УДК 004.4

Козакевич В.А., Собко О.В., Тищенко О.О., Вознюк Л.О., Медведчук В.Ю.

Хмельницький національний університет

МЕТОД АВТОМАТИЗОВАНОЇ ГЕНЕРАЦІЇ ТЕКСТОВИХ ПОВІДОМЛЕНЬ ЗАДАНОЇ СЕМАНТИЧНОЇ СПРЯМОВАНОСТІ З ВИКОРИСТАННЯМ ЛЕКСИЧНИХ N-ГРАМ

Було розглянуто ключові приклади розробки інформаційних систем, які автоматично генерують текстові повідомлення з визначеною семантичною спрямованістю за допомогою лексичних n-грам і були отримані результати роботи цих систем для подальшого аналізу.

The key examples of the development of information systems that automatically generate text messages with a defined semantic orientation using lexical n-grams were considered, and the results of the work of these systems were obtained for further analysis.

Генерація текстових повідомлень – це автоматичне створення або надсилання невеликих обсягів текстової інформації окремим користувачам або групам користувачів з невеликими змінами або без них. Автоматизація такого маркетингу набула популярності серед багатьох брендів, оскільки вона дозволяє їм ефективно взаємодіяти зі своєю новою аудиторією в широкому масштабі. Крім того, автоматична генерація текстових повідомлень часто застосовується в системах сповіщення [1].

У наш час, коли сучасні технології швидко розвиваються і стають все більш інтуїтивно зрозумілими, продуктивними і зручними для користувача, важливо мати засоби ефективного спілкування з ним. Це стає невід'ємною складовою кожного проекту чи роботи. Для підтримки зв'язку з потенційними або постійними клієнтами, укладення контрактів з працівниками компанії та в щоденній робочій атмосфері все частіше використовують засоби для автоматизованої генерації невеликих текстових повідомлень.

Такі системи для автоматичної генерації невеликого обсягу текстових повідомлень застосовуються в різних сферах. Вони можуть бути використані як чат-боти, програми для автоматичного сповіщення користувачів, для модерації активності постійних учасників у групових чатах і соціальних мережах.

Однією з ключових сфер використання є чат-боти. У цьому випадку обмін текстовими повідомленнями між користувачем і сервісом відбувається негайно та нагадує звичайний діалог між двома особами. Однак, з одного боку, одним із учасників є реальна особа, а з іншого - система з автоматизованою генерацією невеликих текстових повідомлень. Також можливий обмін повідомленнями з

більшою кількістю зареєстрованих користувачів. Якщо в бесіді бере участь більше двох осіб одночасно, це вже називається чатом. Також важливою є можливість використання аудіо асистентів [2], які автоматично генерують відповіді у текстовому або аудіо форматі.

Один із прикладів використання цього підходу полягає у використанні типових моделей, який детально розглядається у роботі Ланкгкілде та Найтта. Задача генерації природної мови [3] є важливою складовою роботи системи автоматичної відповіді користувачу.

Моделі, які ґрунтуються на роботі з конкретними випадками, в даному контексті аналізують вхідні дані як набір порівнюваних ситуацій, які відбуваються у конкретному середовищі. Наприклад, задавши однакові запитання великій кількості людей, можна аналізувати їх відповіді, що в свою чергу утворює групу відповідей з однаковою семантикою та змістом. Grounded theory[4] ґрунтується на цьому підході у своїй роботі. Вони використовували модель HALogen [5], яка представлена на рисунку 1.

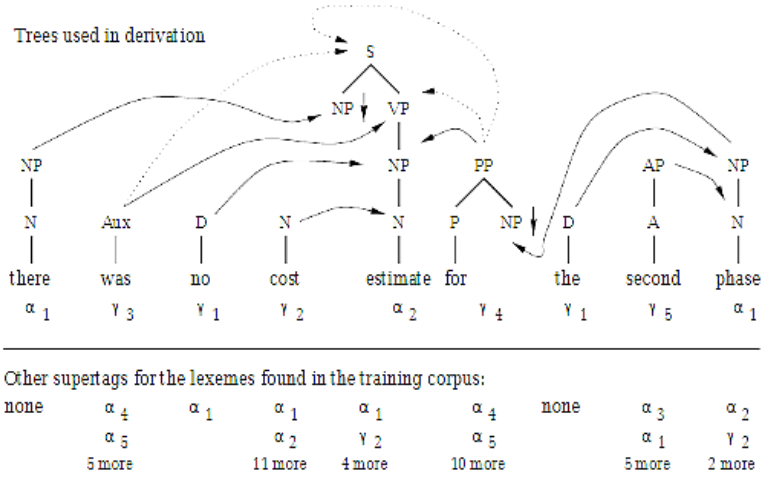

Рисунок 1 – HALogen [5]

Граматична модель у цій реалізації представлена у вигляді структури, що називається "випадковий ліс" (random forest). За допомогою стохастичного алгоритму оцінювання може бути вибраний найбільш підходящий варіант. У даному випадку цей підхід використовує базу у формі n -грам, тоді як сучасні підходи виконують великі експерименти з сучасними та інноваційними моделями для досягнення кращих результатів[6]. Існують різні види n -грам.

