

УДК 004.85

149. ПРИМЕНЕНИЕ AMAZON WEB SERVICES ДЛЯ РАСПРЕДЕЛЕННОГО МАШИННОГО ОБУЧЕНИЯ ПРИ РЕШЕНИИ ЗАДАЧ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

Голубев А.А., студент гр.578105

*Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь*

Ефремов А.А. – канд. эконом. наук, доцент

Аннотация. В статье рассматриваются возможности облачной экосистемы Amazon Web Services (AWS) как инфраструктурной основы для распределенного машинного обучения в задачах компьютерной лингвистики. Актуальность темы обусловлена ростом объема текстовых корпусов, усложнением моделей обработки естественного языка и повышением требований к вычислительной эффективности. Проанализированы сервисы AWS, применяемые для хранения данных, предобработки, распределенного обучения, развертывания и мониторинга лингвистических моделей. Основное внимание уделено управляемым инструментам обучения и масштабирования вычислений. Показано, что использование AWS способствует снижению инфраструктурных издержек, повышению эластичности вычислительной среды и улучшению воспроизводимости решений в области компьютерной лингвистики. Вместе с тем выделены ограничения, связанные со стоимостью вычислений, межузловым взаимодействием, защитой данных и зависимостью от облачного провайдера.

Ключевые слова. Amazon Web Services, распределенное машинное обучение, компьютерная лингвистика, обработка естественного языка, Amazon SageMaker, облачные вычисления, языковые модели.

Введение. Компьютерная лингвистика активно использует нейросетевые модели, обучение которых требует значительных вычислительных ресурсов. Рост объема текстовых данных, усложнение архитектур и необходимость многократных экспериментов делают локальную инфраструктуру недостаточной или экономически неэффективной. Текстовые корпуса отличаются большим объемом, неоднородностью, многоязычием и сложной предобработкой. Дополнительную нагрузку создают трансформерные модели, для которых критичны объем памяти, скорость обмена данными и возможность параллельного обучения. В этих условиях облачные платформы становятся практическим решением. Amazon Web Services предоставляет среду для хранения данных, распределенного обучения, развертывания моделей и их сопровождения.

Применение AWS в распределенном машинном обучении для задач компьютерной лингвистики.

Компьютерная лингвистика опирается на данные, отличающиеся высокой размерностью, зависимостью от контекста и сложной семантической организацией. В отличие от табличных данных текст требует предварительного представления, способного отражать морфологические, синтаксические и семантические закономерности языка. Это определило переход от ручного конструирования признаков и статистических моделей к контекстным представлениям и глубоким нейросетевым архитектурам. Даже прикладные задачи, например классификация обращений или маршрутизация документов, нередко требуют моделей, обученных на крупных корпусах и учитывающих длинный контекст. В машинном переводе, извлечении именованных сущностей и генеративных системах вычислительная сложность возрастает еще сильнее. Она определяется объемом корпуса, числом параметров модели, стоимостью токенизации и необходимостью многократных экспериментов. В результате инфраструктура в современной компьютерной лингвистике перестает играть второстепенную роль и становится одним из условий, определяющих качество и воспроизводимость получаемых результатов.

Распределенное машинное обучение предполагает выполнение одного процесса обучения на нескольких устройствах или вычислительных узлах. К такому подходу обращаются в тех случаях, когда обучение на одном сервере становится слишком длительным или когда объем модели и данных превышает доступные ресурсы памяти. Его развитие связано с исследованиями в области распределенной обработки данных и параллельных вычислений [1; 2]. Основными стратегиями остаются параллелизм по данным и параллелизм по модели. В первом случае модель дублируется на нескольких устройствах, а обучающие данные распределяются между ними; после этого результаты вычислений на каждом узле согласуются. Во втором случае между устройствами распределяются уже

не данные, а сама модель, что особенно востребовано при работе с крупными трансформерными архитектурами [3]. На практике чаще применяют смешанный подход, сочетающий оба механизма с учетом размера модели, характеристик оборудования и параметров сети. Для компьютерной лингвистики это особенно важно, поскольку обучение языковых моделей связано не только с высокими вычислительными затратами, но и с необходимостью устойчиво работать с большими корпусами, токенизированными данными и промежуточными результатами. Именно поэтому вопрос инфраструктуры здесь напрямую связан с качеством и воспроизводимостью результатов, а облачная среда становится естественным продолжением современных методов распределенного обучения.

AWS – это облачная платформа, объединяющая вычислительные мощности, средства хранения данных и управляемые инструменты для разработки и эксплуатации моделей машинного обучения. Для задач компьютерной лингвистики это особенно важно, поскольку работа с моделью включает не только ее обучение, но и ряд сопутствующих этапов. Необходимы хранение и версияция корпусов, воспроизводимая предобработка, сохранение контрольных точек, сравнение запусков и последующее развертывание полученной модели. Облачная инфраструктура позволяет собрать эти этапы в единый контур и уменьшить зависимость от локальных ограничений по памяти, дисковому пространству и вычислительной мощности.

Сильная сторона AWS состоит в сочетании двух режимов работы. С одной стороны, платформа дает доступ к гибкой инфраструктуре, которую можно настраивать под конкретную задачу. С другой стороны, она предлагает управляемые средства, снижающие объем рутинной настройки и позволяющие быстрее переходить от корпуса к обучению модели. Такой подход особенно востребован в компьютерной лингвистике, где исследовательский цикл часто включает серию экспериментов с разными архитектурами и параметрами. Актуальность подобной инфраструктуры усилилась после широкого распространения трансформерных моделей и подходов к предварительному обучению двунаправленных языковых представлений [4]. Тем самым AWS выступает не просто вычислительной площадкой, а средой, в которой работа с языковыми моделями становится более системной, воспроизводимой и масштабируемой.

