ЦИФРОВА ЕКОНОМІКА ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

Digital Economy and Information Technologies

УДК 004.8

Сергій Полуектов,

провідний інженер машинного навчання,
Certivity GmbH (https://certivity.io)
Мюнхен, Німеччина

ВИКОРИСТАННЯ МОДЕЛІ BERT ДЛЯ АВТОМАТИЗАЦІЇ ПОШУКУ В ІНЖЕНЕРНІЙ ДІЯЛЬНОСТІ

Анотація — У сучасному цифровому світі, де щодня створюється та поширюється величезна кількість даних, для інженерів стає дедалі складніше знайти відповідну інформацію, намагаючись вирішити свої технічні проблеми або вдосконалити свою технологію. Поточні програми пошуку та керування знаннями значною мірою покладаються на автоматизацію на основі NLP. Останні досягнення в навчанні трансферу NLP призвели до створення потужних моделей, таких як BERT, які добре виконують завдання текстового пошуку у загальній сфері. У цій роботі ми оцінюємо різні підходи до адаптації BERT до галузі інженерії. Ми порівнюємо кілька предметно-спеціальних моделей щодо їхньої здатності ідентифікувати нові технології та призначати теми інженерним статтям. Наші експерименти показують, що доменно-адаптаційна стратегія подальшого попереднього навчання на предметно-специфічних даних без розширення словника забезпечує найкращу продуктивність у вирішенні цих завдань. Після оцінки ми описуємо проблеми та обмеження нашого підходу та надаємо напрямки для майбутніх досліджень.

Ключові слова— *NLP, BERT, Машинне навчання, Інтелектуальні технології*

I. ВСТУП

Сучасні інженери сьогодні стикаються з величезною кількістю інформації, намагаючись вирішити свої технічні проблеми. Типовий працівник знань витрачає близько 2,5 годин на день на пошук інформації, що призводить до приблизно 18 мільярдів євро щорічних витрат німецьких компаній [1].

Одним із варіантів вирішення проблеми пошуку актуальної інформації в постійно зростаючих обсягах даних є спеціалізація пошукових систем. В останні роки можна спостерігати появу численних нішевих пошукових платформ. Такі вертикальні пошукові системи використовуються в конкретній цільовій області і досягають вищої релевантності результатів пошуку, ніж пошукові системи загального призначення [2]. Однак створення таких пошукових програм є трудомістким процесом, що потребує значних і постійних зусиль людини [3].

Графи знань — це ще один підхід, який можна використовувати разом із пошуковою системою для надання високорелевантних результатів, залежно від контексту. Визначені як «велика мережа сутностей, їхніх семантичних типів, властивостей і зв'язків між ними» [4], графи знань надають інформацію семантично структурованим способом, що дозволяє відкривати та збагачувати результати пошуку в вузьких областях знань [5].

Сучасні методи машинного навчання можуть бути використані для автоматизації багатьох аспектів створення та підтримки спеціальних пошукових систем і для виконання автоматичного вилучення знань як частини процесу побудови графів знань [6].

BERT — це сучасна модель глибокого навчання, яка добре виконує широкий спектр завдань NLP (Natural Language Processing) [7]. Вона використовує парадигму

трансферного навчання в NLP і дозволяє досягти високої продуктивності класифікації тексту та завдань вилучення об’єктів з обмеженими позначеними даними [8].

В цій роботі наведені результати вивчення методів пристосування моделі BERT для розв’язування задачі пошуку в інженерній практиці. Основною метою дослідження було визначення ефективності застосування моделі BERT, яка попередньо навчена на текстах з інженерної сфери, при класифікації вибраного тексту.

II. МОДЕЛІ ТА МЕТОДИ ДОСЛІДЖЕННЯ

Наприкінці 2000 років почалось активне використання нейронних мереж для розв’язування завдань NLP. Всі останні та найуспішніші моделі NLP, такі як BERT [8] або Generative Pretrained Transformer 2 (GPT-2) [9] and 3 (GPT-3) [10] також базуються на методах нейронних мереж.

