Розпізнавання іменованих сутностей при видобуванні інформації з медичних текстів

Андрій Романюк[1], Олена Карпінська[2]

1. Кафедра систем автоматизованого проектування, Національний університет "Львівська політехніка", УКРАЇНА, м. Львів, вул. С. Бандери, 12, E-mail: anrom@polynet.lviv.ua

2. Кафедра систем автоматизованого проектування, Національний університет "Львівська політехніка", УКРАЇНА, м. Львів, вул. С. Бандери, 12, E-mail: olena.karpinsky@gmail.com

This paper deals with specifics of named entity recognition of medical texts. The paper describes approaches on which named entity recognition systems are based.

Ключові слова – видобування інформації, розпізнавання іменованих сутностей, машинне навчання, анотування, словник, корпус текстів.

Вступ

Зі стрімким поширенням інтернету, проблема структуризації інформації є актуальною як ніколи. Це стосується зокрема і медичних текстів: кількість статей та доповідей збільшується в геометричній прогресії, а обробляти їх вручну вже давно стало дорогим та неефективним завданням. Одним з ключових напрямів видобування інформації є розпізнавання іменованих сутностей, що ставить перед собою завдання пошуку та класифікації іменованих сутностей в тексті.

Методи побудови системи розпізнавання іменованих сутностей

Існує кілька підходів до розпізнавання іменованих сутностей в медичній галузі; їх можна класифікувати як такі, що засновані на правилах, словниках та машинному навчанні.

Системи в основі яких лежать словники найкраще застосовувати для розпізнавання хвороб. У цьому випадку система порівнює слова в тексті із словами у словнику і при збігу відносить їх до тих чи інших іменованих сутностей.

Алгоритми розпізнавання іменованих сутностей засновані на правилах застосовують сукупність правил для виділення певної моделі. Ці моделі зазвичай засновані на граматичних, синтаксичних та орфографічних характеристиках сутностей [2].

Метод машинного навчання можна застосовувати до ширшого кола текстів із динамічним набором слів (гени, протеїни). Існують два типи

моделей машинного навчання для алгоритмів розпізнавання іменованих сутностей: наглядові та безнаглядові [2]. Безнаглядовий метод не потребує даних, на яких буде відбуватися навчання. Деякі вчені вважають безнаглядові методи актуальнішими, адже їх легше підтримувати і вони більш універсальні [1], в той час, як серед інших вчених цей метод не є популярним, адже результати його використання є часто значно гіршими, аніж результати наглядового методу [2]. Безнаглядові системи, в свою чергу, потребують заздалегідь анотованих текстових даних – корпусів текстів, розмічених лінгвістами вручну.

Побудова системи розпізнавання іменованих сутностей, яка використовуватиме машинне навчання складається з двох основних етапів: навчання та анотування, які зображені на Рис. 1.

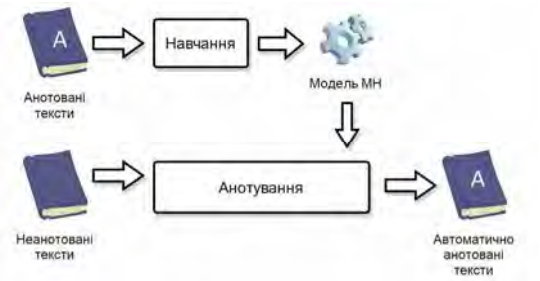

Рис.1 Побудова системи розпізнавання іменованих сутностей з допомогою машинного навчання

Спочатку модель машинного навчання треба навчити, використовуючи анотовані дані. Після цього здійснюється анотування нерозмічених текстів – присвоєння іменованих сутностей на основі інформації, яка виділена з анотованих даних.

Висновок

Існуючі підходи до розпізнавання іменованих сутностей задовольняють різні вимоги, залежно від виду сутностей, які треба видобути з тексту. За наявності необхідних ресурсів системи машинного навчання мають ряд переваг перед іншими методами, тому саме цей підхід обрано для розпізнавання іменованих сутностей в текстах української мови.

Література

1. Шабінський А. С. Змішана тематично-сутнісна онтологія у покращеній тематичній векторній моделі / А. С. Шабінський // Проблеми програмування. – 2014. – № 2-3. – С. 182–187.
2. Malay Named Entity Recognition Based on Rule-Based Approach / R. W. Alfred, L. C. Leong, C. K. On, P. M. Anthony // International Journal of Machine Learning and Computing. – 2014. – No. 3, Vol. 4. – P. 300–306.