УДК 004.8

ОКРЕМІ СПОСОБИ ВИРІШЕННЯ ПРОБЛЕМИ НЕДОСТАТНОСТІ УНІКАЛЬНИХ ДАНИХ ДЛЯ НАВЧАННЯ НЕЙРОННИХ МЕРЕЖ У ЮРИСПРУДЕНЦІЇ

Олексій Шамов

SPECIFIC WAYS TO SOLVE THE PROBLEM OF INSUFFICIENT UNIQUE DATA FOR TRAINING NEURAL NETWORKS IN JURISPRUDENCE

Oleksii Shamov

520 Howard Ct, Clearwater, FL, USA, 33756

Abstract. Neural networks and machine learning are increasingly being integrated into various legal domains, including document analysis, legal research, and predicting court case outcomes. This is driven by the desire to automate routine tasks and improve the efficiency of legal practice. However, a fundamental obstacle to training reliable and accurate models in jurisprudence is the problem of insufficient unique and labeled data. The legal field, with its complexity and need for expert labeling, inherently suffers from data scarcity more than some other fields. Using incorrect or insufficient data can lead to legal risks, biases, and decreased model quality, highlighting the importance of data uniqueness and high quality. This article will discuss several specific methodologies aimed at solving the problem of insufficient unique data for training neural networks in jurisprudence: data augmentation, synthetic data generation, transfer learning, and few-shot learning.

Стратегії аугментації даних для юридичних текстів. Аугментація даних є методом збільшення різноманітності та розміру навчальних даних шляхом застосування трансформацій до існуючих даних без зміни їх основного значення чи мітки. Це пропонує економічно ефективний спосіб розширення наборів даних та покращення узагальнюючої здатності моделей. Існують різні поширені методи, які можна застосовувати до тексту, такі як заміна синонімів, зворотний переклад, перемішування слів, випадкове вставлення та видалення. Однак, ефективність цих базових методів може бути обмежена через спеціалізований характер юридичної мови.

Унікальні виклики юридичної мови, включаючи її складний словниковий запас, семантику, морфологію та синтаксис, вимагають спеціалізованих методів аугментації даних, таких як DALE (Data Augmentation for Low-Resource Legal NLP). Стандартні методи аугментації можуть бути недостатніми для юридичного тексту, що зумовлює необхідність використання специфічних для домену підходів. DALE, наприклад, використовує стратегію селективного маскування, розроблену спеціально для юридичних документів. Ключовим завданням при аугментації юридичного тексту є збереження семантичної цілісності та юридичної релевантності аугментованих даних. Трансформації не повинні змінювати оригінальне значення або вносити оманливу інформацію.

Існують також етичні аспекти аугментації даних у юридичній сфері, включаючи потенційні упередження та необхідність валідації. Хоча аугментація може допомогти зменшити упередження, важливо переконатися, що вона не вводить або не посилює існуючі упередження в юридичних даних. Ефективність аугментації даних у праві залежить від складності використовуваних методів. Прості методи можуть не відображати нюанси юридичної мови, тоді як спеціалізовані методи, такі як DALE, є перспективними, але вимагають значних ресурсів. Крім того, першорядне значення мають етичні міркування щодо збереження цілісності та релевантності юридичної інформації під час аугментації. Розробка

більш просунутих, юридично обґрунтованих методів аугментації даних у поєднанні з надійними механізмами валідації є вирішальною для успішного застосування нейронних мереж у юриспруденції в умовах нестачі даних.

Генерація синтетичних юридичних даних. Генеративні моделі штучного інтелекту, включаючи генеративно-змагальні мережі (GAN), варіаційні автокодувальники (VAE) та великі мовні моделі (LLM), такі як GPT, використовуються для створення штучних юридичних даних, які імітують статистичні закономірності та властивості реальних юридичних даних. Крім генеративного ШІ, існують й інші методи генерації синтетичних даних, такі як рушії правил, клонування сутностей та маскування даних. Вибір методу залежить від конкретного випадку використання та доступних ресурсів.

Якість та корисність синтетичних юридичних даних для навчання моделей ШІ є важливим аспектом. Синтетичні дані мають потенціал для підвищення конфіденційності, покращення справедливості та розширення реальних наборів даних. Однак, синтетичні юридичні дані можуть містити упередження, тому ретельне тестування та валідація є важливими для забезпечення точності та справедливості. Існують юридичні та етичні міркування щодо генерації та використання синтетичних юридичних даних, особливо щодо правил конфіденційності даних, таких як GDPR та CCPA. Хоча синтетичні дані можуть підвищити конфіденційність, їх використання все одно має відповідати законам про конфіденційність даних, і тривають дискусії щодо їх юридичної класифікації.

Доступні різноманітні інструменти для генерації синтетичних даних, кожен з яких має різні можливості та призначений для різних користувачів. Вибір методу генерації синтетичних даних у юридичному ШІ значною мірою залежить від конкретного застосування, чутливості вихідних даних та бажаного рівня точності. Хоча генеративний ШІ пропонує високу реалістичність, системи на основі правил можуть бути більш придатними для конкретних, чітко визначених юридичних сценаріїв. Юридичний статус синтетичних даних як "персональних даних" відповідно до таких правил, як GDPR, залишається важливим питанням. Майбутні дослідження повинні бути зосереджені на розробці надійних метрик для оцінки якості та юридичної обґрунтованості синтетичних юридичних даних, а також на встановленні чітких рекомендацій щодо їх етичного та сумісного використання для навчання моделей ШІ для юриспруденції.

