UOT 004.82 DOI: 10.25045/jpit.v12.i1.05

Qurbanova Ə.M.
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
afruz1961@gmail.com

MƏTNLƏRDƏN TERMİNLƏRİN AVTOMATİK ÇIXARILMASI METODLARININ ARAŞDIRILMASI VƏ MÜQAYİSƏLİ ANALİZİ

Daxil olmuşdur: 19.10.2020 Düzəliş olunmuşdur: 26.10.2020 Qəbul olunmuşdur: 09.11.2020

Məqalədə mətnlərdən terminlərin avtomatik çıxarılmasının beş metodu tədqiq olunmuş və onların müqayisəli analizi verilmişdir. Mətnlərdən terminlərin çıxarılmasının ümumi məqsədi xüsusi sahənin əsas lüğətinin təyin edilməsidir. Terminlərin ənənəvi olaraq əl ilə çıxarılmasından fərqli olaraq avtomatik çıxarılması vaxt aparan bu işi sadələşdirmək üçün kompüterləşdirilmiş bir vasitədir və termin-namizədlərin əvvəlcədən müəyyənləşdirilməsinin avtomatlaşdırılmasına yönəlib. Hazırkı dövrdə bir çox sahələrdə (leksikoqrafiya, terminşünaslıq, informasiya axtarışı və s.) emal olunmalı informasiyanın həcminin artım dinamikası termin və açar sözlərin avtomatik seçilməsi məsələsini xüsusilə aktual edir. Təbii dilin emalı sahəsində qurulan qaydaları təqdim edən mətnlərdən terminlərin avtomatik çıxarılması üçün bir çox fərqli yanaşma və sistem işlənmişdir. Mətnlərdən terminlərin avtomatik çıxarılmasının müxtəlif altməsələləri – korpus kolleksiyası, vahid birliklər, termin və variantların müəyyən olunması və keyfiyyətin qiymətləndirilməsi qaydası təqdim olunmuşdur. Müəyyən predmet sahəsi üçün mətnlərdən terminlərin avtomatik çıxarılmasına tətbiqi yanaşma verilmişdir. "İnformasiya texnologiyaları problemləri" və "İnformasiya cəmiyyəti problemləri" jurnallarının məqalələrinin korpusu üzərində eksperiment aparılmışdır. Ekspert və formal qiymətləndirmə metodikası təklif olunmuş, terminlərin avtomatik çıxarılması metodlarının müqayisəli qiymətləndirilməsinin nəticələri verilmişdir.

Açar sözlər: *terminlərin avtomatik çıxarılması, təbii dilin emalı, korpus kolleksiyası, linqvistik yanaşma, statistik yanaşma.*

Giriş

Mətnlərdən terminlərin çıxarılmasının ümumi məqsədi xüsusi sahənin əsas lüğətinin təyin edilməsinə xidmət edir. Terminlərin ənənəvi əl ilə çıxarılması terminoloq tərəfindən yerinə yetirilir, o, terminə görə termin-namizədləri (TN) (*ing. termin candidate, TC*) siyahıya alır, sonra isə təsdiq edilmiş terminlərin yekun siyahısını almaq üçün predmet sahəsi üzrə ekspertlə məsləhətləşir. Ancaq sürətlə dəyişən dünyada texniki söz ehtiyatının daim artması ilə xidmətin əl ilə yerinə yetirilməsi, yeni texnoloji sahələrin əsas lüğətlərinin ənənəvi dəstəklənməsi, indeksləşməsi və təsviri əməktutumlu işdir. Terminlərin avtomatik çıxarılması (TAÇ) (*ing. Automatic Term Extraction, ATE*) ilk növbədə vaxt aparan bu işi sadələşdirmək üçün kompüterləşdirilmiş bir vasitə kimi nəzərdə tutulmuşdur. Hal-hazırda TAÇ termin-namizədlərin əvvəlcədən müəyyənləşdirilməsinin avtomatlaşdırılmasına yönəlib.

Terminlərin avtomatik çıxarılması "terminologiyanın çıxarılması", "terminologiyanın əldə edilməsi", "terminlərin tanınması", "qlossarinin çıxarılması", "terminlərin identifikasiyası" kimi də tanınır [1]. TAÇ mətn korpuslarının kompüter analizinə əsaslanır, terminlərin əl ilə çıxarılması üçün bir sıra üstünlükləri təklif edir. Belə ki, TAÇ korpusdakı dəlillərə əsaslanan kompüterləşmiş agenti özünə daxil edir, mütəxəssisin mətni əl ilə araşdırması ehtiyacını aradan qaldırır və avtomatlaşdırılmış agent üçün uyğun olan TN-ni əvvəlcədən seçən ilk filtr rolunu oynayır. TAÇ-ın bu üstünlüklərinə baxmayaraq, kompüterlər üçün semantikanın tam avtomatik modelləşdirilməsi hələ də mümkün deyil. Ona görə də terminin statusunun son təsdiqlənməsi predmet sahəsi üzrə mütəxəssis tərəfindən əl ilə yerinə yetirilməlidir.

Mətnlərdən termin və açar sözlərin seçilməsi kitabxana işində, leksikoqrafiyada və terminşünaslıqda, həmçinin informasiya axtarışında meydana çıxır. Hazırkı dövrdə bu sahələrdə emal olunmalı informasiyanın həcminin artım dinamikası termin və açar sözlərin avtomatik seçilməsi məsələsini xüsusilə aktual edir. Bu cür seçilmiş söz və ifadələr terminoloji resursların yaradılması və inkişafı üçün istifadə oluna bilər, həmçinin sənədlərin effektiv emalı – indeksləşmə, referatlaşma, klassifikasiya üçün istifadə oluna bilər.

Mətnlərdən terminlərin avtomatik çıxarılması

TAÇ keçən əsrin 90-cı illərinin əvvəllərindən başlayaraq Təbii Dilin Emalı (*ing. Natural language processing, NLP*) və informasiya axtarışı sahələrində öz təsdiqini tapmışdır [2]. TAÇ ardıcıl yerinə yetirilən bir sıra altməsələlərdən ibarətdir. Aşağıdakı altməsələləri ayırmaq olar:

1. Korpus kolleksiyası – xüsusi sahə korpusunun tərtib edilməsi. Terminlərin çıxarılması üçün ziddiyyətli yanaşmalardan istifadə edilərsə, ümumi dil korpusu da tələb olunur. TAÇ prosesində istifadə olunan metodların tələblərindən asılı olaraq korpus əvvəlcədən emal olunmalıdır: nitq hissələrinin işarələnməsi, fraqmentlərə parçalanması və ya tam sintaksis təhlil və s.;
2. Vahidlərin aşkarlanması (*ing. Detection of units*) – Vahidliyin (UnitHood) aşkarlanması. Çoxsözlü vahidi (*ing. Multiword Unit, MWU*) təşkil edən və bir konseptual vahidə aid olan linqvistik elementlərin identifikasiyası;
3. Terminlərin aşkarlanması (*ing. Detection of terms*) (*ing. TermHood*) – verilən predmet sahəsi üçün həqiqi termini özündə təqdimetmə ehtimalı nöqteyi-nəzərindən çıxarılan vahidləri ranqlaşdıran və ya klassifikasiya edən metod;
4. Termin Variantların (*ing. Term Options*) aşkarlanması – eyni bir predmet sahəsinin spesifik konsepsiyanın müxtəlif dil tətbiqlərinin aşkarlanması;
5. Qiymətləndirmə (*ing. Evaluation*) – predmet sahəsi üzrə ekspert tərəfindən terminin ənənəvi çıxarılması ilə müqayisədə avtomatik çıxarılmasının keyfiyyətinin qiymətləndirilməsi qaydası.

Terminin çıxarılması öz məqsədindən çox terminologiyanın idarə edilməsi sahəsində olan digər məsələlərin yerinə yetirilməsinə xidmət edir. Yuxarıda göstərilən altməsələlərin hər birinin mühümlüyü və dəqiq interpretasiyası TN-nin siyahısının sonradan istifadəsi ehtimalından çox asılıdır.

