Научная статья
УДК 378
DOI 10.52070/2500-3488_2022_4_845_15

Аннотирование данных как объект обучения студентов социально-гуманитарной направленности

Д. В. Алейникова

Московский государственный лингвистический университет, Москва, Россия
festabene@mail.ru

Аннотация. Машинное обучение, искусственный интеллект используют большие массивы данных для решения целого класса профессиональных задач, однако эта практика не получила отражение в современном содержании образования. В статье рассматриваются вопросы актуализации содержания обучения студентов социально-гуманитарного профиля с учетом потребностей взаимодействия «человек-машина». Существенным в данном контексте представляется качественный пересмотр образовательных стратегий, используемых при подготовке современных специалистов.

Ключевые слова: аннотирование данных, искусственный интеллект, социально-гуманитарные науки, цифровизация

Для цитирования: Алейникова Д. В. Аннотирование данных как объект обучения студентов социально-гуманитарной направленности // Вестник Московского государственного лингвистического университета. Образование и педагогические науки. 2022. Вып. 4 (845). С. 15–19. DOI 10.52070/2500-3488_2022_4_845_15

Original article

Data Labeling as an Object of Teaching Social Sciences and Humanities Students

Darya V. Aleynikova

Moscow State Linguistic University, Moscow, Russia
festabene@mail.ru

Abstract. Machine learning and artificial intelligence employ massive amounts of data to solve a wide range of profession-related problems. This practice, however, has not been reflected in the modern educational content. The article addresses issues of „human-machine" interaction while updating the content of teaching social sciences and humanities students. What is crucial in this context is a qualitative revision of the educational strategies used in the training of modern specialists.

Keywords: data labeling, artificial intelligence, social sciences and humanities, digitalization

For citation: Aleynikova, D. V. (2022). Data Labeling as an Object of Teaching Social Sciences and Humanities Students. Vestnik of Moscow State Linguistic University. Education and Teaching, 4(845), 15–19. 10.52070/2500-3488_2022_4_845_15

ВВЕДЕНИЕ

Наращивание «скоростей» – новый тренд, которым руководствуются современные компании и организации. Во всем мире для эффективного и быстрого решения растущего множества задач всё активнее привлекается искусственный интеллект, способный кардинально увеличить производительность труда. Результаты исследования 2021 г,, проведенного компанией McKinsey, показывают, что искусственный интеллект оказывается всё более и более востребованным: 56 % опрошенных (по сравнению с 50 % в 2020 г.) сообщают о внедрении искусственного интеллекта по крайней мере для выполнения хотя бы одной функциональной задачи[1]. Включение искусственного интеллекта в профессиональную деятельность трансформирует ее, тем самым диктуя значительные изменения в осуществлении специалистами своих трудовых функций.

Для повышения эффективности и качества интеграции алгоритмов искусственного интеллекта в бизнес-среду недостаточным оказывается выполнение «заказа» компаний на реализацию технической составляющей деятельности. Важным становится усиление позиций, связанных с подготовкой высококвалифицированных кадров, развитием их специально-предметных и цифровых компетенций на новом методологическом и технологическом основании. Принимая во внимание вызовы времени, требующие качественной трансформации профессиональной деятельности, образовательной сфере необходимо оперативно провести педагогическую конкретизацию поставленных обществом целей, сформулировать задачи, определить пути и средства их достижения.

Важнейшая образовательная задача связана с актуализацией содержания обучения специалистов социально-гуманитарного профиля с учетом контекстов (настоящих и прогнозируемых) взаимодействия человека с машинным интеллектом. Выделение объектов обучения требует глубокого осознания специфики такого взаимодействия, учета целого ряда факторов, влияющих на характер и результативность этой деятельности.

В данной статье предпринята попытка выявить такие объекты на основе анализа профессионального дискурса специалистов социально-гуманитарного профиля, его проецирования на образовательные контексты деятельности.

