Pedaqoji Universitetin Xəbərləri. Riyaziyyat və təbiət elmləri seriyası – 2020, C. 68, №1, s. 244-251
Transactions of Pedagogical University. Series of mathematical and natural sciences – 2020, V. 68, №1, pp. 244-251

UOT 004

G.Y.Məlikov, A.Ə.Hacıyev
Azərbaycan Dillər Universiteti
melikaxmed@mail.ru, a.haciyev58@mail.ru

MƏTNİN KOMPÜTER TƏHLİLİNİN BİR ÜSULU BARƏDƏ VƏ ONUN ALQORİTM VƏ PROQRAMI

***Açar sözlər**: mətn, təhlil, ixtisarlar, cümlə, konkotenasiya, alqoritm, durğu işarəsi*

Məqalədə daxilində ixtisarların istifadə olunduğu təbii dildə yazılmış mətnin kompüter vasitəsilə təhlilinin bir üsuluna baxılır. Bu məqsədlə mətn əvvəlcə boşluq işarələrinin köməyi ilə sözlərə (formal sözlərə) ayrılır və sözlər massivi tərtib olunur. Sonra isə sözün sonunda gələn durğu işarələrindən (əgər varsa) və əvvəlcədən hazırlanmış ixtisarlar massivindən istifadə etməklə, sözlər massivinin elementləri bir-bir araşdırılaraq həmin sözün ixtisar olub-olmaması müəyyənləşdirilir və bundan asılı olaraq cümlələrin reallığa uyğun gələn konkotenasiyası təmin olunur.

Г.Ю.Меликов, А.А.Гаджиев

ОБ ОДНОМ МЕТОДЕ КОМПЬЮТЕРНОГО АНАЛИЗА ТЕКСТА И ЕГО АЛГОРИТМ И ПРОГРАММА

***Ключевые слова:** текст, анализ, сокращения, предложения, конкатенация, алгоритм, знак препинания*

В статье рассматривается метод компьютерного анализа текста, написанного на естественном языке, в котором используются сокращения. Для этого текст сначала делится на слова (формальные слова) с помощью пробелов, составляется массив слов. Затем элементы массива проверяются один за другим с использованием знаков препинания в конце слова (если есть), и заранее подготовленный массив сокращений, чтобы определить, является ли слово сокращенным, и зависимо от этого соответственно обеспечивается реалистическая конкатенация предложений.

G.Y.Melikov, A.A.Hajiyev

ABOUT A METHOD OF COMPUTER ANALYSIS OF TEXT AND ITS ALGORITHM AND PROGRAM

***Keywords:** text, analysis, contractions, sentences, concatenation, algorithm, punctuation mark*

The article deals with a method of computer analysis of a text written in a natural language in which contractions are used. For this purpose, the text is first

divided into words (formal words) with the help of spaces and array of words is compiled. Then using punctuation marks at the end of the word (if any) and a pre prepared array of contractions, the elements of the array are examined one by one and determined whether the word is contracted or not and accordingly a realistic concatenation of the sentences is provided.