BERT, що розшифровується як Bidirectional Encoder Representation from Transformer, є глибокою контекстною моделлю представлення мови. Використовуючи самоконтрольовану процедуру попереднього навчання на величезному наборі даних без міток, вона створює контекстно-залежне представлення слів. Отже, попередньо навчену модель можна налаштувати для виконання широкого діапазону подальших завдань, додавши додатковий вихідний рівень для конкретного завдання та використовуючи набагато менший набір даних для навчання під конкретне завдання.

BERT використовує Transformer, механізм уваги, який вивчає контекстні зв’язки між словами (або підсловами) у тексті. Transformer містить два окремих механізми — кодер, який зчитує введений текст, і декодер, який створює прогноз для завдання.

На відміну від спрямованих моделей, які зчитують введений текст послідовно (зліва направо або справа наліво), кодувальник Transformer зчитує всю послідовність слів одночасно.

Вхідні дані для BERT — це послідовність токенів, які спочатку вбудовуються у вектори, а потім обробляються в нейронній мережі. Вихід - це послідовність векторів розміру H, у якій кожен вектор відповідає вхідному маркеру з таким же індексом.

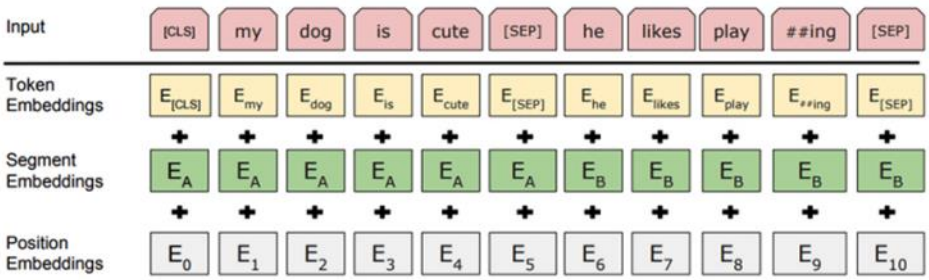

Рисунок 1. Вхідне представлення BERT. Передруковано з [8].

Перед подачею послідовностей слів у BERT 15% слів у кожній послідовності замінюються маркером [MASK]. Потім модель намагається передбачити початкове значення замаскованих слів на основі контексту, наданого іншими незамаскованими словами в послідовності. Цей алгоритм називається моделюванням мови в масках (MLM) (рис.2).

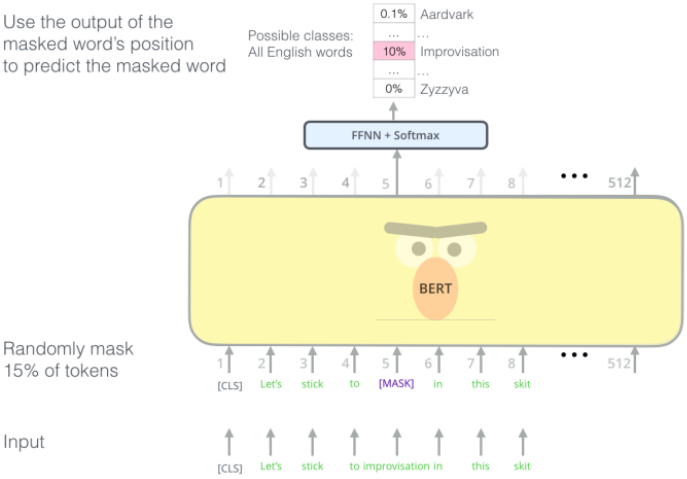

Рис. 2. Ілюстрація механізму (MLM). Передруковано з [8].

Крім того, BERT використовує механізм визначення наступного речення (next sentence prediction (NSP)). BERT модель отримує пари речень як вхідні дані та вчиться передбачати, чи є друге речення в парі наступним реченням у вихідному документі. Під час навчання 50% вхідних даних є парою, у якій друге речення є наступним реченням у вихідному документі, а в інших 50% випадкове речення з корпусу вибирається як друге речення.