[1]URL: https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021

ОБЗОР ЛИТЕРАТУРЫ

Для тренировки алгоритмов искусственного интеллекта важнейшим ресурсом оказывается информация. Искусственный интеллект работает с закономерностями в данных и использует знания, правила и информацию, которые были специально закодированы людьми для их последующей обработки.

Исследователи из разных стран мира стремятся предложить такие технологические решения, которые позволили бы искусственному интеллекту самостоятельно адаптироваться к актуальным условиям и принимать решения на основе больших массивов данных. Обращает на себя внимание тот факт, что в качестве вопросов, ограничивающих возможности применения искусственного интеллекта, мы, в числе прочих, обнаруживаем: отсутствие современных программ подготовки специалистов в сфере искусственного интеллекта; дефицит специалистов на рынке труда; низкую готовность кадров в большинстве компаний к использованию технологий искусственного интеллекта; *отсутствие методологии сбора и разметки данных*[2].

Аннотирование (разметка) данных представляет собой процесс добавления меток к необработанным данным (изображениям, видео, текстам и аудио). Проставленные метки формируют представление о том, к какому классу объектов относится информация, что позволяет моделям машинного обучения классифицировать объекты и делать выводы. Разметка данных лежит в основе контролируемого машинного обучения. Системы искусственного интеллекта обучаются, а не программируются, при этом комплексная сложная задача потребует аккумулирования больших ресурсов для создания обучающей выборки. Получение больших наборов данных может быть затруднено. В некоторых областях они могут быть недоступны, но даже при их наличии усилия по их аннотированию зачастую оказываются трудозатратными. Алгоритмы машинного обучения работают с закономерностями в числах, и если данные нерепрезентативны, то нельзя говорить о корректности полученного результата[3].

Качественное аннотирование данных является необходимым условием для контролируемого машинного обучения. Такое положение дел объясняется зависимостью между производительностью модели в операциях от качества обучающих данных[4]. В некоторых случаях для аннотирования привлекают внешние компании или внешних

[2]URL: https://apr.moscow/content/data/5/Технологии%20искусственного%20интеллекта.pdf

[3]URL: https://habr.com/ru/post/449224/

[4]URL: https://azati.ai/automated-data-labeling-with-machine-learning

участников, специализирующихся на разметке данных (Crowdsourcing) [Polachowska, 2019]. При этом при привлечении внешних агентов для разметки специализированных данных увеличивается производительность при относительно небольших затратах и упрощается процедура поиска (специализированные платформы аккумулируют профили специалистов), но дискуссионным остается вопрос *качества разметки данных*. При использовании crowdsourcing возможны три причины низкого качества разметки: неоднозначность данных, отсутствие комплексного руководства для аннотаторов и нехватка мотивации или предметных знаний у аннотатора [Kilgarriff, 1998]. Разметка «высококонтекстных» данных, таких как, например, классификация юридических контрактов, медицинские изображения или научная литература, настоятельно требует привлечения специалистов из этих предметных областей[1].

Принимая во внимание глобальный рост доступной информации, и, как следствие, возрастающую необходимость обучать все большее количество моделей (в том числе моделей для решения узкоспециализированных задач), мы можем говорить о высокой востребованности и конкурентоспособности специалистов социально-гуманитарного профиля, занимающихся профессиональной деятельностью на «стыке» своей предметной области и искусственного интеллекта.

Так, например, анализ профессиональной деятельности юриста показывает, что специалисты, работающие в предметной области юриспруденции, испытывают значительные трудности при использовании расширенного технологического инструментария при решении задач профессиональной деятельности. По данным исследования, проведенного Bloomberg Law, более половины всех респондентов (54%) заявили, что не используют искусственный интеллект или инструменты машинного обучения; четверть опрошенных сообщили о том, что не знают, применяют ли подобные технологические решения, что, очевидно, свидетельствует о существенных пробелах в знаниях и недостаточном использовании цифровых технологий[2].

