Исследование фреймворка на основе кодера-декодера с длинной краткосрочной памятью для извлекающего резюмирования текста

Билал Сомар

Донской Государственный Технический Университет, Ростов-на-Дону

Аннотация: В данном исследовании мы представляем исследование методов обработки естественного языка и машинного обучения, с особым акцентом на алгоритмах глубокого обучения. В ходе исследования было исследовано применение моделей с длинной краткосрочной памятью с механизмами внимания для задач суммаризации текста. В качестве набора данных для экспериментов использовались новостные статьи и соответствующие им резюме. В статье обсуждаются предварительные этапы обработки данных, включая очистку текста и токенизацию. В рамках исследования также исследуется влияние различных гиперпараметров на производительность модели. Результаты демонстрируют эффективность предложенного подхода в генерации кратких резюме на основе обширных текстов. Полученные результаты способствуют развитию методов обработки естественного языка и машинного обучения для суммаризации текста.
Ключевые слова: извлекающая суммаризация текста, последовательность-последовательность, длинная краткосрочная память, кодировщик-декодировщик, модель суммаризации, обработка естественного языка, машинное обучение, глубокое обучение, механизм внимания.

Введение

Методы машинного обучения предоставляют систематический и данных-ориентированный подход к анализу и прогнозированию результатов, позволяя исследователям принимать обоснованные решения и получать ценные инсайты из сложных наборов данных [1]. Машинное обучение, включая методы глубокого обучения, такие, как Рекуррентные Нейронные Сети (Recurrent Neural Networks (RNN)) и сети с длинной краткосрочной памятью (Long Short-Term Memory (LSTM)), вместе с методами анализа данных, играют важную роль в повышении эффективности обработки данных и извлечении ценных знаний из имеющихся данных [2].

Суммирование текста играет важную роль в управлении огромным количеством информации, доступной онлайн. Техники извлекающего суммирования текста направлены на выделение важных предложений для создания кратких сводок. Модели глубокого обучения, такие как основанная

на LSTM архитектура энкодер-декодер, показали свою эффективность в этой области. Эти модели автоматически выбирают значимые предложения, чтобы предоставить пользователям ключевые моменты без потери исходного контекста.

В данной статье исследуется применение основанного на LSTM-архитектуре энкодера-декодера для извлекающего суммирования текста. Подход заключается в обучении такой архитектуры, где энкодер улавливает контекстуальную информацию из входного документа, а декодер генерирует сводку, выбирая важные предложения. Целью является повышение точности и связности сгенерированных сводок с использованием техник глубокого обучения.

В статье представлена методология, включающая предварительную обработку данных, архитектуру модели и процесс обучения. Экспериментальные результаты и оценка производительности демонстрируют эффективность основанной на LSTM-архитектуры. В заключении подчеркиваются перспективы дальнейших исследований в области извлекающего суммирования текста и потенциал развития моделей глубокого обучения.

Связанные исследования

В [3] рассматривается прогресс в автоматическом извлечении текстовых резюме с использованием методов глубокого обучения. Предлагается подход, основанный на анализе данных, который использует методы глубокого обучения и перефразировки для генерации извлекающих резюме. В экспериментах оценивается предложенная модель на основе набора данных Para Multiling 2015, что показывает высокую точность (>90%), но различную точность (precision), F-меру (F-measure) и полноту (recall).

В [4] представлен новый подход к семантическому извлечению текстовых резюме с целью достижения баланса между степенью сжатия и

сохранением информации. Модель выборочно сохраняет информативные предложения, избегает избыточности и фильтрует несущественную информацию. Она генерирует резюме в хронологическом порядке и показывает результаты, сравнимые с моделью Скрытое распределение Дирихле (Latent Dirichlet Allocation (LDA)).

В [5] статье фокусируется на оценке техники извлечения текстового резюме на основе неконтролируемого обучения для обширных данных обзоров и отзывов. В ней подчеркиваются ограничения контролируемых техник и обсуждаются применение неконтролируемых техник, в частности, алгоритмов К-средних (KMeans), Мини-пакеты К-средних (MiniBatchKMeans) и графового резюмирования.

В [6] рассматривается проблема сжатия объемных текстов путем предложения ансамблевого подхода, объединяющего различные алгоритмы машинного обучения. Исследование приходит к выводу, что ансамблевый подход дает лучшие результаты с использованием показателей Оценка подчиненности с акцентом на полноту для краткого извлечения информации (Recall-Oriented Understudy for Gisting Evaluation (ROUGE)), достигая среднего значения ROUGE-1 в 0.78, ROUGE-2 в 0.66 и ROUGE-L в 0.76.

