Sabinin O.Y.
PhD in Computer Science,
Peter the Great St. Petersburg Polytechnic University
Shabalina Y.V.
Master of Computer Science student,
Peter the Great St. Petersburg Polytechnic University

Сабинин Олег Юрьевич
кандидат технических наук,
Санкт-Петербургский политехнический университет Петра Великого
Шабалина Юлия Владимировна
студент магистратуры,
Санкт-Петербургский политехнический университет Петра Великого

METHOD OF AUTOMATIC TEXT SUMMARIZATION OF TEXTS IN RUSSIAN ON THE BASIS OF NEURAL NETWORKS
МЕТОД АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ НА ОСНОВЕ АППАРАТА НЕЙРОННЫХ СЕТЕЙ

Summary: The use of computers in human activities, especially scientific, not only speeds up the processes of creating and processing documents, but also increases their number and volume. Today, many users regularly face the need to examine a large amount of information in a short time. The way out of the situation is to read through not the whole text, but its summary. The aim of this work is to develop a new method of automatic text summarization of texts in Russian on the basis of neural networks. The neural network is trained to determine whether the sentence should be included in summary or not.

Key words: text summarization, neural networks, natural language processing, machine learning, automated language processing.

Аннотация:
Применение компьютеров в человеческой деятельности, в том числе и научной, не только ускоряет процессы создания и обработки документов, но и чрезвычайно увеличивает их количество и объем. Сегодня многие пользователи регулярно сталкиваются с необходимостью быстрого просмотра большого количества информации. Выходом из ситуации является просмотр не всего документа, а его сжатого описания – аннотации. Целью данной работы является разработка метода автоматического аннотирования текстов на русском языке на основе аппарата нейронных сетей. Производится обучение нейронной сети на определение ключевых предложений текста на основе вектора свойств предложения.

Ключевые слова: аннотирование текста, нейронные сети, обработка естественного языка, машинное обучение, автоматическая обработка текста.

Постановка проблемы.

Задачи автоматического аннотирования текстов приобретают всё большую актуальность с ростом объёмов и потоков данных, при чём не только для сети Интернет, но и для других хранилищ информации, например, библиотек или баз знаний различных организаций.

Огромное число материалов затрудняет быстрое получение кратких ёмких сводок по текстам, так как формирование аннотаций вручную требует колоссальных человеческих ресурсов, и именно поэтому задача реализации эффективных методов автоматического аннотирования приобретает всё большую важность. Аннотирование текстов помогает выделить ключевые части текста и сократить объёмы просматриваемой информации. Некоторые тексты, например, научные статьи сопровождаются аннотациями, которые выделяют основные идеи работы. Однако, большинство источников не обладают подобным преимуществом, а их заголовки (при наличии), зачастую могут значительно отличаться от содержания, что подводит нас к тому, что инструменты автоматического аннотирования текстов могут значительно упростить и уменьшить количество просматриваемой информации.

Анализ последних исследований и публикаций.

С самого начала активного использования электронно-вычислительных машин первого поколения, то есть с середины пятидесятых годов прошлого века, стали предприниматься попытки решения задач обработки текста на естественном языке. Одной из первых задач по обработке естественно-языковых текстов при помощи ЭВМ стало автоматическое аннотирование.

С тех пор было проведено множество исследований по разработке автоматизированных методов и моделей аннотирования. Решением проблемы занимались такие отечественные исследователи как Н.В. Лукашевич, Р.Г. Пиотровский, П.Г. Осминин, С.А. Тревгода, В.А. Яцко и др. Из зарубежных исследователей стоит обратить внимание на: H.P. Luhn, H.P. Edmundson, R. Mihalcea, J. Kupiec, E. Lloret, G. Salton и др. В настоящее время можно выделить два основных подхода к автоматическому аннотированию:

– экстракция – извлекающие методы, основанные на извлечении из первичных документов наиболее информативных фрагментов и включении их в аннотацию в порядке следования в тексте;

– абстракция – генерирующие методы, предусматривающие создание нового текста, обобщающего первичные документы [1].

Извлекающие методы работают путём определения наиболее важных фрагментов текста (предложения, абзацы). При этом данные фрагменты не обрабатывают, а извлекают в таком порядке и виде в каком они приведены в тексте. Основные сложности, связанные с данным подходом, заключаются в определении ключевых предложений текста, и затем связи этих предложений в единый, удобочитаемый текст.

Генерирующие методы, напротив, направлены на создание нового материала, явно непредставленного в тексте исходного документа. Другими словами, они интерпретируют и исследуют текст с помощью передовых методов обработки естественного языка, чтобы создавать новые структурные единицы текста, которые передают самую важную информацию из исходного документа. При использовании генерирующих методов текст аннотации строится на правилах, предполагающих наличие лингвистической базы знаний.

