*1. Анализ предметной области*

*Вводятся основные определения, причём зачастую со ссылками на соответствующие статьи, чтобы показать, что автор не вводит новый термин, а описывает общепринятый. Описывается важность задачи, актуальность задачи.*

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 у*прощением 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.

*Видимо 2 глобальных подхода*

ПОДХОДЫ К УПРОЩЕНИЮ ТЕКСТА

Большая часть ранних работ, посвященных задаче упрощения текстов, использовали экстрактивный (извлекающий) метод обобщения - выделение в документе тех предложений, которые передают больше информации. Этот подход проще всего реализовать, однако он подходит лишь при использовании широкого понятия упрощения текстов, 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. (https://aclanthology.org/P17-1099/). Первоначально этот подход подразумевал упрощение предложений за счет лексических замен на уровне слов или целых фраз, а в последние годы фактически сводится к полной генерации нового предложения благодаря появлению нейронных сетей, в частности, рекуррентных нейронных сетей (RNNs) и сетей долгой краткосрочной памяти (LSTM), которые позволяют решать задачи от-последовательности-к-последовательности (seq2seq).

АБСТРАКТНЫЙ ПОДХОД

Абстрактные подходы в основном сосредоточены на лексических или фразовых заменах для упрощения на уровне предложений (Паетцольд и Специя, 2017). Этот процесс фокусируется исключительно на упрощении словарного запаса текста, а не на дополнительных задачах упрощения грамматического или синтаксического упрощения (Шардлоу, 2014). - плохо, грустно, рассматривать не будем

В отличие от этого, по-настоящему абстрактные подходы к упрощению включают разделение предложений, а также удаление и добавление текста. Наиболее распространенным подходом было моделирование seq2seq с использованием древовидных подходов, RNN и LSTM. Конвейер для этих подходов, хотя и гораздо более универсальный, также довольно сложен. Предварительная обработка данных для моделирования seq2seq включает очистку входного и целевого текста, возможно, удаление знаков препинания и специальных символов, а также сборку словаря общих слов. Затем как входные, так и целевые предложения векторизуются в числовую форму для модели с вектором одинаковой длины либо путем усечения, либо путем заполнения. Как только модель обучена, любой новый тестовый экземпляр сначала векторизуется, преобразуется в однородную длину, затем упрощается моделью, прежде чем быть преобразованным обратно из числовой формы в текстовую форму. Абстрактные подходы к TS можно в широком смысле классифицировать как лексическое упрощение или генерацию нового текста.

4. АБСТРАКТНЫЙ ПОДХОД - ГЕНЕРАЦИЯ НОВОГО ТЕКСТА

Новейшие технологии в TS в значительной степени основаны на данных, используя преимущества сложных структур данных и недавно разработанных нейронных техник. Бингель и Согаард (2016) представляют структурированный подход к TS, используя условные случайные поля для обхода графиков зависимостей сверху вниз, которые совместно предсказывают возможные сжатия и перефразировки (Бингель и Согаард, 2016). Другие подходы к TS, основанные на ИИ, включают глубокое обучение с подкреплением (Чжан и Лапата, 2017), сети с генератором указателей (см. и др., 2017), сети с расширением памяти (Ву и др., 2018), модели кодировщика/декодера LSTM (Ван и др., 2016a), методы нейронного машинного перевода (NMT) (Чо и др., 2014) (Сулем и др., 2018a) и другие методы глубокого обучения (Штайнер и Саггион, 2018) (Нисихара и др., 2019) (Палмеро Апросио и др., 2019). Стейнет и Нисиои (2019) предоставляют подробную оценку нейронных моделей seq2seq для внутридоменных и междоменных TS (Штайнер и Нисиои, 2019). Также предпринимались попытки перейти к неконтролируемым TS, как видно из работ Цяна и Ву (2019) (Цян и Ву, 2019) и Сурьи и др. (2019) (Сурья и др., 2019).). В следующих разделах описываются различные методы упрощения, используемые для создания нового текста.

