotlar bir qancha tillarni avtomatik morfologik, sintaktik va semantik tahlil qilsa-da, o'zbek tiliga samarali qo'llab bo'lmaydi, chunki o'zbek tili morfologik va sintaktik xususiyatlariga ko'ra ingliz va rus tillaridan tubdan farq qiladi. Shu sababli o'zbek tili birliklarini sifatli avtomatik tahlil qilish uchun dasturlarga ehtiyoj mavjud.

Korpus tuzish, interfeys va qidiruv tizimidan foydalanish uchun ham tayyor dasturiy vositalar mavjud:

Rus tili milliy korpusi (NKRYA): Tilshunoslik belgilari bilan rus tili matnlarining ulkan korpusi. Matnni qidirish va tahlil qilish uchun kuchli vositalarni taqdim etadi. Tilshunoslik belgilarining turli darajalarini ko'rish imkonini beradi.

Sketch Engine: Matn korpuslari bilan ishlash uchun onlayn vosita. Qidirish, tahlil qilish va ma'lumotlarni vizuallashtirish funksiyalarini taqdim etadi.

AntConc. AntConc – matn korpuslarini tahlil qilish uchun bepul vosita. U lingvistik tadqiqotlarda, ayniqsa, korpus lingvistikasida so'zlarning chastotasi, kollokatsiyalari, konkordanslar va boshqa til hodisalarini tahlil qilishda keng qo'llaniladi. AntConc-ning asosiy xususiyatlari va qo'llanish sohalari quyidagicha ko'rsatish mumkin:

Asosiy funksiyalari:

Konkordans: korpusdagi muayyan so'z yoki iboralarni qidirish, ularni kontekstda ko'rsatish imkonini beradi.

Chastotali tahlil: korpusdagi so'zlarning chastotasini ko'rsatadigan ro'yxatni taqdim etadi.

Kollokatsiyalar: byerilgan so'z bilan tez-tez uchraydigan so'zlarni ko'rsatadi.

Klasterlar: korpusdagi takrorlanuvchi so'zlar ketma-ketligini aniqlaydi.

N-grammalar: n-so'zlar ketma-ketligini tahlil qiladi.

Asosiy so'zlar: turli korpuslardagi so'zlarning chastotasini solishtirish va asosiy so'zlarni aniqlash imkonini beradi.

AntConc dasturining qo'llanish sohalari sifatida quyidagilarni ko'rsatish mumkin, shuningdek, mazkur ro'yxat uning ahamiyatini ham ko'rsatadi:

Lingvistik tadqiqotlar: lyeksika, grammatika va uslubni tahlil qilish, kollokatsiyalar va frazeologik birliklarni o'rganish, diskursni tadqiq qilishda foydalaniladi.

Xorijiy tillarni o'rganish: horijiy til ta'limida o'quv qo'llanmalari va topshiriqlarni ishlab chiqishda til materialini tahlil qilish, o'rganilayotgan tildagi so'z va kollokatsiyalarning chastotasini kuzatish, tarjimada yordam berishga qaratilgan.

Matnlarni tahlil qilish: matnlardagi asosiy mavzular va tushunchalarni aniqlash, muallif uslubini tahlil qilish, katta hajmdagi matn ma'lumotlarini tahlil qilishda qo'l keladi.

AntConc dasturining afzalliklari sifatida bepul va foydalanish uchun qulayligi, sodda va intuitiv tushunarli interfeysga egaligi, matn ma'lumotlarini tahlil qilish uchun kuchli vosita ekanligi, Windows, Mac va Linuxda ishlay olishini

ta'kidlash zarur. AntConc matn ma'lumotlari bilan ishlaydiganlar, ayniqsa, lingvistika va chet tillarni o'rganish sohasidagilar uchun juda foydali vositadir.

Avtomatik lingvistik teglash vositani tanlashda til, teglash turi va ishlash talablarini hisobga olish lozim. Vositalardan maksimal darajada samarali foydalanish uchun hujjatlarni va foydalanish imkoniyatlarini, yo'riqnomani o'rganish talab etiladi. Yaxshi natijalarga erishish uchun avtomatik va qo'lda belgilash usullarining kombinatsiyasidan foydalanish mumkin.

Turli janrdagi matnlarni avtomatik ekstralingvistik (nolisoniy omillarga ko'ra) tegash murakkab, lekin amalga oshirish mumkin bo'lgan vazifadir. Bunda quyidagi asosiy jihatlarni hisobga olish kerak:

Avtomatlashtirish imkoniyatlari:

1. Metama'lumotlar: ko'p matnlar allaqachon avtomatik ravishda olib tashlanishi mumkin bo'lgan metama'lumotlarni o'z ichiga oladi. Masalan:

Veb-sahifalar: nashr etilgan sana, muallif(lar), sarlavha. Yangiliklar maqolalari: manbalar, kategoriyalar, vaqt belgilari.

