ІДЕНТИФІКАЦІЯ АНТРОПОНІМІВ У ПОВНОТЕКСТОВИХ ДОКУМЕНТАХ

Хайло А. М.
Національний технічний університет
"Харківський політехнічний інститут",
м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60,
e-mail: *alina_khailo@ukr.net*

Автоматична обробка тексту на природній мові дозволяє полегшити пошук і вилучення інформації з метою подальшої аналітичної обробки. Здебільшого потрібен аналіз великих масивів коротких текстів з метою виділення значущої інформації. Автоматичне розпізнавання з подальшим виділенням в текстах власних назв, що позначають людей є особливою проблемою комп'ютерної обробки текстів на природній мові. Власні назви, на відміну від загальних назв, утворюють список, який постійно змінюється та доповнюється. Вирішення цього завдання пов'язане з проблемою ідентифікації власних назв та ланцюжків звичайних слів, які поводять себе в текстах як власні назви.

На сьогоднішній день не існує масштабних систем з автоматизованої обробки природньої мови, які здатні виокремлювати та маркувати антропоніми (за визначенням Виноградова В. С., антропонім – це власна назва (або набір назв), яка офіційно присвоєна окремій людині як її розпізнавальний знак) в текстах українською мовою. Крім цього, опису різноманітних систем ідентифікації антропонімів присвячена досить велика кількість зарубіжних публікацій. Їх автори пропонують різні методи розпізнавання та смислової інтерпретації, які можна умовно розділити на наступні групи:

• статистичний підхід (для створення статистичної моделі використовується корпус розмічених текстів, Sekine S., Eriguchi Y. [2000]);

• обчислювальні методи на основі навчальних моделей (наприклад, в рамках проекту CoNLL-2003);

• метод контекстного аналізу (спирається на правила ідентифікації антропонімів в тексті в залежності від лівого і правого контексту та списки слів відкритих класів, McDonald D. [1996], Kokkinakis D. [2004]);

• гібридний підхід (об'єднує статистичні методи і прийоми контекстного аналізу (Mikheev A. et al. [1998]).

У статті McDonald D. [1996] описується один з ключових компонентів системи розуміння природної мови Sparser – модуль PNF, який призначений не тільки для розпізнавання і класифікації власних назв, а й для виокремлення та інтерпретації антропонімічних груп.

Всього виділяються три етапи роботи, так чи інакше, пов'язаних з виділенням та обробкою антропонімів:

1) визначення меж послідовності слів, з яких утворюється антропонім;

2) віднесення отриманого елемента до тієї чи іншої семантичної категорії з одночасним дозволом неоднозначності;

3) збереження отриманого результату в моделі з метою його подальшого використання при роботі Sparser як з даними, так і з іншими текстами.

Показниками антропонімічних груп являються вже відомі системі власні назви, які зустрічаються в тексті, а також слова-класифікатори (наприклад, spokesman, company), які забезпечують можливість прогнозування: безпосередньо після таких лексичних одиниць велика ймовірність зустріти в тексті антропонім.

У статті Stevenson M., Gaizaukas R. [2000] обговорюється серія експериментів з системою, яка була побудована авторами на основі комплексу LASIE (різні його версії використовувалися в проектах MUC і HUB4). Метою цих експериментів була ідентифікація в текстах антропонімів на основі попередньо побудованих списків слів і словосполучень. Описувана авторами система представляє інтерес, оскільки вона є самонавчальною: користуючись досить простим набором фільтрів, програма поповнює раніше побудовані списки власних назв новими одиницями, в результаті чого вона стає набагато більш точною. У роботі викладаються способи побудови списків, їх поповнення, а також описуються фільтри і методи експериментальної роботи з ними.

Оскільки, на сьогоднішній день основна маса інформації зберігається і обробляється в електронному вигляді, практика показує, що більшість ділових пошукових завдань пов'язані з пошуком власних назв. Правильно виділяти і розпізнавати власні назви необхідно і при комп'ютерному аналізі текстів. До того ж, завдання вилучення антропонімів з тексту є критично важливою технологією для подальшого створення систем інформаційного пошуку і розуміння документів.