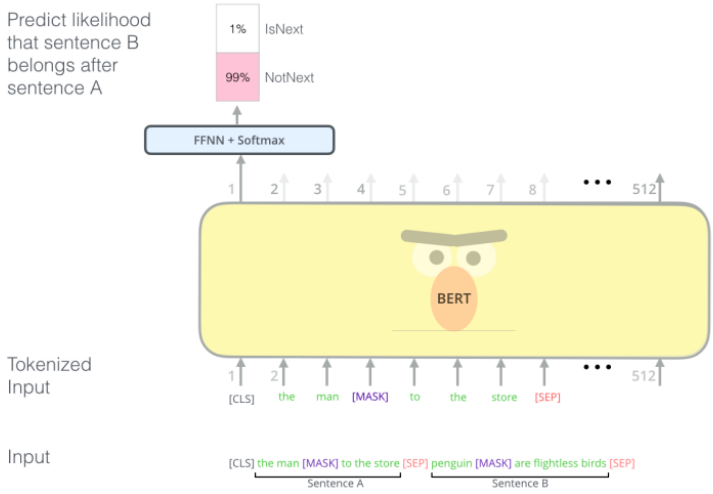

Рис. 3. Ілюстрація механізму (NSP). Передруковано з [8].

BERT навчається одночасно двом самоконтрольованим завданням: моделюванню мови в масках (MLM) і передбаченню наступного речення (NSP). Загальна втрата при навчанні є сумою середньої втрати MLM і середньої втрати NSP.

Шлях до ефективного трансферного навчання в NLP був закладений Говардом і Рудером [11], які представили універсальну мовну модель тонкого налаштування для класифікації тексту (ULMFiT). BERT дотримується подібного підходу: він попередньо навчений на великому наборі даних загального домену, що дозволяє йому охоплювати загальні властивості мови у своїх вагових показниках.Після цього модель може бути налаштована для виконання конкретного подальшого завдання. Парадигма

трансферного навчання пропонує найбільше переваг у ситуаціях, коли розмір набору даних для конкретного завдання невеликий [11].

Тонка настройка для конкретного завдання зазвичай передбачає створення тонкої прямої нейронної мережі (голова) поверх BERT і тренування його параметрів разом із вагами BERT на наскрізному наборі даних для конкретних завдань для кількох епох.

В цьому дослідженні розглядались різні можливості адаптації BERT до інженерної області. Досліджувались наступні стратегії:

• подальше навчання без розширення словникового запасу;
• подальше навчання з розширенням словникового запасу;
• попереднє навчання з нуля.

Таким чином, досліджувались чотири різні моделі BERT для інженерної області:

• BERT-base-nove: модель BERT, ініціалізована зі стандартної базової моделі BASE, яка далі навчалась на корпусі немаркованих інженерних статей без словника розширення;

• BERT-base-ext1000: модель BERT, ініціалізована зі стандартної базової моделі BASE, яка далі попереднє навчалась на наборі немаркованих інженерних статей після його розширення словниковим запасом зі 1000 найпоширеніших слів з цього набору.

• BERT-base-ext5000: модель BERT, ініціалізована зі стандартної базової моделі BASE, яка далі попереднє навчалась на наборі немаркованих інженерних статей після його розширення словниковим запасом зі 5000 найпоширеніших слів з цього набору.

• BERT-base-from-scratch: модель BERT, ініціалізована випадковим чином, а потім попередньо навчена на наборі інженерних статей без маркування.

В якості додаткових текстових даних, на яких началась модель були використані текстові дані з 20 інженерних сайтів (див. рис. 4)

