Введение

Упрощение текста, в узком смысле, – это процесс понижения лингвистической сложности текста при сохранении исходной информации и смысла. В более широком смысле это понятие охватывает и другие операции: концептуальное упрощение содержания и формы текста; уточнение для подчеркивания ключевых моментов; обобщение для исключения второстепенной или не относящейся к теме информации [1].

В последние десятилетия количество неструктурированных текстовых данных резко возросло в связи с появлением Интернета. Как следствие, возросла и потребность в их упрощении, что показано на рисунке 1.

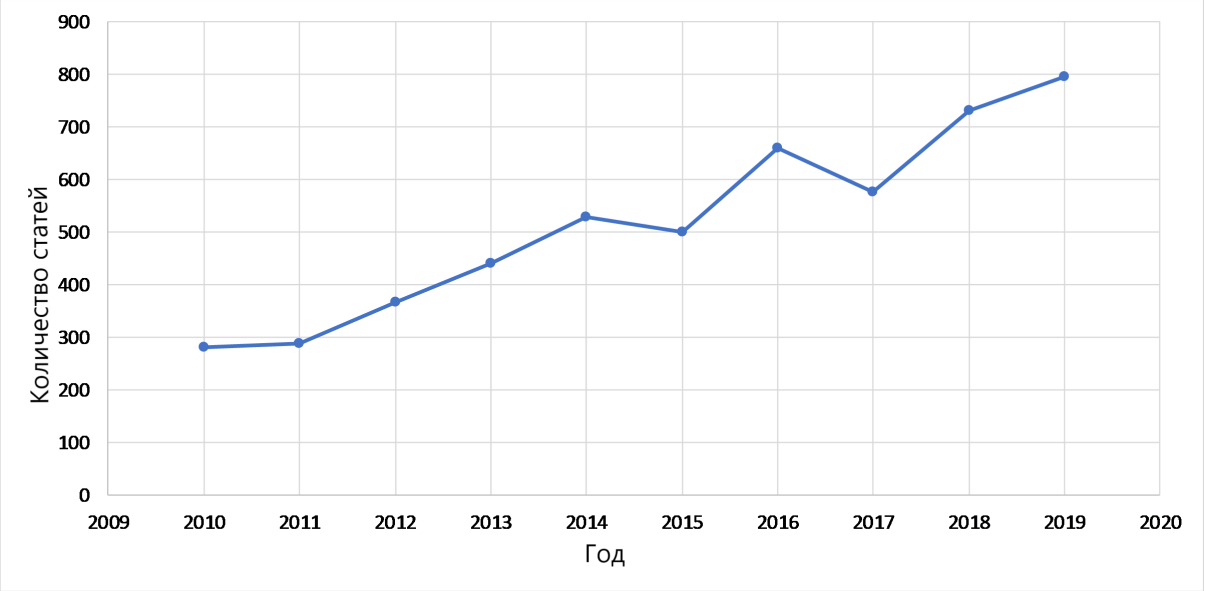

Рисунок 1 – График роста интереса к теме упрощения текста. Создано на основе статистики Google Scholar по поисковым запросам "Text Simplification"("Упрощениие текстов"), "Lexical Simplification"("Лексическое упрощение"), "Syntactic Simplification"("Синтаксическое упрощение")

Упрощение текстов необходимо для различных задач и целевых аудиторий:

- в качестве этапа подготовки текста перед его обобщением [2];
- для людей, изучающих иностранный язык, и детей, учащихся читать (требуется лексическое упрощение для сокращения количества специализированных и нечастотных слов) [3];

- для людей с дислексией и афазией, для которых особенно длинные слова и предложения могут представлять трудности;
- для людей, страдающих аутизмом (необходимо уменьшать количество образных выражений и синтаксическую сложность) [4].

Упрощение предложений привлекло многих исследователей в связи с появлением больших параллельных корпусов. Наиболее известные из них - PWKP и Wiki-large. Последний представляет собой большой параллельный корпус на английском языке, состоящий из сложных предложений, взятых из Википедии, и их выровненных упрощенных версий.