Təbii dildə verilmiş mətnlərin emalı ilə bağlı istənilən süni intellekt məsələsinin həlli mətnlərin kompüter vasitəsilə təhlil və sintezi mərhələlərindən keçir. Kompüter mətnlərinin təhlili ilə əlaqədar kifayət qədər işlər görülsə də [1-3], "Azərbaycan dilinin Orfoqrafiya Normaları"nda qəbul olunmuş qaydalarla əlaqədar olaraq, bu məsələyə yenidən qayıtmaq zərurəti yaranır. Belə ki, [3]-də punktuasiya – durğu işarələrinin (".", "!", "?" və s.) köməyii lə verilmiş mətnin cümlələrə ayrılması məsələsinə (başqa sözlə, kompüter mətninin cümlə səviyyəsində realizəsinə) baxılmışdır. İstinad olunan mənbədə mətnin kompüterlə təhlili aşağıdakı üsulla aparılır. Əvvəlcə mətnin uzunluğu, başqa sözlə, onu təşkil edən simvolların sayı müəyyənləşdirilir. Sonra isə dövrü prosesin köməyi ilə mətndən simvollar bir-bir ayrılaraq cümlələri ayıran punktuasiya işarələri ilə (ayırıcı işarələrlə) tutuşdurulur. Mətndən ayrılaraq götürülən simvollar birinci cümlə üçün nəzərdə tutulan dəyişənə birləşdirilir (konkotenasiya edilir). Müqayisə ayırıcı işarə ilə üst-üstə düşənə qədər davam etdirilir. Sonra həmin qayda ilə ikinci, üçüncü və s. cümlələr formalaşdırılır. Proqramda həmin dəyişənlər birölçülü massivin elementləri olaraq götürülür. Proses mətndəki bütün simvolların bir-bir ayrılaraq nəzərdən keçirilməsi ilə axıra qədər davam etdirilir və nəticədə mətnin bütün cümlələrini özündə saxlayan birölçülü mətn tipli cümlələr massivi yaranmış olur.

Lakin [3]-də cümlə daxilində də həmin punktuasiya işarələrinin işlənə biləcəyi nəzərə alınmamışdır. Həqiqətdə isə, qəbul etmək lazımdır ki, bəzən, cümlənin daxilində sadalanma intonasiyası ilə yazılan sözlərdən sonra yekunlaşdırma məqsədilə işlədilən ixtisarlarda ("və s." yaxud "və b."), izah məqsədilə işlədilən qısaldılmış sözlərdə (məs.), söz birləşməsi formasında olan mürəkkəb şəxs adlarının ilk hərflərindən ibarət ixtisarlarda hərflər arasında nöqtə işarəsi qoyulur (M.Ə.Sabir). Bunlar əsasən dilimizin orfoqrafiya normalarındakı tam və yarımçıq ixtisarlarla bağlı olur. Tam ixtisarlar ümumi isimlərə aid olduqda kiçik hərflə yazılır: *metr – m , cild – c, və sairə – və s., və başqaları – və b.* Yarımçıq ixtisarlar aid olduqları sözlərə uyğun olaraq, böyük və kiçik hərflərlə, mürəkkəb adların tərkib hissələri isə bitişik yazılır: *akademik – akad., doktor – dok., dosent – dos., professor – prof.* və s. Göründüyü kimi, bəzi ixtisarlardan sonra (*metr – m , cild – c*) punktuasiya işarələri yazılmır. Həmin ixtisarlar mətnlərin kompüterlə təhlili prosesində heç bir əlavə çətinlik yaratmır.

Təbii dildə hazırlanmış mətndə punktuasiya işarələri ilə yekunlaşan ixtisarların hər biri ayrılıqda işlənilməli və tərtib olunan alqoritm və proqramda nəzərə alınmalıdır. İlk baxışda elə görünə bilər ki, ixtisardan sonra gələn söz böyük hərflə başlayarsa, mətnin yeni cümləsi başlayır. Əslində isə həmişə belə olmur. Fikrimizi daha yaxşı əsaslandırmaq üçün aşağıdakı mətn fraqmentinə diqqət yetirək:

ictimai mahiyyət daşıyan satiraları onun ümumi yaradıcılığında müstəsna mövqe tutur. Onu öz dövrünün böyük realist şairi kimi tanıdan "Yerdəkilərin göyə şikayət etmələri", "Dəli şeytan", "Məkri-zənan", "Bəlx qazisi və xarrat", "Müctəhidin təhsildən qayıtması", "Elmsiz alim", "Alim oğul ilə avam ata", "Qafqaz müsəlmanlarına xitab" və s. satiralarıdır.

Mətnlərin cümlələr səviyyəsində təhlili üçün [3]-də təklif olunmuş üsuldan istifadə etsək, onda nümunə kimi götürdüyümüz yuxarıdakı mətnin 1-ci cümləsində S.Ə.Şirvaninin sözündəki S. və Ə. simvollar ardıcıllığı, həmçinin həmin mətnin 2-ci cümləsinin "s." ixtisarından sonrakı hissəsi (cari cümlənin sonuna kimi) müstəqil cümlə olaraq qəbul olunacaq.