TAÇ-da üç praktiki tətbiqi ayırmaq olar [3]:

1. Terminoqrafiya. TN-nin siyahısı terminoloji lüğətlərin tərtib edilməsi və ya elektron terminoloji verilənlər bazası üçün giriş verilənləri kimi istifadə olunur. Beləliklə, eyni bir konsepsiyaya aid olan terminlərin variantlarının aşkarlanması və qiymətləndirilməsi mühüm altməsələlərdəndir.
2. Tərcümə dəstəyi – TN-nin siyahısı tərcümə layihəsi üçün xüsusi bir lüğət olaraq fəaliyyət göstərir və tərcüməsinin tapılması zəruri olan naməlum sözləri müəyyənləşdirmək məqsədi daşıyır. Ardıcıl olaraq tərcümə edilməli olan çoxsözlü birləşmələrin aşkarlanması bu tətbiqdə vacib olsa da, terminologiyanın aşkarlanması və yoxlanılması cüzi rol oynayır.
3. İnformasiya axtarışı – TN-nin siyahısı sənədlər toplusunu indeksləşdirmək üçün əsasdır, buna görə istifadəçilər bir predmet sahəsi ilə əlaqəli mövzular üzrə topluları sorğu edə və ya nəzərdən keçirə bilərlər. TN-nin aktuallığı istifadəçilərin axtarış sorğularına əsasən müəyyən edilir. Sənədlər toplusunun tərtib edilməsi informasiya axtarışı üçün mühüm aspektdir.

Yuxarıda qeyd edilən 5 altməsələnin hər birinə nəzər salaq.

Korpus kolleksiyası

TAÇ-ın istənilən metodu xüsusi sahənin mətn korpusuna əsaslanmalıdır. TAÇ-ın bəzi proqram təminatlarında ixtisaslaşmış sahə olduqca məhduddur və analiz üçün müvafiq olan mətnlər sonlu və dəqiq müəyyən olunmuş dəst əmələ gətirir. Məsələn, təşkilat və ya şirkət öz daxili

terminologiyasının inventarizasiyasını aparmaq istəyirsə, təşkilat və ya şirkətin terminoloqa təqdim etdiyi mətn korpusu təbii olaraq sənədlər toplusuna uyğun olur. Amma layihə sahənin terminologiyasının bütövlükdə analizinə istiqamətlənirsə, məsələn, "Biliklər iqtisadiyyatı", "Biometriya" və s., korpuslar toplumuna mütləq həmin sahədən mətnlər seçilib daxil edilir. TAÇ-a yönəlmiş korpusların onlayn tərtibinə bəzi yaxın yanaşmalara qısa şəkildə baxaq.

Başlanğıc korpuslar və terminlər (*ing. Bootstrapping Corpora And Terms, BootCaT*) sistemi İnternetdə sorğu mətnlərinin axtarışı və onların vahid korpusda birləşdirilməsi prosesini avtomatlaşdırır. BootCat proqram dəsti əl ilə seçilən ilkin terminlərdən başlayır, bu terminlər böyük ehtimalla təklif edilən xüsusi sahəni təmsil edir [4]. İlkin terminlər predmet sahəsi üçün spesifik olan korpusun ilkin analizindən də alına bilər. Birinci mərhələdə ilkin terminlərin təsadüfi kombinasiyaları sorğu qismində ümumi təyinatlı axtarış sisteminə (məsələn, Google) göndərilir, məqsəd predmet sahəsi üçün spesifik olan Resursların Vahid Göstəricisi (*ing. Uniform Resource Locator, URL*) – ünvanın əldə olunmasıdır. URL-ünvanlı veb-sayt yüklənir, onun tərkibi ilkin terminlərin siyahısı ilə müqayisə olunur, həmin terminlərin təklif olunan predmet sahəsinə mənsubluğuna əmin olmaq üçün onlar yoxlanılır. Yeni əlavə olunmuş mətnlər TAÇ üçün təqdim olunur, terminlərin ilkin siyahısı isə əlavə terminlərlə doldurulur. Terminlərin bu genişlənmiş siyahısı URL axtarışının ikinci fazası üçün giriş olur. Korpus kifayət qədər böyük olana qədər və ya yeni URL, ya da termin alınana qədər prosedur təkrar olunur. 2011-ci ildən artıq bu yanaşma İnternetdə axtarışa qədər genişləndirildi. Toplanan bu onlayn-korpuslar TAÇ prosesində digər addımlar üçün ilkin verilənlərdir.

Vahidlərin aşkarlanması

Vahidlər "sintaqmatik kombinasiyaların gücü və sabitlik dərəcəsi" kimi təyin olunur [5]. Vahid birliklərin aşkarlanması TAÇ prosesini əhatə edən birinci məsələ idi, 1980-ci illərin sonu 1990-cı illərin əvvəllərində fənn kimi təsdiq olundu. Bunun bir neçə səbəbi var.

Birincisi, təsdiq olunur ki, çoxsözlü vahidlər əsasən texniki sahələrdə geniş yayılmışdır. Bu səbəbdən, onlar TAÇ üçün ən vacib məqsəd hesab olunur. Təsdiq olunub ki, 85% TN iki və daha çox sözdən ibarət texniki ifadələr kimi təyin olunur [6]. İkincisi, bir neçə sözdən ibarət olan terminlər onların analoqu olan bir sözdən ibarət terminlərə nisbətən semantik olaraq daha konkretdir Fransalı alimlər Bourigault Didier və Christian Jacquemin təsdiq edirlər ki, "bir sözdən ibarət olan terminlər daha çoxmənalı və ümumidirlər", bir neçə sözdən ibarət olan terminlər "sahə üzrə daha incə anlayışları təqdim edirlər" [7]. 90-cı illərin əvvəllərində asanlıqla əldə edilə bilən geniş ümumi dil korpusları olmadığından, tezliyə əsaslanan metodlar bir sahə daxilində və xaricində olan korpusları müqayisə edərək termin birləşmələrini seçmək üçün asanlıqla tətbiq edilə bilmirdi. Digər tərəfdən, bir neçə sözdən ibarət blokların aşkarlanması maraqlı tərəflərin təqdim etdiyi texniki sənədlərə əsaslanır. Bu tip tədqiqatlarda termin vahid kimi nəzərdə tutulur [8].

Linqvistik yanaşmalar

TAÇ prosesinə linqvistik yanaşmaya əsasən, çoxsözlü terminlər müəyyən morfosintaksis nümunələrdən (şablonlardan) istifadə edir. Bu nümunələr müəyyən edir ki, ifadə həqiqətən linqvistik vahiddir, əgər elədirsə, deməli, termin-namizəddir. Korpusun əvvəlcədən emalı üçün daha güclü üsulların ortaya çıxması linqvistik informasiyanı yarıavtomatik aşkarlama prosesinə daxil etməyə imkan verir: Nitq hissəsi (*ing.Part-of-speech, POS*) avtomatik olaraq mətnin böyük həcmini emal edir və sözləri öz POS-teqləri ilə təchiz edir [9]. Bu, TAÇ prosesinə POS nümunələri ardıcıllığından ibarət olan sintaksis nümunələri daxil etməyə imkan verir.

Predmet sahəsi üzrə mütəxəssis linqvistik meyarlar əsasında uyğun sintaksis nümunələri müəyyən edir. Bununla yanaşı, etibarlı nümunələrin seçilməsində praktiki mülahizələr də rol oynayır. Tətbiq edilən nümunələrin sayı TAÇ-ın dəqiqliyinə birbaşa təsir edir. Xüsusilə, dəqiqlikdəki fərqlər, açıq və ya qapalı sinif filtrlərin seçilməsini şərtləndirir. Açıq sinif filtrləri bir çox əlavə POS elementlərinin istifadəsinə imkan verir. Bunun üstünlüyü var, belə ki, TAÇ çox

sayda termin-namizədləri daxil edir, çatışmazlığı isə çoxlu sayda səhv namizədlərin alınmasıdır. Ona görə də açıq sinif filtrlərin təqdim etdiyi siyahının əl ilə düzəlişi daha əməktutumludur. Qapalı sinif filtrləri icazəli şablonların seçimini məhdudlaşdırır, bu, dəqiqliyin yüksəlməsi üstünlüyünü verir. Cədvəl 1-də riyaziyyat sahəsindən nümunələrlə POS-filtr şablonları göstərilmişdir.

