УДК 004.89:004.912

ЗАСОБИ АВТОМАТИЧНОГО АНАЛІЗУ І СИНТЕЗУ ТЕКСТУ

Ковальов О.М.
Науковий керівник – к.т.н., старший викладач Бабій В.П.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна
тел.: +38(066) 045-95-93, email: oleksandr.kovalov@nure.ua

This work is devoted to means of automatic text analysis and synthesis. The main purpose of automatic text analysis and synthesis and the possibility of practical application are considered. The principle of automatic text analysis is considered. Modern tools for automatic text analysis and synthesis were studied, namely pymorphy2 and phpmorphy libraries.

Автоматизований аналіз і синтез тексту є важливими завданнями в комп'ютерній лінгвістиці як з точки зору розробки лінгвістичних основ для створення штучного інтелекту, так і з точки зору задоволення практичних потреб людини, наприклад створення ефективних систем машинного перекладу.

Автоматична обробка текстів має широке практичне застосування: створення словників та інтелектуальних пошукових систем, розробка лінгвістичних процесорів для забезпечення спілкування з користувачами природною мовою, автоматична абстракція тексту та ін.

При обробці текстів природною мовою необхідно ідентифікувати певні текстові елементи. Входом у процес є текст природною мовою, а виходом – певні структури даних, що допускають подальшу автоматичну або ручну обробку тексту. Морфологічний аналіз є важливою частиною процесу попередньої обробки тексту.

Морфологія – розділ граматики, що вивчає граматичні властивості слів. Граматичними властивостями слів є граматичні значення, способи вираження граматичних значень. Морфологія вивчає будову частин мови. Основна мета – розбити словоформу на дві частини.

Морфологічна розмітка під час автоматичної обробки текстів природною мовою є основою як для морфологічного аналізу, так і для інших форм аналізу – синтаксичного та семантичного.

У наш час існує вже деяка кількість інструментів та бібліотек, що вирішують проблему автоматичного морфологічного аналізу слів на основі машинного навчання. Деякі з них є більш поширеними, деякі менш популярними, деякі бібліотеки підтримують аналіз слів української мови, а деякі ні.

Одним з найбільш поширених аналізаторів в українському Natural Language processing й не тільки є pymorphy2 [1]. Його код є відкритим й доступним для використання після завантаження з GitHub репозиторію. Під час роботи даний аналізатор використовує словник OpenCorpora. Для

не знайомих слів будуть збудовані гіпотези на основі машинного навчання, що дає можливість аналізу слів, не присутніх у словнику. Сама бібліотека працює досить швидко та може аналізувати від кількох тисяч слів за секунду, до сотні тисяч слів за секунду. Продуктивність роботи аналізатора залежить від виконуваної операції, обраного інтерпретатора та встановлених пакетів. Споживання оперативної пам'яті бібліотекою буде в районі від 10 до 20 мегабайтів. Бібліотека підтримує морфологічний аналіз слів української.

Морфологічний аналізатор pymorphy2 має змогу виконувати наступний перелік функцій:

– приводити слова до нормальної форми. Наприклад, «люди» до «людина», або «гуляв» до «гуляти»;

– перетворювати слова у потрібну форму. Наприклад, перетворити слово у множину, змінювати відмінок слів і тому подібне;

– повертати граматичну інформацію про слово. Наприклад число, рід, відмінок, частину мови тощо.

Ще одним засобом морфологічного аналізу слів є бібліотека phpmorphy [2]. Дана бібліотека реалізована для платформи php й використовується для морфологічного аналізу слів української, англійської та німецької мов. cijic/phpMorphy — це оболонка Laravel для бібліотеки phpMorphy з підтримкою PHP7. Код бібліотеки є відкритим й доступним для використання після завантаження з GitHub репозиторію (див. рис. 2)

Морфологічний аналізатор phpMorphy дозволяє вирішувати наступний перелік завдань:

– лематизація (отримання нормальної форми слова);

– отримання усіх форм слова;

– отримання граматичної інформації про слово (частина мови, відмінок, відмінювання тощо);

– зміна форми слова за заданим зразком.

Бібліотека phpMorphy вміє працювати з українською, англійською, німецькою та естонською мовами.

У морфологічному аналізаторі є можливість додати підтримку інших мов за допомогою myspell словника.

Бібліотека підтримує різноманітні кодування:

– всі однобайтові (windows-1251, iso-8859-* тощо);

– unicode кодування – utf-8, utf-16le/be, utf-32, ucs2, ucs4.

Список використаних джерел:

1. Морфологічний аналізатор pymorphy2 [Електронний ресурс] / pymorphy2. – Режим доступу до ресурсу: https://pymorphy2.readthedocs.io /en/stable/index.html

2. phpMorphy [Електронний ресурс] / phpMorphy. – Режим доступу до ресурсу: https://phpmorphy.sourceforge.net/dokuwiki/