12. 悲歌_百度百科. URL: https://baike.baidu.com/item/悲歌/10723325 (дата обращения : 01.05.2022 г.).
13. 八阵图_百度百科. URL: https://baike.baidu.com/item/八阵图/3707897 (дата обращения : 01.05.2022 г.)
14. 老妪能解. URL: https://bkrs.info/slovo.php?ch=老妪能解 (дата обращения : 01.05.2022 г.).
15. 对雪_百度百科. URL: https://baike.baidu.com/item/对雪/4333 (дата обращения : 01.05.2022 г.)
16. 九章 • 悲回风原文及翻译赏析_屈原古诗大全. URL: https://www.gushimi.org/gushi/39203.html (дата обращения : 01.05.2022 г.).
17. Цюй Юань. Стихотворения. URL: http://www.lib.ru/POECHIN/UAN/uan1_1.txt_with-big-pictures.html (дата обращения : 01.05.2022 г.).

ДРАЙВЕРЫ РАЗВИТИЯ КРУПНОМАСШТАБНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ В США И КНР

Д. О. Мазаник

ГНУ «Центр системного анализа и стратегических исследований НАН Беларуси», ул. Академическая 1, г. Минск, Беларусь, dmitrymazanik42@gmail.com

В статье рассмотрена специфика разработки крупномасштабных языковых моделей в США и КНР. Высокая стоимость разработки приводит к тому, что в США передовые разработки концентрируются в частном секторе, что может повлечь негативные последствия для отрасли. Академическое сообщество оказывается отрезанным от технологического фронтира в области ИИ. В данном контексте интересен опыт КНР, где сформировалась особая институциональная среда. Анализ китайских научных статей показывает, что партнерство между академическим сообществом и частными технологическими компаниями дает доступ научному сообществу к необходимым ресурсам и приводит к конкурентоспособным разработкам.

Ключевые слова: языковые модели; искусственный интеллект; научные организации; технологические компании; технологический фронтир; КНР; США.

DRIVERS FOR THE DEVELOPMENT OF LARGE-SCALE LANGUAGE MODELS IN THE USA AND THE PRC

D. O. Mazanik

State Scientific Institution "Center for System Analysis and Strategic Research of the National Academy of Sciences of Belarus", st. Akademicheskaya 1, Minsk, Belarus, dmitrymazanik42@gmail.com

The article considers the specifics of the development of large-scale language models in the USA and PRC. The high cost leads to the fact that in the US cutting-edge developments are

concentrated in the private sector, which can lead to negative consequences for the industry. The academic community is cut off from the technological frontier in the field of AI. In this context, the experience of the PRC is useful, where a special institutional environment has been formed. An analysis of Chinese scientific articles shows that partnerships between the academic community and private technology companies give the scientific community access to the necessary resources and lead to competitive developments.

Keywords: language models; artificial intelligence; scientific organizations; technology companies; technological frontier; PRC; USA.

Благодаря эффективному применению нейронных сетей в последние несколько лет мир переживает новую волну оптимизма по поводу искусственного интеллекта (ИИ). Современная ситуация примечательна тем, что сейчас влияние ИИ-решений ощущается особенно сильно. Автоматизация и «интеллектуализация» проникают в повседневную жизнь, что создает как выгоду, так и риски. Частные компании активно инвестируют в разработку ИИ-приложений и внедряют их в производственные процессы. Согласно статистике от Stanford Institute for Human-Centered Artificial Intelligence (HAI), в 2021 г. глобальные частные инвестиции в ИИ составили около $93,5 млрд – в два раза больше, чем в 2020 г. [1]. Наибольший интерес к направлению проявляют в США и КНР, о чем говорят объемы частных инвестиций за 2021 г. – $52,9 млрд и $17,2 млрд соответственно [1].

