УДК 004.91

О.В. Бісікало

С.С. Траченко

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

Вінницький національний технічний університет

Анотація

Запропоновано метод отримання лексичної онтології з тексту, який, 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

Ключові слова: NLP, онтологія, WordNet, NLTK, Python.

Abstract

The method of lexical ontology of the text, which is based on determining the characteristics of complex numerical relationships between linguistic units and technological capabilities of modern language packs, allowing you to build ontologies from selected parts of speech, types of connections, as well as on different ways of treatment text.

Keywords: NLP, ontology, WordNet, NLTK, Python.

Вступ

Обсяги вільно доступної сьогодні інформації у мережі Інтернет перевищили найоптимістичніші прогнози початку нинішнього тисячоліття. І хоча в Інтернеті невпинно збільшується питома вага мультимедіа-ресурсів, природно-мовна інформація залишається найбільш важливою з точки зору реалізації глобального пошуку та рекламних функцій для користувачів. Незворотне зростання популярності лінгвістичних Інтернет-технологій вимагає у дослідників підвищення якісних показників розв'язання задач оброблення текстової інформації.

Мета роботи полягає в обгрунтуванні методу побудови лексичної онтології англомовного тексту, що вимагає вибору найбільш зручних інструментальних засобів, аналізу статистики зв'язків між словоформами та synset-ами, програмної реалізації та порівняння отриманих результатів лексичної онтології із експертними зв'язками в існуючій онтології WordNet [1].

Об'єкт дослідження – процеси лексичного та семантичного аналізу англомовного тексту.

Предмет дослідження – методи та інструментальні засоби автоматизованої побудови лексичної онтології англомовного тексту.

Результати дослідження

З метою реалізації запропонованого методу побудови лексичної онтології та з урахуванням обраних у розділі 1 інструментальних засобів потрібно виконати 8 основних операцій, зокрема:

1. Розбиття тексту на речення.
2. Створення лісу ієрархічних дерев із речень, що отримані після кроку 1.

3. Створення списку залежностей між словами, які ідентифікуються з відповідними словоформами за допомогою PyStanfordDependencies.

4. Із отриманого списку створюється частотний словник залежностей між словами.

5. Фільтрування словника, в результаті якого залишаються залежності лише із певним типом зв’язку.

6. Видалення унікальних залежностей із відфільтрованого словника.

7. Групування словника за головним словом – створюється словник, де ключем є деяке слово, а значення – список залежностей, де головним словом є ключ.

8. Проведення аналізу списків залежностей між словами (ключами) отриманого словника, під час якого визначається подібність 2-х слів.

Визначення лексичної онтології з іменників на основі зв’язків типу amod було проведено за текстом роману Г. Мелвілла «Мобі Дік» [2]. Внаслідок обробки тексту програма виявила 50585 різних зв’язків типу amod, що мають більш ніж 1 повторення, з яких найбільшу кількість повторення через спільні прикметники мав зв’язок man - whale (245), а 32217 зв’язків повторювалися двічі. Після порівняння отриманих значень ваги за алгоритмом методу з вагою зв’язків між аналогічними synset’ами у WordNet було визначено онтологічні списки іменників за критеріями абсолютної ваги та точності співпадіння зі значенням у WordNet. На рис. 1 представлено загальний вигляд залежності кількість-вага для всіх отриманих зв’язків.

