GƏNC TƏDQİQATÇILARIN III BEYNƏLXALQ ELMİ KONFRANSI

Fidan Əliyeva

Nərgiz Hüseynova

Nicat Mürsəli

Calal Rəsulzadə

Mündəricat

Bizim müasir zamanımızda həyatımızı asanlaşdırmaq üçün fərqli sahələrdə alqoritmləri öyrənərək müxtəlif növ maşınları düzəltməyə və inkişaf etdirməyə kömək edən bir neçə texnoloji irəliləyişlər var. Məsələn biz artıq developerlər tərəfindən yaradılmış müxtəlif dünya dillərinin morfoloji və leksik analizlərini apara bilən çoxsaylı alqoritmlər tapa bilərik və bu alqoritmlər bizim kimi akademik yazılar yazan insanların işlərini bir xeyli asanlaşdırır. Lakin çox təəssüf ki, belə bir alqoritm bizim dilimiz üçün hələ də yaradılmamışdır. Problematik məsələ hazırki Azərbaycan linqvistikanın morfoloji analizinin çatışmazlığındadır və bu problemi həll etmək üçün Azərbaycan dilindəki hər bir sözün morfoloji analizini təşkil edə biləcək heç bir əlçatan mənbə mövcud deyil. İngilis, rus, alman və s. kimi dillərdə uzun akademik dokument və məqalələr yazan insanlar hazırda mövcud olan demək olar ki mükəmməl dəqiqliyə malik olan müasir proqram təminatlarının koməyi ilə işlərini xeyli asanlaşdıra bilirlər. Lakin əvvəl də qeyd olunduğu kimi bu bizim milli dilimizdə bu demək olar ki imkansızdır və hazırda mövcud olan proqram təminatını hələ mükəmməldən çox uzaqdır.

Çox təəssüf ki, texnologiyaların bu qədər inkişaf etmiş olduğu bu müasir dünyada Azərbaycan dilində məqalə və ya elmi iş yazan insanlar yazılarını inkişaf etdirmək üçün kitablar və lüğətlər kimi artıq çox geridə qalmış vasitələrdən istifadə etmək məcburiyyətində qalırlar. Bu da onların işlərinin keyfiyyətinə və inkişaf müddətinə mənfi təsir göstərir. Bizim layihəmizin başlıca məqsədi bu problemi həll etmək və Azərbaycan dilini digər dünya dilləri ilə eyni səviyyəyə qaldırmaqdır. Potensial proqramlardan biri də söz korreksiyası proqramıdır, hansı ki bizim daxil etdiyimiz alqoritmlərlə öz işini daha yaxşı və dəqiq icra edə biləcək və proqramı istifadə edən hər kəsin bütün ehtiyaclarını ödəyəcəkdir. Hədəf kütlə kimi isə esse və məqalə yazanları, jurnalistləri və ya bu tipli akademik məsələlərlə əlaqəli olan hər kəsi göstərmək olar, hansı ki onlara bizim veb-səhifəmizdən mütləq şəkildə yararlana biləcəklər. Bir çox problemi həll etmək üçün alqoritmimiz bir çox üstünlüyə malikdir, buna görə də bizim komandamız Azərbaycan dilində vəzifələri yerinə yetirmək üçün bu istiqamətdə hərəkətə

keçməyə qərar verdi. Belə ki, Azərbaycan dilinin morfoloji təhlili heç bir mənbə tərəfindən, o cümlədən, Microsoft Word-ün redaktə funksiyası və ya digər proqramlar da daxil olmaq şərti ilə tamamlanmadığından, biz bu problemi həll etmək qərarını aldıq.

Əgər bilmək istiyirsinizsə, niyə məhz bu layihəni seçilib və nə dərəcədə əhəmiyyətlidir, bunun üçün ilk növbədə, dilin nə qədər önəmli olduğundan danışaq. Bildiyiniz kimi, hər bir xalqın milli-mənəvi dəyərini yaşadan onun dilidir və ana dilimizin işlənməsi və inkişaf etdirilməsi, biz gənclərin vətənə olan borcudur. Tarixə nəzər salsaq, müstəqilliyimizin ilk ilində - 1991-ci ildə latın qrafikalı Azərbaycan əlifbası qəbul olundu. Lakin, bu qərarın qəbul olmasına baxmayaraq, bütün sənədləşmələr hələ də öz dilimizdə aparılmırdı. Yalnız ümummilli liderimiz Heydər Əliyevin fərmanından sonra ana dilimizin tətbiqi həyata keçirildi və nəticədə bütün yazılı sənədləşmələr latın qrafikası ilə aparıldı. Bundan başqa, ümummilli liderin 2001-ci ildə imzaladığı fərmana əsasən, avqustun 1-i Azərbaycan Respublikasında Azərbaycan Əlifbası və Azərbaycan dili Günü elan edilmişdir. Bir sözlə, Azərbaycan dilinin dövlət dilinə çevrilməsi və diplomatiya aləminə yol açması ümummilli lider Heydər Əliyevin apardığı siyasət ilə bağlıdır.