Язык – главное средство оперирования знаниями и обмена информацией. Неудивительно, что моделирование языка стало одним из ключевых направлений в ИИ. Сделать тексты доступными для цифровой обработки помогает Natural Language Processing («обработка естественного языка», NLP) – прикладная дисциплина на стыке лингвистики, компьютерных наук и машинного обучения. С появлением необходимых мощностей стало возможным эффективное применение нейронных сетей и глубокого обучения в масштабе. Приложение данного подхода в NLP привело к созданию *крупномасштабных языковых моделей* (large language models, ЯМ) – технологии, которая во многом определяет тренды развития современного ИИ. Самые известные образцы – BERT (Google), GPT-3 (OpenAI), Megatron-Turing NLG (Microsoft).

Статистическая языковая модель – это распределение вероятностей по последовательностям токенов [2]. Как и нейросети с глубоким обучением, языковые модели – не новое изобретение. Однако дискуссии о них вышли за пределы специализированных лабораторий после того, как стали известны их возможности при увеличении масштаба и добавлении новых технологических решений.

С 2020 г. в академической среде и СМИ обсуждают GPT-3 (Generative Pre-Training Transformer 3) – третье поколение генеративной

языковой модели от американской компании OpenAI. GPT-3 поразила способностью к порождению текстов: даже без специальной настройки (fine-tuning) модель может писать эссе, прозу и стихи, отвечать на вопросы и переводить на высоком уровне. Вариант модели DALL-E 2 генерирует уникальные изображения по текстовому описанию. Ключевая особенность GPT-3 и аналогичных моделей – не столько архитектура, сколько размер. Модель включает 175 млрд параметров – в 10 раз больше, чем у любого из предшественников [3]. Обучающие данные составили 570 Гб преимущественно англоязычных текстов.

В августе 2021 г. исследователи из Center Research on Foundation Models и HAI подготовили доклад «On the Opportunities and Risks of Foundation Models» («О возможностях и рисках базовых моделей») [4]. В исследовании, которое стало одним из главных событий в ИИ-аналитике 2021 г., приведен системный обзор крупномасштабных ЯМ, включая их технологическую и лингвистическую специфику, а также социально-экономические последствия от внедрения. Исследователи описывают такие модели как «базовые» (foundation), подчеркивая их центральную, но не исчерпывающую роль в создании AGI и специализированных приложений.

Базовые модели можно причислить к «технологиям общего назначения» (general-purpose technology) [4, с. 148]. Такого рода технологии играют роль «вспомогательных», поскольку открывают новые возможности, но не предлагают окончательные решения. Базовые модели и решения на их основе, вероятно, существенно повлияют на рынок труда. По мере распространения и снижения стоимости ИИ-решений работа, связанная с монотонными и рутинными действиями, может быть полностью автоматизирована – например, в копирайтинге и рекламе.

Для разработки крупномасштабных ЯМ нужны три дорогостоящих компонента:

- Мощное вычислительное оборудование.
- Инвестиции в R&D.
- Обучающие данные.

Исследования также показывают, что стоимость обучения зависит от количества параметров в ЯМ [5, с. 1]:

- $2,5–50 тыс. – для моделей со 110 млн параметров.
- $10–200 тыс. – 340 млн параметров;
- $80 тыс. – 1,6 млн. – 1,5 млрд параметров.

Следствием высокой стоимости ЯМ стало то, что в США они сконцентрированы преимущественно в лабораториях частных корпораций. По этому поводу академическое сообщество США

опасается[1], что оно может оказаться отрезанным от технологического фронтира, а специалисты будут переходить в частный сектор из-за более высоких зарплат и доступа к последним разработкам.

Доминирование корпораций может негативно сказаться на развитии ИИ как технологии, поскольку система принятия решений в бизнесе, ориентированная на получение максимальной выгоды в минимальные сроки, отличается от таковой в науке. Данный подход приводит к доминированию решений, спорных с научной точки зрения, но эффективных с точки зрения рынка. Другая проблема состоит в том, что публикации компаний о своих разработках открывают не всю информацию об устройстве моделей. Без полноценного анализа предложить новые фундаментальные решения становится намного сложнее. Открытость информации необходима для полноценного осмысления и оценки технологического решения со стороны академического сообщества.

В контексте разрыва между частными компаниями и академической наукой в США интересен опыт КНР, которая активно включилась в мировую ИИ-гонку. Как показывает анализ научных публикаций о передовых ЯМ, в Китае научное сообщество и бизнес, наоборот, сближаются.