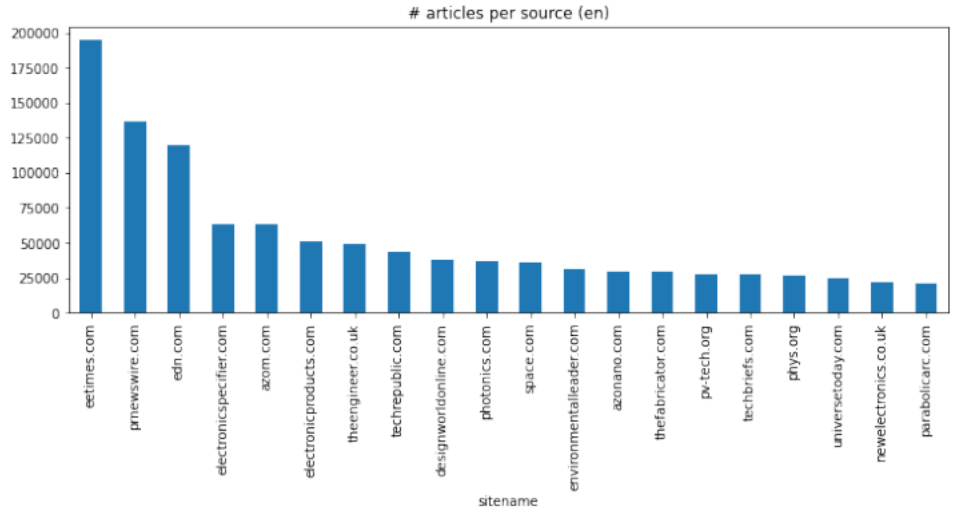

Рис. 4. 20 найпопулярніших джерел інженерних статей

Перше визначене завдання — класифікувати статті на ті, що описують нову технологію і ті, які цього не роблять. Це виконання завдання бінарної класифікації.

Друге визначене завдання — віднести статті до однієї із заздалегідь визначених тем. Серед цих тем: приводи, адитивне виробництво, штучний інтелект, доповнена/віртуальна реальність, автономні транспортні засоби, комунікаційні

технології, електронні компоненти, платформи IoT, IT-безпека, технології виробничих процесів, робототехніка, сенсори, програмне забезпечення для моделювання, відстеження та ідентифікація.

Навчальні дані розподідені за цими мітками згідно з рис. 5.

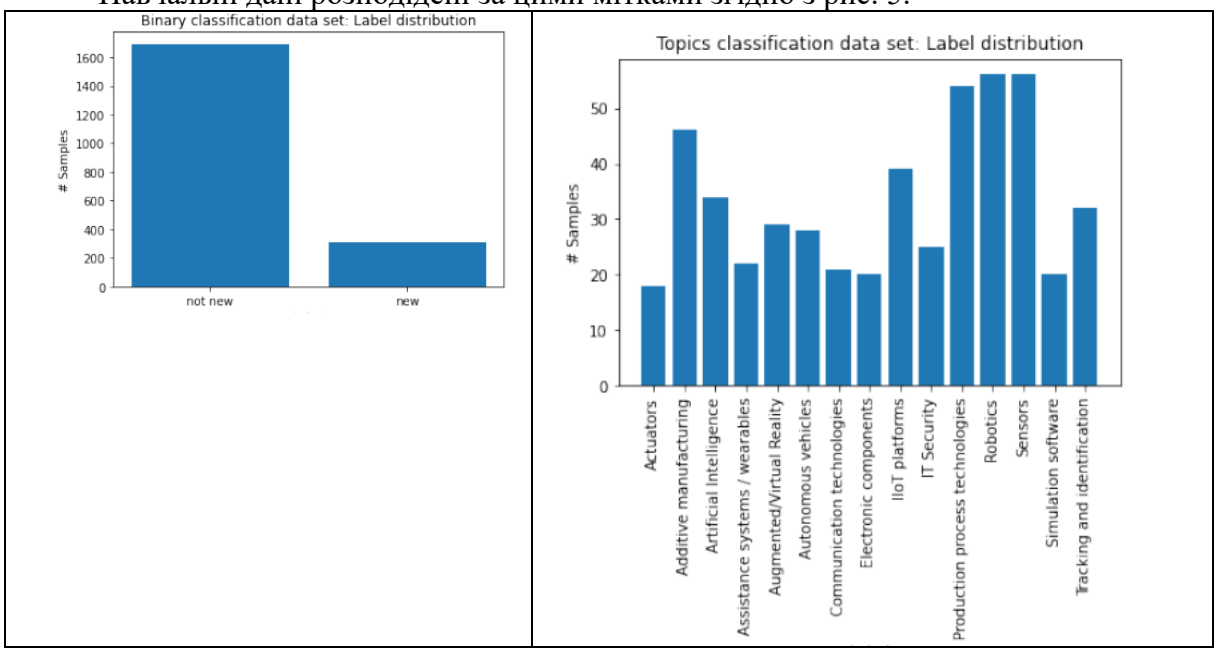

Рис. 5. Розподіл міток в начальному наборі даних

В ході дослідження були використані наступні інструменти.