Digər bir mətn parçasına baxaq: "Sual əvəzlikləri ilə əmələ gələn sual cümlələrində məntiqi vurğu əsasən sual bildirən sözlərin üzərinə düşür; məs.: Polkovnik, bu saat hərəkəti idarə edən kimdir?" (C.Cabbarlı).

Əgər bu mətn parçasının da kompüter təhlilini [3]-də təklif olunmuş üsuldan istifadə edərək aparsaq, "məs." ixtisarından sonra gələn hissə - ": Polkovnik, bu saat hərəkəti idarə edən kimdir?" müstəqil cümlə kimi nəzərdə keçirilməlidir. Bu isə düzgün olmayan nəticəyə gətirib çıxarar. Belə ki, kompüterlə təhlilin sonrakı mərhələlərində (cümlədən sözlərin ayrılması) ":"durğu işarəsi ayrıca söz kimi qəbul olunacaq ki, bu da düzgün deyil. Baxmayaraq ki, belə ixtisarları özündə saxlayan cümlələr mətndə üstünlük təşkil etmir, təbii olaraq, bu cür təhlillə razılaşmaq olmaz.

Vəziyyətdən çıxmaq üçün verilmiş mətndən əvvəlcə sözləri ayırıb sonra isə mövcud durğu işarələrini nəzərə alaraq həmin sözləri araşdırmaqla, onların köməyilə cümlələri sintez etmək olar. Nəzərə almaq lazımdır ki, mətndən sözlərin bir-bir ayrılması üçün yeganə vasitə "probel" ("boşluq") işarəsi olacaq, ona görə də probellərlə əhatələnən istənilən simvollar ardıcıllığı ("s.", "b.", "və", "k." və s.) da müstəqil söz kimi götürüləcək. Araşdırma dedikdə, söz kimi qəbul olunan simvollar ardıcıllığının (formal sözün) axırıncı simvoluna diqqət yetirilməsi nəzərdə tutulur. Bu, ona görə lazımdır ki, mətndən söz olaraq ayrılan fraqment həmin sözlərdən sonra gələn durğu işarələri ilə birgə götürülür (əgər sözdən sonra durğu işarəsi gələrsə). Təhlil prosesində mətn daxilində işlənən ixtisarların problem yaratmasına baxmayaraq, durğu işarələri, ondan əvvəl bə sonra gələn sözlər və onların hansı registirlə daxil edilməsi cümlələrin normal olaraq formalaşdırılması üçün əsas vasitədir.

Qeyd etmək yerinə düşər ki, mətnin kompüter vasitəsilə yığılması prosesində, durğu işarələri birbaşa sözün axırıncı hərfindən sonra daxil edilir. Əgər bu norma pozularsa, işdə təklif olunan üsulun köməyi ilə bu çatışmazlıq aradan qaldırılır. Belə ki, təcrid olunmuş formada daxil edilən durğu işarələri təhlil prosesində ayrı-ayrı "sözlər" kimi götürülsə də sintez prosesinin cüzi də olsa ləngiməsinə baxmayaraq son nəticədə cümlələrin formalaşdırılması düzgün həyata keşirilir. Bunu digər yazılış normalarına da şamil etmək olar. Məsələn, yuxarıda nəzərdən keçirdiyimiz mətn fraqmentində "S.Ə.Şirvani" sözünə diqqət yetirsək görərik ki, müxtəlif mənbələrdə bu söz fərqli şəkildə mətnə daxil edilir: S.Ə.Şirvani və S.Ə.Şirvani kimi. Yuxarıda toxunduğumuz kimi nöqtə işarəsindən sonra boşluq işarəsi daxil edilmiş variantda "S." və "Ə." müstəqil formal sözlər kimi qəbul olunmasına baxmayaraq, son nəticədə cümlələrin formalaşdırılması düzgün həyata keşirilir. Nöqtə işarəsindən sonra boşluq işarəsi daxil edilməmiş variantda isə "S.Ə.Şirvani" tam söz olaraq götürülür və konkotenasiya prosesində əməliyyatların sayının azalmasına səbəb olur. Bu nümunə bir daha göstərir ki, mətn normallaşdırılmamış [4] şəkildə olsa da, alqoritm və ona uyğun hazırlanmış proqram düzgün nəticəyə gətirib çıxaracaq. Bu isə istifadə olunan alqoritmin üstünlüyüdür.

Bunu "məs." ixtisarı üzərində nəzərdən keçirək. Adətən "məs." ixtisarından sonra ":" işarəsi və həmin işarədən sonra ya bir-birindən vergüllə ayrılan sadalanan məlumatlar, ya da izah məqsədi daşıyan mətn fraqmenti gəlir. Dilimizin orfoqrafiya normalarına görə *məsələn* sözünün ixtisarı cümlə daxilində "məs.:" kimi yazılmalıdır. Əgər səhvən "məs. :" kimi daxil edilərsə, alqoritm həmin çatışmazlığı aradan qaldırır. Belə ki, baxdığımız variantda "məs." və ":" probellərlə əhatələndiyi üçün ayrı-ayrı sözlər (formal sözlər) kimi qəbul olunur, konkotenasiyanın nəticəsi olaraq tərtib olunan cümlə real mətndəki vəziyyəti əks etdirəcək.

İndi isə yuxarıda qeyd etdiyimiz üsulla mətndən ayrılmış sözlərin (formal sözlərin) köməyi ilə cümlələrin formalaşdırılması prosesinin vacib məqamlarına diqqət yetirək.

Nümunə kimi nəzərdən keçirilən və digər ixtisarları özündə saxlayan mətnlərə diqqət yetirməklə, ixtisarlardan sonra gələn söz kiçik hərflə başlayarsa, həmin söz əvvəlki cümlənin davamı kimi qəbul olunmalıdır. Bu fikrin təsdiqi olaraq S.Ə.Şirvani haqqında yuxarıda nümunə kimi istifadə etdiyimiz mətn fraqmentinin ikinci cümləsinə diqqət yetirmək kifayətdir. Cümlənin ortasında gələn "b. k." ixtisarı haqqında da həmin fikri söyləmək olar.

Lakin unutmaq olmaz ki, "və s." və "b.k." ixtisarları cümlənin sonunda da gələ bilər. Təbii ki, yeni gələn cümlə böyük hərflə başlayacaq.

Beləliklə, tam əminliklə deyə bilərik ki, "və s." və "b.k." ixtisarlarından sonra kiçik hərflərlə gələn mətn cari cümlənin davamını, böyük hərflərlə gələn mətn isə yeni cümlənin başlandığını göstərir.

Formal təhlildə əgər ixtisardan sonra böyük hərflə başlayan söz gələrsə, bu halda birqiymətli fikir söyləmək olmaz. Məsələn, belə bir cümləyə diqqət yetirək: "Azərbaycan dilinin orfoqrafiya lüğətinin 2004-cü il nəşrinin hazırlanmasında akad. A.Axundovun böyük zəhməti olmuşdur." Bu cümlədə "akad." ixtisarından sonra gələn şəxs adının böyük hərflə başlamasına baxmayaraq ixtisardan sonra gələn hissəni ayrı cümlə kimi qəbul etmək olmaz. Əgər həmin cümlədə "akad." ixtisarının əvəzinə "prof." ixtisarını da işlətmiş olsaq, yenə də yuxarıdakı fikri söyləyə bilərik. Əgər həmin söz *akad., dok., dos., prof.* və s. sözlərindən hər hansı biri ilə üst-üstə düşərsə, həmin ixtisardan sonra gələn böyük hərflə başlayan söz cümlə kimi formalaşdırılacaq əvvəlki hissəyə birləşdirilməlidir.