В последние годы Китай отчетливо демонстрирует свои амбиции в технологическом секторе. Политическая элита декларирует ориентацию на «развитие, движимое инновациями» (创新驱动发展). Руководство КНР стремится привести страну к технологической независимости и доминированию. В этой оптике можно рассматривать программы «Made in China 2025» (中国制造 2025) и «China Standards 2035» (中国标准 2035). Частный сектор также активно вовлечен в развитие технологий, ключевых для технологического развития государства, – включая ИИ

ИИ-экосистема КНР держится на «трех китах» – Baidu, Alibaba и Tencent (BAT, ср. с FAANG в США). Благодаря популярности своих решений (мессенджеров, финансовых технологий, поисковых систем и др.) компании собирают огромное количество данных. Это имеет большое значение для развития ИИ в целом и ЯМ в частности, поскольку информация – это “топливо” для подобных систем.

В апреле 2021 г. Huawei и Recurrent AI (循环智能) выпустили крупномасштабную авторегрессивную предобученную модель PanGu-α (盘古 «Паньгу» – по названию первопредка из китайской мифологии). PanGu-α конкурирует с GPT-3 и обходит ее по числу параметров – 200

[1]См. выступление Дж. Кларка на Workshop on Foundation Models, организованном HAI. URL: https://www.youtube.com/watch?v=dG628PEN1fY&t=3160s (дата доступа: 23.02.2022).

млрд против 175 млрд [6]. Модель, обученная 1,1 Тб китайских текстов, хорошо проявила себя при выполнении ряда NLP-задач, включая суммаризацию, ответы на вопросы, поддержание диалога. Другой пример – ERNIE 3.0 от Baidu [7]. Разработка фокусируется на задачах понимания естественного языка (Natural language Understanding) и генерации текстов. Это шаг в сторону от тренировки гигантских «стохастических попугаев», как иронично назвали подобные модели в своей статье Э.М. Бендер и соавторы [8]. Модель ERNIE 3.0, обученная на 4 Тб текстов и графов знаний, достигла SOTA-результатов при выполнении 54 задач и была встроена в поисковый движок Baidu. Таким образом, один из ключевых игроков китайского «big tech» развивает альтернативные пути в отрасли.

В марте 2022 г. главные эксперты в области ИИ из КНР опубликовали доклад «A Roadmap for Big Model» («Дорожная карта большой модели») [9]. В дорожной карте отмечены основные тенденции в индустрии и предложены будущие направления для исследований. Важно обратить внимание не только на то, *что* написано в докладе, но и *кто* это написал.

Объемный доклад (1638 ссылок на источники) составили 100 китайских исследователей из 19 организаций. Среди них авторитетные китайские университеты (Tsinghua University, Renmin University of China, Peking University, Northeastern University, Shanghai Jiao Tong University, BeiHang University), исследовательские институты (Beijing Academy of Artificial Intelligence; Institute of Computing Technology, Institute of Software, Institute of Automation в составе Китайской академии наук; Harbin Institute of Technology) и лаборатории частных компаний (Wechat, Tencent Inc.; Huawei TCS Lab; JD AI Research; Microsoft Research Asia; ByteDance AI Lab). Именно эти организации стоят в авангарде развития ИИ в КНР. При этом стоит обратить внимание, что большинство – государственные научные организации. В то же время подчеркивается доминирование государственных структур над частными и иностранными лабораториями: после имен исследователей стоит сноска «produced by Beijing Academy of Artificial Intelligence» (BAAI). Самые передовые и известные китайские ЯМ созданы в рамках BAAI.

BAAI (кит. 北京智源人工智能研究院, «Пекинский исследовательский институт ИИ “Чжиюань”») основана в ноябре 2018 г. при поддержке Министерства науки и технологий (科技部), а также горкома и муниципалитета Пекина (北京市委市政府) [10]. Одним из ключевых достижений организации стала модель WuDao 2.0 (от кит. 悟道 «познать истину»/ «просвещение») – китайский конкурент GPT-3, выпущенный в июне 2021 г.

