УДК 004.93:159.95

О.В. Бісікало, С.С. Траченко (Україна, Вінниця)

МЕТОД ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ З ТЕКСТУ НА ОСНОВІ ІНСТРУМЕНТАЛЬНИХ ЗАСОБІВ WORDNET ТА NLTK

Актуальність даної роботи пов'язана із неперервним зростанням мережі інтернет та кількості даних в ній. У зв'язку з цим підвищилась актуальність автоматичного визначення змісту тексту. Для вирішення даної проблеми було розроблено багато методів, зокрема для задачі визначення ключових слів тексту та задачі побудови онтології текстів природної мови. Адекватні рішення даних задач дозволяють користувачу зробити висновок про загальну тему та зміст тексту, що дозволить в подальшому застосовувати отримані напрацювання в таких галузях, як SEO - для визначення найбільш релевантних результатів на пошукові запити користувачів.

Постановка задачі. Дано текст англомовного походження. Необхідно визначити зв'язки між словами, що зустрічаються найчастіше, формально визначити їх "силу зв'язку" та, на основі отриманих результатів, побудувати онтологію вхідного тексту.

З метою **рішення задачі** побудови лексичної онтології потрібно виконати 8 основних операцій. Розглянемо визначені кроки запропонованого алгоритму побудови лексичної онтології на прикладі зв’язку типу amod за Стенфордською класифікацією. Йде мова про створення лексичної онтології між іменниками тексту, де сила зв’язку між будь-якими 2-ма з них визначається виключно на основі статистики зв’язків між ними та спільними прикметниками. Алгоритм передбачає:

1) Розбиття тексту на речення.
2) Створення лісу ієрархічних дерев із речень. Дана операція виконується завдяки інтерфейсу роботи NLTK із Stanford parser’ом.
3) Створення списку залежностей між словами. Операція виконується за допомогою інструменту PyStanfordDependencies, який дозволяє розбити дерево, що отримано на попередньому кроці, на список залежностей між 2-а токенами.
4) Із отриманого списку створюється частотний словник залежностей між словами. Рахуються входження екземплярів власного типу в частотний словник. 2 екземпляри власного типу на даному кроці вважаються однаковими, якщо всі відповідні поля мають однакові значення.
5) Відбувається фільтрування словника, під час якого залишаються залежності лише із певним зв’язком (в нашому випадку – amod).
6) Групування словника за головним словом. Створюється словник, де ключем є деяке слово, а значення – список залежностей, де головним словом є ключ.
7) Проведення аналізу списків залежностей між словами (ключами) отриманого словника, під час якого визначається подібність 2-х слів. На даному кроці визначається подібність ключів із словника, що був отриманий на попередньому кроці. Значення має в собі список зв’язків, де головним словом є ключ із списку, а також кожен елемент списку містить кількість входжень цього зв’язку (із кроку 4 – створення частотного словника).

Схожість між 2-а іменниками S обраховуємо як відношення:

$$S = \frac{SimilarSum}{OverralSum},$$

де OverralSum - кількість входжень всіх зв’язків обох списків;
SimilarSum - однакові залежні слова у двох списках.

Висновки. В роботі уперше запропоновано метод отримання лексичної онтології з тексту, який, на відміну від відомих, базується на визначенні чисельних ознак складних зв’язків між мовними одиницями та технологічних можливостях сучасних лінгвістичних пакетів, що дозволяє будувати онтології з обраних частин мови, типів зв’язків, а також на основі різних способів обробки тексту

Література

1) Бісікало О.В. Формалізація понять мовного образу та образного сенсу природно-мовних конструкцій / О.В. Бісікало // Математичні машини і системи. – 2012. – № 2. – С. 70–73.

2) About WordNet [Електронний ресурс]: – Режим доступу: http://stevenloria.com/tutorial-wordnet-textblob. – Назва з екрану.