Єременко А.І., магістрант
Вакалюк Т.А., д.пед.н., професор
Державний університет «Житомирська політехніка»

ГЕНЕРАЦІЯ ПРИРОДНОЇ МОВИ

Генерація природної мови (англ. Natural Language Generation, NLG), є частиною процесу обробки природної мови, яка займається створенням та вивченням додатків, для відтворення інформації у вигляді зрозумілого для людей тексту. Даний процес приймає вхідні дані у різному вигляді (текст, фото, таблиця, тощо.), після підготовки даних та їх аналізу, готує відповідь у найдоречнішому вигляді. Серед реальних прикладів, де цей процес відбувається, можна навести, як відповіді чатботів, так і переклади текстів з одної мови на іншу. Перевагою даного підходу є збереження сенсу речень та структури тексту загалом [1].

Деякі вчені визначає описують структуру системи, яка спирається на певні правила, чітко розділяючись на 3 рівні. Першим є планування документу, що визначає порядок у якому інформація буде написана. Другим є мікропланування, де генеруються посилання на об'єкти або сутності, разом з вибором слів, які будуть застосовуватися. Також відбувається зіставлення схожих речень для покращення читабельності. Останній, третій рівень, реалізація, тут генерується фактичний текст, використовуючи реальні правила синтаксису та морфології. Надалі починається швидкий розвиток технологій пов'язаних з генерацією природної мови.

Дана сфера спирається на деякі «цілі», для вирішення яких проводиться дослідження та розробка штучного інтелекту, серед них є: підсумовування даних, машинний переклад, генерація діалогу, перефразування, формування запитань, та багато іншого. Фактично всі вони описують різні проблеми генерації текстового відповіді на різні вхідні дані. Також моделі розділяють на текст до тексту та дані до тексту. Текст до тексту, фокусується на генерації відповіді на певну команду (англ. promt). Дані до тексту займаються аналізом даних, наприклад для генерації звітів по продажам на основі результатів запитів до бази даних.

Розглянемо задачі, які виконуються в процесі генерації природної мови, їх кількість та складність алгоритмів можуть відрізнятися від моделі до моделі, спираючись на фінальну ціль розробки.

Першою задачею виділяють визначення змісту. На початку система повинна зрозуміти, яку інформацію потрібно донести до користувача, на чому зосередити увагу. Дані з якими працює модель на даному етапі мають набагато більший об'єм, ніж можливо описати словами. Для прикладу розглянемо ситуацію, користувач надсилає фото з собакою, без коментарів. Штучний інтелект, може надати декілька відповідей, прокоментувати фотографію, провести аналіз кольорової палітри, визначити породу тварини. Ось тут і постає вибір, якій повинен відбутися для генерації відповіді. Звісно, що коли вхідна команда більш детальна та має чіткі інструкції, даний етап спрощується.

Наступною задачею є структурування тексту. Після того, як зміст було визначено, системи повинна вирішити порядок, у якому надавати інформацію. Іноді доречніше почати з загальної інформації, а інколи навпаки, потрібно чітко дати відповідь по суті. Результатом даного процесу є розмітка тексту, з планом, де й що необхідно написати.

Далі відбувається агрегація речень. Тут аналізується інформація, з метою об'єднання декількох схожих повідомлень у одне речення. Наприклад замість того щоб генерувати речення на кожну подію, яка відбувається періодично, можна зробити одне речення, яке буде вказувати на періодичну суть події. Деякі моделі не використовують даний етап, частіше всього, це сфери у яких вимагається точність та повність інформації, наприклад звітність.

Після планування тексту та заготовки структур речень відбувається лексикалізація. Її суть полягає у представленні інформації, яка вже була проаналізована та запланована, у вигляді слів та їх об'єднань. Чим краще навчена модель, там складнішим стає даний етап, деякі події можна описати по різному, використовуючи синоніми. Тому існує декілька підходів для вибору доречних слів, це може бути і вибір прямого варіанту, тобто опис події, так як вона і називається, а може бути проведений додатковий аналіз, на основі якого будуть обрані варіанти.

Наступною задачею постає генерація реферальних виразів. Вона схожа на лексикалізацію, але займається відокремлюванням об'єктів, про які потрібно написати, один від одного. Наприклад, в залежності від контексту, на фото знаходиться певний куб, для нього можна підібрати декілька слів, коробка, ящик, знову-таки, куб. Дана задача є складною, оскільки потрібно дивитися на різні характеристики об'єктів, дослідження та покращення цього процесу активно проводяться на даний момент.