Количество параметров WuDao 2.0 составило 1,75 трлн. – в 10 раз больше GPT-3 [11]. Для обучения модели было использовано 4,9 ТБ текстов на китайском и английском языках, в то время для GPT-3 – 570 Гб в основном англоязычных. Как сообщает официальный аккаунт BAAI в Wechat, китайская разработка преодолела технические ограничения зарубежных моделей. Ряд решений выложен открытый доступ. WuDao 2.0 либо достигла, либо обогнала конкурентов при выполнении контрольных задач в 9 направлениях, включая поиск данных, заполнение пропусков, генерацию текстов и изображений.

Применение WuDao 2.0 не ограничивается сугубо научными изысканиями. BAAI делает акцент на применении своей разработки в реальных сценариях. Например, модель встроена в интеллектуальный помощник Xiaobu (小布) от компании Oppo [12]. В результате затраты на генерацию одного ответа сократились на 99%.

Другой интересный проект – BaGuaLu [13]. Китайское название (八卦炉 «печь восьми триграмм») взято из мифологических сюжетов о волшебной печи, которая могла создавать лекарства. Название отражает суть разработки, которая должна обеспечить эффективную производительность и масштабируемость. Главная цель BaGuaLu – сделать возможным обучение моделей, сопоставимых по масштабу с человеческим мозгом. Речь идет не о миллиардах, а о более чем сотне триллионов (174) параметров сети – по аналогии с количеством синапсов в человеческом мозге. Масштабные вычисления требуют соответствующих мощностей: модель обучена на New Generation Sunway Supercomputer (Sunway TaihuLight, 神威·太湖之光超级计算机) – №4 в списке 500 мощнейших суперкомпьютеров мира на ноябрь 2021 г. [14].

Открыв статью о BaGuaLu, снова стоит взглянуть на список авторов. Помимо ученых из BAAI и Университета Цинхуа в статье указаны исследователи из DAMO Academy и Zhejiang Lab. DAMO Academy (Discovery, Adventure, Momentum and Outlook, кит. 达摩院) – это R&D-центр компании Alibaba [15]. Как и Zhejiang Lab (之江实验室) — но с той разницей, что в создании лаборатории в Чжэцзян участвовали правительство и партийный комитет провинции [16]. Поддержка властей и одного из крупнейших технологических игроков дает хорошее подспорье в разработке – финансирование, оборудование и инфраструктура уже не представляют проблем для исследователей.

Подводя итоги, можно сказать, что в КНР активное государственно-частное партнерство выступает одним из ключевых драйверов развития ЯМ. Благодаря ресурсной поддержке частного сектора китайская наука оказывается способной создавать передовые решения. Более того, именно научная организация (BAAI) берет на себя

лидерство в продвижении к технологическому фронтиру, а не технологические корпорации, как это происходит в США. В КНР созданы условия для мобилизации и объединения ресурсов – интеллектуальных, финансовых и инфраструктурных. Таким образом, в Китае формируется особая институциональная среда, отличная от таковой в США, но способная дать конкурентоспособные результаты.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. The AI Index Report // Stanford Institute for Human-Centered Artificial Intelligence. URL: https://aiindex.stanford.edu/report/м (date of access: 01.03.2022).

2. CS324 – Large Language Models. URL: https://stanford-cs324.github.io/winter2022/lectures/introduction/ (date of access: 12.02.2022).

3. Tom B. Brown et. al. Language Models are Few-Shot Learners. URL: https://arxiv.org/abs/2005.14165 (date of access: 15.03.2022).

4. On the Opportunities and Risks of Foundation Models // Center for Research on Foundation Models. URL: https://arxiv.org/abs/2108.07258v1 (date of access: 15.03.2022).

5. Or Sharir, Barak Peleg, Yoav Shoham. The Cost of Training NLP Models. URL: https://arxiv.org/abs/2004.08900 (date of access: 15.03.2022).

6. Wei Zeng et al. PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation. URL: https://arxiv.org/abs/2104.12369 (date of access: 15.03.2022).