Google Cloud Plattform (GCP) – це набір хмарних обчислювальних служб, які можна використовувати для зберігання даних, обчислення, доставки вмісту тощо.

Бібліотека Huggingface Transformers [12], яка містить ретельно розроблені реалізації всіх найсучасніших моделей NLP, у тому числі BERT, RoBERTa, GPT-2 тощо, для двох найпопулярніших фреймворків глибокого навчання: PyTorch і TensorFlow. Крім того, існує потужний API, який дозволяє навчати та оцінювати моделі на різному обладнанні. Нарешті, вона також надає API для створення наборів даних із готовим кешуванням і швидкою реалізацією всіх токенізаторів, необхідних для моделей.

Weights and Biases 2 - це інструмент для моніторингу експериментів машинного навчання. Він добре інтегрований з бібліотекою Huggingface Transformers і дозволяє автоматично збирати налаштування та показники під час навчання.

Оцінювання продуктивності моделей виконувалось за допомогою традиційних показників якості машинного навчання.

Precision – значення у відсотках, яке вказує, скільки з отриманих результатів є правильними.

Recall – є відсотковим значенням, яке вказує, скільки було знайдено правильних результатів зі всіх, що могли бути знайдені.

F-показник (F1 міра) – це середнє гармонічне значення Precision та Recall показників, яке відповідає такій формулі: 2*[(Precision * Recall) / (Precision + Recall))].

Accuracy – це система оцінки, яка використовується як альтернатива F-оцінці та обчислюється як: [(справжні позитивні + справжні негативні) / (справжні позитивні + справжні негативні + хибні позитивні + хибні негативні результати)].

III. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для кожного завдання та набору даних ми показуємо, як навчені моделі порівнюються одна з одною та з базовою моделлю. Крім того, ми надаємо огляд продуктивності моделей в контрольних точках.

Порівняння продуктивності моделей для визначення новітності технологій (бінарна класифікація) наведено на рис. 6. BERT-base ext5000 досягає найвищого показника у 62,36%, покращуючи базову лінію на 2%. Подальші попередньо підготовлені моделі без розширення словника (BERT-base-nove) і з 1000 додатковими предметно-специфічними словами (BERT-base-ext1000) також працюють краще, ніж базова модель.

Model	Precision	Recall	F1
BERT-*base*	51.79	74.69	60.22
BERT-*base-nove*	52.20	76.89	61.68
BERT-*base-ext1000*	53.27	75.91	62.11
BERT-*base-ext5000*	53.74	75.65	**62.36**
BERT-*base-from-scratch*	44.83	80.00	57.40

Рис. 6. Результати навчання моделі для завдання бінарної класифікації

Як показано на малюнку 7, подальше попереднє навчання моделей до більшої кількості ітерацій покращує їх продуктивність у цьому завданні. Однак цей ефект сповільнюється до кінця розглянутої тривалості попереднього тренування.

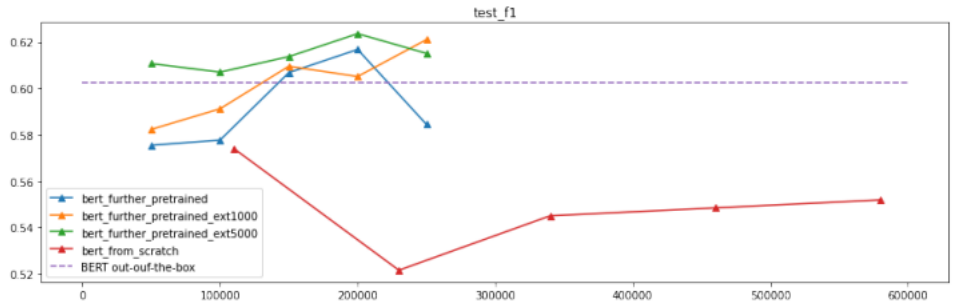

Рис. 7. Залежність продуктивності тренування моделей від кількості кроків тренування

Модель, навчена з нуля (BERT-base-from-scratch), працює гірше, ніж усі інші моделі, і не перевищує базову лінію. Поліпшення його продуктивності завдяки додатковим етапам попереднього навчання вказує на те, що вона потребує додаткового попереднього навчання, щоб конкурувати з іншими моделями.

