

PARALLEL MATNLARDA SEMANTIK MUQOBILLIKNI STATISTIK ANIQLASHDA NLP TEXNOLOGIYALARINING AHAMIYATI

Vasliddinova Kamola Qodirjon qizi

PhD talabasi

kvasliddinova12@mail.ru

O'zbekiston davlat jahon tillari universiteti

Annotatsiya: Maqolada parallel korpuslarda semantik muqobillikni statistik, miqdoriy va zamonaviy neyron NLP modellari (Transformer, multilingual embeddings) orqali aniqlash mexanizmlari, tarjima xotirasi va leksikografik tahlilning amaliy masalalari tadqiq qilinadi. Agglyutinatив o'zbek tili va analitik ingliz tili materiallari misolida gibril metodologiyani samaradorligi asoslab berilgan.

Kalit so'zlar: korpus lingvistikasi, parallel korpus, semantik muqobillik, NLP texnologiyalari, kosinus o'xshashligi, mashina tarjimasi, embedding, o'zbek tili.

Abstract: The article explores the mechanisms of detecting semantic equivalence in parallel corpora using statistical, quantitative, and neural NLP models (Transformers, multilingual embeddings), addressing practical translation memory and lexicographical tasks. The effectiveness of a hybrid methodology is substantiated using Uzbek and English corpus materials.

Keywords: corpus linguistics, parallel corpora, semantic equivalence, NLP technologies, cosine similarity, machine translation, embedding, Uzbek language.

KIRISH

Zamonaviy kompyuter tilshunosligi va amaliy korpus lingvistikasining eng dolzarb vazifalaridan biri ko'p tilli axborot resurslarini integratsiyalash, matnlarni avtomatik qiyoslash hamda semantik mutanosiblikni aniqlashdan iboratdir. Axborot oqimining keskin oshishi sharoitida ikki yoki undan ortiq tildagi matn birliklarining o'zaro mosligini an'anaviy empirik-tavsifiy usullar yordamida qo'lda solishtirish nihoyatda katta vaqt va intellektual resurs talab qiladi. Shu sababli, parallel korpuslar

negizida tabiiy tilni qayta ishlash (NLP – Natural Language Processing) texnologiyalaridan unumli foydalanish amaliy leksikografiya, tarjima xotirasi (TM – Translation Memory) tizimlari va mashina tarjimasini modellarini takomillashtirishda asosiy poydevor bo‘lib xizmat qilmoqda [1:112].

Parallel matnlar shunchaki ikki tildagi jumlalarining yonma-yon keltirilishi emas, balki tilning leksik, morfologik, sintaktik va semantik qatlamlaridagi invariantlikni modellashtirish vositasidir. O‘zbek tili agglyutinativ tabiatga ega bo‘lib, sintaktik qurilishi jihatidan analitik yoki flektiv tillardan (masalan, ingliz tili) tubdan farq qiladi. Bunday sharoitda so‘zma-so‘z yoki qat’iy sintaktik moslashtirish ko‘p hollarda o‘zini oqlamaydi. Binobarin, jumlar va iboralar darajasidagi ma’no yaqinligini – semantik muqobillikni – miqdoriy va statistik ko‘rsatkichlar orqali baholash algoritmlari alohida dolzarblik kasb etadi [5:78].

Parallel korpuslarda semantik muqobillikni aniqlashdan avvalgi dastlabki va hal qiluvchi bosqich tekislash (alignment) jarayonidir. Matnlar bir necha darajada tekislanadi: hujjat darajasida (document-level), xatboshi darajasida (paragraph-level), jumla darajasida (sentence-level) hamda so‘z/ibora darajasida (word/phrase-level) [4:15]. Dastlabki statistik usullar ko‘p hollarda P.Gale va K.Church tomonidan taklif etilgan jumla uzunligiga asoslangan algoritmgaga tayangan [2:20]. Mazkur yondashuv ikki tildagi tarjima jumlarining belgi yoki so‘z soni bo‘yicha proporsional o‘zgarishiga asoslanadi. Masalan, ingliz tilidagi manba matn uzunligi L_{src} va o‘zbek tilidagi tarjima matn uzunligi L_{trg} normal taqsimotga ega deb hisoblanadi. Biroq mazkur sof miqdoriy usul matnning ichki mazmunini, semantik xususiyatlarini hisobga olmaydi. Agglyutinativ o‘zbek tilida bitta so‘z tarkibida bir nechta affikslar yordamida butun bir sintaktik birikma ifodalanishi mumkinligi (masalan: kelolmaganliklaridan – because they were unable to come) faqat uzunlikka tayangan modellarda xatoliklar ulushini oshiradi.

