SCIENCE TIME
Scientific journal
VOLUME:4 ISSUE:2
10.00.00
FILOLOGIYA FANLARI
Ilmiy jurnali
ILM VAQTI

PARALLEL MATNLARDA TERMINLARNI AVTOMATIK ANIQLASH USULLARI

Gafarova Zumrad Zoxirjonovna,
Osiyo xalqaro universiteti Xorijiy til va ijtimoiy fanlar kafedrasi mudiri,
fil.f.f.d.(PhD), dotsent
E-mail: *zumradgafarova8668@gmail.com*

DOI: https://doi.org/10.5281/zenodo.18925837

Annotatsiya. *Mazkur maqolada parallel matnlar asosida terminlarni avtomatik aniqlashning nazariy va amaliy asoslari yoritiladi. Tadqiqotda korpus lingvistikasi, statistik modellar, lingvistik qoidalar hamda mashinaviy o‘rganish yondashuvlari tahlil qilinadi. Parallel korpuslarda terminlarni ekstraksiya qilish bosqichlari, alignment jarayoni va terminologik birliklarni validatsiya qilish mexanizmlari ko‘rib chiqiladi. Tadqiqot natijalari terminologik bazalarni shakllantirish va mashina tarjimasi tizimlarini takomillashtirishda muhim ahamiyat kasb etadi.*

Kalit so‘zlar: *parallel korpus, termin ekstraksiyasi, alignment, statistik model, TF-IDF, mashinaviy o‘rganish, bilingval terminologiya.*

Abstract. *This article examines the theoretical and practical foundations of automatic term extraction based on parallel texts. The study analyzes corpus linguistics approaches, statistical models, linguistic rule-based methods, and machine learning techniques. The stages of term extraction in parallel corpora, the alignment process, and the mechanisms for validating terminological units are discussed. The research findings are of significant importance for developing terminological databases and improving machine translation systems.*

Keywords*: parallel corpus, term extraction, alignment, statistical model, TF-IDF, machine learning, bilingual terminology.*

Аннотация. *В данной статье рассматриваются теоретические и практические основы автоматического выявления терминов на основе параллельных текстов. В исследовании анализируются методы корпусной лингвистики, статистические модели, лингвистические правила и подходы машинного обучения. Освещаются этапы извлечения терминов в параллельных корпусах, процесс выравнивания (alignment), а также механизмы валидации терминологических единиц. Результаты исследования имеют важное значение для формирования терминологических баз данных и совершенствования систем машинного перевода.*

Ключевые слова: *параллельный корпус, извлечение терминов, выравнивание (alignment), статистическая модель, TF-IDF, машинное обучение, билингвальная терминология.*

Kirish. Zamonaviy kompyuter lingvistikasida parallel matnlar asosida terminlarni avtomatik aniqlash dolzarb masalalardan biridir. Ilmiy-texnik matnlar hajmining ortib borishi terminologik birliklarni tez va aniq ajratib olishni talab etadi. An’anaviy qo‘lda tahlil usullari ko‘p vaqt va resurs talab qilganligi sababli, avtomatlashtirilgan metodlarga ehtiyoj ortmoqda. Parallel matnlar – mazmunan mos keluvchi ikki yoki undan ortiq tildagi matnlar majmuasidir. Ular tarjima nazariyasi, terminologiya va mashina tarjimasi sohalarida muhim manba hisoblanadi.

Metodologiya. Tadqiqotda quyidagi metodlar qo‘llanildi:

Korpus tahlili

Statistik modellashtirish

ISSN: 2992-8850
www.uzresearchers.com

Chastota tahlili

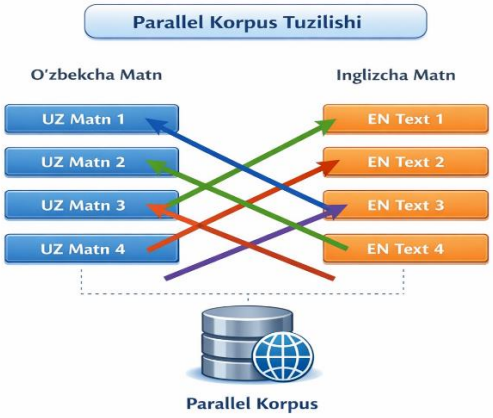
Qiyosiy tahlil

Eksperimental sinov

Tajriba uchun ilmiy-texnik yo‘nalishdagi parallel matnlar tanlab olindi. Natijalar aniqlik (precision), to‘liqlik (recall) va F1-o‘lchov orqali baholandi. Parallel korpus va uning tuzilishi. Parallel korpus quyidagi bosqichlarda shakllantiriladi:

