Оглавление

Введение

Пишется в самом конце работ, является по сути кратким пересказом.

1 Анализ предметной области

В данном разделе вводятся основные определения и описываются важность и актуальность задачи упрощения текстов на русском языке.

1.1 Конвейерная обработка данных

Существует различные формулировки задачи упрощения текста. Так, в статье (https://www.jbe-platform.com/content/journals/10.1075/itl.165.2.06sid) даются определения в двух смыслах:

Упрощение текста в узком смысле - это процесс уменьшения лингвистической сложности текста при сохранении исходной информации и смысла. В более широком смысле упрощение текста охватывает другие операции: смысловое изменение для упрощения как формы, так и содержания; краткое изложение текста для исключения второстепенной или избыточной информации информации.

В статье Muss упрощением предложений называют процесс, целью которого является получение более легкого для чтения и понимания предложения за счет уменьшения его лексической и синтаксической сложности.

При этом задача упрощения текстов относится к области NLP и имеет много общего с другими задачами из этой сферы - машинным переводом, перефразированием и обобщением (резюмированием) текста (Чжу и др., 2010).

Если использовать более широкое понятие задачи упрощения текста, то ее будет сложно отличить от задачи обобщения. Поэтому, чтобы разграничить эти два понятия, в данной работе будет использоваться более узкое определение задачи упрощения.

В таком случае упрощение отличается от обобщения тем, что во втором случае основное внимание уделяется сокращению длины и содержания исходных данных. И хотя обобщенные тексты, как правило, короче, это не всегда так, и обобщение может привести к увеличению длины полученных предложений (Шардлоу, 2014). В рамках же упрощения текста обычно сохраняется все содержание.

1.2 Выводы из аналитиза предметной области

2 Классификация существующих решений (методов решения, алгоритмов)

Описываются существующие решения, даются ссылки. Предлагаются критерии оценки методов. Хорошо, если критерии имеют обоснование. Приводится классификация решений по критериям в виде таблицы, но в отдельных случаях можно и не в виде таблицы (пример мы Вам кидаем отдельно от документа).

В данном разделе описываются существующие решения задчи упрощения текстов, предлагаются критерии оценки методов и приводится классификация решений по этим критериям.

2.1 Существующие подходы

Глобально выделяют два подхода к решению задачи упрощения текстов - экстрактивный (извлекающий) и абстрактный.

Большинство ранних работ, посвященных задаче упрощения текстов, использовали экстрактивный подход - выделение в документе тех предложений, которые передают больше информации. Этот подход достаточно прост в реализации, однако он подходит лишь для решения задачи упрощения текстов в широком смысле, поэтому в дальнейшем в данной работе рассматриваться не будет.

С ростом доступности вычислительных ресурсов и количества исследований в области NLP, для решения задачи упрощения текстов все чаще стал использоваться абстрактный подход, подразумевающий генерацию нового текста [1].

2.2 Абстрактный подход

Решения задачи упрощения текстов, относяшиеся к абстрактному подходу, можно глобально разделить на лексическое упрощение и генерацию нового текста.

Первоначально абстрактный подход подразумевал упрощение предложений за счет лексических замен на уровне слов или целых фраз [2]. Этот процесс фокусируется исключительно на сокращении лексического содержания текста, но не принимает во внимание такие подзадачи, как грамматическое или синтаксическое упрощение [3]. Таким образом, эти решения не учитывают все аспекты решаемой задачи и поэтому в дальнейшем рассматриваться не будут.

Современные решения, основанные на абстрактном подходе, включают в себя разбиение корпуса на предложения, удаление и генерацию текста. Это стало возможным благодаря появлению нейронных сетей, в частности, рекуррентных нейронных сетей (RNNs)[1], которые позволяют решать задачи «от последовательности к последовательности» (seq2seq)[2].

Конвейер (?pipeline, не знаю, удачно ли подобран перевод) для этих подходов является универсальным, также довольно сложен. Предварительная обработка данных для моделирования seq2seq включает очистку входного и целевого текста, удаление знаков препинания и специальных символов, формирование словаря типичных слов. Далее входные и целевые предложения векторизуются в числовую форму для модели с вектором одинаковой длины либо путем усечения, либо путем заполнения. После того как модель была обучена, новая тестовая последовательность также сначала векторизуется, преобразуется к определенной длине и упрощается, а только после этого преобразовывается обратно из числовой формы в текстовую.

2.3 Генерация нового текста

Далее будут рассмотрены различные виды решений задачи упрощения текста с помощью генерации нового текста в рамках абстрактного подхода.

[1]Рекуррентная нейронная сеть, или RNN, - это сеть, которая работает с последовательностью и использут собственные промежуточные выходные данные в качестве входных данных для последующих шагов [4]

[2]Сеть sequence-to-sequence («от последовательности к последовательсности»), или сеть seq2seq, или сеть кодировщика-декодера, представляет собой модель, состоящую из двух RNN, называемых кодировщиком и декодером. Кодер считывает входную последовательность и выдает один вектор, а декодер считывает этот вектор для создания выходной последовательности [4]

2.3.1 Синтаксическое упрощение

3 Построение прототипа метода (или алгоритма, решения)

Сюда помещаются наброски метода – даётся его описание, общая идея

3.1 Выбор средств реализации

Выводы

Как ни странно, описывается не работа уже в третий (потому что второй раз она была описана во введении) раз, а даётся формальный ответ на вопрос, было ли выполнено ТЗ: были выработаны и обоснованы критерии оценки, был предложен метод решения...

Литература

[1] See Abigail, Liu Peter J., Manning Christopher D. Get To The Point: Summarization with Pointer-Generator Networks // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics. С. 1073–1083. URL: https://aclanthology.org/P17-1099.

[2] Paetzold Gustavo H., Specia Lucia. A Survey on Lexical Simplification. Т. 60. С. 549–593. URL: https://www.jair.org/index.php/jair/article/view/11091.

[3] Shardlow Matthew. A Survey of Automated Text Simplification. Т. 4, № 1. Number: 1 Publisher: The Science and Information (SAI) Organization Limited. URL: https://thesai.org/Publications/ViewPaper?Volume=4Issue=1Code=SpecialIssue

[4] NLP From Scratch: Translation with a Sequence to Sequence Network and Attention — PyTorch Tutorials 1.10.1+cu102 documentation. URL: https://pytorch.org/tutorials/intermediate/seq2seq$_t$*ranslation*$_t$*utorial.html*.