УДК 004.85

А. А. Хрустовський[1]
А. В. Кожем'яко[1]

ОГЛЯД ВІДКРИТИХ НАУКОВИХ БАЗ ЗНАНЬ, ЯКІ ВИКОРИСТОВУЮТЬСЯ ДЛЯ НАВЧАННЯ ШІ

[1] Вінницький національний технічний університет

Анотація

У статті досліджено використання штучного інтелекту в різних сферах. Визначено важливість баз знань у навчанні штучного інтелекту. Проаналізовано популярні відкриті бази знань, включаючи якість даних, охоплення та доступність.

Ключові слова: наукові бази знань, штучний інтелект, стандартизація даних, інтероперабельність.

Abstract

The article examines the use of artificial intelligence in various fields. The importance of knowledge bases in artificial intelligence training is determined. Popular open knowledge bases are analysed, including data quality, coverage and accessibility.

Keywords: scientific knowledge bases, artificial intelligence, data standardisation, interoperability.

Вступ

Штучний інтелект (далі - ШІ) в останні роки переживає безпрецедентний ріст і розвиток, і одним з ключових факторів, що зумовлюють цей прогрес, є наявність великих і різноманітних наборів даних для навчання ШІ-моделей. Відкриті бази наукових знань, які є колекціями структурованих і неструктурованих даних, отриманих з наукової літератури, стали цінним ресурсом для навчання моделей ШІ в різних галузях, включаючи медицину, біологію, хімію, фізику та інші. У цій статті ми розглянемо сучасний ландшафт відкритих баз наукових знань, які використовуються для навчання ШІ, їхні переваги, проблеми та потенційні напрямки розвитку.

Відкриті бази наукових знань - це великі колекції структурованих і неструктурованих даних, зібраних з різних джерел, таких як академічні журнали, наукові публікації та бази даних. Ці бази знань використовуються для навчання моделей машинного навчання, які можна застосовувати в різних галузях, таких як обробка природної мови, комп'ютерний зір і робототехніка.

Основна частина

ШІ змінює наш спосіб життя і роботи, застосовуючись у різних сферах - від обробки природної мови до автономних транспортних засобів. Одним із ключових чинників розвитку ШІ є наявність великих обсягів даних для навчання моделей машинного навчання. В останні роки відкриті бази наукових знань стали цінним ресурсом для навчання моделей ШІ, пропонуючи величезне сховище структурованих і неструктурованих даних з різних галузей.

Відкриті бази наукових знань - це всеосяжні сховища наукових даних, відкриті для дослідників і розробників. Ці бази знань створюються шляхом агрегації та кураторства даних з широкого кола джерел, включаючи наукові публікації, бази даних, патенти, клінічні випробування та інші наукові ресурси. Деякі з популярних відкритих наукових баз знань, що використовуються для навчання ШІ, включають PubMed, PubMed Central, Chemical Entities of Biological Interest (ChEBI), Gene Ontology (GO), Protein Data Bank (PDB) і ClinicalTrials.gov, серед інших. Ці бази знань містять величезні обсяги даних з різних галузей науки і є багатими джерелами інформації, які можна використовувати для навчання моделей ШІ [1,2].

Переваги використання відкритих наукових баз даних для навчання ШІ багатогранні. По-перше, вони надають широкий і різноманітний спектр даних, що дозволяє моделям ШІ навчатися на основі широкого спектру наукових знань. Це дає змогу розробляти моделі з розширеними можливостями в таких завданнях, як розуміння природної мови, розпізнавання об'єктів і вилучення зв'язків. По-друге, відкриті бази наукових знань сприяють співпраці та обміну знаннями між дослідниками, що сприяє науковим відкриттям та інноваціям. Вони також сприяють міждисциплінарним дослідженням, оскільки містять дані з різних сфер, що дозволяє навчати моделі ШІ в різних наукових галузях. По-третє, відкриті бази наукових знань сприяють прозорості та відтворюваності досліджень у галузі ШІ, оскільки вони надають доступ до даних, які використовуються в навчанні, що дозволяє іншим

дослідникам перевіряти та відтворювати результати. Нарешті, відкриті бази наукових знань сприяють інклюзивності та доступності, оскільки вони забезпечують вільний і відкритий доступ до наукової інформації, знижуючи бар'єри для дослідників і практиків з різним досвідом [1].

Ще однією перевагою використання відкритих наукових баз знань є потенціал для міжгалузевих і міждисциплінарних досліджень. Ці бази знань охоплюють різні галузі науки, такі як біологія, хімія, фізика, медицина та інші. Це дозволяє дослідникам і розробникам навчати ШІ-моделі, які можуть подолати розрив між різними науковими дисциплінами, що призводить до відкриття нових знань і розробки інноваційних рішень для складних проблем. Наприклад, AI-моделі, навчені на поєднанні біологічних і хімічних даних, можуть допомогти у створенні ліків, передбачаючи потенційну взаємодію між ліками і білками або визначаючи нові мішені для ліків.