Одна из причин, по которой упрощение текстов еще не начали активно применять в приложениях на русском языке, заключается в том, что наиболее значительные общедоступные наборы данных принадлежат английскому домену, тогда как качество моделей сильно зависит от объема и количества тренировочных данных. Более того, во многих решениях имеет место фокусировка на текстах Википедии, что ограничивает исследования и приводит к неадекватности моделей на других типах данных [5].

Другая тема для обсуждения – выбор уровня, на котором текст будет упрощаться. В целом, упрощение текста подразумевает изменения на разных уровнях, включая стилистический, грамматический и лексический уровни. На данный момент одним из наиболее популярных подходов является упрощение текста «предложение за предложением» [6]. (!!Этот подход к какому из названных Вами уровней относится?) В нем в качестве подзадачи применяются перефразирование и генерация текста на естественном языке.

• объект исследования или разработки, цель работы с ее разделением на взаимосвязанный комплекс задач, подлежащих решению

Объектом исследования данной работы является общая задача упрощения текстов на русском языке на уровне предложений. Цель – разработка качественной модели, решающей эту проблему.

В рамках выполнения работы необходимо решить следующие задачи:

1) собрать или найти крупный параллельный корпус на русском языке,

состоящий из сложных предложений и их выровненных упрощенных версий и основанный на разных типах источников;

2) подобрать метрики, по которым можно определить качество упрощения;

3) изучить существующие подходы к решению задачи упрощения текстов на уровне предложений как на русском, так и на иностранных языках;

4) разработать или адаптировать к русскому языку несколько моделей, решающих эту задачу;

5) провести сравнительный анализ разработанных моделей, обученных как на всем собранном корпусе, так и на его отдельных частях (основанных на разных типах источников);

6) выбрать наилучшую модель.

Упрощение предложений можно рассматривать как задачу «от последовательности к последовательности» (sequence-to-sequence, seq2seq) [5], к которой применимы нейронные модели языка. Такая модель получает на вход исходное предложение, а на выходе выдает его упрощенную версию.

(!В статье "A Survey on Text Simplification"выделяется множество подходов. Почему Вы вынесли сюда именно его? Детально анализировать все подходы нужно будет, конечно, не во Введении, а в Аналитическом разделе ВКР (ну или, я так понимаю, в Вашем случае - в самой НИР), но во Введении нужно хотя бы рассказать, как вообще современное научное сообщество понимает "упрощение"текста и кратко, обобщённо пересказать общую суть разных групп подходов к этой задаче.)

Поняте «простота тектса» имеет субъективную природу (особенно при упрощении предложений с помощью моделей seq2seq), поэтому наилучшим образом оценить качество результата может только человек, проанализировав полученное упрощенное предложение с точки зрения грамматики, сохранения смысла и простоты. Однако такая оценка слишком долгая и трудоемкая, поэтому был разработан ряд специальных метрик, из которых SARI на данный момент признана наиболее удачной [7].

(!Опять же, я думаю, стоит обобщить существующие метрики, разделить их на какие-то категории, если у них есть что-то общее, и очень-очень кратко перечислить, на чём в принципе основывается та или иная группа метрик, с указанием примеров конкретных метрик.)

• краткий обзор базы исследования и литературных источников

Задача упрощения текстов на русском языке была предложена в рамках международной конференции по компьютерной лингвистике и интеллектуальным технологиям DIALOGUE 2021 [1]. Организаторы подготовили обучающие и тестовые наборы с использованием краудсорсинговой платформы, а также перевели тексты из Википедии (корпус RuWikiSimple). Другим источником данных могут стать результаты перевода в виде перефразирования [5].

(!Впервые? Или почему Вы именно эту конференцию во Введение вынесли?)

