

BÖYÜK DİL MODELLƏRİNİN AZƏRBAYCAN DİLİNƏ TƏTBİQİ VƏ PROBLEMLƏRİ

M.Heydərova

Azərbaycan Milli Elmlər Akademiyası

Nəsimi adına Dilçilik İnstitutu

Kompüter dilçiliyi şöbəsi, böyük elmi işçi

Аннотация: В статье рассматривается влияние искусственного интеллекта, в частности больших языковых моделей (LLM), на лингвистику и проблемы их применения в тюркских языках. Морфологическое богатство, вариативность аффиксов и синтаксическая сложность тюркских языков влияют на точность моделей. Подчеркиваются трудности автоматической разметки и анализа сентимента для азербайджанского языка. Для решения этих проблем необходимо развитие национальных языковых корпусов и адаптированных моделей.

Annotation: The article examines the impact of artificial intelligence, particularly large language models (LLMs), on linguistics and the challenges of their application in Turkic languages. The morphological richness, affix variability, and syntactic complexity of Turkic languages affect the accuracy of these models. Difficulties in automatic annotation and sentiment analysis for the Azerbaijani language are highlighted. Developing national language corpora and adapted models is essential to address these issues.

Annotatsiya: Maqolada sun'iy intellekt, xususan, katta til modellari (LLM) lingvistikaga ta'siri va ularning turkiy tillarda qo'llanilishi bilan bog'liq muammolar tahlil qilinadi. Turkiy tillarning morfologik boyligi, affikslar xilma-xilligi va sintaktik murakkabligi ushbu modellarning aniqligiga ta'sir ko'rsatadi. O'zbek tilida avtomatik belgilanglash va sentiment tahlil qilishdagi qiyinchiliklar ta'kidlanadi. Ushbu muammolarni hal qilish uchun milliy til korpuslarini va moslashtirilgan modellarni rivojlantirish zarur.

Ключевые слова: Искусственный интеллект, большие языковые модели, тюркские языки, автоматическая разметка, анализ сентимента.

Keywords: Artificial intelligence, large language models, Turkic languages, automatic annotation, sentiment analysis.

Kalit so'zlar: Sun'iy intellekt, katta til modellari, turkiy tillar, avtomatik belgilash, sentiment tahlili.

Son illər texnologiya sahəsində gedən sürətli inkişaf elmin bütün istiqamətlərində olduğu kimi, dilçilik sahəsində də süni intellektə əsaslanan tətbiqlərin meydana çıxmasına səbəb olmuşdur. Süni intellekt (Aİ – artificial intelligence) texnologiyalarının son nailiyyətlərindən biri də istifadəsi geniş yayılmış böyük dil modelləridir (LLM – Large Language Models). Ümumiyyətlə, süni intellektdən danışarkən onun iki əsas növünü qeyd etməliyik: generativ süni intellekt və proqnozlaşdırıcı süni intellekt.

Generativ süni intellekt insan beyninin öyrənmə və qərar vermə proseslərini simulyasiya edən dərin öyrənmə (DL – Deep Learning) alqoritmlərinə əsaslanır.

Bunlara yeni məzmun yaratmaq qabiliyyətinə malik böyük dil modelləri aiddir. Proqnozlaşdırıcı süni intellekt isə verilən məlumatlara əsaslanır və semantik əlaqələri təhlil edərək növbəti hadisələri proqnozlaşdırır. Bu növə əsaslanan süni intellekt proqramları statistika, maşın öyrənməsi, təbii dil emalı metodlarından istifadə edir [1]. Bu proqramlara misal olaraq ChatGPT, Deepseek, Claude, [Jasper](#), Altair AI Studio, Alteryx AI Platform, Dataiku və s.göstərə bilərik.

Böyük dil modelləri kütləvi mətn korpusları əsasında təbii dilin sintaksisini və semantikasını öyrənən süni intellekt proqramlarıdır. Bu tip modellər neyron şəbəkədən ibarətdir və onun işləmə texnologiyası dərin öyrənmə alqoritmlərinə əsaslanır. DL alqoritmləri çoxqatlı neyron şəbəkələrindən istifadə etməklə təbii dili təhlil edir və onu anlamaqla mətndən strukturlaşdırılmış məlumatları çıxarmağa imkan yaradır. Mətn insan biliklərinin ötürülməsi üçün bir vasitə olduğundan, LLM-lər dərin öyrənmə alqoritmləri vasitəsilə təbii dili dərk edə və istehsal edə bilir. Bu da insanlara LLM-lərlə təbii dildə ünsiyyət qurmağa imkan verir. Bu da öz növbəsində LLM-lərə insanın malik olduğundan daha çox biliyə sahib olmağa imkan verir.