Выделение нерешенных ранее частей общей проблемы.

Несмотря на множество проведенных исследований, проблема разработки формальных методов и моделей для автоматического аннотирования еще не решена, ввиду того, что задача формализации естественного языка достаточно трудоемка, а сам язык является неоднозначным, неограниченным и эволютивным.

Вышеупомянутые характеристики естественного языка играют особенно большую роль при исследовании материалов на русском языке, так как именно для русского языка характерен свободный порядок слов, морфологическая сложность, подвижное разноместное ударение и высокая степень сегментной редукции [2].

На текущий момент из методов автоматического аннотирования текстов на русском языке наиболее распространены различные статистические и графовые методы, являющиеся представителями экстрагирующего подхода. Аннотации, полученные с помощью экстрагирующих подходов, зачастую характеризуются недостаточно высоким качеством текста, бессвязностью. Абстрагирующие же подходы потенциально способны обеспечить лучшее качество текста аннотации, но они чрезвычайно трудны для практической реализации и находятся на уровне исследовательских разработок.

Поскольку большинство текстов обладают достаточно выраженной структурой, ключевые части документа можно представить путем выбора предложений на основе их свойств и характеристик. В работе [3] был предложен схожий подход и рассмотрены такие методы машинного обучения с учителем для решения задач автоматического аннотирования, как наивный байесовский классификатор, метод опорных векторов. Исследователем были получены обнадеживающие результаты, поэтому в данной статье было решено использовать вышеупомянутый экстрактивный подход для автоматического аннотирования текстов на русском языке, только в качестве классификатора, в отличие от [3], были выбраны искусственные нейронные сети.

Цель статьи.

Разработка метода автоматического аннотирования текстов на русском языке на основе аппарата нейронных сетей.

Изложение основного материала.

Разрабатываемый нами метод предполагает, что прежде всего исходный документ представляется как набор предложений, а сами предложения рассматриваются как набор свойств и характеристик. Среди этого набора выбираются те предложения, которые нейронная сеть сочтёт более релевантными. Результатом является некоторое подмножество предложений исходного текста.

Изначально необходимо выполнить следующее:

– определить рассматриваемые свойства и характеристики предложений, значения которых будут являться входными данными для нейронной сети;

– создать размеченный тестовый корпус текстов для последующего обучения нейронной сети;

– произвести непосредственно само обучение сети.

Определение рассматриваемых свойств и характеристик предложений.

Каждое предложение аннотируемого текста представляется в виде вектора, состоящего из 6 характеристик $[f_1, f_2, \dots, f_6,]$:

– отношение порядкового номера абзаца, к которому принадлежит предложение, к общему числу абзацев исходного документа (f_1);

– отношение порядкового номера предложения в абзаце к общем числу предложений в абзаце (f_2);

– отношение количества символов рассматриваемого предложения к количеству символов самого длинного предложения текста (f_3);

– отношение количества ключевых слов в предложении к общему количеству тематических слов предложения (f_4);

– отношение количества совпадающих тематических слов данного предложения и предыдущего к общему количеству тематических слов рассматриваемых предложений (f_5);

– отношение количества совпадающих тематических слов данного предложения и предыдущего к общему количеству тематических слов двух рассматриваемых предложений (f_5);

– отношение количества совпадающих тематических слов данного предложения и последующего к общему количеству тематических слов двух рассматриваемых предложений (f_6).

Свойства f_1-f_2основываются на местоположении предложения в документе или в его абзаце. Ожидается, что эти параметры поспособствуют выбору ключевых предложений, так как аннотации, состоящие из первых предложений абзацев, превосходят аннотации, составленные с помощью других методов статьи [4], а предложения, расположенные в начале и конце абзацев, имеют высокие шансы попасть в итоговый текст [5].

Свойство f_3 поможет избавиться от слишком коротких вводных предложений, которые вряд ли попадут в аннотацию [6].

Свойство f_4 зависит от количества ключевых и тематических слов в предложении. Тематические слова получаются следующим образом: из документа производится выборка всех существительных, прилагательных и глаголов, которые в последствии сводятся к их начальной форме. Для получившегося набора слов высчитывается их встречаемость в тексте. Ключевыми словами считаются 25% тематических слов, но не больше 10 что соответствует объему оперативной памяти человеках [7]. Ожидается, что с помощью этого свойства увеличится вероятность выбора ключевых предложений, поскольку термины, которые часто встречаются в документе, вероятно, связаны с его темой [8]. Для выделения тематических слов был использован Томита-парсер компании Яндекс, имеющий встроенную поддержку русского языка, доступную документацию и используемый в таких популярных сервисах как Яндекс.Новости и Яндекс.Работа.

Свойства f_5-f_6 основываются на симметричном реферировании, то есть на определении количества связей между предложениями [9].