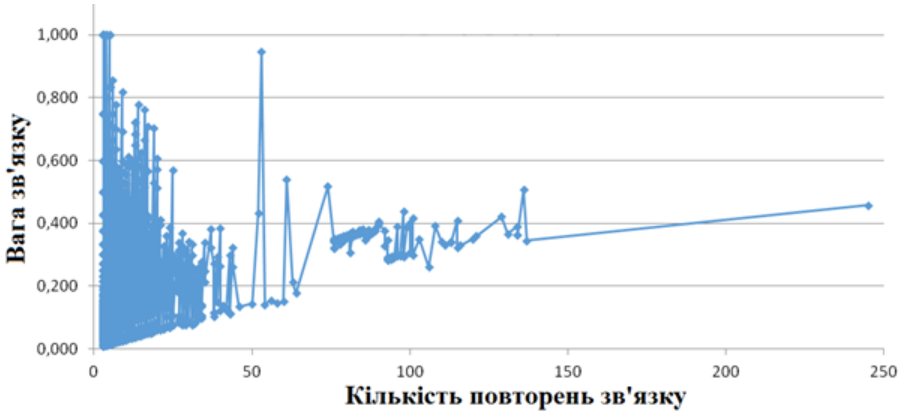

Рисунок 3.2 - Загальний вигляд залежності кількість-вага для всіх отриманих зв’язків

Аналіз отриманої залежності показує: а) її явну парето-подібність[2] у зв’язку з надзвичайно великою кількістю пар, що повторюються 2-3-4 рази та незначною кількістю пар зі статистикою повторення більше 40; б) існування тенденції на пропорційне збільшення середньої ваги зв’язку від кількості його повторення. Останній висновок краще ілюструє деталізація даної залежності на рисунках 3.3, 3.4 та 3.5.

В таблиці 1 наведено список зв’язків між іменниками з вагою, більшою за 0,3 (емпірично визначена межа, що обмежує 1% зв’язків) та точністю, більшою за 0,8 (неспівпадіння менше за 20%).

Таблиця 1 - Список зв’язків між іменниками з вагою, більшою за 0,3 та точністю, більшою за 0,8

1 слово	2 слово	Кількість	Вага наша	Вага WordNet	% неспів-падіння
authorities	man	81	0,367	0,333	9,05%
man	woman	79	0,356	0,333	6,33%
man	women	78	0,351	0,333	5,13%

1 слово	2 слово	Кіль кість	Вага наша	Вага WordNet	% неспів-паління
man	chaps	76	0,347	0,333	3,95%
man	chap	76	0,345	0,333	3,51%
man	traveller	76	0,345	0,333	3,51%
man	king	76	0,342	0,333	2,63%
distance	start	6	0,333	0,333	0,00%

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.

Обґрунтування методу було проведено на основі статистичної оцінки складних залежностей між словоформами та synset'ами, а також статистичного аналізу запропонованого підходу до визначення значимих елементів онтології. Внаслідок цього було висунуто 2 гіпотези про Парето-подібність розподілу для зв'язків між synset'ами з предметної області та доцільність переміщення найбільш вагомих synset'ів та їх зв'язків у початковий пік кривої Парето за допомогою нового методу.

Перспективними задачами подальшого дослідження є отримання та порівняльний аналіз лексичних онтологій для інших частин мови та типів зв'язків, наприклад для дієслів на основі зв'язку з підлеглими прислівниками adverb, а також формальна оцінка якості отриманих онтологій за допомогою різних мір близькості.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. About WordNet [Електронний ресурс]: – Режим доступу: http://stevenloria.com/tutorial-wordnet-textblob. – Назва з екрану
2. Herman Melville. Moby-Dick; or the Whale. [Електронний ресурс] – Режим доступу: https://www.gutenberg.org/files/2701/2701-h/2701-h.htm – Назва з екрану.
3. Rank distributions of words in additive many-step Markov chains and the Zipf law [Електронний ресурс]: – Режим доступу: http://arxiv.org/pdf/physics/0406099.pdf. – Назва з екрану.

Траченко Сергій Сергійович – студент групи 1КСУА-15м, факультет комп'ютерних систем управління, Вінницький національний технічний університет, Вінниця, e-mail: trachenkosergiy@gmail.com;

Бісікало Олег Володимирович – д.т.н., професор, декан ФКСА, Вінницький національний технічний університет

Trachenko Serhii S. – Faculty for Computer systems and Automation, Vinnytsia National Technical University, Vinnytsia, email : trachenkosergiy@gmail.com;

Bisikalo Oleh V. – PHD, Professor, dean of the Faculty of Computer systems and Automation