4.1.СИНТАКСИЧЕСКОЕ УПРОЩЕНИЕ

В отличие от LS, целью которого является уменьшение сложности текста за счет упрощения словарного запаса, синтаксическое упрощение направлено на выявление грамматически сложного текста и переписывание его так, чтобы его было легче понять. Это может включать разделение длинных предложений на более короткие, более удобоваримые куски, изменение пассивного использования голоса на активное и устранение двусмысленностей и анафоры (Шардлоу, 2014).

Основополагающей работой в области синтаксического упрощения была система автоматического создания правил перезаписи для упрощения текста (Chandrasekar, R., & Srinivas, 1997), в которой использовались аннотированные корпуса и изученные правила для упрощения предложений в конкретной области. Хотя эта оригинальная система была разработана как этап предварительной обработки для других приложений на естественном языке, более поздние работы были сосредоточены на применении этого синтаксического упрощения в качестве вспомогательной технологии (Кэрролл и др., 1998) (Клерке и Сегаард, 2013). Более поздние работы по синтаксическому упрощению были сосредоточены на улучшении структуры дискурса (обеспечении того, чтобы предложения появлялись в правильном порядке) (Сиддхартан, 2006), а также на применении синтаксического упрощения в качестве инструмента для распознавания именованных сущностей (NER), особенно в медицинской области (Джонналагадда и Гонсалес, 2010).

Синтаксическое упрощение обычно выполняется в три этапа:

Сначала текст анализируется, чтобы определить его структуру и создать дерево синтаксического анализа. Это может быть сделано на разных уровнях детализации, но, как было показано, лучше всего работает на довольно грубом уровне, где слова и фразы сгруппированы в "супер-теги", которые представляют собой фрагмент основного предложения (Шардлоу, 2014). Эти супер-теги могут быть объединены с обычными правилами грамматики, чтобы обеспечить структурированную версию текста. На этом этапе определяется сложность предложения, чтобы решить, потребуется ли его упрощение. Это может быть сделано с помощью правил автоматического сопоставления, но также было сделано с использованием двоичного классификатора метода опорных векторов (SVM) (Шардлоу, 2014).

На этапе преобразования дерево синтаксического анализа модифицируется в соответствии с набором правил перезаписи, которые выполняют операции упрощения, такие как разделение предложений (Девлин и Тейт, 1998), перестановка предложений (Сиддхартан, 2006) и удаление предложений (Штайнер и др., 2013). Хотя автоматизированные методы применения этих правил существуют (Chandrasekar, R., & Srinivas, 1997), большинство систем синтаксического упрощения используют рукописные правила перезаписи, поскольку это устраняет необходимость в аннотированных корпусах и обычно приводит к повышению точности (Шардлоу, 2014).

После преобразования также может быть проведена фаза регенерации, в ходе которой в текст вносятся дальнейшие изменения для улучшения согласованности, актуальности и удобочитаемости.

Синтаксическое упрощение считалось важным компонентом систем TS и было реализовано в PSET (Девлин и Тейт, 1998) и Porsimples (Алюисио и Гасперин, 2010), двух системах, которые считаются повсеместными в вспомогательных технологиях. Сильные стороны синтаксического упрощения заключаются в его высокой точности и применимости к другим задачам НЛП (Шардлоу, 2014). Однако создание и проверка правил перезаписи - сложный процесс, и последние достижения в области методов глубокого обучения привели к автоматизации обнаружения и применения синтаксических упрощений.

4.2. СТАТИСТИЧЕСКИЙ МАШИННЫЙ ПЕРЕВОД (SMT)

Автоматизированный машинный перевод является устоявшейся техникой в НЛП [88] и включает автоматическое преобразование лексики и синтаксиса одного языка в лексику и синтаксис другого, в результате чего получается переведенный текст. Это было очень успешно применено к TS путем преобразования проблемы упрощения в случай одноязычной генерации текста в текст (Шардлоу, 2014). Мы рассматриваем нашу задачу перевода как перевод с исходного языка сложного английского на целевой простой английский. Недавние исследования в области машинного перевода были сосредоточены на статистических методах, основанных на фразах, которые изучают достоверные переводы из больших выровненных двуязычных корпусов, а затем могут применять их к новым текстам. Эта задача облегчается, когда исходный и целевой языки очень похожи, поэтому требуется немного изменений. И именно этот тип машинного перевода был применен к TS (Шардлоу, 2014).