Останньою задачею, яка повинна бути вирішена, є лексична реалізація. На даному етапі машина має план тексту, вона вирішила, які слова потрібно використати, та на чому необхідно сфокусувати увагу. Залишилося лише скласти речення з слів, які можна представити у вигляді будівельних блоків. Речення повинні бути побудовані правильно, використані доречні форми слів, структури повинні бути зв'язані. Виділяють 3 підходи: шаблонний, на основі граматики та статистичний. Кожен з них підходить до різних сфер. Шаблонний базується на основі шаблонів, які були розроблені людьми, штучному інтелекту потрібно лише підставити слова на відповідні місця. Системи на основі граматики фокусуються на побудуванні граматично вірних речень, вони є складнішими у реалізації та використанні, оскільки вимагають точних вхідних даних. Статистичний підхід базується на основі проаналізованих граматичних правил та прикладів для створення речень. Наприклад деякі слова частіше зустрічаються на початку речення.

Моделі генерації природної мови є модульними, тобто мають чітке розділення між рівнями, що дозволяє додавати або видаляти рівні по необхідності. Архітектура, у свою чергу, може відрізнятися, вона може бути реалізована у вигляді прямої магістралі (англ. pipeline), де кожен модуль передає результати наступному, так і у вигляді циклів, де після виконання кількох модулів, результат повертається на початок циклу для доопрацювання. Також це дозволяє використовувати моделі у більш складних системах, як проміжний етап перетворення інформації, наприклад для подальшого використання у процесах машинного навчання, дані можуть бути класифіковані та структуровані, за

схожим до роботи людини, виглядом.

Одним з найважливіших етапів розробки моделей генерації природної мови є оцінка. Даний процес є складним, через невизначеність результатів, які можна отримати при виконанні завдань. Наприклад, відповідь чат бота на запитання про сьогоднішню погоду, може мати різний вигляд та містити різну інформацію. Оцінка моделей людиною вважається золотим стандартом, але вона є дорогою, особливо у випадках, коли необхідно перевіряти модель майже кожен день під час активної розробки або оптимізації. Тому даний процес окремо досліджується інженерами, які намагаються знайти автоматизоване рішення задачі оцінки моделей.

Серед основних методів, які використовуються на сьогодні виділяють: людино-центричний, автоматичний нетренований та метрики на основі машинного навчання [2].

Людино-центричний метод полягає у тому, що певна людина виступає у ролі судді, та визначає якість моделі. Найвідоміший приклад, це тест Тюрінга, суть якого полягає у тому, що людині надають декілька текстів, один з яких написаний машиною, та запитують відрізнити текст створений моделлю від інших. Головною перевагою даного методу є його гнучкість, оскільки людина може адаптуватися та провести оцінювання з різних сторін, як для граматики так і для загальної плавності.

Другим методом оцінювання виступає нетреноване автоматичне оцінювання. Його суть полягає у підборі алгоритмів, які автоматично порівнюють тексти написанні людьми та машинами, для цього використовуються певні статистичні метрики, такі як відповідність n-грам, накладання рядків та вмісту. Дуже важливо підібрати правильний алгоритм, якій відповідає цілі, для якої розробляється модель, інакше оцінювання не буде коректним. Також даний метод не вимагає проведення додаткового машинного навчання.

Останнім методом оцінки є оцінювання за допомогою метрик на основі машинного навчання. Для проведення даного процесу використовуються моделі, які розроблені з ціллю порівняння текстів написаних людьми та машинами, вони імітують роль людини як судді. Одним з варіантів вирішення даної проблеми є порівняння семантичних подібностей, але даний метод не підходить для випадків, коли відповіді суттєво відрізняються один від одного. Тоді використовують моделі навчені на основі існуючих результатів людино-центричного методу, але даний підхід є дорожчим та більш часозатратним.

Генерація природної мови широко використовується для створення контенту та автоматизації різних процесів. Сучасні моделі демонструють значний прогрес у генерації текстів, схожих до людських, що породжує все більше питань з доречності розвитку даного напрямку, викладачі все частіше бачать згенеровані штучним інтелектом тексти. Це впливає на розвиток сфери додатків для виявлення згенерованого контенту, на що розробники моделей відповідають ще швидшим розвитком. Складно уявити сьогодення без використання штучного інтелекту для пошуку необхідної інформації у зрозумілому вигляді за секунди.

Список використаних джерел

1. Gatt A., Krahmer E. 2018 Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation, Journal of Artificial Intelligence Research, Vol 61, P. 65-170.

2. Evaluation of Text Generation: A Survey. / Celikyilmaz A., Clark E., Gao J. 2021. 1. 911-921. https://arxiv.org/pdf/2006.14799.