Незважаючи на свої переваги, відкриті бази наукових знань також створюють певні проблеми для навчання ШІ. Однією з головних проблем є якість і неоднорідність даних. Відкриті бази наукових знань часто містять дані з різних джерел з різним рівнем точності, узгодженості та повноти. Це може вносити шум і упередженість у навчальні дані, що може вплинути на продуктивність і надійність моделей штучного інтелекту. Тому для забезпечення якості навчальних даних необхідні методи попередньої обробки та очищення даних. Ще однією проблемою є масштабованість і складність даних. Відкриті бази наукових знань можуть містити мільярди записів, що робить їх обробку та аналіз обчислювально дорогими. Для роботи з такими великими масивами даних потрібні ефективні методи обробки та індексування даних. Крім того, міждисциплінарний характер відкритих баз наукових знань може створювати проблеми при розробці моделей ШІ, які можуть ефективно навчатися на основі даних з різних наукових галузей. Забезпечення належного представлення даних і методів інженерії особливостей має важливе значення для відображення багатовимірності наукових знань [3].

Однак відкриті бази наукових знань також мають обмеження, які необхідно враховувати. Одним із суттєвих обмежень є потенційна можливість упередженості даних. Наукова література, як і будь-яка інша форма інформації, може зазнавати впливу різних упереджень, таких як упередженість публікації, упередженість цитування та мовні упередження. Це може вплинути на якість і репрезентативність даних, які використовуються для навчання моделей ШІ, що потенційно може призвести до упереджених або неповних результатів. Тому дуже важливо ретельно враховувати обмеження та упередженість даних з відкритих баз наукових знань і вживати відповідних заходів для їх усунення в процесі навчання моделі [3].

Крім того, відкриті бази наукових знань можуть також стикатися з проблемами, пов'язаними з правами інтелектуальної власності та політикою обміну даними. Деякі джерела даних у цих базах знань можуть мати обмеження на використання даних або вимагати дозволів для певних застосувань. Це може обмежити доступність і зручність використання цих баз знань для навчання ШІ. Крім того, можуть виникнути проблеми, пов'язані з конфіденційністю та безпекою даних, оскільки ці бази знань можуть містити конфіденційну інформацію, наприклад, дані пацієнтів або результати власних досліджень. Необхідно вжити відповідних заходів щодо обробки та захисту даних, щоб забезпечити дотримання правил захисту даних та етичних міркувань [4].

Іншим обмеженням є якість і точність даних у відкритих базах наукових знань. Не всі статті або доповіді в цих базах знань проходять такий самий рівень перевірки, як у рецензованих журналах, і в даних можуть бути невідповідності, помилки або застаріла інформація. Тому важливо перевіряти дані та забезпечувати їхню точність, перш ніж використовувати їх для навчання моделей ШІ.

Висновки

Таким чином, відкриті бази наукових знань необхідні для навчання успішних моделей ШІ. Існує кілька популярних варіантів, кожен з яких має свої переваги та обмеження. Дослідники і практики повинні ретельно враховувати джерело і якість даних при виборі бази знань для своїх потреб у навчанні ШІ.

Незважаючи на ці виклики, відкриті бази наукових знань мають величезний потенціал для розвитку досліджень і застосувань ШІ. У майбутньому можна очікувати подальшого поліпшення якості та доступності цих баз знань, а також розвитку технологій ШІ, які зможуть ефективно використовувати ці бази знань для навчання більш точних і надійних моделей ШІ.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Singh, R., Singh, A., & Kumaraguru, P. OpenKI: an open knowledge base interface for

for improving contextual understanding in artificial intelligence applications. In Proc. of the IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2020, p. 359-368.

2. Перелік науково-технічних баз даних та довідкових ресурсів. Наукова бібліотека ХНУМГ ім. О. М. Бекетова. URL: https://library.kname.edu.ua/e-resursi/elektronni-resursi/perelik-naukovo-tekhnichnykh-baz-danykh-ta-dovidkovykh-resursiv.

3. Mayer, M. A., Skirzynski, M., & Terzic, K. Improving knowledge bases for AI applications: A review of methods and challenges. IEEE Access, 8, 2020.

4. Sun, Y., Han, J., & Yan, X. Intelligent analysis of heterogeneous information networks: Principles and methodologies. Generalized lectures on data mining and knowledge discovery, 2012, p. 1-159.

Хрустовський Анатолій Анатолійович – студент групи КІВТ-21б, факультет інформаційних електронних систем, Вінницький національний технічний університет, Вінниця, e-mail: tolik6566tolik@gmail.com .

Кожем'яко Андрій Вікторович – кандидат технічних наук, доцент кафедри обчислювальної техніки Вінницького національного технічного університету, Вінниця, e-mail: kvantron@vntu.edu.ua.

Khrustovskyi Anatolii – student of the group KIVT-21b, Faculty of Information Electronic Systems, Vinnytsia National Technical University, Vinnytsia, e-mail: tolik6566tolik@gmail.com.

Kozhemiako Andrii – PhD in Engineering, Associate Professor of the Department of Computer Science, Vinnytsia National Technical University, Vinnytsia, e-mail: kvantron@vntu.edu.ua.