Відносно низька точність класифікації, досягнута всіма моделями, пояснюється високою складністю цього завдання. Навіть людині не завжди вдається впевнено визначити статті, в яких описується нова технологія. Крім того, класифікація змінюється з часом, оскільки те, що вважалося новим два роки тому, зараз може вважатися старим.

На рис.8 представлені результати всіх розглянутих моделей, досягнутих за завданням тематичної класифікації. Найкраща з усіх розглянутих моделей, подальша попередньо навчена модель без розширення лексики (BERT-base-nove) досягла 89,8% точності, що дає покращення в порівнянні з базовим рівнем на 4,4%.

Model	Accuracy
BERT-*base*	85.40
BERT-*base-nove*	**89.80**
BERT-*base-ext1000*	89.00
BERT-*base-ext5000*	86.20
BERT-*base-from-scratch*	89.20

Рис. 8. Результати класифікації статей за тематикою

Віднесення статей до тем зводиться до визначення певних ключових слів, характерних для певної теми. На відміну від двійкової класифікації, це завдання не вимагає розуміння більш складних лінгвістичних понять. Це може бути можливим поясненням того факту, що модель, навчена з нуля (BERT-base-from-scratch), досягає хороших результатів продуктивності і в цьому завданні.

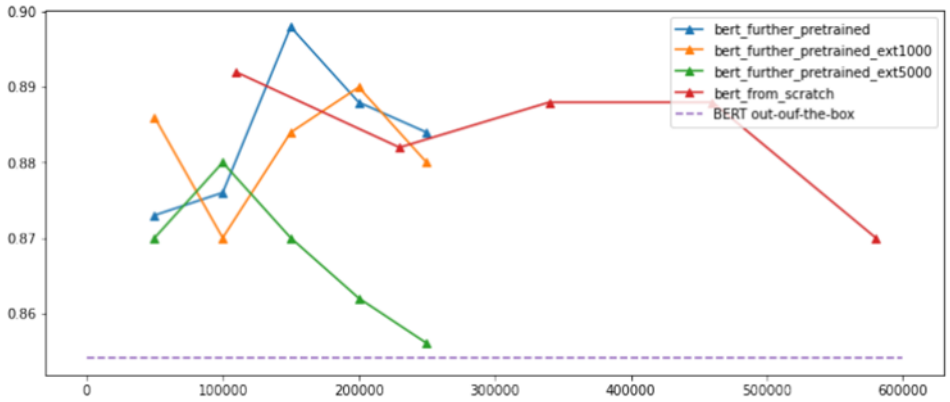

Рис. 9. Результати тематичного класифікаційного завдання для всіх контрольних точок

Як показано на малюнку 9, подальше попереднє навчання навіть для невеликої кількості кроків дає значний приріст порівняно з базовою лінією.

Цікаво, що точність BERT-base ext5000 знижується при подальшому попередньому навчанні. Це залишилося дослідити в майбутньому.

IV. ВИСНОВКИ

Модель BERT додатково попередньо навчена на корпусі немаркованих інженерних статей без розширення словникового запасу (BERT-base-nove) показала найкращі результати в задачах класифікації за темою та бінарної класифікації. Враховуючи її продуктивність і той факт, що це найпростіший і найменш ресурсомісткий підхід до адаптації домену можна зробити висновок, що подальше попереднє навчання без розширення словникового запасу є найбільш доцільним підходом до адаптації BERT до рішення завдань текстового пошуку в інженерної області за заданих умов.

V. ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У ході цього дослідження стали очевидними кілька недоліків запропонованого підходу.

Як було показано в [13], стандартна модель BERT дуже недостатньо навчена, і потрібні сотні гігабайт додаткових даних і десятки годин обчислювального часу на найшвидших процесорах, щоб вивести її на повну потужність мовного моделювання. Під час попереднього навчання це може проявлятися у постійному зниженні втрат як на наборах для навчання, так і на наборах перевірки, так що точку переобладнання неможливо помітити.

