Концепция лаборатории естественного языка

ВШЭ и Яндекс

Создаваемая в НИУ ВШЭ - Санкт-Петербург лаборатория имеет предварительное название «лаборатория естественного языка ВШЭ и Яндекс» (далее — лаборатория). Предполагается, что лаборатория будет входить в состав факультета Санкт-Петербургская школа физико-математических и компьютерных наук НИУ ВШЭ — Санкт-Петербург.

Основное направление научных исследований. Лаборатория будет исследовать фундаментальные теоретико-информационные и статистические свойства дискретных последовательностей. Профильными прикладными направлениями работы лаборатории будут генерация текстов и компьютерная лингвистика, а также разработка программных средств анализа текстов и компонент интеллектуальных информационных систем. Теоретические исследования лаборатории будут связаны с теорией информации и Representation Learning. Кроме того, в рамках работы над генеративными языковыми моделями лаборатория будет неизбежно сталкиваться с дефицитом математического аппарата для описания генерации текстов. Создание математического аппарата для строгого определения таких понятий, как “семантическая информация”, “новизна последовательности” или “сюжетная линия” является вторым направлением теоретических исследований лаборатории. Прикладные и теоретические направления работы лаборатории будут взаимно дополнять друг друга.

Планируемые направления исследований.

1. Анализ статистических свойств текстов на естественном языке.
2. Разработка методов проектирования глубинных нейронных сетей и методов глубинного обучения (deep learning) для генерации текстов.
3. Интерпретация сложных моделей машинного обучения.
4. Анализ данных в компьютерной лингвистике.
5. Исследования теоретико-информационных свойств текстов на естественном языке.

6. Исследования методов представления текстовой информации.
7. Разработка математического аппарата для создания теории семантической информации.

Обоснование актуальности проекта. Колоссальный рост объема разнообразной информации в современном обществе (30% в год), называемый информационным взрывом, настоятельно требует как новых решений в области искусственного интеллекта и анализа данных, так и новых, высококвалифицированных кадров в этой области. Спрос на таких специалистов растет экспоненциально. Особенно остро ощущается потребность в научно-педагогических кадрах в данной области — практически в каждом зарубежном университете ежегодно открываются вакансии на специалистов в этих областях. При этом количество таких специалистов в мире значительно меньше, чем количество открываемых вакансий.

Похожая ситуация наблюдается и в Санкт-Петербурге. Ведущие вузы города (СПбГУ, ИТМО, НИУ ВШЭ СПб, Политех) открывают новые бакалаврские и магистерские программы в области искусственного интеллекта и анализа данных, к ним обращаются ведущие компании (Яндекс, JetBrains, Газпромнефть, Сбербанк и др.) с предложениями о проведении научных исследований в этой области. Однако количество высококвалифицированных специалистов в этой области в городе критически мало.

Понимая эту ситуацию, Яндекс готов помогать организационно и финансово в привлечении ведущих специалистов (как наших бывших соотечественников, так и иностранцев) в области искусственного интеллекта и анализа данных в Санкт-Петербург. При этом ближайшая цель такого проекта - создание точки развития наук в области искусственного интеллекта и анализа данных, которая бы со временем превратилась в полноценную научную лабораторию в этой области, способную как решать современные прикладные и фундаментальные задачи, так и способствовать подготовке кадров высшей квалификации (магистров и аспирантов).

С точки зрения Яндекса, факультет Санкт-Петербургская школа физико-математических и компьютерных наук является идеальной площадкой для создания такой лаборатории. Яндекс активно участвует в подготовке бакалавров и магистров, предоставляя студентам бакалаврских и магистерских программ научно-исследовательские проекты, практики, приглашая их на стажировки, являясь одной из основных компаний, куда по окончании бакалавриата и магистратуры уходят на работу выпускники. При этом компания Яндекс хотела бы углубить и расширить это взаимодействие, участвуя в подготовке аспирантов этой школы, привлекая ведущих ученых и участвуя в совместных научно-исследовательских проектах. Для этого компания готова предложить паритетное (50 на 50) софинансирование для привлечения ведущих ученых на долгосрочной основе в лабораторию под названием «лаборатория естественного языка ВШЭ и Яндекс».

Со своей стороны, факультет Санкт-Петербургская школа физико-математических и компьютерных наук давно и успешно работает в области искусственного интеллекта и анализа данных. Так, на факультете еженедельно проходят 4 содержательных научно-исследовательских семинара в этой области (“Агентные системы и обучение с подкреплением”, “Прикладное машинное обучение и глубокое обучение”, “Машинное обучение в программной инженерии”, “Машинное обучение в биологии и медицине”), на которых выступают как известные в своих областях специалисты, так и студенты бакалаврских и магистерских программ факультета. Появление на факультете новой лаборатории естественного языка, проведение в ней содержательных научных исследований в области искусственного интеллекта и анализа данных существенно усилит научную и проектную составляющие работы со студентами и аспирантами факультета.

Кроме того, вновь создаваемая лаборатория планирует наладить тесное сотрудничество в научной сфере с недавно созданной научно-учебной лабораторией компании Яндекс на ФКН https://cs.hse.ru/big-data/yandexlab/ Тематики этой лаборатории перекликаются с планируемыми темами

исследований вновь создаваемой лаборатории естественного языка ВШЭ и Яндекс.