Рассмотренные свойства могут быть изменены или дополнены. Выбор рассматриваемых свойств предложений определяет, какие предложения попадут в итоговую аннотацию и влияет на работу нейронной сети.

Обучение нейронной сети.

Обучение нейронной сети проводится для изучения типов предложений, которые должны быть включены в аннотацию. Обучение проводится на тестовом корпусе текстов, где каждое предложение вручную отмечено как входящее в аннотацию или же нет.

Нейронная сеть ищет закономерности, присущие предложениям, которые должны быть включены в аннотацию. Используется нейронная сеть прямого распространения с тремя слоями, которая, как было доказано, является универсальным функциональным аппроксиматором [10]. Сеть может обнаруживать паттерны и аппроксимировать функцию любых данных с точностью до 100%, если в наборе данных нет противоречий.

Создание нейронной сети проводилось в NeurophStudio. Входной слой разрабатываемой нейронной сети состоит из шести нейронов, где каждый нейрон соответствует одному из свойств предложения, пяти нейронов скрытого слоя и одного нейрона выходного слоя (рис. 1). В качестве активационной функции используется сигмоида, обучение сети проводится методом обратного распространения ошибки.

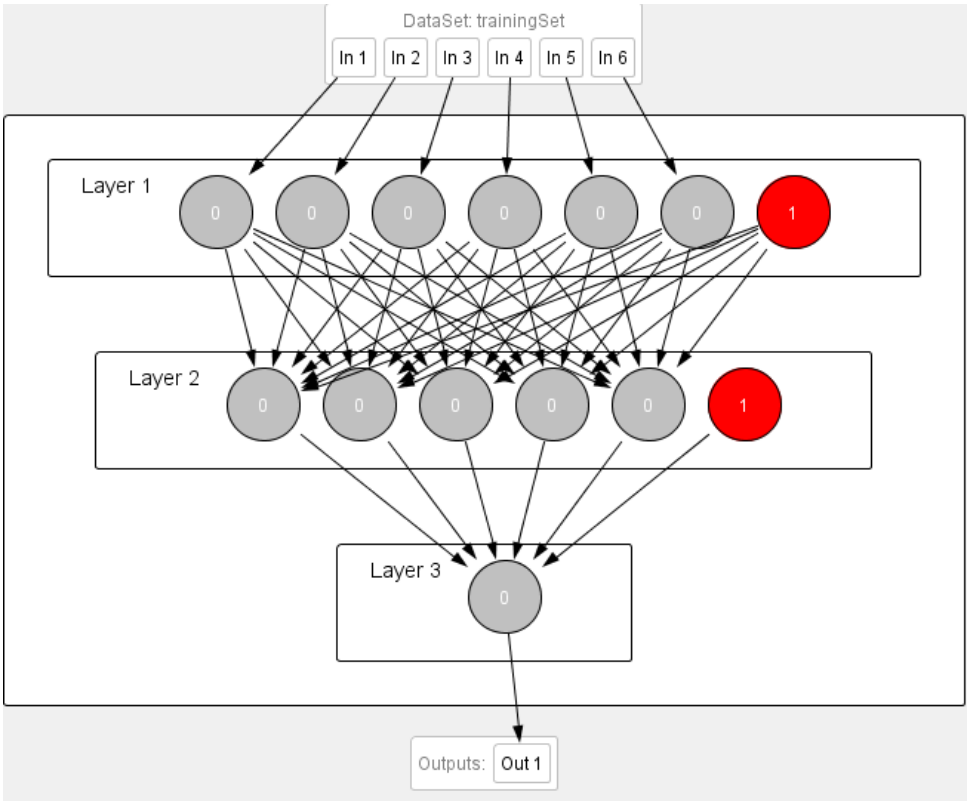

Рис. 1. Модель нейронной сети

Для создания тестового корпуса было использовано 62 статьи на различные тематики, найденные в сети Интернет. Каждый текст состоял от 27 до 102 предложений, в среднем – из 49. Всего было проанализировано 3076 предложений. 565 предложений было помечено, как ключевые, в среднем 9 на текст. Нейронная сеть была успешно обучена за 23 итерации (рис. 2).

