УДК 004.912

Федорович І. – ст. гр. ІП-13мп
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

ПРОЦЕС ОНОВЛЕННЯ СЛОВНИКА УКРАЇНСЬКОЇ МОВИ ДЛЯ МОРФОЛОГІЧНОГО АНАЛІЗАТОРА PYMORPHY2

Науковий керівник: к.т.н., доцент Олійник Ю. О.

Fedorovych I.
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

THE PROCESS OF UPDATING THE UKRAINIAN LANGUAGE DICTIONARY FOR THE MORPHOLOGICAL ANALYZER PYMORPHY2

Supervisor: Oliinyk Y.

Ключові слова: обробка природної мови, українська мова, pymorphy2.
Keywords: natural language processing, ukrainian language, pymorphy2.

Вступ. Pymorphy2 – це морфологічний аналізатор, розроблений на мові програмування Python. У pymorphy2 заявлена підтримка української та російської мови, проте підтримка української мови є експериментальною [1], а словник, використовуваний у аналізаторі – застарілий. Ця робота спрямована на опис процесу оновлення словника української мови актуальною версією Великого словника української мови (ВЕСУМ) [2] для його використання з pymorphy2 [3].

Основна частина. Для генерації файлів ВЕСУМ існує репозиторій dict_uk [4], створений та підтримуваний командою БрУК, які є авторами ВЕСУМ. Код генерації файлів словника розроблений на мові програмування Java. Для отримання файлів потрібно створити локальну копію репозиторію на комп'ютері з встановленим JDK версії 11 або новішої та запустити систему автоматичного збирання Gradle за допомогою команди «./gradlew expand». В результаті виконання команди буде згенеровано декілька файлів, з яких один файл з назвою «dict_corp_lt.txt», що є словником формату LanguageTool.

З огляду на те, що pymorphy2 підтримує лише словники формату OpenCorpora, отриманий словник формату LanguageTool необхідно конвертувати. Одним з рішень є програмне забезпечення, розміщене на репозиторії [5], яке розроблене на мові програмування Python. Окрім програмного забезпечення, репозиторій містить детальну діаграму сумісності тегів LanguageTool та OpenCorpora, а також електронну таблицю з переліком тегів LanguageTool, їх описом, та відповідником з OpenCorpora. Для виконання конвертації необхідно на комп'ютері з локальною копією репозиторію та інтерпретатором Python встановити залежності для програмного забезпечення та виконати команду «python bin/lt_convert.py dict_corp_lt.txt full_uk.xml».

Після отримання словника ВЕСУМ у форматі OpenCorpora необхідно скомпілювати його у локальну Python-бібліотеку за допомогою утиліти pymorphy2-dicts [6]. Проте станом на зараз (pymorphy2 версії 0.9.1), вихідний код морфологічного аналізатору повинен бути модифікований в класі PredictionSuffixesDAWG, а саме збільшення розмірності змінної count з 2 до 4 байт. Це досягається шляхом зміни

символів «>ННН» на «>ІНН». Після модифікації pymorphy2, словник ВЕСУМ у форматі OpenCorpora можна скомпілювати за допомогою команди «python update.py uk compile package» утиліти pymorphy2. Результатом компіляції є згенерована Python-бібліотека pymorphy2-dicts-uk, яку можна встановити за допомогою команди «pip install».

Висновки. Ця робота описує всі необхідні кроки для встановлення актуальної версії Великого словника української мови для морфологічного аналізатора pymorphy2. Оскільки ВЕСУМ з часом оновлюється та доповнюється, це дозволяє підвищити якість обробки української мови в pymorphy2. Також ця робота може бути використана для виведення процесів додавання словників інших мов для pymorphy2, або інтеграції ВЕСУМ в інші засоби обробки природної мови.

Список літератури

1. Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. – 2015. – С. 320-332.
2. Рисін А., Старко В. Великий електронний словник української мови (ВЕСУМ). Вебверсія 5.6.2. 2005-2022 [Електронний ресурс] / Андрій Рисін, Василь Старко – Режим доступу до ресурсу: https://r2u.org.ua/vesum/
3. Pymorphy2 [Електронний ресурс] – Режим доступу до ресурсу: https://github.com/pymorphy2/pymorphy2.
4. Brown-uk/dict_uk: Project to generate POS tag dictionary for Ukrainian language [Електронний ресурс] – Режим доступу до ресурсу: https://github.com/brown-uk/dict_uk.
5. Чаплинський Д. LT2OpenCorpora [Електронний ресурс] / Дмитро Чаплинський – Режим доступу до ресурсу: https://github.com/dchaplinsky/LT2OpenCorpora.
6. Pymorphy2-dicts: Scripts for updating pymorphy2 dictionaries [Електронний ресурс] – Режим доступу до ресурсу: https://github.com/pymorphy2/pymorphy2-dicts.