УДК 004.9:004.65

В. М. Борисюк[1]
А. В. Козловський[1]

ГЕРЕРАЦІЯ SQL ЗАПИТІВ ВИКОРИСТОВУЮЧИ ПІДХОДИ ОБРОБКИ ПРИРОДНЬОЇ МОВИ

[1]Вінницький національний технічний університет

Анотація
A method for executing database queries using a natural language interface has been proposed, addressing a significant problem in database administration for people who lack SQL query writing skills. A module for the interface that can generate SQL instructions from natural language queries has been presented. Semantic grammar is used in the proposed architecture to convert queries in English into SQL queries that can be executed in the database.
Ключові слова: обробка природньої мови, SQL, QCNER, розпізнавання іменованих сутностей, СУБД.

Abstract
In the modern era of information technology, access to relevant data is critically important for business users. However, many users do not possess SQL language skills or find writing long queries challenging. Automating the process of creating SQL queries using natural language processing (NLP) can significantly simplify access to databases. This paper presents the QCNER approach for automatic generation of SQL queries from natural language.
Keywords: natural language processing, SQL, QCNER, named entity recognition, DBMS.

В сучасну епоху інформаційних технологій доступ до релевантних даних є критично важливим для бізнес-користувачів. Проте багато користувачів не володіють мовою *SQL* або вважають складним написання довгих запитів. Автоматизація процесу створення *SQL*-запитів за допомогою природної мови (*NLP*) може значно спростити доступ до баз даних. У даній роботі представлено підхід *QCNER* для автоматичної генерації *SQL*-запитів з природної мови [1].

В якості набору даних було взяти відкритий набір даних *WikiSQL* з платформи Kaggle [2]. *WikiSQL* - великий набір даних, зібраний з різних джерел, для розробки інтерфейсів природної мови для реляційних баз даних. Приклад набору даних зображено в таблиці 1.

Таблиця 1 – Приклад набору даних *WikiSQL*.

Запит	Зрозумілий для людини SQL запит
How many different college/junior/club teams provided a player to the Washington Capitals NHL Team?	SELECT COUNT College/junior/club team FROM table WHERE NHL team = Washington Capitals
What could a Spanish coronel be addressed as in the commonwealth military?	SELECT Commonwealth equivalent FROM table WHERE Rank in Spanish = Coronel

QCNER (*Query Conversion using Named Entity Recognition*) — це підхід, який використовується для автоматичної генерації *SQL*-запитів з природної мови. Цей підхід поєднує техніки обробки природної мови (*NLP*) з методами машинного навчання для перетворення запитів, сформульованих природною мовою, у *SQL*-запити. Основні етапи підходу QCNER включають:

1. Токенізація та попередня обробка тексту.
   a. Токенізація розбиває текст на окремі слова або токени;
   b. Видалення стоп-слів (таких як "*and*", "*the*", "*a*"), які не несуть змістовного навантаження.
   c. Лематизація або стемінг для приведення слів до їх базової форми.
2. Факторизація цільової колонки.
   a. Застосування методів машинного навчання, таких як наївний баєсів класифікатор, для аналізу та обробки попередньо оброблених даних;
   b. Використання методів балансування даних, таких як *SMOTE*, для покращення моделі на незбалансованих наборах даних.

3. Розпізнавання іменованих сутностей (*NER*).
    a. Розробка спеціальної *NER* моделі для витягування сутностей з запитів природною мовою;
    b. Класифікація сутностей у категорії;
    c. Збереження сутностей та відповідних таблиць у словнику для подальшого використання.
4. Генерація *SQL*-запиту.

Підхід *QCNER* дозволяє ефективно перетворювати запити природною мовою у *SQL*-запити, що робить його корисним інструментом для користувачів, які не володіють мовою *SQL*, але потребують доступу до даних у базах даних.

Розглянуто наступній алгоритми обробки природної мови для перетворення природної мови в *SQL* запит:

1. *SMOTE*(*Synthetic Minority Over-sampling Technique*) – техніка надмірної дискретизації синтетичної меншості [3].
2. *Random forest* – алгоритм випадкових лісів [4].
3. *Multinomial Naive Bayes* - наївний баєсів класифікатор [5].
4. *SVM* (*Support vector machine*) – метод опорних векторів [6].
5. *KNN* (*K Neighrest Negihbours*) – метод k-найближчих сусідів [7].
6. *NER* (*Named Entity Recognition*) - розпізнавання іменованих сутностей [8].

SMOTE це метод генерації нових зразків для менш чисельного класу у наборі даних. Він допомагає вирішити проблему дисбалансу класів, що часто виникає при навчанні моделей машинного навчання.

Random Forest — це метод ансамблевого навчання, який об'єднує багато дерев рішень для покращення точності та зменшення перенавчання. Класифікація в *Random Forest* визначається шляхом голосування більшості серед дерев рішень, і кожне дерево навчається на випадково обраній частині навчальних даних. *Random Forest* відомий своєю високою точністю та стійкістю до шуму та викидів у даних.