В [7] предложенный алгоритм сочетает в себе методы наблюдаемого и ненаблюдаемого обучения для экстрактивного и абстрактивного краткого изложения. Он превосходит традиционные методы, достигая повышенной точности в 87.58% и улучшения показателей ROUGE на 38.42%.

В [8] исследуются и оцениваются различные техники текстового краткого изложения на основе параметров, таких, как сжатие, сохранение смысла и грамматические ошибки. ТекстРанк (TextRank) показывает незначительное превосходство и более высокую скорость (примерно на 10%) по сравнению с другими экстрактивными моделями.

В [9] предлагается глубокая модель машинного обучения для краткого изложения фактических отчетов. В этом подходе используются различные характеристики и ограниченная машина Больцмана (Restricted Boltzmann Machine (RBM)) для улучшения характеристик. Предложенный подход показывает лучшую производительность в кратком изложении фактических отчетов, достигая среднего значения точности 0.7 и среднего значения полноты 0.63, что выше, чем у существующего подхода.

В [10] исследование фокусируется на создании функциональной модели краткого изложения документов в качестве базового решения для различных типов документов. Модель показывает разумную производительность по сравнению с референтными краткими изложениями, предоставляя полезный обзор. Оценочные показатели для краткого изложения текста следующие:
ROUGE-1: F-мера - 0.360, точность - 0.288 и полнота - 0.480.
ROUGE-2: F-мера - 0.089, точность - 0.072 и полнота - 0.119.
ROUGE-L: F-мера - 0.329, точность - 0.3 и полнота - 0.365.

В [11] статье рассматривается автоматическое краткое изложение текста, в частности, метод анализа лексических цепочек. Он включает лингвистические предварительные этапы обработки, такие, как сегментация предложений, токенизация, маркировка частей речи (Part of Speech (POS)), обнаружение сущностей и обнаружение отношений.

В [12] систематический обзор рассматривает различные модели автоматического краткого изложения текста на основе графов. Графовые подходы предпочтительны, благодаря своей доступности и независимости от языка. Обзор подчеркивает важность учета языка и области при выборе соответствующей модели.

В [13] предлагается модель автоматического краткого изложения текста с использованием архитектуры Последовательность-последовательность (Sequence-to-Sequence (Seq2Seq)) с рекуррентными нейронными сетями

(Recurrent Neural Networks (RNNs)). Модель генерирует абстрактные краткие изложения и показывает многообещающие результаты для длинных и юридических документов с эффективной генерацией краткого изложения и ROUGE-оценками от 0.6 до 0.7. В связи с чувствительностью юридических или судебных данных, в исследовании использовались новостные данные с аналогичной природой.

Методология

В нашем исследовании мы реализуем модель для автоматической суммаризации текста, используя архитектуру последовательность-к-последовательности (sequence-to-sequence (Seq2Seq)) с использованием фреймворка на основе LSTM для кодирования и декодирования. Вот краткое описание методологии и архитектуры, которые мы использовали:

Модель Seq2Seq состоит из двух основных компонентов: энкодера и декодера.

Энкодер: Входные последовательности (отзывы или текст) проходят через слой эмбеддинга, который отображает слова входных последовательностей на плотные векторы фиксированного размера.

Встроенные входные последовательности затем обрабатываются стеком из трех слоев LSTM. Каждый слой LSTM генерирует последовательные выходы, одновременно улавливая контекст и взаимосвязи между словами. Конечным выходом энкодера является выходная последовательность последнего слоя LSTM, а также конечное скрытое состояние и состояние ячейки LSTM.

Декодер: Целевые последовательности (суммары) также проходят через слой эмбеддинга для получения плотных векторных представлений. Встроенные целевые последовательности подаются на вход единственного слоя LSTM в декодере, который генерирует последовательные выходы.

Начальное состояние LSTM-декодера устанавливается в качестве конечного скрытого состояния и состояния ячейки LSTM энкодера, позволяя декодеру использовать закодированную информацию из входной последовательности.

Выход LSTM-декодера проходит через плотный слой с функцией активации softmax, который предсказывает распределение вероятностей по словарю для каждого слова в целевой последовательности.

В общем, методология включает предварительную обработку данных, создание модели кодировщика-декодировщика Seq2Seq с LSTM-слоями, обучение модели с использованием обучающих и валидационных данных, и генерацию кратких изложений с помощью обученной модели.

Результаты и оценка

Для оценки качества сгенерированных кратких изложений сравниваются справочные краткие изложения с использованием показателей ROUGE. В таблице № 1 представлены результаты оценки ROUGE для метрик текстового изложения.

