TEXT SUMMARIZATION VIA DEEP LEARNING

Nosirov Jaloliddin Azamjon oglu
Tashkent University of Information Technologies 2nd year Master

Abstract
This knowledge these days is stored in various formats in huge repositories mostly in the form of documents, sheets, photos, videos. One finds it difficult to comprehend this whole lot of information. There by, here comes the need of text summarization [2]. Text summarization is a process of extracting the context of a large document and summarize it into a smaller paragraph or a few sentences. Text summarization plays a vital role in saving time in our day to day life. It is also used in many bigger project implementations of classification of documents or in search engines [8]. Text Summarization has become an important and timely tool for assisting and interpreting text information. It is generally distinguished into: Extractive and Abstractive. The first method directly chooses and outputs the relevant sentences in the original document; on the other hand, the latter rewrites the original document into summary using Natural Language Processing (NLP) techniques. From these two methods, abstractive text summarization is laborious task to realize as it needs correct understanding and sentence amalgamation. This paper gives a brief survey of the distinct attempts undertaken in the field of abstractive summarization [5].

Key Words: Text Summarization, Deep Learning, Long Short Term Memory, Natural Language Processing, Recurrent Neural Networks – RNN, Abstractive summary, Extractive summary.

ОБОБЩЕНИЕ ТЕКСТА С ПОМОЩЬЮ ГЛУБОКОГО ОБУЧЕНИЯ

Носиров Джалолиддин Азамжон оглы
Ташкентский университет информационных технологий 2 курс Магистр

Аннотация (Abstract)

Знания сегодня хранятся в различных форматах в огромных хранилищах, в основном в виде документов, листов, фотографий, видео. Человеку сложно осмыслить весь этот объем информации. Вследствие этого возникает необходимость в обобщении текста [2]. Обобщение текста — это процесс извлечения контекста из большого документа и его обобщения в абзаце меньшего размера или нескольких предложениях. Обобщение текста играет важную роль в экономии времени в нашей повседневной жизни. Оно также используется во многих крупных реализациях проектов классификации документов или в поисковых системах [8]. Обобщение текста стало важным и своевременным инструментом для помощи и интерпретации текстовой информации. Обычно его делят на: экстрактивный и абстрактный. Первый метод напрямую выбирает и выводит соответствующие предложения в исходном документе; с другой стороны, последний переписывает исходный документ в резюме, используя (Natural Language Processing - NLP). Из этих двух методов обобщение абстрактного текста является трудоемкой задачей, поскольку требует правильного понимания и объединения предложений. В этой статье дается краткий обзор различных попыток, предпринятых в области абстрактного обобщения [5]

Ключевые слова (Key Words): обобщение текста, глубокое обучение, долговременная кратковременная память (Long Short Term Memory - LSTM), обработка естественного языка

(Natural Language Processing – NLP), рекуррентные нейронные сети (Recurrent Neural Networks - RNN), абстрактное обобщение, экстрактивное обобщение.

1. Введение (Introduction)

В настоящее время доступно огромное количество текстовых данных, включая онлайн-документы, статьи, новости и обзоры, которые содержат длинные строки текста, которые необходимо обобщить. Обобщение текста можно разделить на несколько категорий в зависимости от функции, жанра, контекста, типа реферата и количества документов [3]. Обобщение текста — это процесс создания краткого и лаконичного изложения путем захвата важной информации и всеобъемлющего смысла. Суммирование текста достигается с помощью методов обработки естественного языка с использованием алгоритмов, таких как алгоритмы ранжирования страниц и т. д. Хотя эти алгоритмы выполняют задачу суммирования текста, они не могут генерировать новые предложения, которых нет в документе, как люди. Они также могут иметь грамматические ошибки. Здесь нам на помощь приходит глубокое обучение. Использование глубокого обучения создает эффективную и быструю модель для суммирования текста. Использование методов глубокого обучения помогает нам создавать резюме, которые могут быть сформированы с помощью новых фраз и предложений, а также являются грамматически правильными. Обобщение текста можно разделить на два типа:

1. Абстрактное обобщение - Абстрактное обобщение текста может создавать новые фразы и предложения, которые передают наиболее полезную информацию из исходного текста. Предложения, созданные с помощью этого метода, могут отсутствовать в исходном документе [8]. Вдохновленная успехом нейронных сетей в экспериментах по машинному переводу, парадигма кодировщика-декодера, основанная на внимании, в последнее время широко изучается в абстрактном обобщении [1].