Cədvəl 1

POS-filtr şablon nümunələri

Sifət İsim	Tətbiqi riyaziyyat
İsim İsim	Reqressiya əmsalları
İsim Sifət İsim	Gauss təsadüfi dəyişən
Sifət İsim	Xətti funksiya
Sifət İsim	Paylama funksiyası
İsim İsim İsim	Sinif ehtimal funksiyası
Sifət İsim	Sərbəstlik dərəcələri

POS-qeydiyyatlı korpusun namizədləri birbaşa müqayisə edilir və sintaksis nümunələrin son seçiminə uyğunlaşdırılır. Şablonların istifadəsi ümumi, gündəlik sözlər və texniki sözlər arasında fərq qoymadığı üçün istənilməyən sözlər siyahısı texniki olmayan namizədləri çıxarmaq üçün ikinci dərəcəli filtr kimi istifadə olunur. Ümumi korpusdan əldə edilən tez-tez rast gəlinən sözlərin siyahısından tərtib edilən belə bir siyahı qadağan olunmuş sözlərin siyahısı adlanır. Bu prosedur dəqiqliyi artıra bilər, lakin səhvən etibarlı TN-ləri də silə bilər.

Statistik yanaşmalar

Statistik yanaşmalar bir neçə sözdən ibarət terminlər üçün tipik olan iki xassədən istifadə edir, linqvistik informasiya tələb etmir:

- Bir neçə sözdən ibarət olan termin nisbətən sabit ifadələri özündə təqdim edir;
- Onlar nisbətən böyük tezliklə rast gəlinir.

Bir neçə sözdən ibarət olan terminlərin əksəriyyəti sözlərin nizamını dəyişmədən yüksək səviyyədə sintaqmatik sabitliyi nümayiş etdirdiyi üçün statistik yanaşma *n*-qramın analizi ilə, başqa sözlə, baza dil strukturunu nəzərə almadan sözlərin kəsilməz ardıcıllığı ilə məhdudlaşa bilər. *N*-qramın vahidliyi onların korpusunun bəzi tezlik funksiyaları ilə ölçülür. Bu da linqvistik analiz tələb etmir, yalnız kifayət edəcək ölçüdə korpus tələb edir [10].

Kollokasiya ölçüləri

Baza tezlik informasiyası bir yerdə rast gəlinən sözlərin sayılması yolu ilə korpusdan əldə edilir. İki və daha çox sözün ardıcıl olaraq üst-üstə düşməsi onu göstərir ki, bu sözlər bir-birinə aiddirlər və çoxsözlü termin əmələ gətirirlər. Lakin işlənməmiş tezlik sayğacları, yuxarıda göstərilən POS-filtrləri kimi, yalnız linqvistik filtrlərlə birlikdə istifadə olunur [9]. Statistik yanaşmada rastgəlmə tezliyi müəyyən dərəcədə informativliyə əsasən ölçülür. Bu kollokasiya ölçüləri bir söz birləşməsinin tezliyini həmin birləşməni təşkil edən fərdi sözlərin tezliyi ilə müqayisə edir. İki fərdi sözün (məsələn, "yeni" və "şey") müntəzəm olaraq eyni işlənməsi çox təəccüblü deyil, iki tez-tez birlikdə rast gəlinməyən sözün (məsələn, "maşın" və "mühərrik") müntəzəm olaraq birlikdə işlənməsi göstərir ki, bu birləşmə sabit ifadə və bəlkə də termin ola bilər. Formal olaraq, kollokasiya ölçüləri iki sözün fərdi tezliyi nəzərə alınmaqla müşahidə edilən üst-üstə düşmə tezliyinin təsadüfi gözləmə tezliyindən nə qədər fərqləndiyinin sayını müəyyən etməyə imkan verir. Cədvəl 2-də oyuncaq nümunəsi üçün göstərilir ki, "maşın" və "mühərrik" sözlərinin müşahidə edilən tezliyi (60) təsadüfi gözləmə tezliyindən (24.76) xeyli yüksəkdir. Təsadüfi gözləmə tezliyi (E_{ij}) "maşın yağı" (258) tezliyi (R_i) və "maşın" (96) tezliyinin (Cj) ayrılıqda hasili və bu hasilin korpusun (1000) ölçüsünə (N) bölünməsi ilə hesablanır:

$$E_{ij} = \left(R_i * C_j\right) / N \quad (1)$$

Cədvəl 2

"Maşın mühərriki" kollokasiyasının: a) müşahidə olunan tezliyi; b) gözlənilən tezliyi

a)

	mühərrik	mühərrik	cəm
maşın	60	36	96
maşın	198	706	904
cəm	258	742	1000

b)

	mühərrik	mühərrik
maşın	24,76	71,23
maşın	233,23	670,76

Müşahidə olunan və gözlənilən tezliklər arasındakı uyğunsuzluğun kəmiyyət qiymətləndirilməsinin çoxlu üsulları və kollokasiya ölçüləri var. Tanınmış bir nümunə, terminlərin çıxarılması üzrə təcrübədə istifadə olunan X^2 statistik dəyəridir [11,12]. Ehtimal olunan cədvəlin bir-biri ilə əlaqəli sətir (r) və sütunlarında (c) müşahidə olunan (O_{ij}) və gözlənilən tezliklər (E_{ij}) arasındakı fərq aşağıdakı düstura uyğun olaraq hesablanır:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)}{E_{ij}} \quad (2)$$

Bu nümunə "maşın mühərriki" üçün X^2 74,7156 qiymətini verir. Korpusda hər bir söz birləşməsi üçün bu hesablamanın yerinə yetirilməsi vahid üzrə bütün söz birləşmələrini ranqlaşdırmağa və yalnız müəyyən həddi keçənləri seçməyə imkan verir. Vahid birləşmə üçün digər yerləşdirmə ölçülərinə t-göstəriciləri, loqarifmik ehtimal nisbəti [13] daxildir. Bu istiqamətdə müxtəlif alimlər müxtəlif vaxtlarda müəyyən təkliflər vermişlər. Manning və Schütze [14] ifadələrin ölçülməsinə girişi təklif etmişlər, Evert [15] və Wiechmann [16] daha geniş icmal və riyazi əsaslarını vermişlər. Qeyd etmək lazımdır ki, bütün bu qarşılıqlı rastgəlmə tədbirləri adətən linqvistik yanaşma ilə kombinə edilir və sonra yalnız linqvistik filtrdən keçən ifadələr üçün hesablanır.

Paradiqmatik dəyişiklik

Çöxsözlü terminlərin nisbətən sabit hissələri asanlıqla başqa sözlərlə əvəz edilə bilməz. Wermter və Hahn [17] ifadələrin vahidliyini təyin etmək üçün bu paradiqmatik modifikasiya xassəsindən istifadə edirlər. Linqvistik filtrləmənin başlanğıc mərhələsində əldə edilən bir neçə söz-namizədin hər birləşməsi üçün eyni uzunluğa malik bütün söz birləşmələrinin tezliyi toplanır və ən azı bir söz-namizədlə birlikdə istifadə edilir, eyni zamanda bir və ya bir neçə tərkib hissəsi digər sözlə əvəz edilir. Bu modifikasiya olunmuş versiyanın ümumi tezliyi sonradan bir neçə sözdən ibarət namizədin tezliyi ilə müqayisə olunur, bu da vahidlik üçün paradiqmatik dəyişikliyin *(ing. Paradigmatic Modifiability, P – Mod)* ölçülməsinə gətirir.