Bu mülahizələrdən sonra ümumi bir fikrə gələ bilərik ki, cümlələrin tərtibi üçün mətndən ayrılaraq götürülən sözlərin (formal sözlərin) konkotenasiyası prosesində ixtisarları fərdi formada emal etmək lazımdır. Ona görə də təhlil prosesində cümlələrin reallığı adekvat əks etdirməsi üçün ixtisarlar və onların dilimizin orfoqrafiya normalarına uyğun düzgün emalı mütləq nəzərə alınmışdır.

Bu mülahizələrdən sonra təbii dildə hazırlanmış mətnin kompüterlə təhlilinin alqoritmini aşağıdakı kimi vermək olar:

1) Emal olunacaq mətnin daxil edilməsi. Bu məqsədlə mənimsəmə və ya selection.text operatorundan, həmçinin Selection.Whole Story əmrindən istifadə etmək olar. Selection.text operatorundan istifadə etməklə proqramın daha praktiki olmasını təmin etmək, başqa sözlə proqramın mətninə müdaxilə etmədən mətn redaktorunun pəncərəsindən emal olunacaq mətni əvvəlcədən seçməklə proqramın işini təmin etmək olar. Mətni seçmədən, başqa sözlə bütün mətnin emalını təmin etmək məqsədilə Selection.Whole Story əmrindən istifadə etmək olar. Selection.text və Selection.Whole Story əmrləri proqramın ilkin verilənlərdən asılılığını aradan qaldırır.
2) Mətnin uzunluğunu, başqa sözlə həmin mətndəki simvolların sayını müəyyənləşdirilir.
3) Punktuasiya işarələrinin köməyi ilə mətndəki cümlələrin təxmini sayını müəyyənləşdirməli. Bu saya ixtisarların sonundakı işarələr də daxil olduğu üçün mətndəki cümlələrin real sayından böyük olacaq. Cümlələrin sayını müəyyənləşdirmək üçün dövrü prosin köməyi ilə mətnin simvolları bir-bir mətndən ayrılaraq punktuasiya işarələri ilə tutuşdurulur. Bu say imkan verir ki, proqramın tərtibində yaddaşa qənaət məqsədilə dəyişən ölçülü massivdən istifadə edək.

4) Boşluq işarələrinin (probellərin) mövqelərindən ibarət bir ölçülü massiv tərtib olunur. Bu massivdən istifadə edərək bir ölçülü sözlər massiv tərtib olunur. Sözlər massivinin yaradılmasının əsasında mətndəki hər bir sözün hər iki tərəfdən boşluq işarəsi ilə əhatə olunması durur.
5) Sözlər massivinin elementlərindən istifadə etməklə cümlələr yaradılır. Bu məqsədlə sözlər massivinin ilk elementindən başlayaraq dövrü şəkildə onun bütün elementləri nəzərdən keçirilir və cümlələri bir-birindən ayıran punktuasiya işarələri (!.?) rast gəlinənə qədər konkotenasiya prosesi davam etdirilir və əvvəlcə birinci cümlə, sonra ikinci cümlə və s. formalaşdırılır. Proses sözlər massivinin bütün elementləri nəzərdən keçirilib realizə olunana qədər davam etdirilir. Cümlələr formalaşdırıldıqca bir-bir cümlələr massivinin müvafiq elementlərinə mənimsədilir.
6) İstənilən cümləni ekrana çıxarmaq üçün dialoq qurmalı.

Alqoritmə qədərki təsvir hissəsində qeyd etdiyimiz kimi konkotenasiya prosesində əsas diqqəti cümlənin yaradılmasında iştirak edən müstəqil leksik vahidlərin ixtisarlardan fərqləndirilməsinə yönəltmək lazımdır. Bu məqsədlə də hər bir sözün (formal sözün) sonuncu simvolunun durğu işarəsi olub-olmamasını yoxlamaq və sözün son simvolu durğu işarəsi olarsa, onda emal olunan sözün durğu işarəsinə qədər olan hissəsini, sonrakı sözün ilk hərfini araşdırmaqla cümlənin sonunu bildirən durğu işarəsi ilə ixtisarlardan sonra gələn işarələri bir-birlərindən fərqləndirmək lazımdır.