Bir-birinin ardınca imzalanan fərmanlar, qəbul edilən qaydalar dövlətimizin siyasi həyatında, yazı mədəniyyətimizin tarixində mühüm hadisə oldu. Buna görə də biz məhz belə bir mövzuda layihə seçərək ana dilimizə olan məhəbbətimizi büruzə verməyə çalışdıq.

Məlumdur ki, Azərbaycan dilinin təhlili dedikdə, ilk olaraq, onun morfoloji analizi başa düşülür. Morfologiya bölməsində, əsasən nitq hissələrindən bəhs olunur, hansı ki, ümümi qrammatik əlamətinə görə fərqlənən söz qruplarına deyilir. Azərbaycan dilində olan nitq hissələri 2 qrupa bölünür: əsas və köməkçi nitq hissələri. Əsas nitq hissələri leksik mənaya malik olur, müvafiq suala cavab verir və sintaktik vəzifə daşıyır. Həmin nitq hissələri əşyanın adını, əlamət və keyfiyyətini, miqdarını, hərəkətini, hərəkətin əlamətini bildirir. Əsas nitq

hissələri isim, sifət, say, əvəzlik, fel və zərfdir. Digər qrupa, yəni köməkçi nitq hissələrinə isə qoşma, bağlayıcı, ədat, modal sözlər və nida aiddir. Bu nitq hissələri müstəqil leksik mənaya malik olmur, heç bir suala cavab vermir və buna görə də cümlə üzvü ola bilmirlər.

Bizim hal-hazırda üstündə işlədiyimiz proqram bütün nitq hissələrini əhatə edir; həm əsas nitq hissələrini, həm də köməkçi nitq hissələrini. İndiyə kimi ölkəmizdə bu qədər söz bazasını əhatə edən heç bir proqram olmamışdı. Yalnız Azəri NLP (http://nlp.jsoft.ws/) bu tipdə iş görmüşdü, hansı ki, ancaq və ancaq feilin formalarını/ şəkilçilərini özündə əks etdirir. Halbuki, bizim layihə demək olar ki, Azərbaycan dilində olan bütün verilmiş sözlərə uyğun olan şəkilçiləri tapır, verilən sözdən mümkün olan bütün başlanğıc formalarını tapır. Bunun üçün biz Azərbaycan dilinin lüğətində olan bütün sözləri öz proqramımızın bazasına yükləmişik. Əlavə olaraq, əgər müəyyən bir mətni proqramımıza daxil etsək, o bu zaman, mətndə olan hər bir sözün başlanğıc formasını, hansı nitq hissəsi olduğunu və tərkibindəki şəkilçiləri təyin edəcək.

Bir neçə illərdir ki, dünyanın bir neçə dilləri üçün tədqiqatçılar müxtəlif alqoritmlər tətbiq etmək üçün saysız hesabsız araşdırmalar edibdir. Bu alqoritmləri Azərbaycan dilində də etmək üçün biz digər dillər üçün hazırlanmış alqoritmlərlə bağlı daha dərindən araşdırmalar etdik və onların iş prinsipini anlamağa çalışdıq. Onların güclü və zəif nöqtələrini anlamaq və başa düşmək bizə Azərbaycan dilində daha mükəmməl və işlək alqoritm qurmağa köməklik edirdi. Hər birimizin bildiyimiz kimi, Azərbaycan dili sözlərin zənginliyi və hər hansı bir sözün bir neçə formasını yaratması onu dünyanın indiyə qədər gəlmiş keçmiş ən zəngin dillərindən biri edir. Bizim əlimizdə olan məlumatları əsasən xəbər saytlarından, kitablardan və digər vəsaitlərdən istifadə olaraq əldə etmişik və bu sözlərin və cümlələrin sayı bir milyona yaxındır. Bizim bunu etməyimizdə səbəb bizim alqoritmin dilimizin sözlərindən istifadə edib daha çox söz yaradıb bizim sistemə əlavə etmək olubdur. Bu prosesi başlatdığımız zaman əlimizdə olan

söz bazası çox olmadığından və Python proqramlaşdırma dilində “ə” hərfi olmadığından problemlərlə qarşılaşırdıq amma bu problemləri uğurla aradan qaldırmağı da bacardıq.

