Kompyuter leksikografiyasi va lingvistik ontologiyalar

UDK: 004.42:81`374

AVTOMATIK LUG‘ATLAR HOSIL QILISH ALGORITMI

Karimov Suyun Amirovich
suyun1950@mail.ru
Samarqand davlat universiteti, filologiya fanlari doktori
Qobilov Sami Saliyevich
kobsam@yandex.ru
Samarqand davlat universiteti, texnika fanlari nomzodi
Rabbimov Ilyos Mehriddinovich
ilyos.rabbimov91@gmail.com
Samarqand davlat universiteti, tayanch doktorant

Annotatsiya. Ushbu maqolada o‘zbek tilida yozilgan matnlardan avtomatik lug‘atlar hosil qilish hamda avtomatik lug‘atlar tuzish uchun algoritm va dasturiy ta’minot ishlab chiqish masalalari yoritilgan.

Kalit so‘zlar: *kompyuter leksikografiyasi, alifboli lug‘at, chastotali lug‘at, chappa (ters) lug‘at.*

Hozirgi paytda raqamli texnologiyalarning shiddat bilan rivojlanib borayotgani barcha sohalarda, jumladan tilshunoslik sohasida ham axborot texnologiyalari ahamiyatining oshib borishiga sabab bo‘lmoqda.

Ishlab chiqilayotgan zamonaviy texnika va texnologiyalar ko‘plab masalalarni tez, oson va yuqori sifatda bajarishga imkon bermoqda. Jumladan, lingvistik tadqiqotlarda, tilshunoslikka oid masalalarni hal etish bilan bog‘liq sohalarda ham raqamli texnologiyalardan foydalanish samarali natijalarga olib kelmoqda.

Bunday sohalardan biri kompyuter leksikografiyasi hisoblanadi. Kompyuter leksikografiyasi lug‘atlarni yaratish uchun matnli ma’lumotlarni qayta ishlash usullari va dasturlari majmuasini yaratish bilan shug‘ullanadi[1]. Ushbu jarayonda kompyuter lingvistikasining leksikografiya yo‘nalishi vazifalaridan biri bo‘lgan o‘zbek tilida yozilgan matnlardan avtomatik lug‘atlar hosil qilish masalasi muhokama qilinadi hamda bu masalani hal qiluvchi algoritm va dasturiy ta’minot ishlab chiqiladi.

Avtomatik lug‘atlar, bir tomondan, tabiiy tilda yozilgan matn tahlilining dastlabki bosqichida leksikografik va statistik kuzatishlar olib borish uchun muhim

hisoblansa, ikkinchi tomonidan esa, avtomatik lug‘atlar o‘zbek tilining axborot tavsifini berishga, so‘z turkumlari, morfologik qurilishini sintaktik tahlil qilishga, grafemalar statistikasi, grammatika va leksikaning axborot o‘lchovlarini yaratishga zamin bo‘ladi.

A. Nurmonov va B. Yo‘ldoshevlarning: “Hozirgi o‘zbek adabiy tilining chastotali lug‘atini yaratish juda murakkab muammo sanaladi, shuning uchun hozirgi kunga qadar tilimizning to‘la chastotali lug‘ati yaratilganicha yo‘q. Bu muammoni hal qilish uchun eng avvalo umummilliy adabiy tilning barcha asosiy uslublarini (ilmiy, badiiy, so‘zlashuv, rasmiy, publisistik kabilar), janrlarni (vaqtli matbuot, badiiy adabiyot, xalq og‘zaki ijodi, jonli so‘zlashuv tili kabilarni) statistik metodda to‘liq o‘rganib chiqilishi asosida yuqori chastotali so‘zlarni, grammatik formalarni har bir uslub va janr bo‘yicha aniqlab olish lozim bo‘ladi. Ana shundan keyin o‘zbek adabiy tilining to‘la chastotali lug‘atini tuzish, uning chegarasi, tamoyillari haqida fikr yuritish mumkin”, - degan mulohazalari ham fikrimizni qo‘llab-quvvatlaydi[2].

Bu muammo va masalalarni hal qilish uchun o‘zbek tilshunoslari hamda informatik mutaxassislar hamkorlikda ishlashlari taqozo etiladi. Chunki informatikning til grammatikasini tilshunos mutaxassis darajasida o‘zlashtirmaganligi va tilshunosning algoritmika va dasturlash qoidalarini bilmasligi kabi muammolar mavjud. Shuning uchun bu sohalardagi nazariy, amaliy va texnologik ishlarni bajarishda ushbu ikki soha mutaxassislarning korporativ hamkorligi zarur bo‘ladi.