- Matnlarni yig‘ish;
- Tozalash va normalizatsiya;
- Segmentatsiya;
- Alignment (moslashtirish). [1.12]

1-rasm. Parallel korpus tuzilishi modeli



Mazkur jarayon quyidagicha izohlanadi: dastlab manba tilidagi (L1) va tarjima tilidagi (L2) matnlar alohida yig‘iladi va korpus shakliga keltiriladi. So‘ngra ular segmentatsiya qilinib, gap yoki so‘z darajasida moslashtiriladi (alignment). Moslashtirish natijasida har ikki tildagi ekvivalent birliklar aniqlanadi. Keyingi bosqichda lingvistik va statistik filtrlash mexanizmlari asosida termin nomzodlari ajratib olinadi.

Lingvistik filtr modeli quyidagicha ishlaydi: matn dastlab morfologik tahlildan o‘tkazilib, so‘z turkumlari aniqlanadi. Keyin sintaktik andozalar (masalan, sifat + ot, ot + ot konstruksiyalari) yordamida termin bo‘lish ehtimoli yuqori bo‘lgan birliklar saralanadi. Natijada termin nomzodlari shakllanadi.

Mashinaviy o‘rganish jarayoni esa ketma-ket bosqichlarda amalga oshiriladi: korpusdan xususiyatlar (chastota, n-gram modellari, POS-belgilar) ajratiladi, belgilangan ma’lumotlar asosida model o‘qitiladi va o‘qitilgan model yangi matnlarda terminlarni aniqlaydi. Yakuniy bosqichda model tomonidan prognoz qilingan birliklar terminlar ro‘yxati sifatida shakllantiriladi.[2: 35]

Tahlillar. Bilingval termin ekstraksiyasi jarayonida har ikki tilda alohida ajratilgan termin nomzodlari alignment mexanizmi orqali o‘zaro moslashtiriladi. Moslik ehtimoli

statistik ko‘rsatkichlar asosida hisoblanadi. Ehtimollik darajasi yuqori bo‘lgan juftliklar bilingval termin sifatida qabul qilinadi va validatsiya bosqichidan o‘tkaziladi.

Alignment jarayoni so‘z, gap yoki paragraf darajasida amalga oshiriladi. Ushbu bosqich terminlarni aniqlashning asosiy tayanch nuqtasidir.

2. Terminlarni avtomatik aniqlash usullari

2.1. Statistik yondashuv

Statistik metodlar matndagi birliklarning chastota ko‘rsatkichlariga asoslanadi. Eng ko‘p qo‘llaniladigan usullar:

TF (Term Frequency)

TF-IDF (Term Frequency – Inverse Document Frequency)

Log-likelihood

C-value/NC-value algoritmi

TF-IDF formulasi:

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log(N/df(t))$$

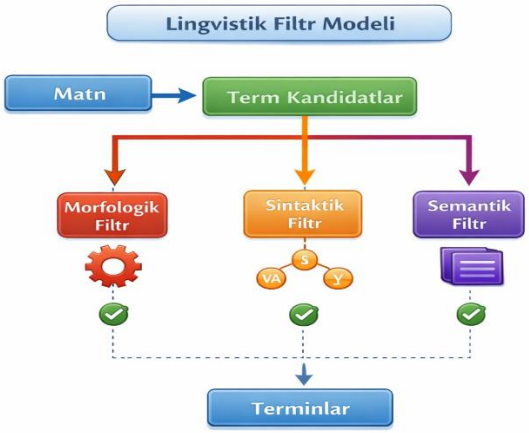
Bu yerda t – termin; d – hujjat; N — hujjatlar soni; df(t) — termin uchragan hujjatlar soni hisoblanadi. Statistik metodlar tezkorligi bilan ajralib turadi, biroq semantik aniqlik past bo‘lishi mumkin. [5:79]

2.2. Lingvistik (qoida asosidagi) yondashuv. Mazkur metod morfologik va sintaktik andozalarga asoslanadi. Masalan:

Sifat + Ot modeli (sun’iy intellekt)

Ot + Ot modeli (ma’lumotlar bazasi)

2-rasm. Lingvistik filtr modeli



Mazkur jarayon quyidagicha izohlanadi: dastlab manba tilidagi (L1) va tarjima tilidagi (L2) matnlar alohida yig‘iladi va korpus shakliga keltiriladi. So‘ngra ular segmentatsiya qilinib, gap yoki so‘z darajasida moslashtiriladi (alignment). Moslashtirish natijasida har ikki tildagi ekvivalent birliklar aniqlanadi. Keyingi bosqichda lingvistik va statistik filtrlash mexanizmlari asosida termin nomzodlari ajratib olinadi.