Leksik bağlı sözlər

Bir sıra yanaşmalar sözlərin daha uzun ardıcıllığının aşkarlanmasına yönəldilir, uzunluğu əvvəlcədən məhdudlaşdırmır və ya ismi ifadələr kimi əvvəlcədən təyin olunmuş POS nümunələrinə məhdudiyyət qoymur. Bu, bəzi sahələrdə (məs. hüquq) frazeoloji ifadələrin tərtibatında (məsələn: həqiqəti, bütün həqiqəti və yalnız həqiqəti söyləyəcəyinizə and içirsinizmi?) xüsusilə vacibdir. Biber və Konrad qeydiyyat üçün spesifik olan ifadələri analiz

edərkən belə uzun söz ardıcıllıqlarına leksik bağlı sözlər kimi müraciət etmiş, kriteriya olaraq seçilmiş milyon sözdə nisbi tezlikdən istifadə etmişlər [18].

Terminlərin aşkarlanması (TermHood)

1990-cı illərin sonlarında "terminlik" anlayışı TAÇ texnologiyasına daxil edildi, "sabit leksik vahidin müəyyən mövzuya aid anlayışla əlaqələndirilməsi dərəcəsi" kimi izah olunur [19]. Termin və vahid termin-namizədin ayrı-ayrı xüsusiyyətləri hesab olunur və vahidin termini ifadə etməsi mütləq deyil: "bir çox zaman" kimi çox sözdən ibarət ifadələr yüksək dərəcədə vahid birləşmədir, lakin istənilən ixtisaslaşmış sahədə aşağı terminliyə malikdir. Digər tərəfdən, bəzi ifadələrdə, məsələn, "premedikasiya" terminində bir neçə sözdən ibarət vahidlik olmasa da, o, tibb sahəsində mühüm termindir.

Terminliyi ölçmək üçün istifadə olunan birinci və sadə yanaşma, bir predmet sahəsinin daxili tezliyini verilən sahədəki TN-in əhəmiyyət göstəricisi və etibarlı bir termin olma ehtimalının göstəricisi kimi istifadə etməkdir. Bununla birlikdə, predmet sahəsinin daxili tezliyi terminlik ilə bir qədər əlaqələndirilsə də, xüsusən, daha uzun çoxsözlü vahidlər iştirak etdikdə – fərdi sözlərin və ya ümumi ifadələrin terminologiyası barədə qərar vermək üçün kifayət qədər məlumatlı deyil: əsas dilin söz və ifadələri ixtisaslaşmış və ya ixtisaslaşmamış istənilən korpusun tezlik elementlərinə aiddir, lakin terminoloji cəhətdən o qədər də maraqlı deyil.

Buna görə də ikinci yanaşma predmet sahəsi daxilində TN-nin paylanma xüsusiyyətlərini, daha dəqiq desək, müxtəlif sənədlər üzrə səpələnməsini nəzərdən keçirir.

Üçüncü yanaşma, TN-nin kontekst istifadəsinə nəzər salaraq, tezlikdən kənara çıxır.

Dördüncü yanaşma xüsusi olaraq birsözlü TN-lər üçün nəzərdə tutulmuşdur və TN-nin daxili morfoloji strukturunu təhlil edir.

Nəhayət, metodların beşinci ailəsi daxili və xarici məlumatları müqayisə edir.

Termin variantın aşkarlanması

Terminologiyaya klassik yanaşma termini predmet-yönümlü bir anlayış kimi təyin edir və termin linqvistik ifadə ilə fərdi münasibətə malikdir. Bununla birlikdə, terminlərin birmənalı ideal vəziyyəti onların dəyişməsinə, yəni anlayışın bir neçə linqvistik formanın köməyilə ifadəsinə görə daha da mürəkkəbdir. Fransa alimi Béatrice Daille təsdiq edir ki, "termin variant – orijinal terminlə semantik və konseptual əlaqəli bir ifadədir" [20]. Daille anlayışların 15-35%-inin bir-birinin variantı olduğunu bildirir [21]. TAÇ-ın alt məsələlərindən biri TAÇ prosesindən sonra termin variantlarının qeydiyyatı və qruplaşdırılmasıdır. Daille terminlərin variantlarının tipologiyasını təklif edir. Eynilə, Bourigault və Jacquemin fransız dili üçün FASTR sistemində termin-variantların aşkarlanması üçün sintaksis informasiyadan istifadə edən çevrilmə qaydalarından istifadə edirlər [7].

Tədqiq olunan metodlar

İxtiyari uzunluğa və struktura malik termintipli ifadələrin seçilməsi üçün istifadə oluna biləcək müxtəlif metodları tədqiq edək. Bu problemin çətinliyi ondadır ki, problemin həlli üçün statistik yanaşma o qədər də effektiv deyil: terminin uzunluğunun artması zamanı onun rastgəlmə tezliyi aşağı düşür, məhdud həcmli xüsusi korpusda termin az rast gəlinə bilər. İlkin ideya tədqiq olunan mətn/korpusda çox az, təsadüfən rast gəlinə bilən uzun terminlərin seçilməsi imkanı və çətinliyinin qiymətləndirilməsidir. Ona görə də metodlar terminlərin seçilməsinin yüksək dolğunluğu və aşağı dəqiqliyi ilə xarakterizə olunur.

İşdə müqayisəli analiz üçün beş metod seçilib və tətbiq edilib.

TF-IDF statistik tezlik metodu

$TF-IDF$ statistik tezlik metodu namizədlərin müəyyən bir sahə üçün mətn korpusunu təşkil edən müxtəlif sənədlərdə terminlərin bölüşdürülməsini nəzərdən keçirir. Güman edilir ki, hər mətndə tapılan söz və ya ifadələr o qədər də spesifik deyil və ehtimal ki, çox vaxt xüsusi korpusda

tapılan ümumi dil elementləridir. Digər tərəfdən, yalnız məhdud sənədlər altçoxluğunda rast gəlinən TN-nin konkret bir sahəyə aid olduğu güman edilir.

Prinsip ondan ibarətdir ki, əgər söz hər hansı bir sənəddə tez-tez rast gəlinirsə, bununla yanaşı digər sənədlər toplusunda hərdən rast gəlinirsə, bu söz həmin sənəd üçün böyük məna kəsb edir [22]. Sözün çəkisi həmin sözün sənəddə istifadəsinin sayı ilə mütənasibdir, digər sənədlər kolleksiyasında həmin sözün istifadə tezliyinə tərs mütənasibdir. *TF–IDF* metodu daha çox mətnin analizi və informasiya axtarışında istifadə olunur.

Maximal Length metodu

TAÇ üçün istifadə olunan ilk metodlardan biri olan maksimum uzunluqlu (*ing. Maximal Length, MaxLen*) ismi ifadələrin çıxarılması metodu [23]-də təsvir olunub. Bu məqsədlə LEXTER sistemi işlənilmişdir. LEXTER – mətnlərdən terminologiyanın çıxarılması üçün proqram paketidir. Hər hansı bir predmet sahəsi üzrə fransız dilində mətn LEXTER-ə daxil edilir və mümkün terminoloji vahidlərin siyahısı tərtib edilir, sonradan bu siyahı yoxlama üçün ekspertə təqdim edilir. Terminoloji vahidləri təyin etmək üçün LEXTER onların formalarını nəzərə alır və iki əsas mərhələni keçir: analiz və təhlil.

