на её обслуживание.

Список литературы

[1] Alabdullah B., Beloff N., White M. Rise of Big Data — issues and challenges // 2018 21st Saudi Computer Society National Computer Conference (NCC). Riyadh, Saudi Arabia, 2018. Vol. 1. P. 1–6.

[2] Gyssens M., Paredaens J., Van den Bussche J. A graph-oriented object database model // IEEE Transactions on Knowledge and Data Engineering. 1994. Vol. 6. P. 572–586.

4.13. *Мезенцева А.А., Бручес Е.П.* Исследование автоматического связывания сущностей в научных текстах на русском языке

Автоматическое связывание сущностей (англ. entity linking) — задача нахождения соотношения между упоминанием в тексте и уникальной сущностью в структурированной базе знаний (в данной работе используется Wikidata). Актуальность исследования заключается в том, что рассмотренные нами методы не требуют большого количества данных, которого для русского языка, насколько нам известно, нет в открытом доступе. Новизна работы состоит в сравнении различных подходов к решению поставленной задачи и анализе результатов, полученных при тестировании на русскоязычном наборе данных.

Нами были проведены эксперименты, в качестве тестового набора данных использовался корпус научных статей RuSERRC [1]. На вход алгоритму подается единичный токен или последовательность токенов, соответствующих термину. Затем входная последовательность предобрабатывалась — приводилась к начальной форме. Для этого были протестированы две библиотеки — Natasha и MyStem, наилучшие результаты показала вторая. Далее выполнялись два основных шага: создание массива кандидатов для связывания и нахождение наиболее подходящей сущности в полученном множестве кандидатов. Для генерации кандидатов использовалось построковое сравнение и расширение за счёт униграмм, биграмм и триграмм. Последний принёс значительный прирост (с 1.95 до 11.73) среднего количества кандидатов для сущностей и увеличение количества множеств кандидатов, которые содержат нужную сущность, но это привело к значительному снижению точности всей системы с 71 до 19%, так как среди большего количества кандидатов сложнее выбрать подходящий. Для второго шага, ранжирования, было протестировано три метода. Первый, выбор сущности, информация о которой наиболее полно представлена в базе знаний, справлялся с задачей неплохо, но не позволял учитывать контекст. Использование второго подхода, рассчёт векторного расстояния между контекстом упоминания и описанием сущности (как например, в статье [2]), привело к повышению точности (с 19 до 38%). Третий подход, взвешенные векторные расстояния, позволил добиться самого высокого значения точности (54%) для тех терминов, у которых есть связь с сущностями в графе знаний в размеченном корпусе среди всех экспериментов.

Таким образом, финальный набор шагов алгоритма: MyStem для предобработки входной последовательности, n-граммы для генерации кандидатов, взвешенные векторные расстояния для ранжирования. Проведенные эксперименты позволили добиться увеличения значений всех метрик, кроме точности, для всех упоминаний. Улучшить это планируется за счёт использования классификатора на основе сиамской сети для ранжирования и расширения списка кандидатов с помощью синонимов и аббревиатур.

Работа выполнена при финансовой поддержке РФФИ (грант № 19-07-01134).

Научный руководитель — к.ф.-м.н. Батура Т. В.

Список литературы

[1] Мезенцева А. А., Бручес Е. П, Батура Т. В. Автоматическое связывание терминов из научных текстов с сущностями базы знаний. // Вестник НГУ. Серия: Информационные технологии. 2021 Т. 19. № 2. С. 65–75.

[2] Winkler W. String comparator metrics and enhanced decision rules in the Fellegi — Sunter model of record linkage // Proceedings of the Section on Survey Research Methods. American Statistical Association. 2020. P. 354–359.

4.14. *Минуллин Д.А.* Сравнительный анализ методов машинного обучения в образовательной аналитике

Данная работа посвящена применению методов анализа больших данных для анализа информации об образовательном процессе в школе. Основу исследования составили данные, собранные через государственную информационную систему «Электронное образование в Республике Татарстан». Система включает в себя базы данных образовательной информации по всем учащимся и всем педагогам общеобразовательных организаций РТ. База данных содержит более двух миллиардов информационных единиц, включая информацию об успеваемости более миллиона учащихся и профессиональной деятельности более 120 000 преподавателей.

Образование является сферой, в которой производится и накапливается большое количество данных. Правильный анализ такой информации может помочь составить более полную картину процесса обучения, выявить полезные и, возможно, неочевидные связи. Методы машинного обучения могут позволить предсказать исход какой-либо ситуации, основываясь на исторических данных. В отличие от традиционных мер измерения результатов учащихся, которые помогают измерять только конечный результат, применение методов машинного обу-

чения может помочь получить ценную информацию о том, как улучшить и персонализировать обучение, составлять прогнозы и рекомендации, проводить изменения в режиме реального времени [1].