7. Yu Sun et al. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation. URL: https://arxiv.org/abs/2107.02137 (date of access: 15.04.2022).

8. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (*FAccT '21*). Association for Computing Machinery, New York, NY, USA, 610–623. URL: https://doi.org/10.1145/3442188.3445922 (date of access: 15.04.2022).

9. Sha Yuan et al. A Roadmap for Big Model. URL: https://arxiv.org/abs/2203.14101v3 (date of access: 15.04.2022).

10. O Beijing Academy of Arteficial Intelligince [关于北京智源人工智能研究院]. URL: https://www-pre.baai.ac.cn/about/ (date of access: 15.03.2022).

11. Открылась конференция BAAI, и выпущена самая большая в мире ИИ-модель «Wu Dao 2.0» // Beijing Academy of Artificial Intelligence [智源研究院]. URL: https://mp.weixin.qq.com/s/NJYINRt_uoKAIgxjNyu4Bw (date of access: 17.03.2022).

12. Форум Чжунгуаньцунь 2021 [2021 中关村论坛] // Beijing Academy of Artificial Intelligence [智源研究院]. URL: https://mp.weixin.qq.com/s/erenpJkMucQGLNEfBciGsg (date of access: 17.03.2022).

13. Zixuan Ma, Jiaao He, Jiezhong Qiu, Huanqi Cao, Yuanwei Wang, Zhenbo Sun, Liyan Zheng, Haojie Wang, Shizhi Tang, Tianyu Zheng, Junyang Lin, Guanyu Feng, Zeqiang Huang, Jie Gao, Aohan Zeng, Jianwei Zhang, Runxin Zhong, Tianhui Shi, Sha Liu, Weimin Zheng, Jie Tang, Hongxia Yang, Xin Liu, Jidong Zhai, and Wenguang Chen. 2022. BaGuaLu: targeting brain scale pretrained models with over 37 million cores. In *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (*PPoPP '22*). Association for Computing Machinery, New York, NY, USA, 192–204.

14. TOP 500. November 2021. URL: https://www.top500.org/lists/top500/2021/11/ (date of access: 20.03.2022).
15. About Damo Academy. URL: https://damo.alibaba.com/about/ (date of access: 21.03.2022).
16. О Zhejiang Lab [之江概况]. URL: https://www.zhejianglab.com/index.php?m=content&c=index&a=lists&catid=11 (date of access: 20.03.2022).

МОДЕЛИ ВАРИАТИВНОСТИ ДЕТЕРМИНАТИВОВ ПРОИЗВОДНЫХ ЛОГОГРАММ КИТАЙСКОГО ЯЗЫКА

Н. В. Михалькова

Минский государственный лингвистический университет,
ул. Захарова, 21, 220040, г. Минск, Беларусь, nadezhdakr@yandex.ru

В статье проводится семантический анализ производных логограмм китайского языка, включающих детерминатив 人 / 亻 ‘человек’, с целью определения возможности варьирования смыслового компонента знака при сохранении семантического ядра значения иероглифа. Устанавливаются модели вариативности детерминативов китайских иероглифов, выявляются их когнитивные основания и структурные особенности. На основе глубокого этимологического и многокомпонентного семантического анализа разрабатываются формулы модификаций детерминативов иероглифических знаков китайского языка.

Ключевые слова: вариативность; детерминатив; логограмма, китайский язык; семантика.

RADICAL VARIATION FORMULAS OF THE CHINESE COMPLEX LOGOGRAMS

N. V. Mikhalkova

Minsk State Linguistic University,
Zakharova Str., 21, 220040, Minsk, Belarus, nadezhdakr@yandex.ru

The article provides the semantic analysis of the derived logograms of the Chinese language, including the radical 人/亻 ‘man’, in order to determine the possibility of varying the semantic component of the sign while maintaining the semantic core of the meaning of the character. The models of variability of the radicals of Chinese characters are established, their cognitive foundations and features are revealed. On the basis of a deep etymological and multicomponent semantic analysis, formulas for modifications of the radicals of Chinese character signs are developed.

Keywords: variation; radical; logogramm; chinese; semantics.