İlk mərhələdə LEXTER, mətnləri analiz etmək və maksimum uzunluqlu ismi ifadələri çıxarmaq məqsədilə sərhəd nişanlarını müəyyənləşdirmək üçün hazırlanmış qaydalar bazasından istifadə edir. İkinci mərhələdə LEXTER terminoloji vahidlər olan altqrupları ayırmaq üçün maksimum uzunluqlu bu ismi ifadələri analiz edir. Bu qaydaların LEXTER-in bütün qaydaları kimi, öz məhdudiyyətləri var və onların "səhv" tətbiq edildiyi hallar da olur. Qəbul olunan ekperimental yanaşmada bu səhvin riski nəzərə alınır və nəzarətdə saxlanılır, belə ki, hər bir qayda korpusdan ayrıca yoxlanılır və o, tətbiq olunduğu halların sayını yoxlayır, müəyyən edir ki, o LEXTER-ə düşür və ya yox. Burada prinsip çox sərt analiz qaydalarının daxil edilməsi deyil. Səhv analizlərin sayı həddən çoxdursa, bir çox hallarda məhsuldar olan qaydanı ləğv etmək lazımdır. Bu prinsip "nisbi sərtlik" adlandırılır, belə ki, texnoloqun həqiqi terminoloji vahidləri tapmaqdan çox ehtimal olunan vahidləri aradan qaldırması daha asan olacaqdır. Bu metodun həyata keçirilməsində durğu işarələri, sabit sözlər, fellər ayırıcı kimi nəzərdən keçirilir; həmin ayırıcılar arasındakı xətlər TN sayılır. Bu, nəzərdən keçirilən üsulların ən sadəsidir.

Hal-hazırda LEXTER müxtəlif korpuslarda müxtəlif formada terminoloji məhsulların hazırlanması üçün istifadə olunur. LEXTER biliyin əldə edilməsi üçün proqram təminatı olan SADE layihəsində istifadə olunub [24].

C-value metodu

Frantzi və digərləri çoxsözlü terminlərin avtomatik seçilməsi üçün *C-value* metodunu təklif edirlər [25]. *C-value* yanaşması əsas vurğunu statistik hissəyə etməklə, linqvistik və statistik informasiyanı birləşdirir. Linqvistik informasiya korpusun nitq hissələri teqindən, linqvistik filtrdən və dayan-koddan ibarətdir. Statistik informasiya termin-namizəd sətirinin statistik xüsusiyyətlərini ölçü formasında (*C-value*) birləşdirir, termin-namizədlərin çıxış siyahısını ranqlaşdırır. *C-value* adlanan terminlik dəyəri aşağıdakı kimi hesablanır:

$$C-Value(a)=\begin{cases}\log_2|a|*freg(a), \\ \log_2|a|*freg(a)-\dfrac{1}{P(T_a)}*\sum\limits_{b\in(T_a)}freg(b)\end{cases} \qquad (2)$$

burada a – termin-namizəd, $|a|$ - sözlərin sayında ölçülən ifadənin uzunluğu, $freg(a)$ – a-nın tezliyi, T_a – a-nı saxlayan ifadələr çoxluğu, $P(T_a)$ – a-nı saxlayan ifadələrin sayıdır.

(2)-dən görünür ki, termin-namizədin tezliyi və onun uzunluğu nə qədər böyükdürsə, onun çəkisi bir o qədər böyükdür. Əgər termin-namizəd çox sayda digər söz birləşmələrinə daxildirsə, onda onun çəkisi azalır.

k-factor metodu

K-factor metodu bir neçə sözdən ibarət terminlərin avtomatik çıxarılması üçün nəzərdə tutulub və *BootCaT* sistemində reallaşıb [26]. *BootCaT* veb-in tematik korpusunun avtomatik formalaşmasına xidmət edir. Korpusun qurulması ilkin terminlər dəstindən başlayır. Axtarış maşınına avtomatik sorğuların köməyi ilə ilkin terminlərdən ibarət sənədlər çıxarılır. Öz növbəsində bu sənədlərdən yeni birsözlü terminlər çıxarılır, həmin terminləri yenidən sorğu qismində istifadə etmək olar. Birsözlü terminlərin son korpusu və siyahısı çoxsözlü terminlərin çıxarılması üçün istifadə olunur. Bu metoda *C-value* metodunun sadələşmiş variantı kimi baxmaq olar: əgər daha qısa TN tam tərkibinə daxil olduğu daha uzun TN-ə nisbətən daha tez-tez rast gəlinirsə, daha uzun variant "əsas" sayılır. Axtarılan çoxsözlü terminlər $k * fq$ -dən yüksək tezliklə daha uzun çoxsözlü terminin hissəsi ola bilməzlər, burada k 0 və 1 arasında sabitdir, fq cari termin tezliyidir. Həmçinin, onlar $\left(1/k\right) * fq$ tezliyindən böyük daha qısa çoxsözlü terminləri ehtiva edə bilməz. Seçim terminlər tezliyinin nisbi həddi (k) ilə idarə olunur.

TERMS metodu

TERMS adlanan metod [27]-də təsvir olunub. Metodun ideyası əvvəlki iki metoda (*C-value, k-factor*) yaxındır – daha qısa termin daha uzun terminin tərkibində tez-tez rast gəlinirsə, ifadəni qurmaq. Amma digər metodlardan fərqli olaraq yalnız əlaqəli halların tezliyi deyil (söz bilavasitə bir-birinin ardınca düzülür), həm də *"pəncərədə"* (*Window)* əlaqəli rastgəlmə nəzərə alınır. Hər iterasiya zamanı, siyahının hər bir elementi üçün yaxın qonşuları və mətn qutusundakı qonşuları yadda saxlanılır. Uyğun cədvəllər qurulur, cütlüklərin rastgəlmə tezliyi hesablanır. Daha sonra, nəzərdə tutulur ki, elementlər cütlüyü eyni bir mətn qutusunda aşkarlandıqları halların yarısında yaxın qonşu kimi rast gəlinirsə, onda bu cütlük termin və ya terminin bir hissəsidir. Cütlük vahid elementə yapışdırılır, bu da terminin daha da genişlənməsinə imkan yaradır. Müəlliflər bu üsulla əldə edilmiş uzun termin nümunələrini misal gətirirlər: nəqliyyat vasitəsi sahiblərinin mülki məsuliyyətinin icbari sığortası haqqında qanun, yerli özünüidarəetmə orqanı. Bu tətbiqdə qutunun ölçüsü 9 sözdür. Yapışdırılan elementlərin meydana gəlməsi tezliyinə məhdudiyyət qoyulmasa, metod unikal (1 tezliyi ilə) etibarlı sözlərin zəncirlərini birləşdirəcək (yəni, *MaxLen* metodunun nəticələrini təkrarlayacaq).

Sintaksis analiz (Mətnlərin Avtomatik Emalı)

Məlumdur ki, terminlərin çoxu ismi ifadələrdir. Bu metod çərçivəsində TN qismində sintaksis analizatorun köməyilə seçilən ismi ifadələrə baxılır. Metod istifadə olunan analizator üzrə – Mətnlərin Avtomatik Emalı (*ing. Automatic Text Processing*) adını almışdır [28]. Sintaksis analizə əsaslanan metod baxılan digər metodlardan fərqli olaraq termin-namizədlərin tərkibində sözönlərinin olmasına imkan verir.

Qiymətləndirmə

TAÇ-ın son altməsələsi qiymətləndirmə mərhələsidir, terminlərin əl ilə çıxarılmasına nisbətən TAÇ metodunun nə dərəcədə yaxşı işlədiyi qiymətləndirilir. TN-lərin siyahısı sahə terminologiyasının lüğətinin qızıl standartına uyğun olaraq və ya xüsusi olaraq bir sahə mütəxəssisi tərəfindən və ya müəyyən bir layihədə iştirak edən terminoloq tərəfindən qiymətləndirilir [29]. Bu qızıl standarta və ya ekspert rəyinə əsasən TAÇ prosesi bir neçə üsulla qiymətləndirilir. TN siyahısının dəqiqliyi, təklif olunan bütün TN-lərin ümumi sayının içindən dəqiq müəyyən edilmiş terminlərin faizidir.

