УДК 004.4:004.934

ДОСЛІДЖЕННЯ КОМБІНОВАНИХ МЕТОДІВ ВИЛУЧЕННЯ СТРУКТУРОВАНИХ ДАНИХ ДЛЯ МОВ З ВІЛЬНИМ ПОРЯДКОМ СЛІВ

Орлова Г.О.
Науковий керівник – к.т.н., доц. Кириченко І.В.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна
тел.: +38(067) 252-10-18, email: hanna.orlova.cpe@nure.ua

This work is devoted to the development of a technique for extracting facts and events from texts, which is based on combining approaches to the use of rules on context-free grammars together with the analysis of syntactic trees of dependency grammars. A combined approach for the problem of fact extraction for languages with free word order is considered. The idea of the technique is to combine an approach to extracting information based on the fusion of results obtained using context-free grammars with the results of pattern selection during text parsing. Based on the proposed approach, a software library has been developed, which includes own templates for extracting subgraphs from dependency trees, considering morphological and syntactic features and the results of extraction using context-free grammars.

На сьогодні в епоху Великих даних обсяги вироблюваної людством інформації більші ніж коли-небудь і її кількість росте з кожним днем. Людина вже не в змозі вручну обробляти, аналізувати, витягати знання з неструктурованих даних, переважно текстових, і передавати їх по різних каналах. У зв'язку з цим особливої актуальності набула задача перетворення текстів, написаних на природній мові в структуроване представлення для застосування в прикладних завданнях.

До найбільш поширених задач обробки природних мов (Natural Language Processing – NLP) можна віднести машинний переклад, автоматичне реферування тексту, розпізнавання та синтез мови, генерацію тексту [1-2]. Також обробка природних мов використовується для задач аналізу тексту, таких як класифікація (сентимент-аналіз, фільтрація спаму та ін.), створення пошукових та питально-відповідних систем, включаючи віртуальних співрозмовників, та задачі з витягання інформації.

Під витяганням інформації мається на увазі пошук в слабо структурованих документах або фрагментах документів/текстів окремих фактів, що представляють певний інтерес для шукачів. Сьогодні вже існують рішення, що дозволяють витягати з текстів різноманітні іменовані сутності з прийнятним рівнем якості, проте задача по витяганню їх відносин, фактів і подій є відносно новою і найбільш актуальною зважаючи на величезний зріст обсягів даних [4].

У цій доповіді пропонується підхід до вирішення задачі витягання фактів для мов з вільним порядком слів. На рисунку 1 представлена таблиця з покроковим описом розробленої методики вилучення фактів з тексту.

Попередня обробка тексту	**Побудова синтаксичного дерева**	**Застовування правил КВГ**	**Злиття результатів**	**Парсинг дерева**
Токенізація на слова та речення. Морфологічний аналіз. Лематизація.	Застосування моделі UDPipe 2.0, навченої на корпусі SyntaxRus	Написання і застосування правил по витягуванню окремих сутностей і атрибутів з використанням бібліотеки Yargy	Парсинг отриманої структури (ConLL - U) в структуру даних tree. Додавання в елементи отриманого дерева витягнутих за допомогою КВГ даних.	Написання шаблонів для опису фактів. Парсинг синтаксичного дерева по написаних шаблонах.

Рисунок 1 – Методика пропонованого підходу

Ідея методики полягає в комбінуванні підходу щодо вилучення інформації, заснованого на злитті результатів, отриманих за допомогою контекстно-вільних граматик з результатами виділення шаблонів при синтаксичному розборі тексту.

На основі запропонованого підходу було розроблено програмну бібліотеку, що включає власні шаблони для вилучення підграфів з дерев залежностей з урахуванням морфологічних та синтаксичних ознак та результатів вилучення за допомогою контекстно-вільних граматик. Розроблена бібліотека була успішно застосована на прикладі завдання щодо вилучення інформації з резюме кандидатів та були отримані шаблони для збору всієї необхідної інформації, яка включає контактні дані кандидатів, опис досвіду роботи, перелік їхніх навичок, компетенцій, досвід роботи з технологіями.

Список використаних джерел:

1. Moens, M.-F. (2009). Information Extraction: Algorithms and Prospects in a Retrieval Context. Netherlands: Springer.

2. Tunstall, L., & von Werra, L. (2022). Natural Language Processing with Transformers. Revised Edition. O’Reilly Media.

3. Thomas, A. (2020). Natural Language Processing with Spark NLP: Learning to Understand Text at Scale. 1st edition. O’Reilly Media.