Використання трансферного навчання в юридичній обробці природної мови. Трансферне навчання є методом машинного навчання для адаптації попередньо навчених моделей до нових, пов'язаних завдань, особливо корисний при роботі з обмеженою кількістю розмічених даних. Цей підхід може значно зменшити потребу у великих юридичних наборах даних, використовуючи знання, отримані з попередньо навчених моделей. Загальноцільові попередньо навчені мовні моделі (наприклад, BERT, GPT) можуть бути донавчені на менших, специфічних для завдання юридичних наборах даних. Також існують попередньо навчені мовні моделі, спеціально навчені на великих юридичних корпусах, такі як Legal-BERT.

Ефективне донавчання включає вибір відповідних шарів для донавчання та налаштування гіперпараметрів. Застосування трансферного навчання в юридичній сфері має свої виклики та потенційні обмеження, такі як зміщення домену, потреба навіть у невеликій кількості розмічених юридичних даних та потенційні упередження в попередньо навчених моделях. Однак, зростаюча доступність юридичних моделей та бібліотек НЛП з відкритим кодом полегшує використання трансферного навчання. Вибір між загальноцільовими та юридично-специфічними попередньо навченими моделями, а також стратегія донавчання, повинні ретельно розглядатися залежно від конкретного юридичного завдання та характеристик доступних даних. Потенціал упереджень у попередньо навчених моделях також потребує ретельного дослідження. Майбутні дослідження повинні бути зосереджені на розробці більш ефективних методів адаптації попередньо навчених моделей до юридичної сфери, включаючи методи пом'якшення упереджень та усунення специфічних для домену

нюансів, можливо, за допомогою багатовузлового навчання або антагоністичного навчання, адаптованих для юридичного НЛП.

Застосування методів навчання з малої кількості прикладів. Навчання з малої кількості прикладів є парадигмою машинного навчання, яка дозволяє моделям навчатися на дуже обмеженій кількості розмічених прикладів. Цей підхід є перспективним рішенням, коли розмічених юридичних даних вкрай мало. Навчання з малої кількості прикладів показало свою ефективність та застосовність для конкретних юридичних завдань, таких як класифікація юридичних документів, розпізнавання іменованих сутностей та відповіді на запитання.

У сценаріях навчання з малої кількості прикладів важливими є інженерія промптів та вибір даних для максимізації продуктивності. Однак, навчання з малої кількості прикладів має свої обмеження, включаючи чутливість до якості прикладів, потенційні упередження та труднощі з обробкою складних завдань міркування. Важливо розуміти відмінності між навчанням з малої кількості прикладів, навчанням з нульової кількості прикладів та традиційними підходами навчання з учителем. Існують бібліотеки та фреймворки для навчання з малої кількості прикладів, такі як TaiChi, що робить цей підхід більш доступним для дослідників та практиків. Навчання з малої кількості прикладів є привабливим підходом для юридичних завдань ШІ, де розмічених даних вкрай мало. Однак, його продуктивність значною мірою залежить від якості та репрезентативності наданих небагатьох прикладів та ретельного розроблення промптів. Для складних юридичних міркувань може знадобитися донавчання більших моделей. Майбутні дослідження повинні бути зосереджені на розробці більш надійних та ефективних методів навчання з малої кількості прикладів, спеціально адаптованих до складностей юридичного міркування та мови, можливо, шляхом включення специфічних для домену знань у стратегії промптів або розробки мета-навчальних підходів, адаптованих для юридичної сфери.

Висновки та майбутні напрямки. У цій статті було розглянуто декілька методів вирішення проблеми недостатності унікальних даних для навчання нейронних мереж у юриспруденції: аугментація даних (включаючи спеціалізовані методи), генерація синтетичних даних (з різними методами та інструментами), трансферне навчання (з використанням загальних та юридично-специфічних попередньо навчених моделей) та навчання з малої кількості прикладів (та його залежність від промптів). Вибір та впровадження цих рішень повинні базуватися на конкретному юридичному завданні, кількості та характері доступних даних, а також бажаній продуктивності.

Першорядне значення мають етичні міркування, включаючи потенційні упередження в даних та моделях, необхідність прозорості та пояснюваності, а також дотримання правил конфіденційності даних та юридичної відповідності при розробці та розгортанні юридичних рішень на основі ШІ. Майбутні напрямки досліджень можуть включати вивчення гібридних підходів, що поєднують кілька рішень для подолання нестачі даних, розробку кращих метрик оцінки для юридичних моделей ШІ, навчених на обмежених даних, та дослідження використання активного навчання для стратегічного вибору даних для розмітки.

Бібліографія

1. The Rise of AI in Legal Practice: Opportunities, Challenges, & Ethical Considerations, By Joely Williamson / March 21, 2025, https://ctlj.colorado.edu/?p=1297
2. Overview of Synthetic Data Generation Methods, By Paul Pokotylo, Oct 21, 2024, https://keymakr.com/blog/overview-of-synthetic-data-generation-methods/
3. Zero-Shot and Few-Shot Learning with LLMs, by Michał Oleszak, 25th September, 2024, https://neptune.ai/blog/zero-shot-and-few-shot-learning-with-llms
4. Bringing transparency to the data used to train artificial intelligence, by Beth Stackpole, Mar 3, 2025, https://mitsloan.mit.edu/ideas-made-to-matter/bringing-transparency-to-data-used-to-train-artificial-intelligence