Sənədlər toplusu üçün əl ilə tərtib edilmiş terminlərin siyahısı və ya bir sahə üçün qızıl standartın tam bir lüğəti varsa, bu da rəyi hesablamaq üçün, yəni xüsusi korpusda aşkarlanan bütün terminlərdən müəyyən edilmiş terminlərin nisbətini hesablamaq üçün mümkündür. Ümumiyyətlə, yüksək dəqiqlik rəy vasitəsilə əldə edilir və əksinə. Təcrübədə TAÇ-ın layihələndirilməsi prosesində rəyə və ya yüksək dəqiqliyə üstünlük verilməlidir və çox vaxt sonuncuya üstünlük verilir.

Verilənlər və alətlər

Təcrübə "İnformasiya texnologiyaları problemləri" [30] və "İnformasiya cəmiyyəti problemləri" [31] jurnalları əsasında aparılmışdır. Korpusa 2017-ci ildən 2019-cu ilə qədər adıçəkilən jurnallarda nəşr olunan müxtəlif müəlliflərin informasiya cəmiyyəti və informasiya texnologiyaları üzrə 140 məqaləsi daxildir. Korpusun xarakteristikası: 358750 söz, stop-sözlər olmadıqda, 247320 söz. Terminlərin avtomatik emalı sistemi korpusda 51250 cümlə ayırıb (nəticələrə əsasən, cümlənin sonu üçün qayda çox sadədir: "nöqtə+aralıq" birləşməsi həmişə cümlənin sonu kimi təfsir edilir).

Nəticələrin formal qiymətləndirilməsi üçün "İnformatika" terminlər lüğətindən istifadə edilib. Lüğət təxminən 5000 girişi (sətir) özündə saxlayır. Hər sətir bir neçə yaxın termini saxlaya bilər: bir düymədə olan nöqtələrin sayı; bir düymədəki baytların sayı; bir düymədəki bitlərin sayı; bir düymədəki simvolların sayı.

Bütün terminlərə bərabərhüquqlu kimi baxılır (cəmi 5000 termin), lüğət terminlərinin uzunluğunun paylanması belədir: 1894 bir sözdən ibarət termin, 2316 iki sözdən ibarət termin, 538 üç sözdən ibarət termin, 156 dörd sözdən ibarət termin və 96 beş və daha çox sözdən ibarət termin. Qeyd etmək lazımdır ki, lüğətə abreviatura, xüsusi strukturlu terminlər, məs. latın hərfləri ilə yazılan, həmçinin mürəkkəb terminlər də daxildir. Korpus ayrı-ayrı məqalələrə bölünmədən bir bütöv sənəd kimi emal olunmuşdur.

Qiymətləndirmə metodikası

"Etalon siyahı" üzrə ekspert qiymətləndirmə ilə formal qiymətləndirmə kombinə edilir. Ekspert qiymətləndirməsi üçün hər bir metodun işinin nəticələrindən 100 namizəd götürüldü. *C*-value üçün çeşidlənmiş siyahıdan ən yüksək həddi (100), digər metodlar üçün siyahıda 100 sətir: 35 – üç sözdən, 35 – dörd sözdən, 30 – beş sözdən ibarət termin götürüldü.

Ekspert qiymətləndirməsi üçün əvvəlcə predmet sahəsinin qısa təsviri, həmçinin verilən sahə üçün bir neçə müsbət və mənfi termin nümunəsi təqdim olunur. Sonra ekspert sadə interfeysdən istifadə etməklə, siyahının hər bir elementi üçün "Verilən ifadə predmet sahəsinin terminidir?" sualına cavab verir. Ekspertin cavab variantları: "hə", "yox", "cavab verməyə çətinlik çəkirəm", "qismən" (təqdim olunan ifadə termini özündə saxlayır və ya daha uzun terminin hissəsidir). Ekspertə siyahı 10 ifadə olmaqla hissələrlə verilir. Hər bir TN həmin sahədən olan iki ekspertlə birlikdə müstəqil olaraq qiymətləndirilir. Güclü qiymətləndirmə halında hər iki ekspertin termin kimi tanıdığı ifadə termin sayılır; zəif qiymətləndirmə halında yalnız ekspertlərdən biri ifadəni termin kimi qiymətləndirmişdir.

Formal qiymətləndirmə "qısa", "uzun" və "orta" siyahı üçün aparıldı. Sonuncuya yalnız "uzun siyahıdan" vahiddən böyük tezliklə rast gəlinən sətirlər düşdü. İki tip formal qiymətləndirmə aparıldı: 1) səlis və 2) qeyri-səlis müqayisə. Birinci halda üç parametri nəzərə alırıq: 1) ayrılan terminlərin lüğət terminləri ilə dəqiq üst-üstə düşməsi, 2) lüğət terminlərinin seçilmiş ifadələrə daxil edilməsi, 3) seçilmiş ifadənin lüğətin daha mürəkkəb (dörd və daha çox söz) terminə daxil edilməsi. Qeyri-səlis qiymətləndirməyə normal formaya salınmış çoxsözlü ifadə kimi baxılır, iki sətirin yaxınlığını isə iki ifadədə üst-üstə düşən sözlərin sayının unikal sözlərin ümumi sayına nisbəti kimi təyin edirik: $sim(S_1, S_2) = |S_1 \cap S_2| / |S_1 \cup S_2|$. Qiymətləndirmədə $\geq 0{,}5$ yaxınlığı ilə lüğətdə heç olmasa bir terminin uyğun gəldiyi TN-nin sayı hesablanır. Bu ölçü üzrə yaxın terminlərə nümunə cədvəl 3-də göstərilmişdir.

Cədvəl 3

İki sətirin yaxınlıq dəyəri

Namizəd sətri	Lüğət termini	*sim*
qlobal informasiya cəmiyyəti	informasiya cəmiyyəti	0.67
informasiya kommunikasiya texnologiyaları	informasiya texnologiyaları	0.67
Süni intellekt	Süni intellekt	1.0
terminoloji informasiya sistemi	informasiya sistemi	0.67
maşın tərcüməsi	maşın tərcüməsi	1.0
şəbəkə fayl sistemi	şəbəkə əməliyyat sistemi	0.5
dinamik proqramlaşdırma dili	proqramlaşdırma dili	0.67
proqram mühəndisliyi	proqram təminatı lisenziyası	0.4
informasiya axtarış sistemi	axtarış sistemi	0.67
informasiya təhlükəsizliyi	informasiya təhlükəsizliyi	1.0

Qiymətləndirmənin nəticələri

Qiymətləndirmə üçün ekspertlərə təqdim edilən sətir nümunələri cədvəl 4-də verilir. 450 TN-dan ibarət qiymətləndirmə siyahısının (qısa siyahı) nəticələri cədvəl 5-də verilir. Ekspertlərin qiymətləndirməsi 53% üst-üstə düşdü. 25 qiymətdə (5,5%) ekspertlərin fikirləri bir-birinə əksdir: "termin" əvəzinə "termin deyil".

