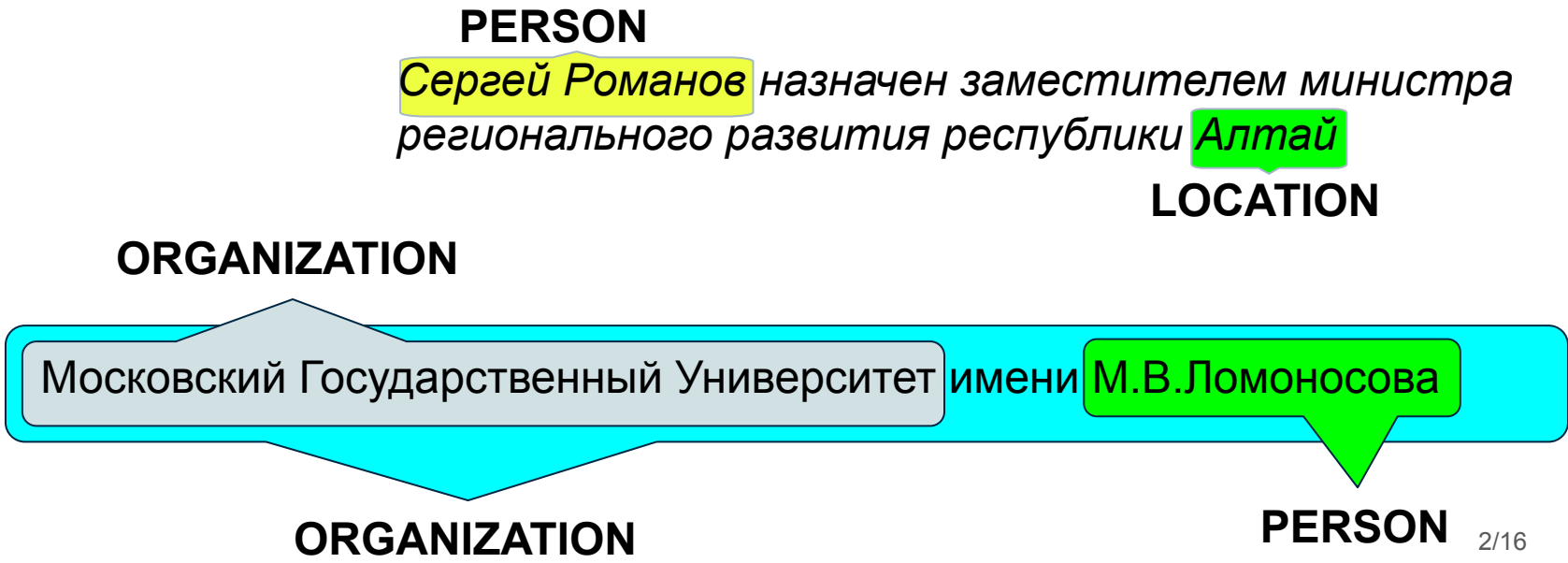
ИЗВЛЕЧЕНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ ПРЕДМЕТНОЙ ОБЛАСТИ ИЗ НЕОДНОРОДНОГО ВХОДНОГО ПОТОКА

Н. В. Лукашевич, И. С. Рожков, Б. В. Добров

АЯ ВМК МГУ, НИВЦ МГУ,
Москва

Извлечение вложенных именованных сущностей

Именованная сущность — (обычно) объект реального мира, такой как люди, местоположения, организации, продукты и т. д., который может быть обозначен неким (собственным) именем.



Предметные области vs Поток общего назначения

- Хорошо исследована для новостного потока общего назначения
- Для конкретных предметных областей размеченных данных существенно меньше
- Примеры:
  - медицина (много данных)
  - IT-область (нехватка датасетов)

Актуальность задачи

- При обучении модели извлечения именованных сущностей для предметной области:
  - Собирается совокупность специализированных текстов
  - Проводится категоризация типов значимых имен
  - Обучаются специализированные модели
- Проблема: реальный входной поток часто более разнообразен
- Тексты выходят за рамки предметной области
- Качество извлечения имен резко деградирует

Проблема Data Shift

Data Shift (или Data Drift) — изменение вероятностных распределений данных в процессе работы

1. Классификатор обучен на узкой предметной области
2. На вход подаются тексты более широкой тематики
3. Модель должна извлекать целевые сущности и игнорировать остальные

Проблема недостаточно изучена в литературе

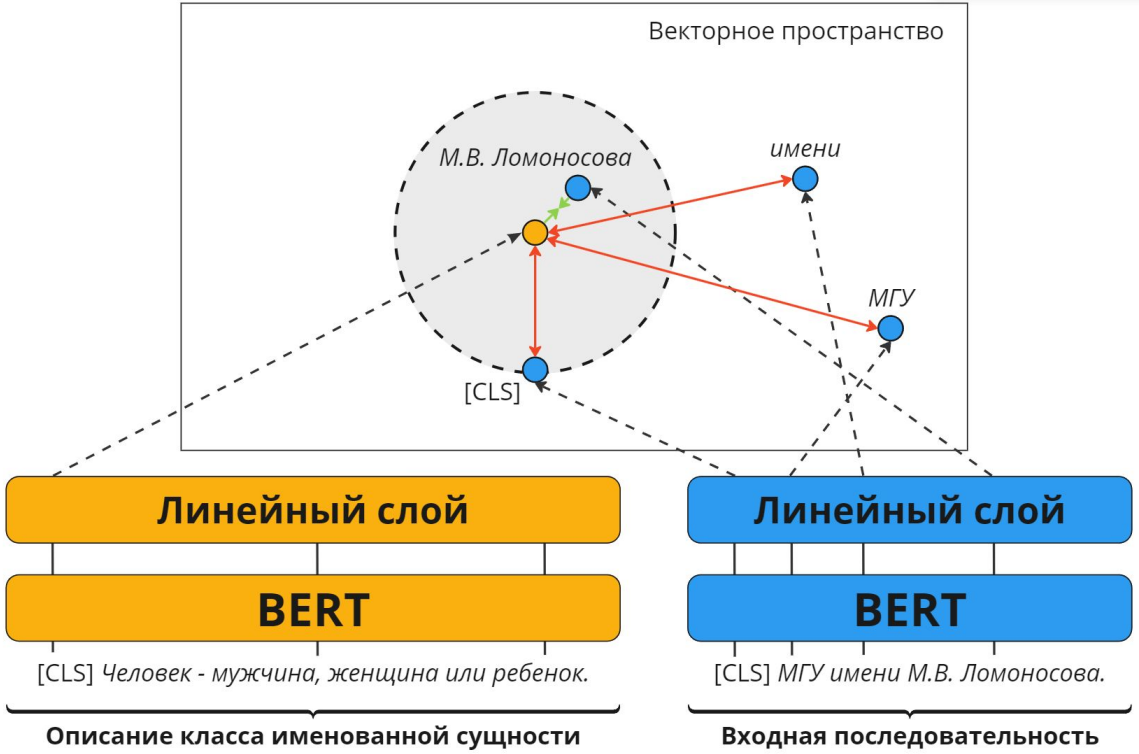
Цели и задачи

Цель: анализ работы классификатора именованных сущностей, обученного для компьютерных текстов, на неоднородном входном потоке

- Применить IT-классификатор к общим новостным текстам (датасет NEREL)
- Проанализировать типичные ошибки классификации
- Предложить метод автоматизированного порождения датасета для обучения
- Оценить эффективность предложенного метода

Модель **Binder**

- Единое векторное пространство для сущностей и других последовательностей слов
- Описание класса сущности как идеальный пример сущности в пространстве – «центр»
- Контрастивное обучение: сущности нужного класса – ближе к «центру», всё остальное – дальше
- Предсказание по обучаемому порогу-радиусу от центра



Zhang S. et al. Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning //The Eleventh International Conference on Learning Representations. – 2022.