Работа по выполнению TS с помощью SMT была выполнена для английского языка (Костер и Каучак, 2011) (Чжу и др., 2010) (Вуббен и др., 2012), бразильского и португальского языков (Specia, 2010) и была предложена для немецкого (Клапер и др., 2013), китайский (Чен и др., 2012) и шведский (Стим и др., 2013), среди прочих. На практике системы часто используют и модифицируют стандартный инструмент SMT, такой как Moses (Koehn и др., 2007), который был применен к задаче TS по английскому языку (Костер и Каучак, 2011). Мозес был дополнен модулем удаления фраз, который удалял ненужные части сложного исходного текста с многообещающими результатами (Шардлоу, 2014).

Встраивание слов в последнее время используется в качестве жизнеспособного подхода к TS (Nisioi и др., 2017). Цян и Ву сосредоточены на проблеме неконтролируемого упрощения без использования параллельных простых и сложных предложений, используя таблицы фраз, собранные из вложений слов и частот слов, из сложных и простых корпусов (Цян и Ву, 2019). Nisioi и др. (2017) предложили модель нейронной последовательности в последовательности, адаптирующую доступную архитектуру OpenNMT для решения задачи упрощения (Nisioi и др., 2017).

Производительность моделирования seq2seq для SMT в значительной степени зависит от набора данных, используемого для обучения модели. Например, если входные данные содержат слишком длинные предложения или упрощения, которые структурно не похожи на входные данные, этот подход не позволяет отслеживать и выравнивать различные части предложений. Кроме того, этот метод не учитывает знаки препинания и разделение предложений, что делает его неспособным понять основной контекст.

4.3. НОВЕЙШИЕ МЕТОДЫ ГЛУБОКОГО ОБУЧЕНИЯ

Алгоритмы глубокого обучения были успешно применены к SMT в начале этого десятилетия, о чем свидетельствует работа Cho и др. (2014), которые предложили кодер–декодер RNN для задачи перевода с английского на французский (Cho и др., 2014). Они показали, что их подход к подсчету пар фраз с помощью кодера–декодера RNN улучшает производительность перевода и что их модель изучает непрерывное пространственное представление фразы, которое сохраняет как семантическую, так и синтаксическую структуру фразы (Cho и др., 2014)

Ван и др. (2016) предложили использовать модель NMT на основе RNN для задачи TS (Ван и др., 2016b). Однако они сослались на отсутствие каких-либо согласованных пар сложного и простого предложений, позволяющих разработать такую модель, и сосредоточили свои усилия на применении NMT только к LS. Ван и др. (2016) также предложили использовать модель кодировщика-декодера LSTM для изучения правил работы, таких как реверсирование, сортировка и замена из пар последовательностей, которые аналогичны правилам упрощения, которые изменяют структуру предложения, заменяют слова и удаляют слова (Ван и др., 2016a).

Примерно в то же время Бингель и Согаард (2016) предложили подход к упрощению предложений, в котором использовались условные случайные поля линейной цепочки над графами зависимостей для совместного прогнозирования сжатия и перефразирования целых синтаксических единиц с целью удаления или перефразирования целых поддеревьев в графах зависимостей, в качестве стратегии, позволяющей избежать неграмматического вывода (Бингель и Согаард, 2016). Они новаторски использовали трехкратный параллельный одноязычный корпус, в котором представлены заголовки и сжатия для изучения перефразировок и удалений соответственно, демонстрируя, что их подход приводит к показателям читабельности, сопоставимым с предыдущими современными подходами к более простой задаче сжатия предложений, основанной на оценке человека (Бингель и Сегаард, 2016).

В ранних моделях seq2seq было две существенные проблемы:

Воспроизводить неточный вывод: В то время как кодер/декодер в моделях seq2seq способен извлекать уроки из исходного текста, его производительность сильно зависит от встраивания слов (расположение слова в источнике относительно других слов). Для нечастых слов модель не может разместить слово в правильном месте на выходе. Nisioi и др. (2017) предлагают способы решения этой проблемы путем добавления указателя на исходный текст, в то время как декодер выдает упрощенный вывод (Nisioi и др., 2017). Это показало многообещающие результаты в поддержании точного контекста вновь созданного упрощенного текста.