Qeyd etdiyimiz kimi LLM-lər böyük həcmdə mətn məlumatı üzərində qurulur. Burada milyardlarla söz və cümlədən söhbət gedir. Mətn korpusları böyük dil modelləri vasitəsi ilə dil qaydalarını, təhlilini və sözlərin məna əlaqələrini öyrənərək yüksək keyfiyyətli mətn yaratmağa və bir çox digər tapşırıqları yerinə yetirməyə imkan verir. Böyük dil modelləri mətn korpuslarında verilən cümlələrdən hansı sözlərin bir-biri ilə daha tez-tez əlaqəyə girərək istifadə edildiyini və kontekstdə necə işləndiyini anlayaraq bu sözlər və onların sinonimləri əsasında yeni mətn yarada bilir.

Bütün dillərdə olduğu kimi böyük dil modelləri türk dilləri ailəsi üçün də eyni prinsiplə işləyir. Bu modellərin arxitekturası transformator əsaslıdır və hər bir dilin özünəməxsus xüsusiyyətlərinə uyğunlaşdırılıb. Neyron şəbəkə arxitekturasının bir növü olan transformator özünə diqqət (self-attention) mexanizmi ilə mətndəki sözlər arasındakı əlaqələri öyrənir və konteksti başa düşməyə kömək edir. Türk dillərində sözlərin cümlə içərisindəki yeri və morfoloji dəyişiklikləri əsas olduğundan bu mexanizm xüsusi əhəmiyyət daşıyır.

Qeyd etdiyimiz kimi, böyük dil modelləri əsasən ingilis dili kimi yüksək resurslu flektiv dillər üçün hərtərəfli effektiv nəticələrlə təmin etsə də, türk dilləri kimi morfoloji cəhətdən zəngin və aqlutativ dillərin unikal dil mürəkkəbliklərini nəzərə almaqda çətinlik çəkir. Türk dilləri mürəkkəb morfolojiyası və sintaktik quruluşu ilə NLP üçün fərqli problemlər yaradır. Məsələn, türkcə tək feil kökü, şəkilçi, zaman, şəxs və əhval-ruhiyyəni yığcam, lakin mürəkkəb şəkildə kodlaşdırmaqla çoxsaylı söz formaları yarada bilir. Bu dəyişkənlik tokenləşməni və semantik təhlili çətinləşdirir, modellərin həm səthi nümunələrini, həm də dərin linqvistik strukturları qavramasını tələb edir [2].

Eyni problem ilə biz milli dil korpuslarının hazırlanması zamanı avtomatik işarələnmə məsələsində qarşılaşırıq. Buna misal olaraq Azərbaycan dilinin milli korpusunun hazırlanması zamanı yaranan çətinlikləri göstərə bilərik. Azərbaycan dilinin iltisacı dillərdən biri kimi avtomatik işarələnməsi texniki cəhətdən kifayət qədər ciddi problemlər yaradır. Əsas problemlər isə morfoloji sistemdə özünü