Bu sohalar kesimidagi masalalar yechimi va texnologiyasi quyidagilardan iborat:

- kompyuter lingvistikasining ko‘plab yo‘nalishlari, muammo va masalalari mavjud. Ularni o‘rganish va hal qilish uchun kompleks yondashish lozim;
- tabiiy til sun’iy tildan farqli o‘laroq murakkab tuzilishga ega. Uni formallashtirish va algoritmik tasvirlash zarur;
- matnni tahlil qilish va sintez jarayonlarini modellashtirish, matematik mantiq va matematik statistikaning zamonaviy usullarini qo‘llashni talab qiladi;
- yaratilgan texnologiyadan kompyuter lingvistikasi masalalarini yechish uchun dasturiy ta’minot yaratishda tizimli ravishda foydalanish maqsadga muvofiq.

Shu nuqtai nazardan, tilshunos mutaxassislar va informatik-dasturchilar hamkorligida avtomatik lug‘atlar yaratish uchun algoritm va dasturiy ta’minot ishlab chiqish masalasini ko‘rib chiqish lozim bo‘ladi.

O‘zbek tilida yozilgan matnlardan avtomatik lug‘atlarni hosil qilish dasturiga quyidagi talablar qo‘yildi:

1. Fayldan (.rtf, .doc, .docx kengaytmali) matnni o‘qish, matnni qayta ishlash hamda hosil bo‘lgan lug‘atni faylga yozish, shuningdek lug‘atni boshqa dasturiy muhitga (MS Excel, MS Access) eksport qilish.

2. Kirish matnini kitob ko‘rinishida rasmiylashtirish va sahifalarni raqamlash.

3. Matndagi har bir so‘zni W S P(F) formatda chop qilish. Bu yerda W - so‘z, S - so‘zning chastotasi (matnda uchrash soni), P - sahifa nomeri, F - so‘zning P sahifadagi chastotasi.

4. Takroriy, ketma-ket so‘zlarni hisobga olish. Masalan, “kamdan-kam”, “yuzma-yuz”, “birin-ketin” kabi so‘zlar bitta so‘z sifatida qaraladi.

Avtomatik lug‘at hosil qilish algoritmiga qo‘yilgan talablarni bajaradigan algoritm quyidagicha loyihalandi.

Boshlash.

Qadam 1. MS Word dasturidan foydalanib faylni yuklash.

Qadam 2. Matndan maxsus simvollar (!, @, $, %, ^, &, va h.k.) va ortiqcha bo‘sh joylarni o‘chirish.

Qadam 3. Alifboli lug‘atni loyihalash.

3.1. Matndan so‘zlarni ajratish.

3.2. Saralash va alifboli lug‘at yaratish.

3.3. Lug‘atni saqlash va chiqarish.

Qadam 4. Chastotali lug‘atni loyihalash.

4.1. Alifboli lug‘atni o‘qish.

4.2. Saralash va chastotali lug‘at yaratish.

4.3. Lug‘atni saqlash va chiqarish.

Qadam 5. Chappa lug‘atni loyihalash.

5.1. Alifboli lug‘atni o‘qish.

5.2. Saralash va chappa lug‘at yaratish.

5.3. Lug‘atni saqlash va chiqarish.

Qadam 6. Agar so‘zni qidirish kerak bo‘lsa, Search funksiyasiga murojaat qilish.

Qadam 7. Agar statistik ma’lumotlarni chop qilish kerak bo‘lsa, Statistics funksiyasiga murojaat qilish.

Qadam 8. Agar natija faylini eksport qilish kerak bo‘lsa, Export funksiyasiga murojaat qilish.

Qadam 9. Agar lug‘atlar formasini o‘zgartirish kerak bo‘lsa, Settings funksiyasiga murojaat qilish.

Tamom.

