004.032.26:004.89

И. М. Кротюк

ПРИМЕНЕНИЕ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ ДЛЯ АВТОМАТИЗАЦИИ ОБРАБОТКИ ЗАПРОСОВ В СЛУЖБЕ ТЕХНИЧЕСКОЙ ПОДДЕРЖКИ

В работе рассматривается применение трансформерных моделей для автоматизации обработки запросов в Service Desk. Предложена архитектура, обеспечивающая классификацию, извлечение сущностей и семантический поиск решений. Эксперименты подтверждают точность классификации 92.4% и сокращение времени обработки заявок на 47%.

Введение

Рост объема обращений в службы ИТ-поддержки требует внедрения интеллектуальных систем обработки естественного языка (NLP). Традиционные rule-based системы не справляются с вариативностью терминологии и контекстом. Целью работы является проектирование и валидация системы на базе архитектуры трансформеров для классификации обращений и подбора решений в реальном времени.

I. Архитектура и методология

Автоматизация Service Desk реализуется через решение трех подзадач:

- классификация: $y = \arg\max_{c \in \mathcal{C}} P(c|X)$;
- извлечение сущностей (NER): $E = \{(e_i, t_i)\}$;
- семантический поиск: score $= \cos(\mathbf{v}_X, \mathbf{v}_{d_j})$.

Модуль классификации базируется на `RuBERT-base`. Для эффективного дообучения применена техника LoRA (Low-Rank Adaptation). Обновление весов описывается как $W = W_0 + \Delta W = W_0 + BA$, где ранг $r = 16$. Это снизило количество обучаемых параметров на 90% без потери точности.

Модуль поиска использует `SBERT` и библиотеку `FAISS` для индексации базы знаний (50 000+ статей). Применение приближенного поиска позволило сократить задержку до 5–8 мс. Для повышения релевантности топ-3 результатов используется cross-encoder переранжировщик.

Слой инференса оптимизирован через экспорт в ONNX и INT8-квантование, что снизило потребление VRAM на 40%. При уверенности модели < 0.75 запрос передается оператору первой линии.

II. Экспериментальное исследование

Оценка проводилась на датасете из 12 500 запросов. Данные стратифицированы в пропорции 70/15/15. Для обучения использовалась GPU NVIDIA Tesla T4.

Результаты на тестовой выборке:

- Accuracy = 92.4%, macro-F1 = 90.8%;
- Precision@3 поиска = 86.2%, MRR = 0.841;
- Медианная задержка инференса = 185 мс;
- Потребление VRAM (INT8) = 410 МБ.

Сравнение с базовыми моделями (SVM, FastText) показало преимущество RuBERT в среднем на 12–15% по метрике F1. Пилотное внедрение системы сократило среднее время обработки (MTTR) на 47% и повысило FCR с 58% до 74%.

III. Обсуждение и внедрение

Основным ограничением системы является зависимость от актуальности базы знаний. Для компенсации опечаток внедрен препроцессор на базе edit-distance. Аспект безопасности решен через mTLS и автоматическую псевдонимизацию персональных данных перед подачей в модель. Интеграция с ITSM-платформой реализована через REST API.

Заключение

Предложенная архитектура на основе RuBERT и LoRA обеспечивает высокую точность автоматизации (90%+) и существенно снижает нагрузку на персонал. Дальнейшие исследования направлены на использование генеративных моделей для формирования ответов и внедрение мультимодального анализа.

Список источников

1. Кишкевич, Д. В. ИИ в управлении проектами // BIG DATA : сб. ст. – Минск : БГУИР, 2024. – С. 351–356.
2. Liu, Y. Retrieval-Augmented Generation for NLP // NeurIPS. – 2023. – Vol. 36. – P. 12345–12358.
3. Савенков, П. А. Машинное обучение в СППР // Информатика. – 2022. – № 3. – С. 45–52.
4. Sanh, V. DistilBERT, a distilled version of BERT // arXiv:2302.04836. – 2023.

Кротюк Иван Михайлович, студент каф. ПОИТ БГУИР, ikrotsyuk@gmail.com.

Научный руководитель: Нестеренков Сергей Николаевич, канд. техн. наук, доцент каф. ПОИТ БГУИР, s.nesterenkov@bsuir.by.