Таблица 1 – Возможности и ограничения использования AWS в задачах компьютерной лингвистики.

Критерий	Возможности AWS	Практическое значение для компьютерной лингвистики
Масштабирование вычислений	Гибкое увеличение числа вычислительных ресурсов	Ускоряет обучение крупных языковых моделей и проведение серии экспериментов
Хранение данных	Централизованное хранение корпусов, артефактов и результатов	Упрощает работу с большими текстовыми коллекциями и повышает воспроизводимость
Организация экспериментов	Объединение этапов подготовки данных, обучения и развертывания	Сокращает число ручных операций и делает исследовательский цикл более управляемым
Ограничения по стоимости	Оплата зависит от объема и длительности вычислений	Требует контроля бюджета, особенно при распределенном обучении
Межузловое взаимодействие	Эффективность зависит от обмена данными между узлами	Может снижать прирост производительности при неудачной конфигурации
Защита данных	Требуются меры контроля доступа и безопасного хранения	Особенно важно при работе с чувствительными текстовыми данными

Использование AWS в задачах компьютерной лингвистики дает заметные практические преимущества. Облачная платформа дает возможность изменять объем используемых ресурсов в соответствии с текущим этапом работы – от подготовки корпуса до обучения и развертывания модели. Это особенно важно, поскольку современные языковые модели обычно требуют длительных запусков и регулярной корректировки конфигурации. Кроме того, объединение хранения данных, обучения и мониторинга в одной среде сокращает число ручных операций и делает экспериментальный цикл более управляемым. На фоне роста больших языковых моделей такая инфраструктурная гибкость становится все более значимой [5].

Однако эффективность AWS не является автоматической. Использование облачной платформы связано с затратами на вычисления, риском зависимости от конкретного провайдера и повышенными требованиями к защите чувствительных текстовых данных. Существенное значение имеют и накладные расходы, возникающие при межузловом обмене в распределенной среде. Поэтому применение AWS оправдано прежде всего тогда, когда архитектура проекта заранее соотнесена с объемом корпуса, типом модели, допустимой стоимостью экспериментов и требованиями к инференсу. В таком случае облачная платформа выступает не просто внешним источником вычислительных мощностей, а полноценным инструментом решения задач компьютерной лингвистики.

Заключение.

В рамках работы проведен анализ, который показывает, что AWS можно рассматривать как одну из наиболее удобных платформ для организации распределенного машинного обучения в данной области. Ее значимость связана не только с возможностью масштабировать вычисления, но и с объединением в единой среде хранения корпусов, подготовки данных, обучения, развертывания и сопровождения моделей. Вместе с тем сама по себе облачная инфраструктура не гарантирует эффективности. Практическая результативность определяется тем, насколько выбранная архитектура соответствует объему корпуса, типу модели, бюджету экспериментов и требованиям к защите данных. Поэтому использование AWS наиболее оправдано в тех случаях, когда облачная платформа изначально включается в исследовательский и прикладной контур как продуманный инфраструктурный элемент, а не только как источник вычислительных ресурсов.

Список использованных источников:

1. Dean, J. MapReduce: Simplified Data Processing on Large Clusters / J. Dean, S. Ghemawat // *Communications of the ACM*. – 2008. – Vol. 51. – No. 1. – P. 107–113.
2. Li, M. Scaling Distributed Machine Learning with the Parameter Server / M. Li, D. G. Andersen, J. Park et al. // *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation*. – 2014. – P. 583–598.
3. Vaswani, A. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar et al. // *Advances in Neural Information Processing Systems*. – 2017. – Vol. 30.
4. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // *Proceedings of NAACL-HLT*. – 2019. – P. 4171–4186.
5. Brown, T. B. Language Models are Few-Shot Learners / T. B. Brown, B. Mann, N. Ryder et al. // *Advances in Neural Information Processing Systems*. – 2020. – Vol. 33. – P. 1877–1901.

UDC 004.85

APPLICATION OF AMAZON WEB SERVICES FOR DISTRIBUTED MACHINE LEARNING IN COMPUTATIONAL LINGUISTICS PROBLEMS

Golubev A.A.

Belarusian State University of Informatics and Radioelectronics¹, Minsk, Republic of Belarus

Efremov A.A. – PhD in Economic Sciences, Associate Professor

Annotation. This paper examines the capabilities of the Amazon Web Services (AWS) cloud ecosystem as an infrastructure foundation for distributed machine learning in computational linguistics. This topic is relevant given the growing size of text corpora, the increasing complexity of natural language processing models, and increasing demands on computational efficiency. We analyze AWS services used for data storage, preprocessing, distributed training, deployment, and monitoring of linguistic models. We focus on managed training tools and computational scalability. We demonstrate that using AWS reduces infrastructure costs, increases the elasticity of the computing environment, and improves the reproducibility of computational linguistics solutions. However, limitations related to computational costs, inter-node communication, data security, and cloud provider dependency are highlighted.

Keywords. Amazon Web Services, distributed machine learning, computational linguistics, natural language processing, Amazon SageMaker, cloud computing, language models.