УДК 004.021
С. А. Лупенко, докт. техн. наук, професор, В.О. Васьков
Тернопільський національний технічний університет імені Івана Пулюя, Україна

АНАЛІЗ МЕТОДІВ ДЛЯ ЗАДАЧ ОПРАЦЮВАННЯ ПРИРОДНОЇ МОВИ

S.A. Lupenko, Dr., Prof., V.O. Vaskov

ANALYSIS OF METHODS FOR NATURAL LANGUAGE PROCESSING TASKS

Застосування методів машинного навчання до текстів на природній мові може дати багато цікавих і корисних результатів, наприклад – автоматичне сортування текстів за темами (завдання класифікації), пошук схожих текстів (задача кластеризації), автоматичний переклад та ін. Для того, щоб застосувати математичні методи до текстів необхідно певним чином формалізувати дані. У разі класифікатора текстів формалізація виконується з допомогою частотного аналізу (g), у цьому випадку кожному тексту T ставимо у відповідність точку в просторі ознак $X \subset Rn$ [1].

Цей метод досить добре працює для текстів середнього розміру, однак для коротких повідомлень частотна характеристика може виявитися неінформативною. Не підходить цей метод і для завдань машинного перекладу, де цікавою є не загальна характеристика тексту, а кожне слово окремо і послідовності зі слів. Таким чином, виникає необхідність побудувати ефективний метод кодування окремих слів.

Слова можна кодувати різними методами. Напевно найпростіший спосіб це їх пронумерувати, тобто складаємо повний словник з тексту, збираємо всі можливі форми слів, використані в тексті, і нумеруємо всі ці слова. Але такий спосіб кодування не несе ніякого смислового навантаження, тобто за кодом не можна сказати наскільки близькими за змістом є слова.

В 2013 році команда дослідників компанії Google розробила метод побудови «осмисленого» простору для слів – Word2Vec [2].

Word2Vec ґрунтується на тому, що слова, які мають подібний контекст, мають також подібні смислові значення. В основі технології Word2Vec лежить представлення слів у вигляді векторів заданої розмірності, розташовуючи схожі слова близько один до одного [3].

Для створення бази відповідностей «слово – вектор», алгоритм спочатку переглядає весь виданий йому текст, формулюючи, таким чином, «словник», який в наступних етапах роботи алгоритму, буде використаний для визначення відповідних векторів.

Результатом роботи Word2Vec є набір векторів (матриця) – кодів слів, яку отримуємо в результаті навчання певної нейронної мережі на деякому тексті (впорядкованій множині слів), й перше, що потрібно зробити, це підготувати набір навчальних даних.

Для навчання мережі Word2Vec застосовують два основних методи: CBOW (Continuous Bag of Words) і Skip-gram (рис. 1).

CBOW – модельна архітектура, яка передбачає поточне слово, виходячи з навколишнього його контексту. Архітектура типу Skip-gram діє інакше: вона використовує поточне слово, щоб передбачати оточуючі його слова [4].

Реалізація Word2Vec складається з трьох частин:

– кодування – на вхід надходить відфільтрований текст, на виході отримуємо кодовані набори слів тексту P і контекстів Q;

– навчання мережі – на вхід надходить навчальний набір P, Q, на виході

маємо матрицю уявлень Vi;

– тест результату – на вхід надходить словник і матриця уявлень Vi, на виході отримуємо випадково вибрані слова зі словника і найбільш близькі до них слова Word2Vec.

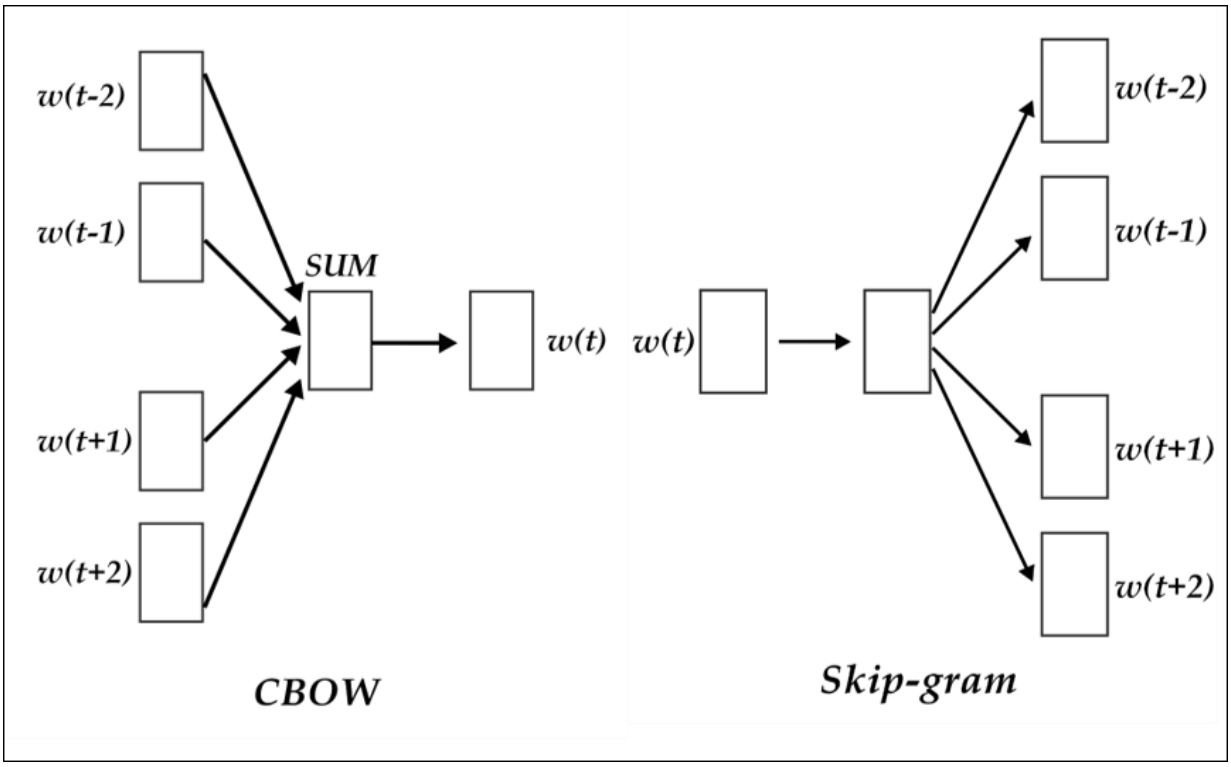

Рисунок 1. Структура моделей CBOW та Skip-gram

Література

1. Rong X. word2vec Parameter Learning Explained [Електронний ресурс] / Xin Rong. – 2016. – Режим доступу до ресурсу: https://www.researchgate.net/publication/268226652_word2vec_Parameter_Learning_Explained.

2. Lakhey M. Word2vec Made Easy [Електронний ресурс] / Munesh Lakhey. – 2016. – Режи доступу до ресурсу: https://towardsdatascience.com/word2vec-made-easy-139a31a4b8ae.

3. Nayak M. An Intuitive Introduction of Word2Vec by Building a Word2Vec From Scratch [Електронний ресурс] / Manish Nayak. – 2019. – Режим доступу до ресурсу: https://medium.com/towards-artificial-intelligence/an-intuitive-introduction-of-word2vec-by-building-a-word2vec-from-scratch-a1647e1c266c.

4. Karani D. Introduction to Word Embedding and Word2Vec [Електронний ресурс] / Dhruvil Karani. – 2018. – Режим доступу до ресурсу: https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa.