



**ВЯТСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ**

Разработка системы анализа выпускных квалификационных работ

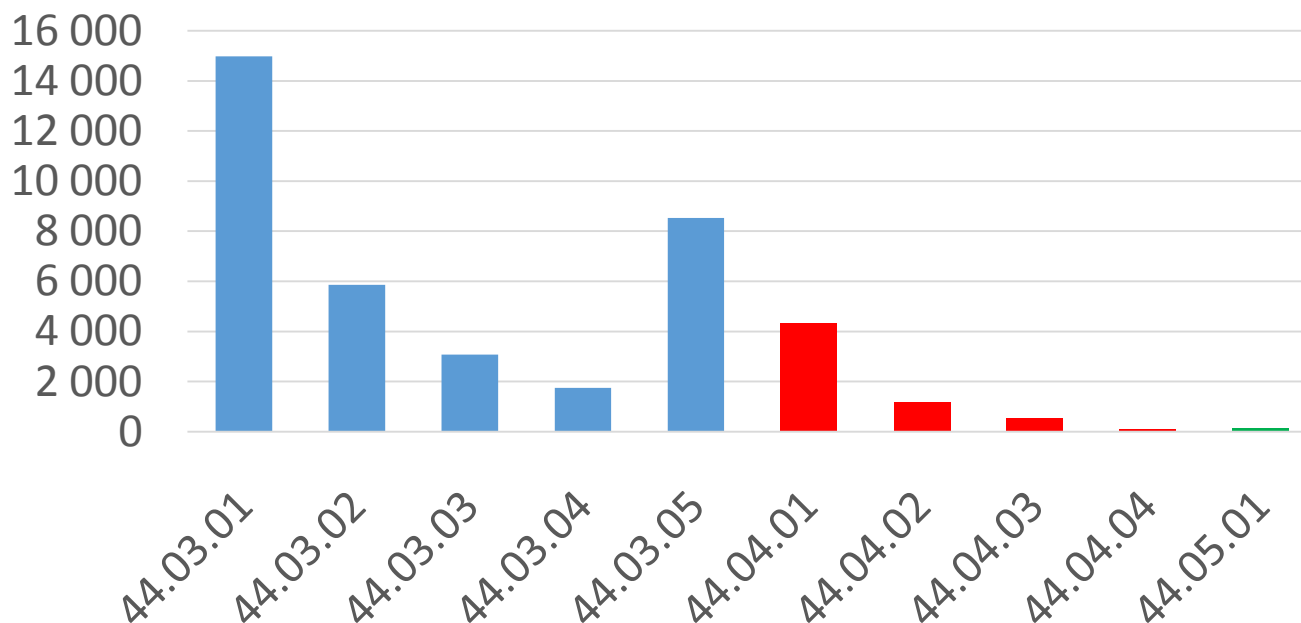
к.т.н. Татаринова Александра
каф. ПМИ

Актуальность

- Оценивание качества ВКР является трудной задачей:
Членам ГЭК необходимо в течение ограниченного времени оценить проделанную в ВКР работу на основании доклада студента и достаточно объемного текста
- Например: анализ 60 случайно выбранных ВКР из 11 вузов по направлениям подготовки УГСН 44 «Образование и педагогические науки», проведённый экспертами ВятГУ и НГПУ им. К.Минина, показал, что наличие более полной информации по тексту ВКР способствует объективной оценке

Сбор ВКР

Количество ВКР



Около 40 тыс. ВКР выпускников 2017 по УГСН 44
Форматы файлов: **PDF, DOC, DOCX, RTF, ODT**

Основные модули системы

Работа велась совместно специалистами ВятГУ и
НИВЦ МГУ имени М.В.Ломоносова

**Поиск
заимствований**

**Проверка
тематической
связности**

**Проверка
используемой
терминологии**

**Анализ
новизны и
актуальности**

**Проверка
орфографии**

**Проверка
грамматической
связности**

**Наличие
обценной
лексики**

**Контроль
научного стиля**

Морфологический анализ

Конвертирование в SimpleHTML-файлы

Обнаружение и замена некириллических букв в русских словах

Обнаружение и удаление способов обхода антиплагиата

Предобработка

- Этапы предобработки направлены на то, чтобы максимально извлечь «реальный» текст ВКР из файлов
- **Пример #1** обхода антиплагиата

Текст в PDF-файле (как видит читатель):

Образы воображения – это образы предметов и явлений, которых мы раньше не воспринимали.

Выявление обхода антиплагиата:

Образы воображения – это образы **плане** предметов и явлений, **я** которых **вма епрльнизв** мы раньше не воспринимали.

Предобработка

- **Пример #1** обхода антиплагиата

Отчет Антиплагиат.ВУЗ для исходного PDF-файла:
83% уникальности (предупреждений НЕ выдается)

Отчет Антиплагиат.ВУЗ для преобразованного нами
HTML-файла: **23%** уникальности

* данные на 2018 год

Предобработка

- **Пример #2** обхода антиплагиата

Текст в PDF-файле (как видит читатель):

«Таким образом, мы выяснили, что важную роль в системе коррекционного обучения играет изучение семантической стороны речи у детей с умственной отсталостью»

Выявление обхода антиплагиата:

«Тѡакоимъ ообрѣзѡмъ, мы въяснѡйшѡю,
чтѡ ѡбѡжѡнуѡю рѡлѡю в сѡсѡтѡемѡе
кѡррѡекцѡй ѡнѡѡгѡ обѡучѡенѡя
и гѡрѡает изѡучѡенѡе
сѡмѡантѡичѡескѡй сѡбрѡнѡы рѡчѡю у
дѡетѡей с умѡсѡвѡенѡѡю
ѡтѡсѡтѡалѡѡсѡтъѡю».