Таблица № 1

Результаты оценки ROUGE для метрик текстового изложения

Metric	Recall (R)	Precision (P)	F1-Score (F)
Rouge-1	0.542	0.765	0.618
Rouge-2	0.096	0.157	0.115
Rouge-L	0.542	0.765	0.618

Результаты оценки ROUGE свидетельствуют о производительности модели в выявлении важной информации из справочных кратких изложений. Для показателя ROUGE-1 модель достигла значения полноты (recall) равного 0.542, что означает, что она улавливает примерно 54.2% униграмм, присутствующих в справочных кратких изложениях. Значение точности (precision) равное 0.765 указывает на то, что 76.5% предсказанных униграмм были правильными и соответствующими. Значение F1-меры (F1-score)

равное 0.618, которое является гармоническим средним полноты и точности, предоставляет общую меру производительности модели. Относительно высокая точность свидетельствует о том, что модель генерирует краткие изложения с достаточным количеством правильных униграмм. Однако есть возможность для улучшения полноты, что указывает на то, что некоторые важные униграммы могли быть упущены.

Для показателя ROUGE-2 производительность модели в выявлении биграмм ниже. Значение полноты (recall) равное 0.096 указывает на то, что только 9.6% биграмм из справочных кратких изложений были уловлены, а значение точности (precision) равное 0.157 показывает, что 15.7% предсказанных биграмм были правильными и соответствующими. Значение F1-меры (F1-score) равное 0.115 отражает общие сложности, с которыми сталкивается модель при генерации точных и содержательных последовательностей биграмм.

Rouge-L измеряет общую схожесть наибольших общих подпоследовательностей между предсказанными и референсными резюме. Оценка полноты 0.542 указывает на то, что модель уловила примерно 54.2% наибольших общих подпоследовательностей (Longest Common Subsequence (LCS)), а оценка точности 0.765 предполагает, что 76.5% предсказанных LCS являются правильными и соответствующими. Значение F1-меры 0.618 дает общую оценку производительности модели в улавливании LCS.

В целом, модель демонстрирует приемлемую производительность в уловлении соответствующих отдельных слов (униграмм) и общей последовательности наибольших общих подпоследовательностей LCS в кратком изложении. Однако она испытывает трудности в уловлении смысловых биграмм, что может быть некоторым ограничением. Улучшение генерации точных и содержательных биграмм значительно повысит качество сгенерированных кратких изложений.

Полученные результаты могут быть обусловлены несколькими факторами, включая архитектуру модели, продолжительность обучения, качество и количество данных, а также настройку гиперпараметров. Продолжительность обучения и количество эпох могут оказывать влияние на производительность, и более длительное обучение может привести к лучшим результатам, но это следует сбалансировать с ресурсами и проблемами переобучения. Качество и количество тренировочных данных являются значимыми факторами, и наличие шумных или несогласованных данных может негативно сказаться на производительности модели. Гиперпараметры, такие, как скорость обучения, размер пакета, выбор оптимизатора и методы регуляризации, следует оптимизировать для улучшения результатов.

Заключение

В данном исследовании мы разработали модель экстрактивной текстовой суммаризации на основе архитектуры sequence-to-sequence с использованием LSTM-основанного кодировщика-декодировщика.

Модель продемонстрировала многообещающие результаты в генерации кратких резюме путем выбора важных предложений из исходного текста. Наши эксперименты на разнообразном наборе данных показали, что модель достигает конкурентоспособных оценок ROUGE, что указывает на ее эффективность в улавливании ключевой информации и генерации связных резюме.

Однако есть еще место для улучшений. В дальнейшей работе можно исследовать альтернативные архитектуры, такие как модели на основе Transformer, чтобы усилить способность модели улавливать долгосрочные зависимости. Увеличение размера и разнообразия тренировочных данных также может привести к дальнейшему улучшению производительности модели. Кроме того, можно рассмотреть возможность тонкой настройки

предварительно обученной языковой модели, специфичной для задач суммаризации, для более хорошей обобщенности.

В целом, данное исследование вносит вклад в область экстрактивной текстовой суммаризации, демонстрируя потенциал моделей Seq2Seq в генерации точных и информативных резюме. Наши результаты предоставляют ценные инсайты для исследователей и практиков, работающих над автоматическими методами суммаризации.

Литература

1. Горлатов Д.В. Машинное обучение прогнозных моделей на несбалансировнных данных по опасным астероидам // Инженерный вестник Дона, 2023, №5. URL: ivdon.ru/ru/magazine/archive/n5y2023/8394.

2. Федутинов, К. А. Машинное обучение в задачах поддержки принятия решений при управлении охраной природы // Инженерный вестник Дона, 2021, №9. URL: ivdon.ru/ru/magazine/archive/n9y2021/7186.

3. Bhargava, R. and Sharma, Y., 2020. Deep extractive text summarization. Procedia Computer Science, 167, pp.138-146.

4. Fatima, Z., Zardari, S., Fahim, M., Andleeb Siddiqui, M., Ibrahim, A.A.A., Nisar, K. and Naz, L.F., 2022. A novel approach for semantic extractive text summarization. Applied Sciences, 12(9), p.4479.