2. Экстрактивное обобщение - Экстрактивное обобщение текста включает в себя извлечение ключевых фраз из исходного документа и их объединение для создания резюме. Мы определяем важные слова или фразы из текста и извлекаем только те из них, которые нужны для резюме.

2. Сопутствующая работа или реализация (Related Work or Implementation)

Мы реализуем абстрактный метод, используя технику глубокого обучения, называемую долговременной кратковременной памятью (Long Short Term Memory - LSTM), которая является типом алгоритмов рекуррентной нейронной сети. Данные, используемые для этого проекта, представляют собой набор данных CNN_dailymail.

Данные - используемые данные представляют собой набор данных CNN_dailymail. Он имеет две функции: статьи и основные моменты. В статью включен документ, который необходимо обобщить. Это новостная статья. Highlights — это заголовки соответствующих новостей, которые используются в качестве сводок.

Методы - используемая модель представляет собой абстрактный метод, который реализуется с использованием методов глубокого обучения.

Алгоритм - используемый алгоритм представляет собой модель LSTM или Long Short Term Memory, которая является типом модели рекуррентной нейронной сети.

Модель - Используемая модель представляет собой модель от последовательности к последовательности. Обучение от последовательности к последовательности — это модель обучения, которая может преобразовывать последовательности одного входного домена в

последовательности другого выходного домена. Обычно он используется, когда вход и выход модели могут иметь переменную длину [8].

3. Методы (Methods)

CleanData() - Он используется для очистки данных с помощью шагов предварительной обработки, упомянутых ранее.

BuildDataSet() - Он используется для создания обучающих и тестовых наборов данных.

BuildDict() - Он используется для создания словаря, где ключи — это слова, а значения — случайные и уникальные числа. Он также создает еще один словарь с ключами в виде уникальных чисел и значениями в виде слов. Они используются при токенизации слов, так что входными данными для модели является набор чисел, а не слов, чтобы упростить вычисления с использованием векторов.

Tokenize() - Он используется для токенизации данных и отправки их в модель. Токенизация данных важна, поскольку для работы сетям нужны числовые данные, а не необработанные данные с символами.

3.1. Алгоритм. В настоящее время мы пытаемся создать алгоритмы, которые могут помочь нам воспроизвести человеческий мозг и достичь его функциональности. Это было достигнуто с помощью нейронных сетей. Нейронные сети — это набор алгоритмов, которые распознают закономерности в данных. Они очень похожи на человеческий мозг и способны создавать модели, которые могут работать или функционировать как человеческий мозг. Рекуррентная нейронная сеть (RNN) представляет собой тип нейронных сетей. Это нейронные сети с прямой связью, которые имеют внутреннюю память. В традиционной нейронной сети входная и выходная последовательности не зависят друг от друга. Но чтобы предсказать последовательность или предложение, нам нужно знать предыдущие слова, чтобы предсказать следующее слово. Следовательно, нам нужно внутренняя память. RNN помогает нам хранить предыдущую память с помощью скрытых состояний, которые запоминают информацию о предыдущих последовательностях.

RNN назван так, поскольку он периодически выполняет одну и ту же функцию для всех входных данных и скрытых слоев. Это уменьшает сложность хранения различных параметров для каждого из слоев в сети, тем самым экономя память. Выход текущего входа также зависит от прошлых выходов. После создания вывода он отправляется обратно в ту же сеть, чтобы его можно было сохранить и использовать для обработки следующего вывода в той же последовательности. Чтобы сгенерировать вывод в RNN, мы рассматриваем текущий ввод и вывод, который был сохранен из предыдущего ввода.

RNN отлично работают, когда речь идет о коротких контекстах. Но когда мы хотим создать сводку всей статьи, нам нужно зафиксировать контекст всей входной последовательности, а не только результат предыдущего ввода. Следовательно, нам нужна сеть, которая может захватывать весь контекст, как человеческий мозг. К сожалению, простой RNN не может уловить контекст или долгосрочную связь данных, то есть он не может запомнить или вызвать данные во входных данных, которые произошло задолго до этого и, следовательно, не может сделать эффективного предсказания. RNN может запоминать данные или контекст только на короткий срок. Это называется проблемой исчезающего градиента. Эта проблема может быть решена с помощью немного другой версии RNN — The Long Short Term Memory Networks.