СОВРЕМЕННЫЕ ПОДХОДЫ К АННОТИРОВАНИЮ ДАННЫХ

Исследователи, занимающиеся вопросами аннотирования данных, сходятся во мнении, что при привлечении специалистов определенной области для выполнения проектов, требующих экспертных знаний, необходимо уделять особое внимание обучению этой деятельности [Tseng, Stent, Maida, 2020]. Обычно процессы аннотирования подробно не описываются, а их качество не оценивается, что часто делает общедоступные наборы данных и предоставляемые ими аннотации непригодными для использования [Hein и др., 2014]. В этой связи вопросы создания обучающей выборки затрагивают ряд существенных моментов.

Прежде всего отметим разные подходы к *суммаризации*. Первый подход предполагает суммаризацию текста через «понимание». Суммаризация в области искусственного интеллекта рассматривает процесс автоматического реферирования как эквивалент человеческой деятельности, при этом суммаризация основывается на частичном или полном понимании текста. В то же время исследователи данного вопроса особо подчеркивают, что подобный подход к суммаризации довольно сложен в реализации, так как требует *автоматического понимания*, представления и формирования текста [Atanassova, Blais, Descles, 2008]. Существующее технологические способы решения подобных задач не удовлетворяют актуальным требованиям, что на данном этапе является весомым аргументом в пользу их успешной реализации непосредственно человеком.

Второй подход связан с созданием простого информативного резюме представленного фрагмента и не предполагает глубокого анализа текста; он скорее связан с непосредственным извлечением релевантных смысловых единиц. При этом критерии релевантности часто определяются машиной на основе заведомо неактуальных для человека параметров (частота предъявления, выделение заголовка и др.), если речь идет об использовании численных методов статистики, или в контексте лингвистических методов – определение лингвистических маркеров, лежащих на поверхности [там же]. В рамках проводимого исследования предметом особого интереса является *дискурсивная суммаризация* текста (автореферирование). Дискурсивная суммаризация текста направлена на автоматическое создание речевого произведения, которое объясняет, как дискурсивные единицы соотносятся друг с другом, и какую роль они играют в общем дискурсе [Wolf, Gibson, 2005]. При этом дискурсивная суммаризация может быть направлена на классифицирование дискурсивных единиц или на поиск и установление смысловых связей между дискурсивными единицами [Putra и др., 2021]. Несмотря на внедрение передовых

[1]URL: https://medium.com/syncedreview/data-annotation-the-billion-dollar-business-behind-ai-breakthroughs-d929b0a50d23

[2]URL: https://news.bloomberglaw.com/us-law-week/lawyers-arent-taking-full-advantage-of-ai-tools-survey-shows

технологических решений [Putra и др., 2021], процессы, связанные с поиском смыслов, образуют поле неопределенности для машины.

В рамках еще одного подхода при разметке текста также значимым оказывается стиль произведения. Так, в речевых произведениях разной стилевой направленности релевантная информация может быть расположена в разных структурных частях. Обучение алгоритмов искусственного интеллекта обнаружению существенных условий типового договора купли-продажи не будет вызывать трудностей. При этом нестандартизированные юридические документы могут ввести машину в заблуждение. Приведем пример.

> В канцелярию вошел рыжий бородатый милиционер… в форменной фуражке, тулупе с косматым воротником. Под мышкой милиционер осторожно держал маленькую разносную книгу в засаленном полотняном переплете. Застенчиво ступая своими слоновыми сапогами, милиционер подошел к Ипполиту Матвеевичу и налег грудью на тщедушные перильца.
>
> – Здорово, товарищ, – густо сказал милиционер, доставая из разносной книги большой документ, – товарищ начальник до вас прислал, доложить на ваше распоряжение, чтоб зарегистрировать.
>
> Ипполит Матвеевич принял бумагу, расписался в получении и принялся ее просматривать. Бумага была такого содержания:
>
> «Служебная записка. В загс. Тов. Воробьянинов! Будь добрый. У меня как раз сын народился. В 3 часа 15 минут утра. Так ты его зарегистрируй вне очереди, без излишней волокиты. Имя сына – Иван, а фамилия моя. С коммунистическим пока Замначальника Умилиции Перервин».
>
> Ипполит Матвеевич заспешил и без лишней волокиты, а также вне очереди (тем более что ее никогда и не бывало) зарегистрировал дитя Умилиции [Ильф, Петров, 2000 с. 381–382].

