

УДК 004.89

Цимбалюк Г. – ст. гр. СНм-61, Цвігун В. – ст.гр. СБ-11

Тернопільський національний технічний університет імені Івана Пулюя

РОЛЬ ТА ОСНОВНІ ВИКЛИКИ ЛЕМАТИЗАЦІЇ ТА СТЕММІНГУ В ОБРОБЦІ ПРИРОДНОЇ МОВИ

Науковий керівник: к.т.н., доцент Загородна Н.В.

Tsybaliuk G.O., Tsvihun V.

Ternopil Ivan Puluj National Technical University

THE ROLE AND MAIN CHALLENGES OF LEMATIZATION AND STEMING IN NATURAL LANGUAGE PROCESSING

Supervisor: PhD, associated professor Zagordna N.

Ключові слова: ЛЕМАТИЗАЦІЯ, СТЕММІНГ, ОБРОБКА ПРИРОДНОЇ МОВИ

Key words: LEMATIZATION, STEMING, NATURAL LANGUAGE PROCESSING

Обробка природної мови (NLP, англ. Natural Language Processing) — це галузь штучного інтелекту, що займається взаємодією між комп'ютерами та людськими мовами. Автори [1] поділяють NLP на 2 основних напрямки: генерація природної мови (NLG, англ. Natural Language Generation) і розуміння природної мови (NLU, англ. Natural Language Understanding). Дана стаття хоч і виділяє ці два основні напрямки та наводить сфери застосування NLP, проте не співвідносить їх. В даному дослідженні здійснено розподіл задач NLP між двома базовими напрямками.

До задач NLU можна віднести:

- Аналіз тональності текстів (Sentiment Analysis): займається автоматичним визначенням емоційного забарвлення тексту з метою розуміння ставлення автора до тієї чи іншої ситуації чи особи.
- Обслуговування клієнтів (Customer Service) допомагає здійснювати аналіз дзвінків, настроїв, очікувань клієнтів з метою розуміння намірів та емоцій.
- Керування рекламою (Managing the advertisement) відповідає за виявлення ключових слів для таргетингу.
- Виявлення спаму (Spam Detection) належить до типових задач класифікації текстів.

Хоча існують «чисті» задачі напрямку NLG, на практиці значно частіше використовується комбінація двох методів обробки природної мови:

- Машинний переклад (Machine Translation) - система має зрозуміти зміст вхідного тексту, розпізнати граматичну структуру, контекст тощо, а потім згенерувати відповідний текст іншою мовою, з дотриманням граматики та стилю цільової мови.
- Автоматичне резюмування тексту (Automatic Summary) - система повинна проаналізувати текст, виявити основні ідеї, структуру, важливі речення та ключові слова. Потім на основі аналізу система генерує короткий текст, який передає суть початкового повідомлення в узагальненій формі.
- Чат-боти (Chatbots) – виконують дві задачі – розуміння тексту запиту та формулювання відповіді.

Як бачимо розуміння природної мови має широкий спектр застосування і застосовується як для вирішення окремих задач, так і в комбінованих підходах. NLU охоплює розпізнавання, аналіз та інтерпретацію природної мови, яку люди використовують у повсякденному спілкуванні.

Стемінг і лематизація — це споріднені методи в обробці природної мови, обидва належать до етапу попередньої обробки (preprocessing) та нормалізації тексту, а отже є одним з етапів напрямку розуміння природної мови. Обидва методи намагаються привести слова до "основної форми", проте роблять це по різному. Стемінг урізає слова до короткої основи без розуміння значення, натомість лематизація приводить слово до лема – початкової форми.

З одного боку, техніка стемінг реалізується шляхом відсікання будь-якого кінця або початку слова, відстежуючи загальноживані префікси і суфікси, що дозволяє нормалізувати слова та зменшити їх кількість для аналізу. Проте з іншого боку ця «груба» техніка може давати наступний результат:

Слово=studies, суфікс = -es, стема = studi

Слово =studying, суфікс =-ing, стема=study.

Як бачимо, стемінг утворив з двох однакових за змістом слів два різних об'єкти в зв'язку з тим, що ця техніка просто обрізає закінчення слів, незалежно від того, чи має отриманий результат якесь змістовне значення. Крім того, як показують дослідження результат стемінгу важко прогнозувати, оскільки він буде різнитися в залежності від обраного алгоритму. Популярна бібліотека nltk в Python містить низку стемерів, які можна використовувати. Експериментальні дослідження показали, що RegexpStemmer очікувано трансформує «studying» до «study», проте «studies» стає «studie». Натомість PorterStemmer та SnowballStemmer дають результат для двох слів «studi».

Натомість лематизація, яка шукає основну форму слова, завжди приведе ці одного початкового слова:

Слово=studies, лема = study

Слово =studying, лема =study.

Хоча виконання стемінгу є простішим, а отже швидшим ніж процес лематизації, остання все ж є кращим вибором для задач, що вимагають кращої точності. Для лематизації потрібне глибоке лінгвістичне розуміння, щоб сформувати словник, який дозволяє алгоритму знаходити значущу частину слова. Лематизація використовує словниковий запас і морфологічний аналіз слів, щоб отримати результат.

В [2] зазначено, що є обмежений набір інструментів для реалізації стемінгу та лематизації українською мовою і запропоновано програмний модуль на PHP для стемінгу текстів українською мовою. Проте для задач NLP зручно використовувати бібліотеки Python, а отже необхідно дослідити можливості бібліотеки nltk для україномовних текстів.

Підсумовуючи, можна сказати, що стемінг, зазвичай використовують у задачах, де потрібна швидка обробка тексту, таких як пошукові системи або кластеризація текстів. Лематизація застосовується у задачах, що вимагають більш високої точності, таких як машинний переклад, аналіз тональності чи лінгвістичний аналіз.

Література.

1. Khyani Divya. An Interpretation of Lemmatization and Stemming in Natural Language Processing/ Khyani Divya , Siddhartha B.S., Niveditha N.M., Divya Y M, Manu Y.M. // Journal of University of Shanghai for Science and Technology. -2021. – 22(10).P. 350-357.

2. Глибовець А. М. Алгоритм токенизації та стемінгу для текстів українською мовою / Глибовець А. М., Точицький В. В. // Наукові записки НаУКМА. Комп'ютерні науки. - 2017. - Т. 198. - С. 4-8.