Уніграми (1-грами) – це найпростіші н-грами, які представляють окремі слова. Вони не містять контексту та розглядаються ізольовано. Наприклад, у реченні "Сьогодні гарний день", уніграмами будуть "Сьогодні", "гарний", "день".

Біграми (2-грами) – це n-грами, які складаються з двох слів. Вони дозволяють враховувати контекст, але тільки в обмеженому обсязі. Наприклад, для того ж речення, біграмами будуть "Сьогодні гарний", "гарний день".

Триграми (3-грами) – це n-грами, які мають три слова в послідовності. Вони враховують більший контекст, ніж біграми. Наприклад, для того ж речення, триграмами будуть "Сьогодні гарний день".

Чотириграми (4-грами) – це n-грами з чотирма словами в послідовності.

Що стосується використання н-грамів, вони використовуються для створення статистичних моделей мови, які визначають ймовірність входження певної послідовності слів у текст. Автоматичний переклад, використовуються для визначення найбільш ймовірних перекладів.

Розпізнавання мови в основному використовуються для визначення найімовірніших слів або фраз в аудіозаписах. Автокоректори та системи підказки. Використовують н-грами для визначення найбільш ймовірного продовження введеного тексту.

Важливо враховувати, що зі збільшенням значення n (кількість слів в n-грамі), зростає складність обробки та обсяг затребуваної статистики для навчання моделі. З іншого боку, більші значення n дозволяють враховувати більший контекст та покращувати якість результатів.

N-грами знаходять застосування у багатьох галузях, де важливо аналізувати текстову інформацію та розуміти її контекст. Використовуються для прогнозування ймовірностей появи певних слів чи фраз у тексті. Це важливо для завдань, таких як автоматичний переклад, розпізнавання мови, генерація тексту тощо. N-грами використовуються для визначення найбільш ймовірних перекладів для конкретних слів чи фраз. Використовують н-грами для визначення найбільш ймовірного продовження введеного тексту, що допомагає вказувати на можливі помилки або надавати рекомендації. N-грами використовуються для визначення найбільш ймовірних слів чи фраз в контексті генерування тексту. N-грами можуть бути використані для прогнозування подій чи трендів на основі аналізу текстових даних.

N-грами, хоч і є потужним інструментом для аналізу тексту та розуміння контексту, мають кілька недоліків. Для багатьох завдань, особливо в глибоких аналізах, n-грами обмежені в тому, що вони не можуть врахувати дуже великий контекст. Це особливо стає проблемою у випадках, коли значення слова залежить від довшого контексту. Велика кількість n-грам може бути рідкісними, тобто вони можуть взагалі не зустрічатися в тренувальних даних. Це ускладнює побудову надійних статистичних моделей. Для великих значень n (наприклад, 4-грами та вище), розмір словника та потрібна кількість обчислень для побудови моделі можуть бути вкрай великими. N-грами враховують тільки послідовність слів, але не

їхній фактичний порядок. Це може призводити до неправильних або незрозумілих висновків.

Інший приклад цього може бути підхід, що не вимагає складних обчислень для сортування та автоматичної генерації. Він використовує статичну інформацію для генерації відповіді у момент вибору та прийняття рішення. Цей метод часто використовується в системі PCRU, яка була вперше запропонована в 2007 році [7]. Робота цієї системи показана на рисунку 2. Вона може генерувати потенційне закінчення речення, яке має найбільшу ймовірність.