Lug‘atlar kompyuter xotirasida quyidagicha strukturada saqlandi.

```
Type Pnode = ^Tree;
Tree = record
word: WideString;  {so‘z}
                count: Integer; {Soni}
                total_count: WideString; {har bir sahifada uchrash soni}
                left, right: Pnode; {Chap va o‘ng daraxtlar}
 end;
        Var
alptree: Pnode; {Alifboli lug‘at daraxti}
fretree: Pnode; {Chastotali lug‘at daraxti}
revtree: Pnode; {Chappa lug‘at daraxti}
```

Ushbu algoritm va ma’lumotlar qurilishi asosida ishlab chiqilgan dasturiy ta’minot quyidagi funksiyalarni bajaradi[3, 4]:

- uch turdagi lug‘atni tayyorlaydi: alifboli, chastotali, chappa(ters);
- matndan harflar chastotasini aniqlaydi;
- katta hajmli matnlarni qayta ishlaydi;
- turli alifbolarda (krill alifbosi, o‘zbek krill alifbosi va lotin alifbosi asosida) tayyorlangan matnlarni qayta ishlaydi;
- statistik ma’lumotlarni tayyorlaydi;
- sodda va qulay foydalanuvchi interfeysi qo‘llangan.

Dasturiy ta’minot interfeysi menyu va qism menyulari quyidagi ko‘rinishda tashkillashtirilgan.

Fayl (ochish, saqlash, eksport, chiqish).

Lug‘at (alifboli lug‘at, chastotali lug‘at, chappa lug‘at).

Statistika (kirish fayli, matnni qayta ishlash vaqti, lug‘atlar).

Lug‘at ko‘rinishini sozlash.

Yordam (foydalanish yo‘riqnomasi, dastur haqida).

Dasturiy ta’minotni testlash natijalari quyidagi jadvalda keltirilgan:

Matn sahifalari soni	1	5	10	50	100	200	300	400
Xotira hajmi(Kb)	34	67	163	278	552	2064	3090	4119
Xotira hajmi(Mb)	0,03	0,07	0,16	0,27	0,54	2,02	3,02	4,02

Qayta ishlash vaqti(sek)	7,09	13,42	15,72	48,77	88,69	194,28	333,53	411,38
Qayta ishlash vaqti(min)	0,12	0,22	0,26	0,81	1,48	3,24	5,56	6,86

Kompyuter leksikografiyasi masalalarining ba’zi bir murakkab va hali hal qilinmagan kamchiliklarini ham qayd etish kerak. Masalan: lug‘atda ayrim so‘zlar qo‘shtirnoq ichida keltiriladi, kompyuter qo‘shtirnoq ichidagi har bir so‘zni alohida o‘qimoqda; ayrim harfiy ifodalar bir so‘z sifatida joy olgan; ba’zi so‘zlar matnda ikki xil berilgan (olam-jahon, ro‘y-rost, ro‘yirost) ikki holatni ikki so‘z sifatida e’tirof etishga to‘g‘ri keladi.

Shularga qaramasdan ishlab chiqilgan dasturiy ta’minot va algoritmlar majmuasini avtomatik lug‘atlar yaratishda, badiiy matnlarni statistik tahlil qilishda, matndagi harflar chastotasini aniqlashda, informatsion qidiruv tizimlarini yaratishda, matnlarni imloga tekshirish tizimini ishlab chiqishda, tilni o‘qitish uchun dasturiy ta’minot yaratishda, o‘zbek tilining milliy korpusini ishlab chiqishda qismiy modul sifatida qo‘llash mumkin.

Foydalanilgan adabiyotlar:

1. Пўлатов А. Компютер лингвистикаси. -Тошкент: Академнашр, 2011. - 518 б.

2. Нурмонов А., Йўлдошев Б. Тилшунослик ва табиий фанлар. - Т.: Шарқ, 2001. - 125 б.

3. Кобилов С., Раббимов И. Разработка программного обеспечения для решения одной задачи компьютерной лингвистики//Наука и Мир, 2015, № 6(22), Т. 1. - С. 21-23.

4. *Каримов С.А., Қобилов С.С., Раббимов И.М. «Ўзбек тилида ёзилган бадиий матнларни лексикографик ва статистик таҳлил қилиш» дастурий мажмуаси. № ДГУ 04885. 24.11.2017*

5. Rabbimov I., Kobilov S. and Mporas I. “Uzbek News Categorization using Word embeddings and Convolutional Neural Networks” 2020 IEEE 14th International Conference on Application of Information and Communication Technologiyes (AICT), Tashkent, Uzbekistan, 2020, pp. 1-5.