

ІНТЕЛЕКТУАЛЬНИЙ МОДУЛЬ ОНОВЛЕННЯ СЕМАНТИЧНОГО СХОВИЩА НА ОСНОВІ МІКРОБЛОГІНГУ

Островська М.В.

Науковий керівник – проф., к.т.н. Месюра В.І.

В сучасних умовах важливого значення набуває семантичний аналіз інформації, а саме таке автоматизоване опрацювання природної мови, завдяки якому з'являється можливість зі зрозумілих людині речень вилучати важливу інформацію, поміщати її в спеціальні семантичні сховища або онтології та опрацьовувати відповідним чином як базу даних. Головною задачею розроблюваного інтелектуального модуля є виявлення у вхідних повідомленнях користувачів нових зразків класів предметної області, що ще не містяться в онтології, та занесення їх до семантичного сховища. Цей модуль може бути використаний як основа системи прийняття рішень на основі соціальних медіа для підприємств, громадських установ, наукових проектів тощо.

На даний момент існує не так багато створених засобів для обробки природних мов і працювати зі звичним людині текстом програмним засобом ще важко, тому для спрощення розуміння було вирішено використовувати так звані контрольовані мови, що отримані за рахунок обмеження граматики та лексики для того, щоб зменшити або усунути неоднозначність і складність. В якості такої мови було обрано формат Twitter Data, що є відкритим, доступним і нескладним. В якості засобу для обробки природних мов було обрано GATE - набір Java-інструментів, розроблених для вилучення інформації на багатьох мовах. До переваг GATE відносяться його доступність і інтеграція з Ontotext OWLIM Lite, який дозволяє роботу з онтологіями.

Алгоритм вирішення поставленої задачі містить наступні основні кроки:

1. Створення початкової опорної онтології предметної області засобами GATE.

2. Створення додатку в середовищі GATE, який виконає первинний аналіз тексту (розбиття на речення, слова, виділить цифри, пробіли), а також анотує текст в відповідності з створеною онтологією та відповідно до розмітки Twitter. Для виконання цього пункту було створено код на мові JAPE – спеціальна мова специфікації шаблонів, що дозволяє розпізнавати регульовані вирази в анотації документів.

3. Перевірка наявності отриманих після анотації документа текстових значень суб'єкта та об'єкта в онтології, і в разі відсутності створення зразків відповідних класів.

Вирішення задачі забезпечує постійне оновлення онтології новою інформацією з повідомлень користувачів системи та можливість виконання подальших запитів до вже поповненої бази.