Предобработка

- Обход антиплагиата на основе вставки невидимых элементов:
 - Встретилось в 2 623 ВКР (6,5%)
 - Встретилось в 72% вузов
 - Максимальное количество вставок – 115 122

Грамматическая связность

Проверка согласованности слов в предложениях

Примеры несогласованности:

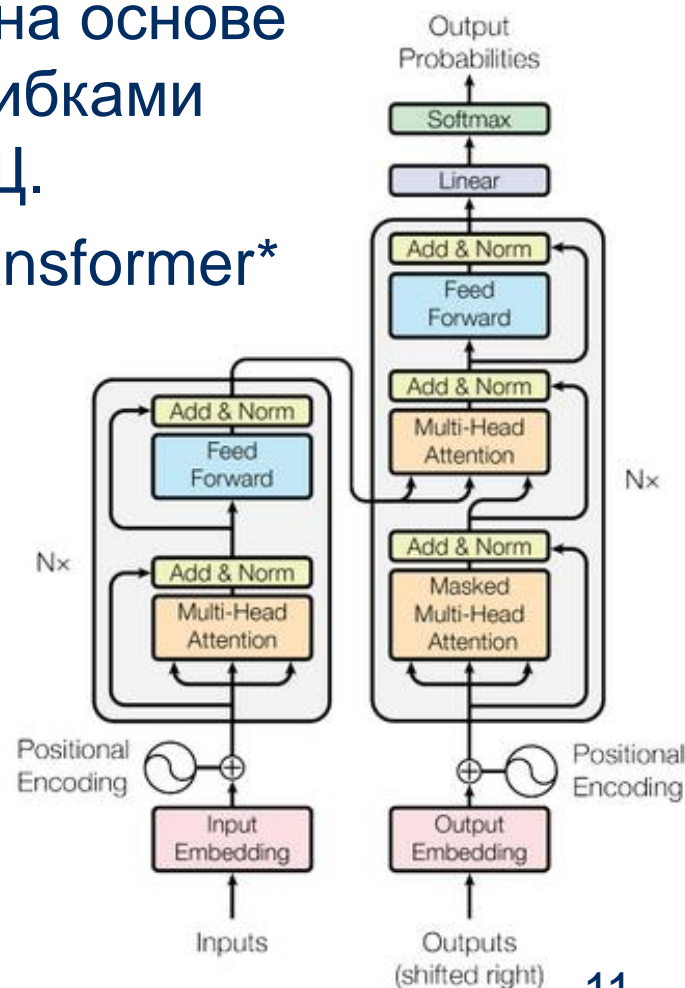
- Непосредственно согласно себя **данный процедура** способен прерваться только лишь в таком случае, если все без исключения станет истреблено.
- **Общество заражаются** и погибают с заболеваний, каковых в обычных обстоятельствах, присутствие здоровом роли существования, постоянно функционирующей медицине и твердо регулируемых автосанитарных нормах, возможно исключить

Грамматическая связность

- Алгоритм выделения словосочетаний на основе шаблонов:
 - «Подлежащее + глагол (сказуемое)»
 - «Прилагательное в полной форме + существительное»
 - «Причастие в полной форме + существительное»
 - «Глагол и предлог»
- Синтаксический анализ предложений
- Нейронная сеть архитектуры Transformer

Грамматическая связность

- Генерация обучающего корпуса на основе «зашумления» предложений ошибками несогласованности ПРИЛ. + СУЩ.
- Нейронная сеть архитектуры Transformer* (механизм self-attention)
- Tensorflow 2.0
- В планах:
 - Transfer Learning на основе BERT
 - Расширить до области GEC (Grammatical Error Correction)



* <https://arxiv.org/abs/1706.03762>

Грамматическая связность

Зашумленное предложение:

Федеральный архив Германии предоставил в распоряжение **электронного** библиотеки Wikipedia порядка 100 тысяч исторических фотографий.

Исправленное предложение:

Федеральный архив Германии предоставил в распоряжение **электронной** библиотеки Wikipedia порядка 100 тысяч исторических фотографий.

Поиск заимствований

- Формальное и смысловое сегментирование ВКР:
 - представление предложений на основе word2vec
 - вычисление мер связности и разрыва между соседними предложениями
- Определение связности фрагментов с темой ВКР
- Использование лингвистической онтологии
- Поиск по внутренним базам данных (сформированных на основе Wikipedia и РИНЦ)
- Поиск по внешним базам данных на основе публичных поисковых машин
- Анализ общих слов в анализируемом тексте и сравниваемом тексте

Выводы

- Анализ коллекции ВКР 2017 года по УГСН 44 «Образование и педагогические науки» показал наличие несамостоятельности выполнения ВКР
- Присутствует маскирование заимствований
- А также частое использование заимствований без указания источников
- Необходимы вспомогательные технические средства лингвистической обработки текстов для оценивания работ студентов

Спасибо за внимание!