Ilmiy maqolalar: muallif(lar), jurnallar nom(lar)i, nashr etilgan sanalar.

Veb-skreypping va matnni qayta ishlash vositalari ushbu ma'lumotni avtomatik ravishda olib tashlash uchun ishlatilishi mumkin.

2. Janrni aniqlash. Mashina o'rganish modellari matnning lingvistik xususiyatlari va metama'lumotlari asosida matn janrini avtomatik ravishda aniqlash uchun o'rgatilishi mumkin. Matn janrini yetarli darajada aniqlik bilan topa oladigan matn tasnifi algoritmlari mavjud.

3. Bo'yoqdorlik va emotsionallikni tahlil qilish. Zamonaviy NLP vositalari matnning bo'yoqdorgini, shuningdek, uning emotsional xususiyatini avtomatik ravishda aniqlashga imkon beradi. Bu muallifning matnga yoki hikoya mavzusiga bo'lgan munosabati haqida xulosa chiqarishga imkon beradi.

4. Nomlangan obyektlarni aniqlash. Nomlangan obyektlarni aniqlash vositalari odamlar, joylar, tashkilotlar va ekstralingvistik teglash uchun ahamiyatli bo'lgan boshqa obyektlar haqidagi ma'lumotlarni avtomatik ravishda olib tashlashi mumkin.

5. Ijtimoiy tarmoqlarni tahlil qilish. Ijtimoiy tarmoqlar ma'lumotlari foydalanuvchilar, ularning qiziqishlari va ijtimoiy munosabatlari haqidagi ma'lumotlarni olish uchun avtomatik ravishda tahlil qilinishi mumkin.

Matnning ekstralingvistik xususiyatlarini avtomatik tahlil qilish va aniqlashning ayrim qiyinchilik va cheklovlari mavjud bo'lib, ular quyidagilar:

Janrlar xilma-xilligi. Turli janrdagi matnlar turli xususiyatlarga va ekstralingvistik teglash talablariga ega. Barcha janrlarni qayta ishlay oladigan universal vositani yaratish murakkab vazifadir.

Metama'lumotlarning yetishmasligi. Hamma matnlar ham to'liq va aniq metama'lumotlarni o'z ichiga olmaydi. Ba'zi hollarda metama'lumotlar mavjud bo'lmasligi yoki noto'g'ri bo'lishi mumkin.

Kontekst va talqin. Ekstralingvistik tellashning ba'zi jihatlari kontekstni chuqur tushunishni va talqin qilishni talab qiladi. Avtomatik tizimlar bunday vazifalarda qiyinchiliklarga duch kelishi mumkin.

Subyektivlik. Ekstralingvistik teglashning bir qismi subyektiv bo'lishi mumkin. Masalan, hissiyo bo'yoqdorlik va sentiment tahlil bunga misol bo'ladi.

Kuzatishlardan ma'lum bo'ldiki, ekstralingvistik teglash avtomatlashtirish usullari va yondashuvlari quyidagicha tasniflanadi:

1. Gibrid tizimlardan foydalanish, ya'ni avtomatik va qo'lda teglash usullaridan aralash holda foydalanish eng samarali usul bo'lishi mumkin. Avtomatik tizimlar asosiy ma'lumotlarni olish uchun ishlatilishi mumkin, keyin esa ekspertlar natijalarni tekshirishi va tuzatishi mumkin.

2. Mashinali o'rganish (Machine learning) yondashuvidan foydalanish, ya'ni ekstralingvistik ma'lumotlarni avtomatik ravishda olish va tasniflash uchun mashinali o'rganish modellaridan foydalanish bugungi kunda samalari natijalar berayotgani kuzatiladi.

3. Ma'lumotlar bazalari va ontologiyalarni qo'llash, ya'ni ekstralingvistik ma'lumotlarni saqlash va tashkil qilishda ma'lumotlar bazalari va ontologiyalardan foydalanish mumkin.

XULOSA

Xulosa sifatida aytish mumkinki, turli janrdagi matnlarni to'liq avtomatik ekstralingvistik teglash murakkab, lekin rivojlanayotgan sohadir. Zamonaviy vositalar va usullar bu jarayonning ba'zi jihatlarini avtomatlashtirishga imkon beradi, lekin yuqori aniqlik va to'liqlikka erishish uchun ko'pincha avtomatik va qo'lda teglash usullari kombinatsiyasidan foydalanish talab qilinadi.

Foydalanilgan manbalar:

1. <https://yandex.ru/dev/mystem/>
2. <https://pymorphy2.readthedocs.io/>
3. <https://github.com/stanfordnlp/CoreNLP>
4. <https://spacy.io/>
5. <https://wordnet.princeton.edu/>
6. <https://ruwordnet.ru/>
7. <https://ruscorpora.ru/>
8. <https://www.sketchengine.eu/>
9. <https://www.laurenceanthony.net/software/antconc/>