IT-классификатор и датасет NEREL

- IT-классификатор: обучен для извлечения именованных сущностей в компьютерной области
- Типы IT-сущностей: DEVICE, FILE, ATTACK, HACKER и др.
- Датасет NEREL:
  - Русскоязычные новостные тексты произвольной тематики
  - 29 типов именованных сущностей (без компьютерных)
  - Небольшая доля текстов с IT-сущностями
- Задача: извлекать IT-сущности, игнорировать все другие типы

Проблема применения IT-классификатора

- IT-классификатор находит «похожие» на компьютерные сущности
- Типичные ошибки:

  Преступления (CRIME) → компьютерные атаки (ATTACK)
  Преступники (PERSON) → хакеры (HACKER)
  Произведения искусства (WORK_OF_ART) → файлы (FILE)
  Здания (FACILITY) / продукты (PRODUCT) → устройства (DEVICE)

- При этом веб-сайты и программы извлекаются правильно
- Качество извлечения резко деградирует

Примеры неправильной классификации

Класс NEREL	Класс IT	Примеры
CRIME	ATTACK	злоупотребления служебными такси, убийство
EVENT		скандал, инцидент, смерть, пикет, ДТП
DISEASE		африканская чума свиней, свиной грипп
FACILITY	DEVICE	Саяно-Шушенская ГЭС, АЭС «Фукусима-1»
PRODUCT		Toyota RAV4, Cadillac, Ан-72
WORK_OF_ART	FILE	Рисунки Бродского, кантаты Баха
PERSON	HACKER	Степан Бандера, Дмитрий Фирташ
IDEOLOGY		сепаратист, радикал, нацисты

Метрики до улучшения

Применение IT-классификатора для извлечения компьютерных сущностей на NEREL

IT сущности	Кол-во	FP	FN	Точность, %	Полнота, %	F1, %
HACKER	4	256	0	1.54	100	3.03
FILE	5	24	0	17.24	100	29.41
ATTACK	4	224	0	1.75	100	3.45
DEVICE	7	101	5	6.48	58.33	11.67

- Очень низкая точность
- Высокое число ложноположительных (FP)
- Много лишних сущностей размечается

Предложенный метод — NEREL-IT

Цель: снизить предсказание лишних сущностей при сохранении качества на целевом датасете

Метод автоматизированного создания датасета NEREL-IT:

1. Идентификация релевантных IT-документов в NEREL
2. Для остальных документов — *фильтрация*:
   — Удаление автоматических разметок FILE, ATTACK, DEVICE
   — Замена типа HACKER на PERSON
3. Создание образца разметки IT-классификатора на новостных текстах

Тексты NEREL-IT подмешиваются к обучающему множеству IT-классификатора

Результаты

До

IT сущности	Кол-во	FP	FN	Точность, %	Полнота, %	F1, %
HACKER	4	256	0	1.54	100	3.03
FILE	5	24	0	17.24	100	29.41
ATTACK	4	224	0	1.75	100	3.45
DEVICE	7	101	5	6.48	58.33	11.67

Обучение на смешанном датасете: IT + 200 текстов NEREL-IT

NEREL-IT

IT сущности	Кол-во	FP	FN	Точность, %	Полнота, %	F1, %
HACKER	4	15	4	0	0	0
FILE	5	5	0	100	50	66.67
ATTACK	4	9	3	25	10	14.29
DEVICE	7	8	8	33.33	33.33	33.33

- Точность существенно повышается (FP значительно снижается)

Анализ результатов

- Точность предсказания существенно повышается
- Резко снижается предсказание некорректных сущностей
- Проблема с типом HACKER:
  - Модели трудно отличить хакеров от других преступников
  - FN = 4 (все хакеры пропущены в этом эксперименте)
- Эксперименты показали неустойчивость процедуры обучения
- Различные типы и размеры подмешивания дают нестабильные результаты
- Требуются дальнейшие исследования

Заключение

- Рассмотрена проблема извлечения именованных сущностей в условиях неоднородного входного потока
- Данная проблема — подвид известной проблемы Data Shift
  - В задачах извлечения информации ей уделялось мало внимания
- Предложен метод автоматизированного порождения датасета NEREL-IT
- Показано улучшение точности при подмешивании к обучающей выборке
- Направления будущей работы:
  - Исследование устойчивости процедуры обучения
  - Оптимизация размеров подмешивания
  - Применение к другим предметным областям

Спасибо за внимание!