Упрощение текстов

Студент: Зайцева А. А., ИУ7-52Б

Руководитель: Кивва К. А.

Москва, 2022 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.).

Критерии

- Понимание задачи:
  - в узком смысле (как процесс, целью которого является получение более легкого для чтения и понимания текста за счет уменьшения его лексической и структурной сложности);
  - широком смысле (как процесс, охватывающий и другие операции: смысловое изменение для упрощения как формы, так и содержания; краткое изложение текста для исключения второстепенной или избыточной информации).
- Учет двух составляющих «простоты»:
  - лексической -
    - «сложное» предложение: «Около 5 способов предварительно апробированы.»;
    - упрощенное предложение: «Около 5 способов заранее проверены.»;
  - структурной -
    - «сложное» предложение: «Листок, сорванный Петей, упал на землю.»;
    - упрощенное предложение: «Петя сорвал листок. Листок упал на землю.».

Метрики

Название метрики	**Достоинства**	**Недостатки**
Индексы удобо-читаемости	• простоты в вычислении;	• не распознают грамматически неправильные результаты; • не следят за сохранением смысла исходного предложения.
SARI	• высоко коррелирует с оценками качества упрощения человеком; • напрямую оценивает корректность выбора слов для добавления, удаления и сохранения;	• слабо способна оценивать структурную сложность; • требует нескольких «справочных» примеров упрощения.
SAMSA	• высоко коррелирует с оценками качества упрощения человеком; • способна оценивать структурную сложность;	• требует нескольких «справочных» примеров упрощения; • мало изучена.

Подходы к решению задачи

Экстрактивный

Выделение в документе тех предложений, которые передают больше информации.

Абстрактный

Переписывание передаваемого текста предложение за предложением.

Текстовые замены

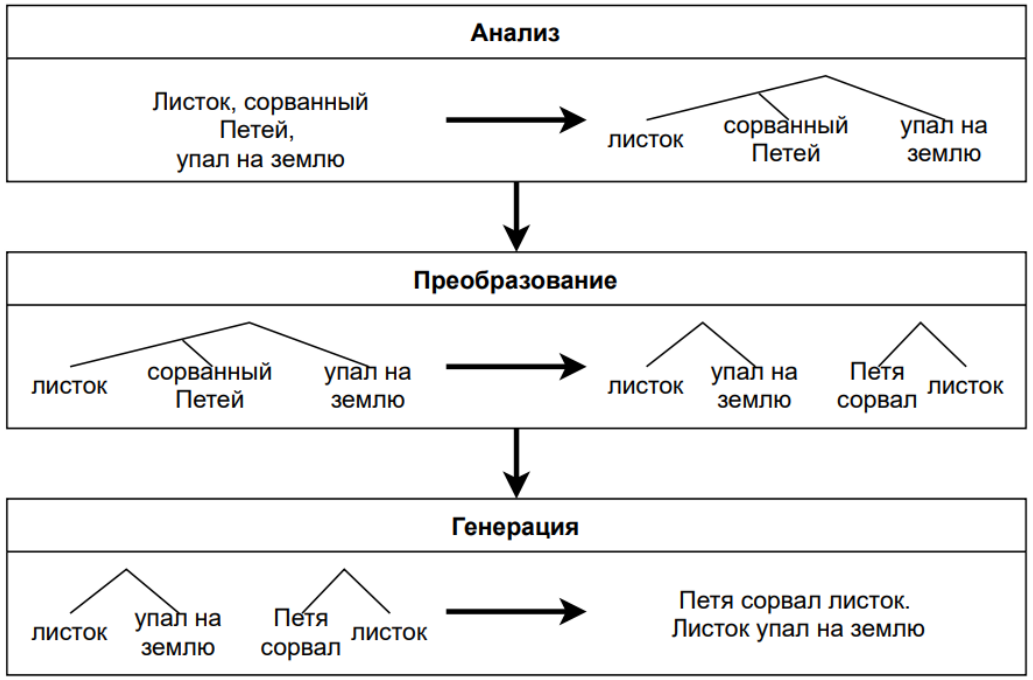
Фокусировка на сокращении лексического содержания текста.

Генерация нового текста

Разбиение сложных предложений на более простые, удаление редко употребляемых слов и генерация на этой основе нового текста.

Генерация нового текста

- Синтаксическое упрощение - выявление грамматически сложных частей текста и их переписывание для облегчения понимания.
- Статистический машинный перевод - сведение задачи упрощения к переводу с исходного «сложного» языка на целевой «простой» язык.
- Использование методов глубокого обучения – формулировка задачи упрощения в терминах моделирования seq2seq.



Синтаксическое упрощение

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: лексическая сложность, глубина дерева зависимостей, длина предложений, легкость чтения, косинусное сходство, сохранение именованных сущностей, ROUGE-L

Заключение

- 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.

Выбор из различных методов глубокого обучения слабо влияет на результат упрощения. Однако значительное влияние на результат оказывает использование дополнительных показателей для фильтрации обучающих данных или для выбора наиболее подходящего упрощения из созданных. Их влияние на результат методов глубокого обучения требует дальнейшего исследования.