Современный искусственный интеллект хорошо справляется с задачей поиска элементов – *search and find task*, соответствующих критериям, определенных человеком, и выявляет закономерности в данных[1]. Другими словами, в юриспруденции искусственный интеллект, обученный на наборах размеченных данных, может классифицировать акты / договоры и др. и распределять их в компании юристам соответствующего профиля.

Отметим, что для обучения нейронных сетей используется многократное предъявление примеров из обучающей выборки, что позволяет сгруппировать сходные образы в классы [Кабалдин и др., 2019]. В рассмотренном примере высокочастотным оказываются слова «миллиционер», «Умилиция», что, вероятно, приведет машину к неправильным выводам и позволит классифицировать приведенный документ как оформленный в милиции или милиционером. При этом собственно все существенные характеристики, необходимые для документов такого рода, соблюдены. Тем не менее для правильной классификации нейронной сетью речевое произведение требует серьезной доработки – экспликации скрытых в нем смыслов человеком.

ЗАКЛЮЧЕНИЕ

Вероятно, уровень технологического прогресса в скором времени позволит искусственному интеллекту работать с речевыми произведениями, уникальными по форме и стилю. Однако вопросы *подготовки обучающей выборки*, а именно – *первичная подготовка текста*, вопросы *выделения дискурсивных категорий / подкатегорий*, необходимых для разметки, *установление значимых позиций* и *их экспликация* останутся частью профессиональной деятельности человека. С нашей точки зрения, значимым для проектирования образовательной стратегии в данном контексте оказывается положение об изоморфизме деятельности учебной и профессиональной [Яроцкая, 2021], где последняя на современном этапе развития общества имеет ярко выраженную технологическую составляющую. Это обусловливает выбор актуальных объектов обучения специалистов социально-гуманитарного профиля, подкрепление этих объектов на междисциплинарном уровне, их последующую практическую реализацию в предмете профессиональной деятельности.

[1]URL: https://jolt.law.harvard.edu/digest/a-primer-on-using-artificial-intelligence-in-the-legal-profession

СПИСОК ИСТОЧНИКОВ

1. Polachowska K. AI in education: can AI improve the way we teach and learn? – Neoteric. Software House That Helps You Innovate, 2019. URL: https://neoteric.eu/blog/ai-in-education-can-ai-improve-the-way-we-teach-and-learn/
2. Kilgarriff A. Gold standard datasets for evaluating word sense disambiguation programs // Computer Speech and Language. 1998. № 3. Vol. 12. P. 453–472.
3. Tseng T., Stent A., Maida D. Best Practices for Managing Data Annotation Projects, 2020. URL: https://www.researchgate.net/publication/344343968_Best_Practices_for_Managing_Data_Annotation_Projects
4. Hein A. [et al.]. WOAR 2014: Workshop on Sensor-based Activity Recognition, chapter Towards causally correct annotation for activity recognition / A. Hein, A. F. Kruger, K. Yordanova, T. Kirste. Fraunhofer. 2014. P. 31–38.