Задачей данного исследования являлся анализ процесса перехода учащегося из 9-го класса в 10-й. Для первичной обработки данных с целью группировки учащихся и расчёта средних оценок за четверть был разработан программный комплекс на языке программирования Python с использованием библиотеки для параллельных и распределённых вычислений Dask [2]. Для прогнозирования перехода учащегося использовались методы машинного обучения (Таблица 1) [3], которым на вход подавались оценки ученика и на выход параметр отвечающий за переход ученика в следующий класс.

Метод	**Точность**, %
LogisticRegression	70.20
LinearDiscriminantAnalysis	70.45
KNeighborsClassifier	71.12
GaussianNB	70.47
DecisionTreeClassifier	70.12
RandomForestClassifier	70.79
SupportVector(linear)	70.12
SupportVector(rbf)	71.08
Simple neural network	70.14

Таблица 1. Точность прогнозирования.

В результате проведённого анализа видно, что процент точности прогнозирования невысок, все методы показали приблизительно одинаковый результат (в районе 70%), но даже основываясь на таком показателе удалось выделить некоторые взаимосвязи. Из полученных результатов видно, что есть возможность прогнозирования, но одних только оценок для точного ответа недостаточно. В дальнейшем данное исследование будет продолжаться с расширением количества параметров, характеризующих образовательный процесс.

Работа выполнена при финансовой поддержке РФФИ (грант № 19-29-14082).

Научный руководитель — к.ф.-м.н. Гафаров Ф. М.

Список литературы

[1] Romero C., Ventura S. Educational data mining: A review of the state of the art // IEEE Transactions on Systems Man and Cybernetics. Part C (Applications and Reviews). 2010. Vol. 40. N. 6. P. 601–618.

[2] Rocklin M. Dask: Parallel computation with blocked algorithms and task scheduling // Proc. 14th Python In Science Conf. (SCIPY 2015). P. 126–132.

[3] Sarker Iqbal H. Machine learning: Algorithms, real-world applications and research directions // SN Computer Science. 2021. Vol. 2. N. 3. P. 160.

4.15. *Ондар С.К.* Моделирование геодинамических данных о сейсмическом режиме сильных землетрясений на территории Республики Тыва

Представлено информационное и алгоритмическое обеспечение для решения основных задач геомониторинга и оценки геодинамической опасности территории Республики Тыва. На основе данных сейсмического мониторинга в территории республики Тыва выполнена разработка методики анализа данных комплексного геомониторинга геодинамических полей для оценки напряженно-деформированного состояния (НДС) геосреды и повышения точности прогноза сильных землетрясений.

Геодинамический мониторинг является обязательным элементом государственной системы обеспечения геодинамической безопасности в сейсмически активных регионах России. Начиная с 2000 г. получили развитие и региональные наблюдательные геодинамические сети в различных субъектах федерации (Красноярский край, Кемеровская область, республика Тыва и др.). При этом используются как сейсмологические, так и комплексные сети, регистрирующие различные геолого-геофизические поля и их параметры. Вместе с тем, несмотря на длительное использование комплекса геолого-геофизических методов, применяемых при геодинамическом мониторинге, нормативно-методическая основа упомянутого комплекса не разработана.

В настоящей работе анализируются данные комплекса геолого-геофизических методов (такие, как сейсмология, естественное импульсное электромагнитное поле Земли (ЕИЭМПЗ) и эмиссия радона на геодинамических полигонах в Сибири), пригодные для оценки изменения НДС геологической среды и прогноза сильных сейсмических событий в республике Тыва.

Геодинамический мониторинг комплексом геолого-геофизических методов (сейсмология, ЕИЭМПЗ, радон) обеспечивает не только оценку изменения НДС геологической среды, но также среднесрочный (1–3 месяца) и краткосрочный (1–10 суток) прогноз сильных землетрясений с $M \geq 5.0$. В тоже время уровень комплексирования (низкая плотность сетей регистрации ЕИЭМПЗ и измерения уровня концентрации радона в подземных водах) не обеспечивает в регионе надёжный прогноз положения эпицентра. Для повышения надёжности определения положения эпицентров землетрясений, необходимо увеличить плотность сетей регистрации ЕИЭМПЗ и радона, а также дополнить применяемый геолого-геофизический комплекс данными спутниковых съёмок (инфракрасной и геохимической (CO_2, метан)). В связи с распространением в геологической среде, наря-