KORPUS ASOSIDA TERMINLARNI AVTOMATIK AJRATIB OLISH METODOLOGIYASI

Asqarova Madina Boymurod qizi
O'zbekiston Davlat Jahon Tillari Universiteti
Tayanch doktoranti
https://doi.org/10.5281/zenodo.17830015

Annotatsiya: Maqola parallel korpuslardan terminologik birliklarni avtomatik ajratib olish metodologiyasini o'rganadi. Tadqiqotda statistik, lingvistik va gibrid yondashuvlar tahlil qilinadi hamda 4,400 gap va 1,468 termindan iborat ingliz-o'zbek korpusi asosida Sketch Engine vositasining imkoniyatlari ko'rsatiladi. Natijalar gibrid yondashuv (precision=89.7%, recall=86.2%) terminologiya ajratishda eng samarali ekanligini tasdiqlaydi.

Kalit so'zlar: korpus lingvistikasi, avtomatik terminologiya ajratish, parallel korpus, Sketch Engine, terminologik ekvivalentlik.

Kirish: Zamonaviy korpus lingvistikasida terminlarni qo'lda aniqlash juda ko'p vaqt talab etadi, shu sababli avtomatik terminologiya ajratish (ATE) usullari keng qo'llanilmoqda (Frantzi et al., 2000). Parallel korpuslar tarjima tadqiqotlari va kontrast lingvistika uchun asos bo'lib xizmat qiladi (Hunston, 2002). Ingliz-o'zbek parallel korpuslarida lingvistik terminlarning semantik ekvivalentligini aniqlash uchun avtomatik usullardan foydalanish zarurati tobora oshmoqda. Ushbu tadqiqot 4,400 gap juftligi va 1,468 terminologik juftlikdan iborat korpus asosida olib borildi. Terminologiya ajratishning nazariy asoslari korpus va kompyuter lingvistikasining kesishmasida joylashgan. Cabré (1999) terminlarni umumiy leksikadan ajratuvchi asosiy xususiyatlarni ta'kidlab o'tdi: mutaxassislik, mazmuniy aniqlik va kontekstga bog'liqlik. Kageura va Umino (1996) metodlarni uchta guruhga ajratdilar: statistic, lingvistik, va gibrid yondashuvlar. Frantzi et al. (2000) ishlab chiqqan C-value va NC-value metodlari ko'p so'zli terminlarni aniqlashda samarali ekanligini isbotladi.

Zamonaviy korpus vositalari orasida Sketch Engine eng keng tarqalgan bo'lib, word sketch va kollokatsiya tahlilini taqdim etadi. Parallel korpuslar uchun Tiedemann (2011) alignment-based extraction usulini ishlab chiqdi. Machine learning yondashuvlari terminologiya ajratishda yuqori natijalar ko'rsatmoqda (Zhang et al., 2018).

Tadqiqot metadologiyasi: Tadqiqot korpusi ingliz-o'zbek parallel matnlaridan tuzildi: 4,400+ parallel gap, lingvistika sohasidagi ilmiy maqolalar va darsliklar, 1,468 terminologik juftlik, TMX va TXT formatda. Korpus alignment jarayonidan o'tkazildi va parallel konkordans uchun tayyorlandi.

Terminologiya ajratish 6 bosqichda amalga oshirildi:

1. Preprocessing – tokenizatsiya, lemmatizatsiya, POS tagging,
2. Candidate extraction – POS naqshlar: (Adj)*(Noun)+, (Noun)+ Prep (Noun)+;
3. Statistical filtering – chastota (min=3), TF-IDF, log-likelihood;
4. Collocation analysis – Sketch Engine word sketch, MI>3, t-score>2;
5. Term alignment – Giza++ (accuracy=94.3%);
6. Expert validation – precision=89.7%, recall=86.2%.

Natijalar: Korpusda eng yuqori chastotali terminlar: 'phoneme' (237), 'morpheme' (189), 'syntax' (156), 'semantics' (142). Ko'p so'zli terminlar: 'speech act' (78), 'phonetic transcription' (64), 'case system' (59). O'zbek korpusida 'fonema' (234), 'morfema' (186),

'sintaksis' (153) dominant. Chastota farqi 2-3% dan oshmaydi, bu yaxshi alignment natijasi. TF-IDF tahlili maxsus terminlarni aniqladi: 'allophone' (0.87), 'derivational morphology' (0.82).

Kollokatsiya tahlili barqaror birikmalarni ko'rsatdi: 'phonetic transcription' phonetic alphabet' 'phonetic feature'. Sketch Engine word sketch funksiyasi grammatik munosabatlarni vizualizatsiya qildi: 'bound morpheme', 'free morpheme', 'inflectional morpheme'.

Terminologik ekvivalentlik va morfologik trasformatsiyalar. Parallel korpus tahlili uchta ekvivalentlik turini aniqladi: to'liq (63.4%) – 'phoneme–fonema', 'syntax–sintaksis'; qisman (28.9%) – 'case' → 'kelishik/hol/shakl', 'speech act' → 'nutq harakati/nutqiy harakat'; nol (7.7%) – 'speech community' → 'bir xil tilda so'zlashuvchi jamiyat'.

O'zbek tilining agglutinativ xususiyati tufayli terminlarning 42.7%i kelishik shaklida ishlatilgan: joylashuv (18.3%), qaratqich (14.6%), tushum (9.8%). Bu lemmatizatsiyaning muhimligini ko'rsatadi.

Xulosa: Tadqiqot korpus asosida terminlarni avtomatik ajratish metodologiyasini ingliz-o'zbek parallel korpusida amaliy tatbiq etdi. Gibrid yondashuv – statistik chastota, lingvistik POS naqshlar va kollokatsiya metodlarining kombinatsiyasi – eng yuqori samaradorlikka erishadi. Sketch Engine zamonaviy korpus tahlilini osonlashtiradi. Parallel korpuslarda cross-lingual validation aniqlikni oshiradi. O'zbek tilining agglutinativ xususiyatlari lemmatizatsiya va morfologik tahlilni majburiy qiladi. Metodologiya boshqa til juftliklari va sohalarga tatbiq etilishi mumkin.

Adabiyotlar, References, Литературы:

1. Cabré, M. T. (1999). *Terminology: Theory, methods, and applications*. John Benjamins.
2. Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms. *International Journal on Digital Libraries*, 3(2), 115-130.
3. Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.
4. Kageura, K., & Umino, B. (1996). Methods of automatic term recognition. *Terminology*, 3(2), 259-289.
5. Kilgarriff, A., et al. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7-36.
6. Tiedemann, J. (2011). *Bitext alignment*. Morgan & Claypool.
7. Zhang, Z., Gao, J., & Ciravegna, F. (2018). SemRe-Rank: Improving automatic term extraction. *ACM Transactions on Knowledge Discovery*, 12(5), 1-41.