В последнее время достигнуты хорошие результаты в решении проблемы упрощения предложений. Модель-трансформер mBART, первоначально разработанная для машинного перевода, оказалась эффективной для решения задач такого рода, в особенности при добавлении специальных маркеров управления [8].

В основе большинства современных моделей лежит подход encoder-decoder (кодировщика-декодера), зачастую дополненный и другими инструментами. Например, модель DRESS [9] добавляет в архитектуру encoder-decoder подход обучения с подкреплением: выполняется обзор возможных упрощений и выбирается наилучший из них.

Решение, победившее в соревновании, организованном в рамках конференции DIALOGUE 2021, в значительной степени основано на системе MUSS (Multilingual Unsupervised Sentence Simplification, система мультиязычного упрощения, обучающаяся без учителя) [8]. Модель состоит из mBART, настроенного на корпусах ParaPhraserPlus[2] (пары предложений на русском языке, разделенные на 3 класса по степени "перефразированности") и RuWikiSimple, в которую добалены управляющие токены (рас-

[1] http://www.dialog-21.ru/dialogue2021/results/

[2] https://metatext.io/datasets/paraphraser-plus

стояние Левенштейна, доля совпадающих символов между оригинальным и упрощенным предложениями, сходство лексем).

Модели, занявшие следующие места - генеративные модели на основе GPT, настроенные на отфильтрованном корпусе RuWikiSimple.

(!Не хватает какого-то завершения Введения. Например, в статье "Система автоматического аннотирования текстов с помощью стохастической модели"Введение заканчивается словами: "Таким образом, на сегодняшний день оценка качества аннотирования не обходится без работы экспертов, что безусловно дорого и требует существенных временных затрат."

Надо как-то подвести итог, в каком состоянии находится сейчас решаемая задача и обозначить место Вашей работы в текущем положении вещей.)

[8] [?] [?] [?] [?] [?]

Упрощение текстов на русском языке

Литература

[1] Sikka Punardeep, Mago Vijay. A Survey on Text Simplification. URL: http://arxiv.org/abs/2008.08612.

[2] Finegan-Dollak Catherine, Radev Dragomir R. Sentence simplification, compression, and disaggregation for summarization of sophisticated documents. Т. 67, № 10. С. 2437–2453. URL: https://onlinelibrary.wiley.com/doi/10.1002/asi.23576.

[3] Liu Jun, Matsumoto Yuji. Simplification of Example Sentences for Learners of Japanese Functional Expressions // Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016). The COLING 2016 Organizing Committee. С. 1–5. URL: https://aclanthology.org/W16-4901.

[4] Evans Richard, Orăsan Constantin, Dornescu Iustin. An evaluation of syntactic simplification rules for people with autism // Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR). Association for Computational Linguistics. С. 131–140. URL: https://aclanthology.org/W14-1215.

[5] RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian / Kazan Federal University, Kazan, Russia, Andrey Sakhovskiy, Alexandra Izhevskaya [и др.]. С. 607–617. URL: http://www.dialog-21.ru/media/5558/sakhovskiyaplusetal161.pdf.

[6] ruBTS: Russian Sentence Simplification Using Back-translation / Farit Galeev, Marina Leushina, Vladimir Ivanov [и др.]. С. 259–267. URL: http://www.dialog-21.ru/media/5510/galeevfplusleushinamplusivanovv144.pdf.

[7] Optimizing Statistical Machine Translation for Text Simplification / Wei Xu, Courtney Napoles, Ellie Pavlick [и др.]. Т. 4. С. 401–415. URL: https://doi.org/10.1162/tacl_a_00107.

[8] MUSS: Multilingual Unsupervised Sentence Simplification by Mining

Paraphrases / Louis Martin, Angela Fan, Éric de la Clergerie [и др.]. URL: http://arxiv.org/abs/2005.00352.

[9] Fang Meng, Li Yuan, Cohn Trevor. Learning how to Active Learn: A Deep Reinforcement Learning Approach // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. C. 595–605. URL: https://aclanthology.org/D17-1063.