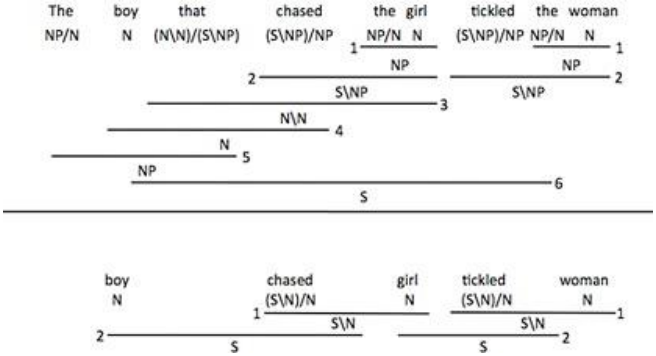

Рисунок 2 – Робота системи PCRU

Вагомим прикладом є фреймворк OpenCCG[8]. Цей інструмент є потужним засобом для аналізу великих повідомлень різних форматів та складності. Він використовує граматичний підхід [9] і базу початкових даних, побудовану на основі бібліотеки Penn Treebank (рисунок 3), яка широко використовується для статистичної оцінки мовних моделей.

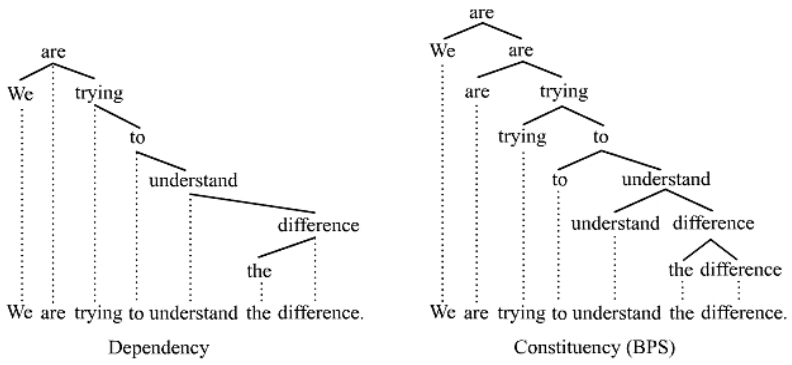

Рисунок 3 – Penn Treebank

Щоб розпочати роботи з цією системою автоматичної генерації текстових повідомлень потрібно підготувати корпус (див. рисунок 4), на якому ми будемо навчати нашу модель, а також визначити необхідну кількість слів у вихідному тексті. Генератор видасть "очищену" послідовність слів та всіх розділових знаків. Також доданий ще один генератор, який надає три токени поспіль. У цьому випадку токенами можуть бути слова або символи.

Рисунок 4 – Загальна схема роботи методу побудови текстів за допомогою n-грам

Наступним кроком є додавання функції, яка виводить кілька слів поспіль. Це полегшує вибір першого слова у фразі, що генерується. Узагалі, метод працює так: він повертає три токени поспіль, кожен наступний зрушується на один рівень після кожної ітерації.

Для початку ініціалізуємо генератори. Потім обчислюємо n-грами та визначаємо ймовірність кожного слова в залежності від попередніх. Далі, це слово та його ймовірність додаються у словник. Важливо зауважити, що цей метод може не бути найбільш оптимальним, оскільки він може вимагати значних ресурсів пам'яті. Проте для невеликих корпусів він є досить ефективним.

Цей метод ґрунтується на поступовому виборі найбільш ймовірних слів та розділових знаків до того моменту, коли виявляємо початок наступної фрази.

Основною метою є інтеграція всіх компонентів та етапів в одній системі машинного навчання. Це може призвести до створення оптимізованих та зручних систем, які не потребують попередньої обробки вхідних даних або редагування та форматування тексту.

Тому були ретельно проаналізовані ключові приклади розробки інформаційних систем, які використовують лексичні n-грами для автоматичної генерації текстових повідомлень з визначеною семантичною спрямованістю. Результати цієї роботи свідчать про значущий прогрес у сфері автоматизованої генерації контенту та його аналізу.

Перелік посилань

1. Text Message Automation. URL:https://www.slicktext.com/text-message-automation.php
2. Google Assistant. URL: https://www.techrepublic.com/article/google-assistant-the-smart-persons-guide/
3. Natural language generation. URL: https://research.aimultiple.com/nlg/
4. A design framework for novice researchers. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6318722/.
5. Model for Generation. URL: https://www.researchgate.net/publication/2473335_Exploiting_a_Probabilistic_Hierarchical_Model_for_Generation;
6. OpenCCG. URL: http://openccg.sourceforge.net/;
7. Combinatory categorial grammar. URL: https://en.wikipedia.org/wiki/Combinatory_categorial_grammar;
8. Chart Generation grammar. URL: https://www.inf.ed.ac.uk/teaching/courses/nlg/readings/KayACL96.pdf;
9. Grammar-Based Approach to Microplanning. URL: https://www.aclweb.org/anthology/J17-1001.pdf