Lingvistik filtr modeli quyidagicha ishlaydi: matn dastlab morfologik tahlildan o‘tkazilib, so‘z turkumlari aniqlanadi. Keyin sintaktik andozalar (masalan, sifat + ot, ot +

ISSN: 2992-8850
www.uzresearchers.com

ot konstruksiyalari) yordamida termin bo‘lish ehtimoli yuqori bo‘lgan birliklar saralanadi. Natijada termin nomzodlari shakllanadi.[4:18]

Mashinaviy o‘rganish jarayoni esa ketma-ket bosqichlarda amalga oshiriladi: korpusdan xususiyatlar (chastota, n-gram modellari, POS-belgilar) ajratiladi, belgilangan ma’lumotlar asosida model o‘qitiladi va o‘qitilgan model yangi matnlarda terminlarni aniqlaydi. Yakuniy bosqichda model tomonidan prognoz qilingan birliklar terminlar ro‘yxati sifatida shakllantiriladi.

Bilingval termin ekstraksiyasi jarayonida har ikki tilda alohida ajratilgan termin nomzodlari alignment mexanizmi orqali o‘zaro moslashtiriladi. Moslik ehtimoli statistik ko‘rsatkichlar asosida hisoblanadi. Ehtimollik darajasi yuqori bo‘lgan juftliklar bilingval termin sifatida qabul qilinadi va validatsiya bosqichidan o‘tkaziladi.

Bu usul aniqligi yuqori bo‘lsa-da, til resurslari va grammatik modellarni talab etadi. Supervised va unsupervised modellar qo‘llaniladi:

Naive Bayes

Support Vector Machine (SVM)

Neyron tarmoqlar

Transformer modellari

Jarayon quyidagi bosqichlarda amalga oshiriladi:

Belgilangan (annotatsiyalangan) korpus yaratish

Xususiyatlar ajratish (POS-tag, n-gram, chastota)

Modelni o‘qitish terminlarni klassifikatsiya qilishdan iborat bo‘lib, mazkur jarayon quyidagicha izohlanadi:

- dastlab manba tilidagi (L1)
- tarjima tilidagi (L2)
- matnlar alohida yig‘iladi va korpus shakliga keltiriladi.

So‘ngra ular segmentatsiya qilinib, gap yoki so‘z darajasida moslashtiriladi (alignment). Moslashtirish natijasida har ikki tildagi ekvivalent birliklar aniqlanadi. Keyingi bosqichda lingvistik va statistik filtrlash mexanizmlari asosida termin nomzodlari ajratib olinadi.

Lingvistik filtr modeli quyidagicha ishlaydi: matn dastlab morfologik tahlildan o‘tkazilib, so‘z turkumlari aniqlanadi. Keyin sintaktik andozalar (masalan, sifat + ot, ot + ot konstruksiyalari) yordamida termin bo‘lish ehtimoli yuqori bo‘lgan birliklar saralanadi. Natijada termin nomzodlari shakllanadi. [10: 30]

Muhokama. Mashinaviy o‘rganish jarayoni esa ketma-ket bosqichlarda amalga oshiriladi: korpusdan xususiyatlar (chastota, n-gram modellari, POS-belgilar) ajratiladi, belgilangan ma’lumotlar asosida model o‘qitiladi va o‘qitilgan model yangi matnlarda terminlarni aniqlaydi. Yakuniy bosqichda model tomonidan prognoz qilingan birliklar terminlar ro‘yxati sifatida shakllantiriladi.

Bilingval termin ekstraksiyasi jarayonida har ikki tilda alohida ajratilgan termin nomzodlari alignment mexanizmi orqali o‘zaro moslashtiriladi. Moslik ehtimoli statistik ko‘rsatkichlar asosida hisoblanadi. Ehtimollik darajasi yuqori bo‘lgan juftliklar bilingval termin sifatida qabul qilinadi va validatsiya bosqichidan o‘tkaziladi.

Bu usul katta hajmdagi ma’lumotlarda yuqori samaradorlik beradi.

3. Parallel matnlarda bilingval terminlarni aniqlash

Parallel korpuslarda terminlar quyidagi usulda aniqlanadi:

Har ikki tilda alohida termin nomzodlarini ajratish

Alignment orqali mos birliklarni topish

Ehtimollik koeffitsientini hisoblash

Validatsiya va filtrlash

4-rasm. Bilingval termin ekstraksiyasi sxemasi

Mazkur jarayonda dastlab manba tilidagi (L1) va tarjima tilidagi (L2) matnlar alohida yig‘iladi va korpus shakliga keltiriladi. So‘ngra ular segmentatsiya qilinib, gap yoki so‘z darajasida moslashtiriladi (alignment). Moslashtirish natijasida har ikki tildagi ekvivalent birliklar aniqlanadi. Keyingi bosqichda lingvistik va statistik filtrlash mexanizmlari asosida termin nomzodlari ajratib olinadi. [8: 22]

Lingvistik filtr modeli quyidagicha ishlaydi: matn dastlab morfologik tahlildan o‘tkazilib, so‘z turkumlari aniqlanadi. Keyin sintaktik andozalar (masalan, sifat + ot, ot + ot konstruksiyalari) yordamida termin bo‘lish ehtimoli yuqori bo‘lgan birliklar saralanadi. Natijada termin nomzodlari shakllanadi.

Mashinaviy o‘rganish jarayoni esa ketma-ket bosqichlarda amalga oshiriladi: korpusdan xususiyatlar (chastota, n-gram modellari, POS-belgilar) ajratiladi, belgilangan ma’lumotlar asosida model o‘qitiladi va o‘qitilgan model yangi matnlarda terminlarni aniqlaydi. Yakuniy bosqichda model tomonidan prognoz qilingan birliklar terminlar ro‘yxati sifatida shakllantiriladi.

Bilingval termin ekstraksiyasi jarayonida har ikki tilda alohida ajratilgan termin nomzodlari alignment mexanizmi orqali o‘zaro moslashtiriladi. Moslik ehtimoli statistik

ISSN: 2992-8850
www.uzresearchers.com

ko‘rsatkichlar asosida hisoblanadi. Ehtimollik darajasi yuqori bo‘lgan juftliklar bilingval termin sifatida qabul qilinadi va validatsiya bosqichidan o‘tkaziladi.

Word alignment uchun statistik modellardan foydalaniladi (masalan, IBM Models konsepsiyasi asosidagi yondashuvlar). Tadqiqot natijalari shuni ko‘rsatdiki, statistik metodlar tezkor, ammo shovqin darajasi yuqori. Lingvistik metodlar aniqligi yuqori, biroq qamrovi cheklangan. Mashinaviy o‘rganish usullari eng optimal natija berdi. Kombinatsiyalangan (gibrid) model eng yuqori F1 ko‘rsatkichini ta’minladi.

Xulosa. Parallel matnlarda terminlarni avtomatik aniqlash ko‘p bosqichli jarayon bo‘lib, statistik, lingvistik va mashinaviy yondashuvlarni uyg‘unlashtirish eng samarali natija beradi. Ushbu tadqiqot natijalari terminologik bazalarni yaratish, avtomatik lug‘atlar tuzish va mashina tarjimasi tizimlarini rivojlantirishda qo‘llanishi mumkin.

Foydalanilgan adabiyotlar ro‘yxati:

1. Ahmad, Kh. Corpus Linguistics and Terminology Extraction. London: Continuum. 2000. – P. 12.
2. Bowker, L., Pearson, J. Working with Specialized Language: A Practical Guide to Using Corpora. London: Routledge. 2002. – 242 p.
3. Cabré, Maria T. Terminology: Theory, Methods and Applications. Amsterdam: John Benjamins. 1999. – 248 p.
4. Daille, B. Study and implementation of combined techniques for automatic extraction of terminology. In J. Klavans, P. Resnik (Eds.), The Balancing Act: Combining Symbolic and Statistical Approaches to Language (pp. 49–66). Cambridge, MA: MIT Press. 1996. – P. 18.
5. Evert, S. Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), Corpus Linguistics: An International Handbook (Vol. 2, pp. 1212–1248). Berlin: Mouton de Gruyter. 2008. – P. 37.
6. Jurafsky, D., Martin, James. H. Speech and Language Processing (3rd ed., draft). Stanford University. 2023. – 600 p.
7. Manning, Christopher D., Raghavan, Prabhakar, & Schütze, Hinrich (2008). Introduction to Information Retrieval. Cambridge: Cambridge University Press. 482 p.
8. Mikolov, Tomas et al. (2013). Efficient estimation of word representations in vector space. Proceedings of ICLR. 2013. – P. 1-12.
9. Och, Franz Josef, & Ney, Hermann (2003). A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1), 19–51. 33 p.
10. Salton, Gerard, & Buckley, Christopher (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523. 11 p.
11. Vaswani, Ashish et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30, 5998–6008. 11 p.
12. Church, Kenneth Ward, & Hanks, Patrick (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22–29. 8 p.