Yuxarıda verilmiş alqoritmə uyğun olaraq Visual Basic dilində tərtib olunmuş proqram aşağıdakı kimi olar:

```
sub 'cumlenin tehlili
  Dim metn As String
  Dim n, k, l, m As Integer
  Dim t As Boolean
  metn = selection.text
   n = Len(metn)
  k = 0
  For i = 1 To n
  If Mid(metn, i, 1) = " " Then k = k + 1
  Next i
  ' Mətndəki cümlələrin sayıj
i = 1: j = 0
While i <= n
h = Mid(metn, i, 1)
t = (h = ".") Or (h = "!") Or (h = "?")
If t = True Then j = j + 1
i = i + 1
```

```
        Wend
        ReDim nn(k) As Integer, sozler(k + 1), cumle(j) As String
         l = 1: m = 1
         While l <= k
         nn(l) = InStr(m, metn, " ")
        m = nn(l) + 1
         l = l + 1
         Wend
        sozler(1) = Mid(metn, 1, nn(1) - 1)
         For i = 2 To k
        sozler(i) = Mid(metn, nn(i - 1) + 1, nn(i) - nn(i - 1) - 1)
        Next i
        sozler(k + 1) = Mid(metn, nn(k) + 1)
        zen = "!?."
        jj = 1: cumle(jj) = " "
        For i = 1 To k + 1
         krit = Right(sozler(i), 1)
         krit1 = InStr(1, zen, krit)
         t1 = (krit1 = False) Or ((krit = ".") And (Asc(sozler(i + 1)) > 90))
         t2 = ((krit = ".") And (sozler(i) = "mes."))
         t3 = ((krit = ".") And (sozler(i) = "akad.")) Or ((krit = ".") And (sozler(i)
= "dok."))
         t4 = ((krit = ".") And (sozler(i) = "prof.")) Or ((krit = ".") And (sozler(i)
= "dos."))
                    t5 = ((krit = ".") And (Len(sozler(i)) = 2)) And (Left(sozler(i), 1)
        <> "s") And    (Left(sozler(i), 1) <> "b")
         t = t1 Or t2 Or t3 Or t4 Or t5
         If t = True Then cumle(jj) = cumle(jj) + " " + sozler(i): GoTo 10
         cumle(jj) = cumle(jj) + " " + sozler(i)
          jj = jj + 1
        10   Next i
         cumle(jj) = cumle(jj) + " " + sozler(k + 1)
        'İxtiyari cümlənin ekrana çıxarılması
        m=inputbox("Hansı cümləni ekrana çıxarmalı?")
        i=val(m)
        MsgBox cumle(i)
        End Sub
```

ƏDƏBİYYAT

1. *Mahmudov M.* Kompüter dilçiliyi. Bakı: Elm və təhsil, 2013, 356 s.
2. *Fətullayev Ə.B.* Azərbaycan-ingilis maşın tərcüməsi sistemi üçün rəqəmsal modelləşdirmə metodunun işlənib hazırlanması və tətbiqi: Texnika elm.nam. dis. avtoref. Bakı, 2006, 20 s.
3. *Məlikov G.Y.* Tətbiqi dilçilik məsələlərinin MS Office-də həlli. / G.Y.Məlikov, V.B.Müslümov, A.H.Novruzov. Bakı: UniPrint, 2012, 256 s.
4. *Məlikov G.Y., Səmədova S.A.* Kompüter mətninin normallaşdırılması barədə / İnformasiya sistemləri və texnologiyalar. Nailiyyətlər və perspektivlər. Beynəlxalq elmi konfransının materialları. Sumqayıt: SDU, 15-16 noyabr, 2018, s.418-419

Redaksiyaya daxil olub 29.05.2020