5. Verma, J.P. and Patel, A., 2017. Evaluation of unsupervised learning based extractive text summarization technique for large scale review and feedback data. Indian Journal of Science and Technology, 10(17), pp.1-6.

6. Singh, P., Chhikara, P. and Singh, J., 2020, February. An ensemble approach for extractive text summarization. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-7). IEEE.

7. Meena, S.M., Ramkumar, M.P., Asmitha, R.E. and G SR, E.S., 2020, September. Text summarization using text frequency ranking sentence prediction.

In 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP) (pp. 1-5). IEEE.

8. Dehru, V., Tiwari, P.K., Aggarwal, G., Joshi, B. and Kartik, P., 2021, March. Text summarization techniques and applications. In IOP Conference Series: Materials Science and Engineering (Vol. 1099, No. 1, p. 012042). IOP Publishing.

9. Verma, S. and Nidhi, V., 2017. Extractive summarization using deep learning. arXiv preprint arXiv:1708.04439.

10. Mehta, Ashwinee & Sanikommu, Bhargavi. (2022). Extractive Document Summarization. 10.13140/RG.2.2.23249.40804. URL: researchgate.net/publication/362903296_Extractive_Document_Summarization.

11. Patel, S.M., Dabhi, V.K. and Prajapati, H.B., 2017. Extractive Based Automatic Text Summarization. J. Comput., 12(6), pp.550-563.

12. Bichi, A.A., Keikhosrokiani, P., Hassan, R. and Almekhlafi, K., 2022. Graph-based extractive text summarization models: a systematic review. Journal of Information Technology Management, 14(Special Issue: 5th International Conference of Reliable Information and Communication Technology (IRICT 2020)), pp.184-202.

13. Prasad, C., Kallimani, J.S., Harekal, D. and Sharma, N., 2020, October. Automatic Text Summarization Model using Seq2Seq Technique. In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 599-604). IEEE.

References

1. Gorlatov D.V. Machine learning of predictive models on imbalanced data for hazardous asteroids. Inzhenernyj vestnik Dona, 2023, No. 5. URL: ivdon.ru/ru/magazine/archive/n5y2023/8394.

2. Fedutinov, K.A. (2021). Machine learning in decision support tasks for nature conservation management. Inzhenernyj vestnik Dona, 2021, №9. URL: ivdon.ru/ru/magazine/archive/n9y2021/7186.

3. Bhargava, R. and Sharma, Y., 2020. Deep extractive text summarization. Procedia Computer Science, 167, pp.138-146.

4. Fatima, Z., Zardari, S., Fahim, M., Andleeb Siddiqui, M., Ibrahim, A.A.A., Nisar, K. and Naz, L.F., 2022. A novel approach for semantic extractive text summarization. Applied Sciences, 12(9), p.4479.

5. Verma, J.P. and Patel, A., 2017. Evaluation of unsupervised learning based extractive text summarization technique for large scale review and feedback data. 10(17), pp.1-6.

6. Singh, P., Chhikara, P. and Singh, J., 2020, February. An ensemble approach for extractive text summarization. International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-7). IEEE.

7. Meena, S.M., Ramkumar, M.P., Asmitha, R.E. and G SR, E.S., 2020, September. Text summarization using text frequency ranking sentence prediction. 4th International Conference on Computer, Communication and Signal Processing (ICCCSP) (pp. 1-5). IEEE.

8. Dehru, V., Tiwari, P.K., Aggarwal, G., Joshi, B. and Kartik, P., 2021, March. Text summarization techniques and applications. In IOP Conference Series: Materials Science and Engineering (Vol. 1099, № 1, p. 012042). IOP Publishing.

9. Verma, S. and Nidhi, V., 2017. Extractive summarization using deep learning. arXiv preprint arXiv:1708.04439.

10. Mehta, Ashwinee & Sanikommu, Bhargavi. (2022). Extractive Document Summarization. 10.13140/RG.2.2.23249.40804. URL: researchgate.net/publication/362903296_Extractive_Document_Summarization.

11. Patel, S.M., Dabhi, V.K. and Prajapati, H.B., 2017. Extractive Based Automatic Text Summarization. J. Comput., 12(6), pp.550-563.

12. Bichi, A.A., Keikhosrokiani, P., Hassan, R. and Almekhlafi, K., 2022. Graph-based extractive text summarization models: a systematic review. 5th International Conference of Reliable Information and Communication Technology (IRICT 2020)), pp.184-202.

13. Prasad, C., Kallimani, J.S., Harekal, D. and Sharma, N., 2020, October. Automatic Text Summarization Model using Seq2Seq Technique. Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 599-604). IEEE.