göstərir. Belə ki, dilimizdə şəkilçilərin çoxvariantlılığı və omonimliyi bu cür problemlərin həllini çətinləşdirir. Məsələn, Azərbaycan dilində “-in⁴” şəkilçisi həm mənsubiyyət, həm də yiyəlik hala aid olduğundan kompüter proqramının onu ayırmasında problem yaradır. Bundan başqa, dilimizdə şəxs adlarında cins fərqlərinin olmaması, məsələn qız və oğlan adlarının eyni olması problem yaradan məsələlərdəndir. Misal üçün, Azərbaycan dilində bir cümləyə diqqət edək: Arzu kitablarını götürüb evdən çıxdı. Burada “Arzu” sözünün qadın və kişi cinsinə aid olduğunu müəyyənləşdirmək proqram üçün çox çətinidir. “Kitabın cildi” və ya “Sənin kitabın məndədir” birləşmələrində “-ın”, “-in” şəkilçilərinin avtomatik işarələnməsi məsələsi ciddi problem kimi qarşıda duran məsələlərdəndir. Eyni zamanda bildiyimiz kimi, dilimizdə bəzi sözlərdə söz sonunda şəkilçilər artırılarkən söz kökündə son samit dəyişir. Məsələn, “külək” sözünə saitle başlayan şəkilçi artırısaq söz “küləyin” olacaq. Məlumdur ki, dilimizdə “k”, “q” samitləri ilə bitən çoxhecalı və milli mənşəli sözlərə saitle başlayan şəkilçi qoşulduqda söz kökündəki son samit dəyişir. Məsələn, papaq – papağa, külək – küləyə və s. Amma təkhecalı sözlərdə bu hal baş vermir. Məsələn, kök – kökə və s. Eyni zamanda bəzi alınma sözlərdə də bu hal baş vermir. Məsələn, bank – banka və s. Bu tip hallar isə avtomatik işarələnmə üçün ciddi problem olaraq qalmaqdadır. Bu problem eyni zamanda böyük dil modellərində də özünü göstərir [3].

Böyük dil modellərində yaranan bu tipli problemlərdən biri söz formalarının çoxluğuudur ki, bu zaman bir sözün bir neçə formada yazılması məsələn, “oxu” (oxumaq) felinin formaları kimi “oxudum”, “oxuyaram”, “oxuyacağam”, “oxuyurdum”, “oxusaydım” və s. problem yaradada bilər. Buna görə də model hər bir sözün formasını tanımalı və kontekstə uyğunlaşdırmalıdır. Digər tərəfdən şəkilçilərin çoxluğu, yəni kökə bir neçə şəkilçi əlavə olunması modellərin mətni düzgün şəkildə emal etməsini çətinləşdirir. Sözün kökünə əlavə olunan hər bir şəkilçinin sözün qrammatik rolunu və mənasının dəyişməsi modellərin daha dəqiq təhlilini tələb edir. Burada modelin performansına təsir edən əsas cəhət tokenləşmədir, yəni mətnin modelin anlaya biləcəyi tokenlərə bölməsi prosesidir.

Çoxdilli böyük dil modellərinin problemlərində biri də söz sırasında olan dəyişkənlikdir. Belə ki, ingilis dilində söz sırası SVO (subject – verb – object) olsa da türk dillərində SOF (subyekt – obyekt – feil) formasında olur. Böyük dil modelləri əsasən ingilis dili üzərində qurulduğundan bəzən SOF quruluşlu cümlələri anlamaqda çətinlik çəkə bilər. Bəzən də eyni cümlə vurğuya və ya durğu işarəsinə görə fərqli mənalar verə bilər. Bu zaman konteksti anlamaq üçün mətnin sentiment təhlili əsas rol oynar (məsələn, Gəl, evə gedək. Gəl evə gedək). Digər tərəfdən Azərbaycan dilində söz sırası sərbəst olur vurğuya görə mənanı anlamaq olur (məsələn, Mən evə getdim. Evə mən getdim. Getdim mən evə). Yaranan problemləri aradan qaldırmaq üçün Azərbaycan dili üçün böyük dil modellərinin hazırlanması məsələsi bu problemlərin həllində kömək ola bilər. Burada əsas məsələ kimi sentiment təhlil problemidir. Sentiment təhlil mətndə əhval-ruhiyyəni, duyğusallığı müəyyənləşdirən, mətni müsbət, mənfi, neytral kateqoriyaya bölən təhlil üsuludur.

Türk dilinin sentiment analizi zamanı yaranan çətinliklərə nəzər salsaq Azərbaycan dili üçün də eyni qaydada olduğunu qeyd edə bilərik:

- sentiment təhlilin prosesinin dildən asılılığı;

- türk dilinin mürəkkəb strukturu;
- türk dilində mövcud olan idiomatik ifadələrin olması;
- hisslərin təhlili hər bir söz üzərində fərdi aparılması;
- türk hiss leksikonlarında resursların məhdud olması;
- müxtəlif dillər arasında mədəniyyət fərqləri [4].