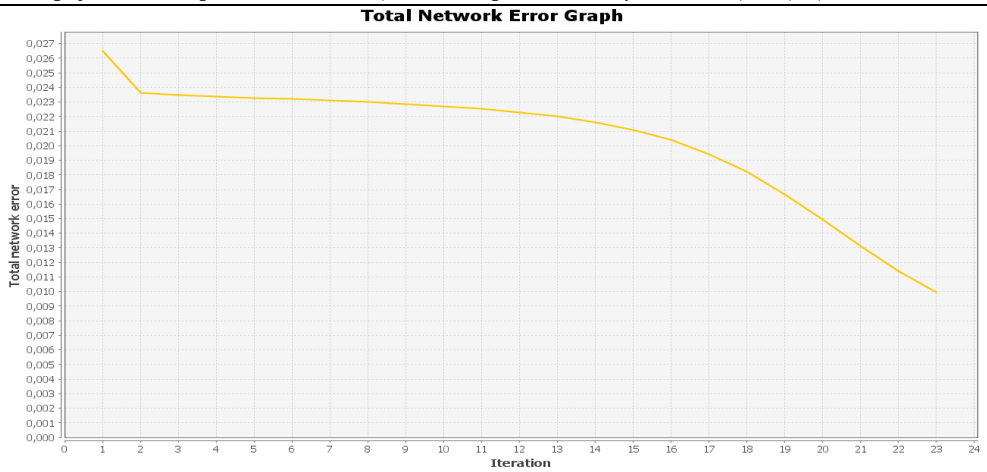

Рис. 2. График обучения нейронной сети

Итоговая среднеквадратичная ошибка для тестового корпуса составила 0.16733 (рис. 3). Точность определения является предложение ключевым или не является для нейронной сети составила 88,76% по сравнению с ручной разметкой для тестового корпуса. Для корпуса из 10 новых текстов точность составила 82,31%.

```
Input: 0,75; 0,1111; 0,4348; 0,7918; 0,0256; 0,1579;  Output: 0,8415; Desired output: 1;  Error: -0,1585;
Input: 0,1667; 0,5; 0,1333; 0,7951; 0,0526; 0;  Output: 0,7811; Desired output: 1;  Error: -0,2189;
Input: 0,5; 0,3333; 0,625; 0,2567; 0,2727; 0;  Output: 0,7373; Desired output: 1;  Error: -0,2627;
Input: 0,0625; 0,5; 0,125; 0,4481; 0,0769; 0;  Output: 0,706; Desired output: 1;  Error: -0,294;
Input: 0,0833; 0,1667; 0; 0,1337; 0; 0;  Output: 0,7228; Desired output: 1;  Error: -0,2772;
Input: 1; 0,5; 0,8; 0,3005; 0; 0,0833;  Output: 0,5936; Desired output: 1;  Error: -0,4064;
Input: 0,625; 0,25; 0,6; 0,3648; 0; 0;  Output: 0,7618; Desired output: 1;  Error: -0,2382;
Input: 0,875; 0,5333; 1; 0,3316; 0,0833; 0,1111;  Output: 0,6675; Desired output: 1;  Error: -0,3325;
Total Mean Square Error: 0.16733012877903028
```

Рис. 3. Среднеквадратичная ошибка для тестового корпуса текстов

Выводы и предложения.

Точность результатов предложенного метода автоматического аннотирования текстов на тестовой выборке составила 88,76%. Результаты оказались вполне удовлетворительными, что позволяет проводить дальнейшие исследования. Выбранные для анализа предложений свойства, а также выбранные независимым читателем ключевые предложения для тестового корпуса текстов имеют большое влияние на работу нейронной сети. Сеть обучается в соответствие со стилем читателя и в соответствие с предложениями, которые именно этот читатель считает ключевыми. Можно рассматривать данную особенность как преимущество данного подхода, так как любой человек может обучить нейронную сеть в соответствие со своими личными предпочтениями.

Список литературы:

1. Осминин, П. Г. (2016). Построение модели реферирования и аннотирования научно-технических текстов, ориентированной на автоматический перевод (Doctoral dissertation, Челябинск).
2. Большакова, Е. И., Клышинский, Э. С., Ландэ, Д. В., Носков, А. А., Пескова, О. В., & Ягунова, Е. В. (2011). Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 272, 3.
3. Wong, K. F., Wu, M., & Li, W. (2008, August). Extractive summarization using supervised and semi-supervised learning. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 (pp. 985-992). Association for Computational Linguistics.
4. Brandow, R., Mitze, K., & Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. Information Processing & Management, 31(5), 675-685.
5. Baxendale, P. B. (1958). Machine-made index for technical literature—an experiment. IBM Journal of Research and Development, 2(4), 354-361.
6. Kupiec, J., Pedersen, J., & Chen, F. (1995, July). A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 68-73). ACM.
7. Ванюшкин, А. С., & Гращенко, Л. А. (2016). Методы и алгоритмы извлечения ключевых слов. Новые информационные технологии в автоматизированных системах, (19).
8. Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.
9. Ступин, В. С. (2004). Система автоматического реферирования методом симметричного реферирования. In Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог (pp. 579-591).
10. Гализдра, В. И., & Бабаев, Ш. Б. (2011). Нейронные сети и аппроксимация данных. Научные и образовательные проблемы гражданской защиты, (3).