Cədvəl 4

Qiymətləndirmə üçün ekspertlərə təqdim edilən sətir nümunələri

MaxLen
milli telekommunikasiya və informasiya texnologiyaları sənayesi
təhlükəsizliyin təmin edilməsi
informasiya infrastrukturunun təkmilləşdirilməsi
Azərbaycan dilinin terminoloji verilənlər bazası
maşın təlimi alqoritmləri
mobil rabitə üçün qlobal sistem
ikinci yaddasaxlama qurğusu
ikiölçülü model
elektron kənd təsərrüfatı
böyük həcmdə şəbəkə trafiki
C-*value*
informasiya təhlükəsizliyinin təmini
Milli terminoloji informasiya sistemi
"eMotion Software" proqram məhsulu
informasiya texnologiyaları profilli təhsil
informasiya kommunikasiya texnologiyaları sektoru
proqram mühəndisliyi
müasir proqramlaşdırma dilləri
verilənlərin emalı mərkəzləri
qlobal informasiya cəmiyyəti quruculuğu
süni intellekt texnologiyaları
TERMS
proqram mühəndisliyinin koordinasiya komitəsi
Süni İntellektin İnkişafı Assosiasiyası
şəbəkə verilənlər bazası
tətbiqi proqramların sistem arxitekturası

infrastruktur mühitinin strukturunun deformasiyası
enerji təhlükəsizliyinin təmin olunması
texnologiya və elmi işləmələrin ixracına nəzarət
inklüziv inkişaf səviyyəsi
mətn tipli məlumatlarda emosional tonallığın analizi
enerjidən asılı olmayan yaddaş
k-factor
mətnlərin avtomatik emalı
terminlərarası semantik əlaqələr
kompüter şəbəkələrinin monitorinqi
şəbəkə təhlükəsizliyinin intellektual monitorinqi
AzScienceNet elm-kompüter şəbəkəsi
virtual yaddaş resursları
kontrafakt proqram məhsullarının istehsalı
antiplagiat sistemləri
İT-mütəxəssislər seqmenti
strukturlaşdırılmış sorğu dili
ATP
informasiya cəmiyyəti iqtisadiyyatı
faylların sonlu sayda səviyyələr arasında dinamik paylanması
informasiya resursları
şəbəkə texnologiyaları vasitəsi ilə piratçılıq
əmək bazarının konyukturası
kəmiyyət disbalansı göstəricisi
intellektual mülkiyyət hüquqlarının qorunması
elmin kommersiyalaşdırılması
insan resurslarının idarə olunması sistemləri
böyük həcmli elmi verilənlər

Cədvəl 5

"Qısa siyahı"nın qiymətləndirilməsinin nəticələri

Qiymət		*MaxLen*	*C-value*	*k-factor*	*TERMS*	*ATP*
Ekspert, "termin"	zəif	30	62	30	25	20
	ciddi	8	24	6	2	5
Ekspert, "qismən"	zəif	44	38	59	53	47
	ciddi	14	7	21	13	12
Formal	dəqiq	0	5	0	0	0
	Daxil etmə, qoşma	66	70	70	71	63
	Daxil olma, girmə	0	6	0	0	0
	Qeyri-səlis	6	35	8	8	8

Cədvəl 6-da "İnformasiya cəmiyyəti problemləri" və "İnformasiya texnologiyaları problemləri" jurnallarının korpuslarının emalının bütün nəticələrinin (uzun siyahı) müxtəlif metodlarla formal qiymətləndirilməsinin nəticələri verilmişdir. Cədvəl 7-də "orta siyahının" (korpusda minimum iki dəfə rast gəlinən sətirlərdən ibarətdir) formal qiymətləndirilməsinin nəticələri, cədvəl 8-də namizədlərin uzunluqları nəzərə alınmaqla qeyri-səlis qiymətləndirilməsinin nəticələri verilir.

Cədvəl 6

"Uzun siyahı"nın formal qiymətləndirilməsinin nəticələri

Qiymət	*MaxLen*	*C-value*	*k-factor*	*TERMS*	*ATP*
Siyahının ölçüsü	14 970	34 370	16 986	13 845	18 772
Dəqiq	23	34	27	14	33
Daxil etmə	10 309	23 322	11 663	9 300	10 579
Daxil olma	11	29	8	6	17
Qeyri-səlis	1 613	3 640	1 836	1 382	1 712

Cədvəl 7

"Orta siyahı"nın formal qiymətləndirilməsinin nəticələri

Qiymət	*MaxLen*	*C-value*	*k-factor*	*TERMS*	*ATP*
Siyahının ölçüsü	743	2 466	1 949	1 352	1 190
Dəqiq	10	20	21	13	18
Daxiletmə	492	1 643	1309	883	726
Daxilolma	4	15	5	3	6
Qeyri-səlis	150	501	420	267	196

Cədvəl 8

Terminə namizədlərin uzunluğu nəzərə alınmaqla "orta siyahı"nın qeyri-səlis formal qiymətləndirilməsinin nəticələri

Uzunluq	*MaxLen*		*C-value*		*k-factor*		*TERMS*		*ATP*	
	cəmi	*yaxın*	*cəmi*	*yaxın*	*cəmi*	*yaxın*	*cəmi*	*yaxın*	*cəmi*	*yaxın*
3 söz	597	118	1 963	409	1 609	342	1 075	208	987	160
4 söz	114	27	375	82	273	69	230	54	171	34
5 və daha çox söz	32	5	128	10	67	9	47	5	32	2
Cəmi	743	150	2 466	501	1 949	420	1 352	267	1 190	196

Nəticələrin analizinə əsasən, demək olar ki, müqayisə olunan metodlar ümumilikdə oxşar nəticələr verir. İfadəyə daxil olan terminləri nəzərə alan metodlar (*C-value*, *k-factor*) daha yaxşı nəticələr verir. Sintaksis analiz əsasında (*ATP*) ad qruplarının ayrılması daha zəif nəticə verir. Ekspert və formal qiymətləndirmənin nəticələrinin müqayisəsi (cədvəl 5) göstərir ki, formal metodlar TN-in böyük siyahısının müqayisəsi üçündür.

İnformatika üzrə lüğətdən terminlərin uzunluğunun paylanması, həmçinin uzun terminlərin (beşsözlü və s.) seçilməsi keyfiyyətinin aşağı düşməsi göstərir ki, keyfiyyətin yüksəldilməsi və tamlığın təmin olunması üçsözlü və dördsözlü terminlər üçün şablonların köməyi ilə mümkündür.

Nəticə

Terminlərin avtomatik çıxarılması təbii dilin emalı sahəsində qurulmuş bir qaydadır və bir çox fərqli yanaşma və sistem işlənmişdir. Buna baxmayaraq, hamısı üçün bir sıra təkrarlanan altməsələləri ayırd etmək olar: korpuslar toplusu, vahidlik, terminlik və termin variantın seçilməsi və qiymətləndirmə. İlk sistemlər terminləri müəyyənləşdirmək üçün yalnız linqvistik məlumatlardan istifadə edirdi, lakin iri korpuslardan terminlərin çıxarılması üçün tədricən daha çox və daha mürəkkəb statistik metodlar işləndi. Müasir sistemlərin əksəriyyəti hər iki məlumat növünü birləşdirən hibridlərdir [32].

Ekspert və formal qiymətləndirmələrin nəticələrini analiz etməklə (cədvəl 5) belə qənaətə gəlmək olar ki, müqayisə edilən metodlar oxşar nəticələr verir. *C-value* və *k-factor* metodları sintaksis analiz tələb edən metoda (*ATP*) nisbətən daha yaxşı nəticələr verir. İnformatika

lüğətindən olan terminlərin uzunluğunun paylanması, eyni zamanda uzun terminlərin (5 və daha çox söz) seçilməsi keyfiyyətinin kəskin pisləşməsi göstərir ki, keyfiyyətin yüksəldilməsi və tamlığın təmin edilməsi üçsözlü və dördsözlü terminlər üçün şablonların köməyilə mümkündür.

Böyük həcmdə tədqiqatlara baxmayaraq, yaxşı səviyyədə terminlərin avtomatik çıxarılmasına nail olan ümumi qəbul edilmiş bir standart yoxdur. Bu, həm terminlərin çıxarılması üçün nəzərdə tutulan xüsusi proqram təminatlarından, həm də konkret dil, predmet sahəsi və korpusdan asılıdır.