Sentiment təhlil üsulu üç kateqoriyaya bölünür, bunlar lüğətə əsaslı, maşın öyrənməsi əsaslı və hibrid əsaslıdır. Böyük dil modellərində əsasən hibrid əsaslı sentiment təhlildən istifadə edilir.

Lüğət əsaslı sentiment təhlil əvvəlcədən hazırlanmış sentiment söz bazası əsasında işləyir. Mətndəki sözlərin mənfi, müsbət və ya neytral olduğu bu baza əsasında müəyyən olunur. Cümlənin tonallığını tapmaq üçün cümlədəki sentiment sözlərin sayı hesablanaraq təyin edilir. Məsələn, Konsert çox maraqlı idi, amma tez bitdi. Cümlədə maraqlı – müsbət mənəli, amma – sözünün əksər lüğətlərdə sentiment dəyəri yoxdur, lakin kontekstə görə mənə dəyişə bilər, tez – sözü neytral dəyərləndirilir. Son nəticə olaraq cümlənin tonallığı müsbət qiymətləndirilir. Lüğət əsaslı yanaşma sadədir, az mənbə tələb edir və mürəkkəb modellər olmadan işləyə bilər. Bu da hər zaman doğru nəticə verməyə bilər. Məsələn, “yaxşı deyil” ifadəsində “yaxşı” sözü müsbət olsa da ifadənin özü mənfi anlam verir. Sentiment təhlili dildən asılı proses olduğundan, müxtəlif dillərdə sentiment təhlilinin aparılması ilə bağlı çətinlik dərəcəsi dəyişir. Buna görə də lüğət əsaslı yanaşmaya qaydalarla bağlı təkmilləşdirmə tələb olunur.

Maşın öyrənməsi əsaslı sentiment təhlili, mətnin tonallığını, əhval-ruhiyyəni müəyyən etmək üçün statistik və alqoritmik üsullardan istifadə edir. Bu metod lüğət əsaslı yanaşmadan fərqli olaraq məlumatlardan öyrənir və daha mürəkkəb dil xüsusiyyətlərini anlamaq bilər.

Hibrid əsaslı yanaşmada hisslərin təhlili performansını və effektivliyini artırmaq üçün həm lüğət əsaslı, həm də maşın öyrənməsi əsaslı metodların elementlərindən istifadə edilir.

Nəticə olaraq qeyd edə bilərik ki, süni intellektin dilçilikdə tətbiqi, xüsusilə böyük dil modellərinin inkişafı ilə yeni imkanlar yaratsa da, Azərbaycan dili kimi aqlutativ və morfoloji cəhətdən mürəkkəb dillər üçün hələ də müəyyən texniki və metodoloji problemlər qalmaqdadır. Bu problemlərin həlli üçün milli korpusların yaradılması, sentiment təhlil sistemlərinin inkişaf etdirilməsi və dilin spesifik xüsusiyyətlərini nəzərə alan modellərin hazırlanması zəruridir.

İstifadə olunmuş ədəbiyyat siyahısı

1. Guillaume Desagulier (25 novembre 2024). Corpus linguistics in the LLM era – the changing nature of language data. *Around the word*. Consulté le 23 janvier 2025 à l'adresse. Electronic resource (date of use: 01.04.2025): <https://doi.org/10.58079/12qwh>
2. Setting Standards in Turkish NLP: TR-MMLU for Large Language Model Evaluation M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Banu Diri, Savaş Yıldırım, Öner Aytaş. Electronic resource (date of use: 01.04.2025): <https://arxiv.org/html/2501.00593v2>

3. Süni intellektin linqvistik problemləri. M.Mahmudov, Bakı, “Elm və təhsil” 2024, s. 376.

4. Turkish sentiment analysis: A comprehensive review. Ayşe Berna ALTINEL GİRGIN, Gizem GÜMÜŞÇEKİÇÇİ , Nuri Can BİRDEMİR. Sigma J Eng Nat Sci, Vol. 42, No. 4, pp. 1292–1314, August, 2024 s. 1294. Electronic resource (date of use: 01.04.2025): <https://sigma.yildiz.edu.tr/storage/upload/pdfs/1722849273-en.pdf>