Bizim yaratdığımız proqram üçün biz bütün nitq hissələrini əlavə etdik və gözəl nailiyyətlər əldə etdik. Bizim bundan sonra edəcəyimiz şeylər də çoxdur və buna sözlərin daha çox mümkün formalarını çıxarıb onların doğruluğunu yoxlamaq və bunu cümlələrə tətbiq etmək olacaqdır.

Bir çox dünyaca məşhur alimlər, ingiliscə desək “Data Scientists” özlərinin böyük həcmdə olan vaxtlarını məlumatı təmizləmək və onları işləyə biləcəkləri formaya gətirmək olur. Bir çox alim də hər hansı bir proyektdə işin ən önəmli cəhəti həmin məlumatı təmizləmək olur və bu ən başlanğıc addımdır. Bizim qarşımıza çıxan ən mühüm problem isə bizdəki məlumat toplusunun içərisində bir sıra lazımsız şeylərin olması idi. Məsələn, bizdəki Excel fayllarının içində Javascript, rusca və ərəbcə sözlər, səhv yazılmış sözlər və digər problemlərlə qarşılaşırdıq. Bu səbəbdən biz “Data Cleaning”, yəni əlimizdə olan məlumat toplusunu təmizləmək mühüm idi. Bunu üçün komandamızın bir üzvü, Nicat Mürsəli, özünün Python proqramlaşdırma dilində yazdığı kodlarla bizim əlimizdəki məlumatı təmizləmək öhdəliyi verildi.

Bəs belə bir sual yaranır: “həmin məlumat toplusunu necə təmizləmək mümkündür?”. Bunun üçün bir sıra alqoritmlər var və biz bir neçəsini qeyd etməyə çalışacağıq. İlk öncə məlumat toplusundan nələr silinməlidir onlar təyin olunmalıdır. Əsasən “eyni cümlələr” və “lazımsız cümlələr” seçilib məlumat bazasından silinir. Eyni cümlələr dedikdə bir neçə məlumatları birləşdirdikdə eyni cümlələr olma ehtimalı çoxdur və bunun üçün onlar silinməlidir. Lazımsız cümlələr dedikdə isə, xəbər saytlarında olan xarici dillərdə (əsasən rus, ərəb, ingilis dillərində) yazılmış cümlələr nəzərdə tutulur.

Sözün morfoloji cəhətdən təhlil etmək üçün biz xüsusi alqoritm yaratmışıq. Bu alqoritm Söz və Şəkilçi siniflərindən istifadə edib sözü düzgün təhlil edir və bütün mümkün cavabları qaytarır.

Şəkilçi sinifinin hər bir nümayəndəsinin tərkibində şəkilçinin adı, mümkün formaları, və özündən sonra gələ biləcək bütün şəkilçilər üçün məlumat strukturları yaradılmışdır. Hər bir şəkilçi yalnız özündən sonra gələ biləcək şəkilçilər barəsində məlumatlıdır. Şəkilçini sözə artırmaq üçün alqoritm şəkilçinin müvafiq formasını seçir, ehtiyac olduğu halda bitişdirici samiti (n, y, s) silir və siniflin yeni yaranmış nümayəndəsini qaytarır. Nümunə olaraq, şagird sözünə birinci növ şəxs şəkilçisini (-yam, -yəm) gəldikdə, saitlərin ahənginə əsasən şəkilçinin birinci forması (-yam) seçilir və y bitişdirici samiti düşür.

Alqoritmimiz digər mühüm obyekti olan söz sinifinin atributları sözün başlanğıc forması, nitq hissəsi və şəkilçilərdir. Bunu da qeyd etmək lazımdır ki, sözün şəkilçiləri dedikdə yalnız və yalnız sözün başlanğıc formasına birbaşa artırıla bilən şəkilçilər nəzərdə tutulur.

İstənilən sözü morfoloji təhlil etmək üçün proqram həmin sözün başlanğıc forması rolunda iştirak edə bilən bütün sözləri və onların nitq hissələrini lüğətdən seçir. Seçilmiş hər bir söz üçün bütün mümkün şəkilçi növləri sözə artırılır. Yeni yaranmış sözlərdən təhlil olan sözə bərabər olan söz tapıldıqda həmin söz massivə əlavə olunur. Hesablama əməliyyatları bitdikdən sonra massiv qaytarılır. Qaytarılmış sözlərin içərisində, tərkibində olan şəkilçilər və onların ardıcıllığı barəsində məlumat var.

Sözü düzgün şəkildə morfoloji təhlil etmə qabiliyyətinə malik olduğuna görə, alqoritm Azərbaycan dilində olan istənilən mətni asanlıqla lemmalaşdıra bilir.