Однак продуктивність моделі визначається не лише можливостями представлення мови, але й складністю самого завдання. Таким чином, ми стверджуємо, що точне налаштування та оцінку всіх доступних завдань слід виконувати під час попереднього навчання або подальшого попереднього навчання, щоб виявити точку та можливість ранньої зупинки. Це дозволить заощадити час і, відповідно, дорогі обчислювальні ресурси.

Крім того, набір технічних статей без міток, який використовується моделлю, попередньо навченою з нуля (BERT base-from-scratch), містить близько 2 мільйонів статей, що еквівалентно 6,7 ГБ. Стандартна модель BERT була попередньо навчена на 16 ГБ даних, а автори RoBERTa використовували набір даних 160 ГБ. Тому слід докласти певних зусиль для розширення корпусу немаркованих інженерних статей. Це можна зробити, зібравши нові статті зі списку визначених джерел.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] R. STARK, H. Bedenbender, P. Müller, F. Pasch, R. Drewinski, and H. Hayka. “Kollaborative produktentwicklung und digitale Werkzeuge”. In: Defizite heute-Potenziale morgen (2013).

[2] C. L. Giles, Y. Petinot, P. B. Teregowda, H. Han, S. Lawrence, A. Rangaswamy, and N. Pal. “eBizSearch: A niche search engine for e-business”. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. 2003, pp. 413–414.

[3] A. McCallumzy, K. Nigamy, J. Renniey, and K. Seymorey. “Building domain-specific search engines

[4] M. Kroetsch and G. Weikum. “Special issue on knowledge graphs”. In: Journal of Web Semantics 37.38 (2016), pp. 53–54

[5] X. Zou. “A survey on application of knowledge graph”. In: Journal of Physics: Conference Series. Vol. 1487. 1. IOP Publishing. 2020, p. 012016

[6] Z. Zhao, S.-K. Han, and I.-M. So. “Architecture of knowledge graph construction techniques”. In: International Journal of Pure and Applied Mathematics 118.19 (2018), pp. 1869–1883.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: arXiv preprint arXiv:1810.04805 (2018).

[8] J. Alammar. The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning). url: http://jalammar.github.io/illustrated-bert/ (visited on 03/11/2021).

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. “Language models are unsupervised multitask learners”. In: OpenAI blog 1.8 (2019), p. 9.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. “Language models are few-shot learners”. In: arXiv preprint arXiv:2005.14165 (2020).

[11] J. Howard and S. Ruder. “Universal language model fine-tuning for text classification”. In: arXiv preprint arXiv:1801.06146 (2018).

[12] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al. “Transformers: State-of-the-art natural language processing”. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020, pp. 38–45.

[13] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. “GLUE: A multitask benchmark and analysis platform for natural language understanding”. In: arXiv preprint arXiv:1804.07461 (2018).

отримано 01.11.2022 р.

USING THE BERT MODEL FOR SEARCH AUTOMATION IN ENGINEERING ACTIVITIES

Sergii Poluektov,

Lead Machine Learning Engineer,
Certivity GmbH (https://certivity.io)
Munich, Germany

Abstract — In today’s digital world, where the amount of data created and shared daily is overwhelming, it becomes more and more challenging for engineers to find relevant information when trying to solve their technical problems or improve their technology. Current search and knowledge management applications rely heavily on NLP-based automation. The recent advances in NLP transfer learning have resulted in powerful models, such as BERT, which perform well on NLP tasks in the general domain. In this work, we evaluate different approaches to adapting BERT to the domain of engineering. We compare multiple domain-specific models in their ability to identify new technologies and assign topics to engineering articles. Our experiments show that the domain-adaptation strategy of further pre-training on domain-specific data without vocabulary extension leads to the best performance in this tasks solutions. After the evaluation, we describe the challenges and the limitations of our approach and provide directions for future research.

***Keywords*—*NLP, ML, BERT, Cognition, Intelligence*.**

Received 01.01.2022