5. Atanassova I., Blais A., Descles J.-P. A Cross-Lingual Approach to the Discourse Automatic Annotation: Application to French and Bulgarian // Proceedings of the Twenty- First International FLAIRS Conference. 2008. P. 450–455.
6. Wolf F., Gibson E. Representing discourse coherence: A corpus-based study // Computational Linguistics. 2005, № 2. Vol. 31 C. 249–288.
7. Putra J. W. G. [и др.]. TIARA 2.0: an interactive tool for annotating discourse structure and text improvement / J. W. G. Putra, K. Matsumura, S. Teufel, T. Tokunaga // Language Resources & Evaluation. 2021. DOI 10.1007/s10579-021-09566-0 URL: https://link.springer.com/article/10.1007/s10579-021-09566-0
8. Ильф И., Петров Е. Двенадцать стульев: роман. М. : ВАГРИУС, 2000.
9. Кабалдин Ю. Г. [и др.]. Искусственный интеллект, интернет вещей, облачные технологии и цифровые двойники в современном механообрабатывающем производстве: монография / Ю. Г. Кабалдин, Д. А. Шатагин, П. В. Колчин, М. С. Аносов. Нижегородский государственный технический университет им. Р. Е. Алексеева. Нижний Новгород, 2019.
10. Яроцкая Л. В. Иностранный язык как инструмент формирования современной профессиональной личности в условиях неязыкового вуза // Вестник Московского государственного лингвистического университета. Образование и педагогические науки. 2021. Вып. 1(838). С. 193–201. DOI 10.52070/2500-3488_2021_1_838_193.

REFERENCES

1. Polachowska, K. (2019). AI in education: can AI improve the way we teach and learn? – Neoteric. Software House That Helps You Innovate. https://neoteric.eu/blog/ai-in-education-can-ai-improve-the-way-we-teach-and-learn/
2. Kilgarriff, A. (1998). Gold standard datasets for evaluating word sense disambiguation programs. Computer Speech and Language, 12(3), 453–472.
3. Tseng, T., Stent, A., Maida, D. (2020). Best Practices for Managing Data Annotation Projects. https://www.researchgate.net/publication/344343968_Best_Practices_for_Managing_Data_Annotation_Projects
4. Hein, A. et al. (2014). WOAR 2014: Workshop on Sensor-based Activity Recognition (pp. 31–38). Chapter Towards causally correct annotation for activity recognition. Fraunhofer.
5. Atanassova, I., Blais, A., Descles, J.-P. (2008). A Cross-Lingual Approach to the Discourse Automatic Annotation: Application to French and Bulgarian. Proceedings of the Twenty-First International FLAIRS Conference. (pp. 450–455).
6. Wolf, F., Gibson, E. (2005). Representing discourse coherence: A corpus-based study. Computational Linguistics, 1(2), 249–288.
7. Putra, J.W.G. et. al. (2021). TIARA 2.0: an interactive tool for annotating discourse structure and text improvement. Language Resources & Evaluation. 10.1007/s10579-021-09566-0. https://link.springer.com/article/10.1007/s10579-021-09566-0
8. Ilf, I., Petrov, E. (2000). The twelve chairs: a novel. Moscow: VAGRIUS. (In Russ.)
9. Kabaldin YU. G. et. al. (2019). Iskusstvenny`j intellekt, internet veshhej, oblachny`e texnologii i cifrovy`e dvojniki v sovremennom mexanoobrabaty`vayushhem proizvodstve: monografiya = Artificial intelligence, Internet of things, cloud technologies and digital twins in modern metal processing industry: monograph. Nizhny Novgorod State Technical University n. a. R. E. Alekseev. (In Russ.)
10. Yarotskaya, L. V. (2021). Foreign language as a tool for developing a modern professional at non-linguistics university. Vestnik of Moscow State Linguistic University. Education and Teaching, 1(838), 193–201. 10.52070/2500-3488_2021_1_838_193 (In Russ.)

ИНФОРМАЦИЯ ОБ АВТОРЕ

Алейникова Дарья Викторовна

кандидат педагогических наук, доцент кафедры лингвистики и профессиональной коммуникации в области права Института международного права и правосудия Московского государственного лингвистического университета

INFORMATION ABOUT THE AUTHOR

Aleynikova Darya Viktorovna

PhD (Pedagogy), Associate Professor of the Department of Linguistics and Professional Communication in the Field of Law, Institute of International Law and Justice, Moscow State Linguistic University

Статья поступила в редакцию 10.02.2022
одобрена после рецензирования 25.02.2022
принята к публикации 26.09.2022

The article was submitted 10.02.2022
approved after reviewing 25.02.2022
accepted for publication 26.09.2022