Упрощение текстов

Студент: Зайцева А. А., ИУ7-52Б

Руководитель: Кивва К. А.

Москва, 2021 г.

Цели и задачи

Цель: выбрать метод, который наиболее полно решает задачу упрощения текстов

Задачи:

- провести анализ предметной области;
- рассмотреть существующие метрики оценки качества упрощения;
- провести анализ существующих решений задачи упрощения текстов;
- сформулировать критерии выбора решения;
- на основе этих критериев провести классификацию решений;
- определить, какой метод или методы являются лучшими по совокупности критериев.

Термины предметной области

- Упрощение предложений — процесс, целью которого является получение более легкого для чтения и понимания текста за счет уменьшения его лексической и структурной сложности.
  - Лексическая сложность – сложность текста с точки зрения используемых в нем слов (их длина или частотность употребления).
  - Структурная сложность – сложность текста с точки зрения сложности грамматики его предложений (количество простых предложений в составе сложного, наличие сравнительных, причастных и деепричастных оборотов и т. д.).

Метрики

- Индексы удобочитаемости
- SARI (System output Against References and against the Input sentence)
- SAMSA (Simplification Automatic evaluation Measure through Semantic Annotation)

Критерии

- Понимание задачи: в узком или широком смысле
- Учет двух составляющих «простоты»:
  - лексической
  - структурной

Классификация решений

Подход	Класс	Решение	Понимание задачи	Учет лексического упрощения	Учет структурного упрощения
Экстрактивный	-	-	Шир.	-	-
Абстрактный	Текстовые замены	-	Узк.	Да	Нет
	Генерация нового текста	Синтаксическое упрощение	Узк.	Частично	Да
		Статистический машинный перевод	Узк.	Да	Нет
		Глубокое обучение	Узк.	Да	Да

Дополнительные показатели для отбора данных

- Лексическая сложность
- Глубина дерева зависимостей
- Длина предложений
- Легкость чтения
- Косинусное сходство
- Сохранения именованных сущностей
- ROUGE-L

Выводы

- Проведен анализ предметной области.
- Рассмотрены существующие метрики оценки качества упрощения.
- Сформулированы критерии выбора решения задачи упрощения текстов.
- Проведены анализ и классификация существующих решений
- Выбран наиболее подходящий метод.