Кадровое обеспечение лаборатории

Предполагается привлекать в лабораторию в год от 3 до 5 студентов, аспирантов и научных сотрудников с тем, чтобы в перспективе за три года выйти на следующий кадровый состав лаборатории: 1-2 ведущих ученых, 3-5 научных сотрудников, 3-5 аспирантов, а также 8-10 студентов бакалавриата и магистратуры.

На данный момент к научной работе формируемого в настоящее время коллектива уже привлечен внешний научный сотрудник Шарвин Резаголи — перспективный специалист в области theoretical computer science. В ближайшие несколько лет он планирует работать над исследованием клеточных автоматов в контексте обработки потоковой информации динамическими обучающимися системами. Его результаты могут быть востребованы в рамках работы ЛЕЯ для исследований вопросов обработки и генерации естественного языка. В данный момент господин Резаголи руководит исследовательской группой в частной компании, но он заинтересован в публикации научных результатов, не связанных с прикладными задачами компании. Уже достигнута договоренность, что свои научные результаты господин Резаголи будет публиковать с аффилиацией Высшей школы экономики (кампус в Санкт-Петербурге) без привлечения дополнительного финансирования со стороны ВШЭ. Он также будет выступать в качестве соруководителя ряда теоретических исследований сотрудников лаборатории.

Кроме того, у потенциального руководителя лаборатории — Ивана Ямщикова, есть кадровый резерв из молодых специалистов, которые могли бы быть привлечены к работе лаборатории в качестве научных сотрудников, стажеров и лаборантов. Это студенты бакалаврской программы “Прикладная математика и информатика” и магистерской программы “Программирование и анализ данных”, с которыми Иван Ямщиков уже ведет научную работу, а также

аспиранты других российских и зарубежных университетов, планирующие прийти в лабораторию в качестве научных сотрудников. Кроме того, к работе планируется привлечь несколько молодых специалистов из российских технологических компаний, заинтересованных в исследовательской деятельности и возможной смене своей карьерной траектории с отраслевой на научную.

Требуемые помещения и оборудование

Дополнительных помещений и оборудования на данный момент вновь создаваемой лаборатории не требуется.

Примерные количественные планы функционирования лаборатории

	Наименование показателя	**2021 г.**	**2022 г.**	**2023 г.**
1.	Количество сотрудников подразделения, включая руководителя	4	7	9
1.1.	Студентов/аспирантов НИУ ВШЭ в штате подразделения	2	4	6
1.2.	Постдоков в штате подразделения	0	1	2
2.	Количество статей сотрудников подразделения с аффилиацией НИУ ВШЭ, в международных журналах, индексируемых WoS (ед.), а также с докладами на конференциях	1	6	9
2.1.	Из них статей в журналах уровня Q1/Q2 и с докладами на конференциях уровня A* (ед.)	1	2	4
3.	Объем привлеченного в НИУ ВШЭ внешнего финансирования (млн. руб.)	5	6	12

3.1.	Из них средства компании Яндекс	5	5	5

Целевые конференции лаборатории

В современных компьютерных науках приоритетными являются не журнальные публикации, а участие в крупных тематических конференциях. В контексте исследований лаборатории следующие конференции класса A* по версии ВШЭ являются для неё целевыми: NeurIPS, ICML, AAAI, KDD, ICLR.

Компьютерная лингвистика и обработка естественного языка — более узкое направление исследований в рамках компьютерных наук, поэтому, помимо указанных конференций, целесообразно ориентироваться на наиболее цитируемые конференции по данным платформы Google Scholar (https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics), в особенности на конференции ACL, EMNLP, HLT-NAACL, EACL, COLING, CoNLL.

Профильные журналы из первого квантиля Scopus или WebOfScience также являются целевыми направлениями для публикаций лаборатории, однако не являются первым приоритетом.

Описание финансовой модели

На старте предполагается софинансирование лаборатории компанией Яндекс в размере пяти миллионов рублей в год. Конкретный объем средств в дальнейшем будет зависеть от количества и качества привлекаемых сотрудников, и будет уточнён в 2023 году по итогам оценки эффективности работы лаборатории.

Объем финансирования лаборатории из различных источников, млн.руб.

Источник финансирования	2021	2022	2023
1. Яндекс	5	5	5
2. Центральный бюджет	5	5	5
3. Академические надбавки 3-го уровня	0.6	2.5	4.2
4. Питерский кампус	-	5	5
5. Гранты РФФИ, РНФ	-	1	7
Итого	**10.600**	**18.500**	**26.200**

Основные расходы лаборатории будут состоять в выплате заработной платы сотрудникам лаборатории. Кроме того, в расходной части планируются командировки, связанные с поездками на конференции, а также расходы, связанные с публикацией результатов исследований.

Примерный объем расходов лаборатории

	2021	2022	2023
Количество ставок	5	8	9
Средняя заработная плата сотрудника лаборатории, тыс.руб.	128	133	180
Суммарные расходы на заработную плату, включая ЕСН, тыс.руб.	10 000	16 690	25 310