O‘ZBEK TILI MILLIY KORPUSIDA SEMANTIK BOG‘LIQLIKLARNI NEYRON TARMOQLARI VOSITASIDA MODELLASHTIRISH

Murtazoyeva Robiya No‘monovna
Axborot texnologiyalar va menejment universiteti
“Boshlang‘ich ta‘lim” yo‘nalishi 2-kurs BT-29-24-guruh talabasi

Annotatsiya

Ushbu maqolada o‘zbek tili milliy korpusini rivojlantirishda neyron tarmoqlari asosidagi Word Embedding modellarining o‘rni nazariy jihatdan tahlil qilinadi. O‘zbek tilining agglutinativ xususiyatlari hisobga olingan holda, so‘zlar o‘rtasidagi semantik bog‘liqliklarni matematik ifodalash va ularni vektor fazoda modellashtirishga oid ilmiy yondashuvlar ko‘rib chiqiladi. Maqolada mavjud NLP metodlari va ularning o‘zbek tili korpusiga tatbiq etish imkoniyatlari yoritiladi.

Kalit so‘zlar: *Kompyuter lingvistikasi, neyron tarmoqlar, NLP, o‘zbek tili korpusi, Word Embedding, FastText, semantik tahlil.*

Аннотация

В данной статье проводится теоретический анализ роли моделей Word Embedding, основанных на нейронных сетях, в развитии национального корпуса узбекского языка. С учетом агглютинативных особенностей узбекского языка, его морфологической сложности и семантических слоев рассматриваются вопросы математического моделирования скрытых связей между словами. Также анализируются возможности применения современных методов обработки естественного языка (NLP) к корпусу узбекского языка и существующие проблемы в данной области.

Ключевые слова: *Компьютерная лингвистика, нейронные сети, NLP, корпус узбекского языка, Word2Vec, FastText, семантический анализ, когнитивное моделирование, векторные модели.*

Abstract

This article presents a theoretical analysis of the role of neural network-based Word Embedding models in the development of the Uzbek National Corpus. Considering the agglutinative nature, morphological complexity, and semantic layers of the Uzbek language, the study explores the mathematical modeling of hidden semantic relationships between words. Additionally, the paper examines the applicability of modern Natural Language Processing (NLP) approaches to the Uzbek language corpus and highlights existing challenges in this field.

Keywords:*Computational linguistics, neural networks, NLP, Uzbek language corpus, Word2Vec, FastText, semantic analysis, cognitive modeling, vector space models.*

Kirish

Hozirgi raqamli davrda til faqat aloqa vositasi emas, balki axborotni saqlash, uzatish va qayta ishlashning murakkab tizimi sifatida ham qaralmoqda. Ayniqsa, sun’iy intellekt va katta ma’lumotlar (big data) rivojlanishi natijasida tilni avtomatik tahlil qilishga bo‘lgan ehtiyoj sezilarli darajada ortib bormoqda. Shu nuqtai nazardan qaraganda, til birliklarini faqat grammatik qoidalar asosida emas, balki matematik va statistik usullar yordamida ham o‘rganish dolzarb masalaga aylangan. Bu esa kompyuter lingvistikasi va NLP (Natural Language Processing) sohalarining jadal rivojlanishiga sabab bo‘lmoqda.

O‘zbek tili ham ushbu jarayondan chetda emas. Aksincha, o‘zbek tilining boy morfologik tuzilishi, so‘z yasash imkoniyatlarining kengligi va agglutinativ tabiati uni avtomatik tahlil qilish nuqtai nazaridan qiziqarli, lekin murakkab tizimga aylantiradi.Shu o‘rinda savol tug‘iladi: til birliklari o‘rtasidagi ma’no bog‘liqliklarini kompyuter tizimlari qanday aniqlaydi? Aynan shu savol zamonaviy neyron tarmoqlari va Word Embedding modellarining paydo bo‘lishiga asos bo‘lgan. Ushbu modellar so‘zlarni faqat alohida birlik sifatida emas, balki kontekst ichidagi munosabatlar orqali vektor ko‘rinishida ifodalash imkonini beradi. Shu sababli, o‘zbek tili milliy korpusida semantik bog‘liqliklarni modellashtirish masalasi nafaqat lingvistik, balki zamonaviy sun’iy intellekt tadqiqotlari uchun ham muhim yo‘nalishlardan biri hisoblanadi.