Shu munosabat bilan, zamonaviy NLPda leksik bog‘liqlik va ehtimollik modellari integratsiya qilingan gibrid usullardan foydalaniladi. IBM Models (1 dan 5 gacha bo‘lgan statistik modellar) hamda HMM (Hidden Markov Model) arxitekturasi asosida ishlovchi GIZA++ yoki fast_align vositalari jumlar orasidagi leksik o‘xshashlik ehtimolini shartli ehtimollik funksiyalari orqali hisoblaydi [5:85]. Ushbu statistik miqdorlar orqali korpusdagi har bir leksemaning tarjima ehtimoli jadvali shakllanadi va bu keyingi bosqichda semantik tahlil uchun tayanch zamin yaratadi. Statistik leksikografiyada semantik bog‘liqlik darajasi ko‘pincha birgalikda uchrash chastotalari (co-occurrence metrics) yordamida aniqlanadi. Ikkita til birligi o‘rtasidagi semantik muvofiqlikni baholash uchun quyidagi statistik mezonlar keng qo‘llaniladi. Biroq, an’anaviy statistik mezonlar polisemantik (ko‘p ma’noli) so‘zlar, frazeologizmlar va metaforik ko‘chimlarni aniqlashda cheklangan imkoniyatga ega. Ular so‘z shaklining invariantligiga bog‘liq bo‘lib, sintaktik o‘zgarishlar sharoitida kontekstual semantikani to‘liq ifodalay olmaydi.

Zamonaviy kompyuter tilshunosligida statistik usullar o‘rnini chuqur o‘rganish (Deep Learning) va taqsimlangan semantik modellar (Distributional Semantics) egallamoqda. Word2Vec, FastText modellarining paydo bo‘lishi so‘zlarni ko‘p o‘lchamli fazoda zich vektorlar (dense vectors) ko‘rinishida tasvirlash imkonini berdi. Keyinchalik, kontekstual semantikani to‘liq modellashtirish imkonini beruvchi ko‘p tilli Transformer arxitekturalari (mBERT, XLM-RoBERTa, LaBSE) korpus lingvistikasidagi semantik tahlil sifatini yangi bosqichga olib chiqdi [5:210].

O‘zbek tili uchun keng qamrovli milliy va parallel korpuslarni ishlab chiqish davlat tilining raqamli muhitdagi o‘rnini mustahkamlashning eng muhim shartlaridan biridir. Xususan, o‘zbek-ingliz parallel korpuslarida semantik muqobillikni aniqlash algoritmlarini qo‘llash bir qator amaliy natijalarni ta’minlaydi:

Tarjima xotirasi (CAT tools) bazalarini avtomatlashtirish: Lokalizatsiya va tarjima jarayonida professional tarjimonlar foydalanadigan bazalarda har bir yangi kiritilgan jumla semantik tekshiruvdan o'tkaziladi. Kosinus o'xshashligi 0.85 dan yuqori bo'lgan tayyor segmentlar avtomatik tavsiya etiladi, bu esa tarjima tezligi va birxilligini 40–50% ga oshiradi.

Mashina tarjimasi ma'lumotlar to'plamini (Dataset) tozalash: Neyron tarjima modellarini (NMT) o'qitish uchun foydalaniladigan parallel korpuslarda noto'g'ri tarjima qilingan yoki chalkashgan jumalarning mavjudligi modelning yakuniy sifatini keskin tushirib yuboradi. Semantik yaqinlik chegarasi (threshold) orqali sifatsiz ma'lumotlar avtomatik filtrlanadi [5:215].

Ikki tilli avtomatlashtirilgan terminografiya: Parallel korpuslar orqali fan va texnikaning muayyan sohalariga oid ko'p so'zli terminologik birikmalarning muqobillari statistik PMI va kontekstual embeddinglar asosida ajratib olinadi va elektron lug'atlar bazasiga kiritiladi.

XULOSA

Xulosa qilib aytganda, bugungi korpus lingvistikasida parallel matnlarning semantik muqobilligini statistik va miqdoriy jihatdan aniqlash an'anaviy filologik tahlil doirasidan chiqib, tabiiy tilni qayta ishlashning (NLP) eng ilg'or matematik-algoritmik sohalari bilan uzviy birlashmoqda. Uzunlikka asoslangan dastlabki algoritmlardan tortib, ehtimollik modellari (IBM Models), statistik assotsiatsiya ko'rsatkichlari (PMI, Dice) hamda zamonaviy ko'p tilli vektorli fazoviy modellar (LaBSE, mBERT) parallel ma'lumotlar bazasining yuqori semantik aniqligini kafolatlaydi.

O'zbek tili morfemik va sintaktik jihatdan o'ziga xos agglyutinatив tuzilishga ega bo'lganligi sababli, amaliyotda faqat bitta usul bilan cheklanmasdan, statistik ko'rsatkichlar va neyron embeddinglarni birlashtirgan gibrid tizimlarni qo'llash eng maqbul yechim hisoblanadi. Bunday texnologik konveyerlar O'zbek tili milliy

korpusining parallel qismini shakllantirish, mashina tarjiması sifatini oshirish va raqamli leksikografiya resurslarini yaratishda muhim ilmiy-amaliy omil bo'lib xizmat qiladi.

Foydalanilgan adabiyotlar ro'yxati

1. Brown P.F., Pietra V.J., Pietra S.A., Mercer R.L. The Mathematics of Statistical Machine Translation: Parameter Estimation // Computational Linguistics. – 1993. – Vol. 19, No. 2. – pp. 263–311.
2. Gale W.A., Church K.W. A Program for Aligning Sentences in Bilingual Corpora // Computational Linguistics. – 1993. – Vol. 19, No. 1. – pp. 75–102.
3. Koehn P. Statistical Machine Translation. – Cambridge: Cambridge University Press, 2010. – 433 p.
4. Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing. – Cambridge, MA: MIT Press, 1999. – 680 p.
5. Yang F., Cer D., Ahmad A. et al. Multilingual Sentence Representations using Universal Encoder and LaBSE // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2020. – pp. 8771–8780.