Сети с долговременной кратковременной памятью (LSTM) являются улучшенной версией RNN. Они могут легко вспомнить прошлые данные, решив проблему исчезающего градиента. LSTM использует обратное распространение для обучения модели. LSTM хорошо подходит для прогнозирования и классификации последовательностей данных неизвестной

продолжительности. Их также можно использовать в методах языкового перевода и реферирования текста.

4. Заключение (Conclusion)

Растущий рост Интернета сделал доступным огромное количество информации. Людям трудно обобщать большие объемы текста. Таким образом, в наш век информационной перегрузки существует огромная потребность в автоматических инструментах обобщения. Международная корпорация данных (IDC) прогнозирует, что общий объем цифровых данных, ежегодно циркулирующих по всему миру, вырастет с 4,4 зеттабайт в 2013 году до 180 зеттабайт в 2025 году. Это огромный объем данных, циркулирующих в цифровом мире. Существует острая потребность в алгоритмах, которые можно использовать для автоматического сокращения объема данных с точными сводками, отражающими суть предполагаемых сообщений. Кроме того, применение реферирования текста сокращает время чтения и ускоряет процесс поиска информации, играющей важную роль в нынешнюю эпоху быстрого развития и цифровизации.

В этой работе мы применяем декодер кодировщика внимания для задачи абстрактного суммирования с очень многообещающими результатами, значительно превосходящими современные результаты на двух разных наборах данных. Каждая из предложенных нами новых моделей решает конкретную проблему абстрактного обобщения, обеспечивая дальнейшее улучшение производительности. Мы также предлагаем новый набор данных для обобщения нескольких предложений и устанавливаем на нем эталонные числа. В рамках нашей будущей работы мы планируем сосредоточить свои усилия на этих данных и построить более надежные модели для суммирования, состоящего из нескольких предложений.

Список литературы (References)

[1] Ekaterina Zolotareva, Tsegaye Misikir Tashu and Tomáš Horváth. Abstractive Text Summarization using Transfer Learning. CEUR-WS.org/Vol-2178/paper28 (2020).

[2] Soma Shrenika, GouriPriyaRamini, LunsavathBhagavathiDevi, B Geetavani. Abstractive Text Summarization By Using Deep Learning Models. *Turkish Journal of Computer and Mathematics Education*. Vol.12 No.14 (2021).

[3] Dima Suleiman and Arafat Awajan. Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges. Volume 2020, Article ID 9365340. Published 24 August 2020. https://doi.org/10.1155/2020/9365340.

[4] Panagiotis Kouris, Georgios Alexandridis, Andreas Stafylopatis. Abstractive text summarization based on deep learning and semantic content generalization. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5092 Florence, Italy, July 28 - August 2, 2019. c 2019 Association for Computational Linguistics.

[5] Neha Rane, Sharvari Govilkar. Recent Trends in Deep Learning Based Abstractive Text Summarization. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019.

[6] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çaglar Gulçehre, Bing Xiang. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 280–290, Berlin, Germany, August 7-12, 2016. c 2016 Association for Computational Linguistics.

[7] Chandra Khatri, Gyanit Singh and Nish Parikh. Abstractive and Extractive Text Summarization using Document Context Vector and Recurrent Neural Networks. KDD'18 Deep Learning Day, August 2018, London, UK.

[8] Kasimahanthi Divya, Kambala Sneha, Baisetti Sowmya, G Sankara Rao. Text Summarization using Deep Learning. International Research Journal of Engineering and Technology (IRJET). Volume: 07 Issue: 05 | May 2020. www.irjet.net.

[9] Julien Romero, Thomas Hofmann, Jason Lee. Abstractive Text Summarisation with Neural Networks (2017).

[10] Rana Talib Al Timimi, and Fatma Hassan Al Rubbiay. Multilingual Text Summarization using Deep Learning. International Journal of Engineering Research and Advanced Technology (IJERAT). Volume.7, No. 5 May – 2021. https://ijerat.com/index.php/ijerat.