O‘zbek tilida so‘z ma’nosi ko‘pincha kontekst va qo‘shimchalar orqali o‘zgaradi, bu esa uni semantik modellashtirish nuqtai nazaridan murakkab, ammo qiziqarli tilga aylantiradi.

Tadqiqot metodologiyasi

Mazkur ish nazariy-tahliliy yondashuv asosida olib borilgan bo‘lib, unda mavjud ilmiy manbalarda qo‘llanilgan metodlar o‘rganilgan. Tahlilda quyidagi yondashuvlar asos qilib olingan: Vektorli modellashtirish. So‘zlar ko‘p o‘lchovli fazoda vektor sifatida ifodalanadi va ularning semantik yaqinligi matematik masofalar orqali aniqlanadi[3]. Bu yondashuv til birliklarini formal modelga o‘tkazish imkonini beradi. Word2 Vec va Fast Text modellarida so‘zlar kontekst asosida o‘rganiladi. Ushbu modellar so‘zlarning ma’nosini faqat lug‘aviy emas, balki kontekstual asosda ham ifodalash imkonini beradi[4]. Ayniqsa, Fast Text modeli so‘zlarni subword birliklarga ajratgani sababli agglutinativ tillar uchun samarali hisoblanadi.So‘zlar o‘rtasidagi ma’no yaqinligi kosinus o‘xshashligi orqali baholanadi[5]. Bu usul semantik klasterlash va yashirin bog‘liqliklarni aniqlashda keng qo‘llaniladi.

Natijalar va muhokama

Ilmiy manbalar tahlili shuni ko‘rsatadiki, neyron tarmoqlari asosidagi modellar o‘zbek tili kabi murakkab morfologik tizimga ega tillarda ham samarali qo‘llanilishi mumkin. Word Embedding yondashuvi so‘zlarni vektor fazoda yaqinlashtirish orqali ularning semantik munosabatlarini aniqlash imkonini beradi[6]. O‘zbek tilidagi qo‘shimchalar so‘zning ma’nosini sezilarli darajada o‘zgartiradi va bu holat model uchun alohida subword yondashuvni talab qiladi[7]. Shu sababli Fast Text modeli ushbu til uchun nisbatan mos yechimlardan biri hisoblanadi. Mavjud tadqiqotlar shuni ko‘rsatadiki, o‘zbek tili korpuslari hozircha asosan ma’lumot yig‘ish darajasida bo‘lib, to‘liq semantik tahlil qiluvchi “aqlli tizim” darajasiga yetmagan[8]. Shu bois kelajakda korpuslarni faqat matn bazasi emas, balki kontekstni tushunuvchi dinamik tizim sifatida rivojlantirish zarur[9].

Xulosa

O‘zbek tili milliy korpusida semantik bog‘liqliklarni neyron tarmoqlari yordamida modellashtirish zamonaviy kompyuter lingvistikasining muhim yo‘nalishlaridan biri hisoblanadi. Word Embedding va Fast Text kabi modellar til birliklarini chuqur semantik darajada tahlil qilish imkonini beradi. Shu bilan birga, mavjud yondashuvlar hali to‘liq mukammal emas va ularni o‘zbek tilining lingvistik xususiyatlariga moslashtirish ustida qo‘shimcha ilmiy izlanishlar talab etiladi. Bu esa milliy til korpusining raqamli texnologiyalar bilan integratsiyasini kuchaytiradi. Bundan tashqari, ushbu yo‘nalishning rivojlanishi kelajakda o‘zbek tilida avtomatik tarjima, matnni tushunish va sun’iy intellektga asoslangan tizimlarni takomillashtirish uchun muhim ilmiy asos bo‘lib xizmat qiladi.

Foydalanilgan adabiyotlar

1. Mo‘minova L. R. O‘zbek tili nazariyasi. Toshkent, 2020.
2. Karimov S. Kompyuter lingvistikasi asoslari. Toshkent: Fan, 2023.
3. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016.
4. Bojanowski P. va boshq. 2017. Enriching word vectors with subword information.
5. Zokirov F. Matematik tilshunoslik. Toshkent, 2021.
6. Mikolov T. va boshq. 2013. Word Representations in Vector Space.
7. Abdullayev A. O‘zbek tili korpusi. Toshkent, 2022.
8. Jurafsky D., Martin J. Speech and Language Processing. 2023.
9. Mirzayev M. Raqamli lingvistika istiqbollari. 2022.
10. Devlin J. va boshq. BERT modeli. 2019.