Multinomial Naive Bayes використовується для виконання завдань класифікації. Основою класифікатора є теорема Баєса. При вирішенні задач класифікації документів, таких як визначення, чи належить документ до певної категорії таблиці, широко використовується багатономінальний наївний Баєс. Однією з характеристик або предикторів, які використовує класифікатор, є частота термінів у документі.

SVM використовується для задач класифікації та регресії, називається методом опорних векторів (*SVM*). Гіперплощина, яка найкраще розділяє класи в даних, знаходиться за допомогою *SVM*. *SVM* може обробляти нелінійні дані шляхом їх перетворення у вищий вимірний простір, де класи можуть бути розділені лінійною гіперплощиною. SVM відомий своєю високою точністю та здатністю обробляти складні набори даних, але вибір функції ядра та гіперпараметрів може впливати на його продуктивність**.**

KNN використовується як для класифікації, так і для регресії. Знаходження k найближчих точок даних у навчальному наборі та присвоєння найбільш поширеного класу серед них дозволяє *KNN* класифікувати нову точку даних. *KNN* простий у використанні та часто добре працює на невеликих наборах даних.

NER відповідає за розпізнання та класифікацію іменованих сутностей процесу розпізнавання іменованих сутностей в обробці природної мови (NLP). Використовуючи сирий та структурований текст, визначені сутності поділяються на людей, організації, локації, гроші, час тощо. По суті, іменовані сутності ідентифікуються та сортуються у кілька груп. Системи NER розробляються з використанням різноманітних лінгвістичних стратегій, а також статистичних та машинних методів навчання.

Використовуючи вище наведені підходи методів машинного навчання та набір даних *WikiSQL* отримано порівняльну таблицю 2. Підчас порівняння підходів було обраховано наступні метрики: точність, чутливість, специфічність.

Чутливість або повнота — це метрика, яка показує, наскільки добре модель розпізнає справжні позитивні випадки серед усіх фактичних позитивних зразків.

Специфічність — це метрика, яка показує, наскільки добре модель розпізнає справжні негативні випадки серед усіх фактичних негативних зразків.

Таблиця 2 – порівняння точності різних підходів для перетворення природної мови в SQL запит.

Алгоритм	Точність	Чутливість	Специфічність
Random forest	94.6%	93.2%	95.4%
SVM	92.7%	91.5%	93.8%
Multinomial Naive Bayes	89.4%	87.9%	90.2%
KNN	88.1%	86.3%	89.0%

Запропонований підхід *QCNER* для автоматичної генерації *SQL*-запитів з природної мови демонструє високу ефективність і точність, спрощуючи доступ до баз даних для користувачів, які не володіють мовою *SQL*. Використання методів обробки природної мови (*NLP*) та машинного навчання дозволяє ефективно перетворювати запити природною мовою у структуровані *SQL*-запити. Результати дослідження показують, що алгоритми, такі як *Random Forest, SVM, Multinomial Naive Bayes* та *KNN*, досягають високої точності, чутливості та специфічності, що робить їх придатними для застосування у реальних системах.

Використання набору даних *WikiSQL* дозволило провести порівняння ефективності різних підходів та алгоритмів. Результати показують, що підхід *QCNER* має значний потенціал для подальшого розвитку та впровадження у різних сферах, де необхідний доступ до баз даних без знання *SQL*. Це може включати бізнес-аналітику, освітні платформи та інші галузі, де користувачі часто потребують швидкого та зручного доступу до інформації.

У майбутньому можливі подальші дослідження для покращення роботи з більш складними запитами, які включають з'єднання, обмеження та групування, а також розширення функціоналу для роботи з різними мовами та діалектами. Це дозволить ще більше розширити застосування підходу QCNER і зробити його ще більш універсальним та ефективним інструментом для роботи з базами даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Chaudhari, M.S., Hire, M.A., Mandale, M.B., & Vanjari, M.S. Structural Query Language Question Creation by using Inverse Way, 2021.
2. Офіційна сторінка набору даних *WikiSQL* [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/shahrukhkhan/wikisql
3. Ling, Charles X., and Chenghui Li. "Data mining for direct marketing: Problems and solutions." Kdd. Vol. 98. 1998.
4. Tin Kam Ho, Random Decision Forests, Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, 1995.
5. McCallum, Andrew. Graphical Models. Lecture2: Bayesian Network Representation, 2019.
6. Cortes, Corinna; Vapnik, Vladimir Support-vector networks, 1995.
7. Fix, Evelyn; Hodges, Joseph L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties, 1951.
8. Elaine Marsh, Dennis Perzanowski. MUC-7 Evaluation of IE Technology: Overview of Results, 1998.

Борисюк Володимир Миколайович – аспірант кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, e-mail: *volodymyr.borysiuk0@gmail.com*

Козловський Андрій Володимирович – канд. техн. наук, доцент, доцент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, e-mail: akozlovskyi@vntu.edu.ua.

Volodymyr Borysiuk M. – *PhD. Student of Computer Science Department, Vinnytsia National Technical University, Vinnytsia, e-mail:* volodymyr.borysiuk0@gmail.com

Andrii Kozlovskyi V. — *Cand. Sc. (Eng), Assistant Professor of Computer Science Department, Vinnytsia National Technical University, Vinnytsia.*