Ədəbiyyat

1. Heylen K., Hertog D. Automatic Term Extraction. In Hendrik J. Kockaert and Frieda Steurs (eds.) / Handbook of Terminology. John Benjamins Publishing Company, 2015, vol.1, pp.203–221.
2. Cabré M. Teresa, Estopà R., Vivaldi J. Automatic term detection: a review of current systems // In Recent Advances in Computational Terminology, edited by Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme, John Benjamins Publishing Company. Natural Language Processing, 2001, vol.2, pp.53–88
3. Gregor T. Making Term Extraction Tools Usable / In Proceedings of the Joint Conference of the 8th Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop. Dublin: European Association for Machine Translation, 2003. https://www.aclweb.org/anthology/2003.eamt-1.20.
4. Baroni M., Bernardini S. BootCaT: Bootstrapping Corpora and Terms from the Web / Proceedings of LREC 2004. Lisbon: ELDA, 2004, pp.1313–1316.
5. Kageura K. Computing the potential lexical productivity of head elements in nominal compounds using the textual corpus // Progress in Informatics, 2009, no.6, pp.49–56.
6. Nakagawa H., Tatsunori M. A simple but powerful automatic term extraction method / In Proceedings of the Second International Workshop on Computational Terminology, Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp.1–7.
7. Didier B., Jacquemin Ch. Term extraction + term clustering: An integrated platform for computer-aided terminology" / In Proceedings of the ninth conference on European Chapter of the Association for Computational Linguistics (EACL), Bergen, Stroudsburg, PA, USA: Association for Computational Linguistics, 1999, pp.15–22.
8. Chunyu K. Corpus tools for retrieving andderiving termhood evidence / In 5th East Asia Forum of Terminology, Haikou, China, 2002, pp.69–80.
9. Justeson J.S., Slava M.K. Technical terminology: some linguistic properties and an algorithm for identification in text // Natural Language Engineering, 1995, vol.1, issue 1, pp.9–27.
10. Patrick P., Dekang L. A Statistical Corpus-Based Term Extractor / In Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of intelligence: Advances in Artificial intelligence, edited by Eleni Stroulia and Stan Matwin, Lecture Notes In Computer Science, London: Springer-Verlag, 2001, vol.2056, pp.36–46.
11. Drouin P. Termhood: Quantifying the Relevance of a Candidate Term / Modern approaches to terminological theories and applications, 2006, pp.375–391.
12. Yutaka M., Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information // International Journal on Artificial Intelligence Tools, 2003, vol.13, issue 1, pp.157–169.
13. Dunning T. Accurate methods for the statistics of surprise and coincidence // Computational Linguistics, 1993, vol.19, issue 1, pp.61–74.
14. Manning Ch., Hinrich Sc. Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999, 720 p.
15. Evert S. The Statistics of Word Cooccurrences: Word Pairs and Collocations, PhD diss., 2004, 353 p. elib.uni-stuttgart.de/bitstream/11682/2573/1/Evert2005phd.pdf

16. Wiechmann D. On the Computation of Collostruction Strength: Testing Measures of Association as Expressions of Lexical Bias // Corpus Linguistics and Linguistic Theory, 2008, vol.4, issue 2, pp.253–290.
17. Wermter J., Udo H. Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms / In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2005, pp.843–850.
18. Susan C., Douglas B. The Frequency and Use of Lexical Bundles in Conversation and Academic Prose // Lexicographica, vol.20, issue, 2005, pp.56–71.
19. Kageura K., Umino Bin. Methods of automatic term recognition: a review // Terminology, International Journal of Theoretical and Applied Issues in Specialized Communication, 1996, vol.3, issue 2, pp.259–289.
20. Béatrice D. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology // In The Balancing Act: Combining Symbolic and Statistical Approaches to Language, edited by Philip Resnik and Judith L. Klavans, Cambridge, MA, USA: MIT Press, 1996, pp.49–66.
21. Béatrice D. Variations and application-oriented terminology engineering // Terminology, 2005, vol.11, issue 1, pp.181–197.
22. Salton G., Wong A., Yang Chung-Su. A vector space model for automatic indexing // Communications of the ACM, 1975, vol.18, pp.613–620.
23. Bourigault D. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases / Proc. of COLING-92, Nantes, France, August 23–28, 1992, pp.977–981.
24. Aussenac-Gilles, N., Bourigault D., Condamines A., Gros C. How can Knowledge Acquisition benefit from Terminology? / Proceedings of the 9th Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW'95), 1995, pp.11–16.
25. Frantzi K., Ananiadou S., Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method // Int J Digit Libr, 2000, vol.3, pp.115–130.
26. Bernardini S., Ferraresi A. Old needs, new solutions Comparable corpora for language professionals //, Publisher: Springer, Editors: Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, Pascale Fung, 2013, pp.303–319.
27. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических сочетаний по текстам предметной области / Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды пятой Всероссийской научной конференции, С.-Петербург, 2003, с.201–210.
28. Синтаксический анализ. Проект АОТ. http://www.aot.ru/docs/synan.html
29. Terryn A. R., Hoste V., Lefever E. A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structureand Translation Equivalents / Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, ELRA, 2018, pp.1803–1808. https://www.aclweb.org/anthology/L18-1284.pdf
30. "İnformasiya Texnologiyaları Problemləri" jurnalı. www.jpit.az
31. "İnformasiya Cəmiyyəti Problemləri" jurnalı. www.jpis.az
32. Francesco S., Velardi P. Termextractor: a web application to learn the common terminology of interest groups and research communities / In Proceedings of the 7th Conference on Terminology and Artificial Intelligence (TIA-2007), Sophia Antipolis, 2007, pp.85–94.

УДК 004.82
Гурбанова Афруз М.
Институт Информационных Технологий НАНА, Баку, Азербайджан
afruz1961@gmail.com
Сравнительный анализ методов автоматического извлечения терминов из текстов
В статье исследованы пять методов автоматического извлечения терминов, дан их сравнительный анализ. Общая цель извлечения терминов – это разработка базового словаря конкретной области. В отличие от традиционного ручного извлечения терминов, автоматическое извлечение − это компьютеризированный инструмент, упрощающий эту трудоемкую задачу и направленный на автоматизацию предварительного определения терминов-кандидатов. В настоящее время динамика роста объема информации, которая должна обрабатываться во многих областях (лексикография, терминология, информационный поиск и т.д.), делает вопрос автоматического выбора терминов и ключевых слов особенно актуальным. Для автоматического извлечения терминов из текстов, которые представляет собой правила, установленные в области обработки естественного языка, было разработано множество различных подходов и систем. Представлены различные подзадачи автоматического извлечения терминов – сборник корпусов, единство, определение терминов и вариантов, оценка системы. Приведен прикладной подход к автоматическому извлечению терминов из текстов для конкретной предметной области. Был проведен эксперимент по корпусу статей в журналах «Проблемы Информационных Технологий» и «Проблемы Информационного Общества». Предложена экспертная и формальная методика совместной оценки, приведены результаты сравнительной оценки методов автоматического вывода терминов.
Ключевые слова: *автоматическое извлечение терминов, обработка естественного языка, сборник корпусов, лингвистический подход, статистический подход, терминология.*

Afruz M. Gurbanova
Institute of Information Technology of ANAS, Baku, Azerbaijan
afruz1961@gmail.com
Comparative analysis of methods of automatic term extraction from texts
This article provides an applied approach to the automatic terms' extraction from the corpus for a particular subject area. Terms' extraction is commonly focused on the determination of the basic vocabulary of a particular field. Unlike the traditional manual terms' extraction, automatic extraction is a computerized tool to simplify this time-consuming task and is aimed at automating the pre-determination of term-candidates. Currently, the dynamics of the growth of the volume of information that must be processed in many areas (lexicography, terminology, information retrieval, etc.) makes the issue of automatic selection of terms and keywords especially relevant. Automatic Term Extraction is well-established discipline within Natural Language Processing and many different approaches and systems developed. Various sub-issues of automatic terms' extraction that is corpus collection, unity, definition of terms and variants, and system evaluation are presented. Five methods for automatic terms' extraction are studied and comparatively analyzed. An experiment is conducted on the corpus of articles included into the journals "Problems of Information Technology" and "Problems of the Information Society". An expert and formal joint assessment methodology is proposed, and the results of the comparative assessment of the automatic terms' extraction methods are presented.
Keywords*: automatic term extraction, Natural Language Processing, corpus collection, linguistic approaches, statistical approaches, termhood.*