ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ, СИСТЕМНИЙ АНАЛІЗ ТА КЕРУВАННЯ

DOI: 10.20535/1810-0546.2018.4.141286
УДК 004.8

О.В. Іванишин*, А.Є. Батюк
Національний університет "Львівська політехніка", Львів, Україна

МЕТОД АВТОМАТИЧНОГО ЕКСТРАКТИВНОГО УЗАГАЛЬНЕННЯ ТЕКСТУ НА ОСНОВІ РЕКУРЕНТНИХ НЕЙРОННИХ МЕРЕЖ

Проблематика. У статті розглядається вирішення завдання автоматичного екстрактивного узагальнення тексту на основі рекурентної штучної нейронної мережі з використанням графової інтерпретації тексту й алгоритму оцінки важливості текстової одиниці. Абстрактний підхід набагато складніший, ніж екстрактивний, оскільки він вимагає створення власного вектора думки мережі, який не обов'язково має містити слова з вхідного тексту, а також повинен бути граматично правильно побудований. Алгоритм оцінки важливості текстової одиниці полягає у використанні принципу рейтингу рекомендацій, який збалансовує ваги графу залежно від популярності текстових одиниць. Принцип навчання без учителя є набагато ближчим до навчання біологічного інтелекту і не вимагає позначених підготовлених даних.
Мета дослідження. Метою статті є аналіз методу автоматичного екстрактивного узагальнення тексту на основі рекурентних штучних нейронних мереж з використанням навчання без учителя.
Методика реалізації. Пропонується алгоритм досягнення глибшого абстрактного опрацювання тексту за допомогою інтерпретації тексту у вигляді графу. Використано елементи теорії графів, теорію та методи проектування алгоритмів. Алгоритм оцінки важливості текстової одиниці використовує принцип рейтингу рекомендацій.
Результатати дослідження. У відносному порівнянні продуктивність нейронної мережі на базі орієнтованого графу майже на 5 % перевищує неорієнтовану версію. За допомогою графової інтерпретації продуктивність мережі на 15 % вища, ніж звичайне лексичне семантичне *n*-грам представлення.
Висновки. Метод відрізняється тим, що враховує власну структуру тексту замість того, щоб обробляти текст як прості ряди лексичних семантичних термінів. Саме перетворення тексту на багатовимірний орієнтований граф відкриває потенціал значно абстрактнішої обробки. Практичне застосування своєю чергою охоплює велику область безперервного опрацювання не тільки соціальних мереж і новин, блогів, статей чи повідомлень, а й сферу освіти, генетики та медицини.

Ключові слова: рекурентна штучна нейронна мережа; екстрактивне узагальнення тексту; глибоке навчання; обробка природної мови; навчання без учителя.

Вступ

Галузь автоматичного узагальнення (підсумовування) тексту викликала інтерес ще з кінця 50-х років. Це завдання охоплює безмежне поле практичного застосунку завдяки можливості опрацьовувати безперервний потік тексту будь-якого змісту. Побудова висновків, узагальнених тез, виділення теми або ідеї у вигляді ключових текстових одиниць може використовуватись як у моніторингу фондових бірж, ресурсів новин, соціальних мереж, месенджерів, так і у шпигунських та воєнних цілях. Саме побудова стислих і максимально наповнених інформацією текстових сигнатур у вигляді конспекту чи логу з великих об'ємів текстових даних грає ключову роль у розвитку машинного інтелекту.

Основні дослідження в цій галузі підкреслюють екстрактивні підходи до узагальнення з використанням статистичних лексичних методів [1]. Традиційні методи підсумовування тексту аналізують частоту слів або речень у тексті, щоб визначити найбільш важливі лексичні елементи. Кілька статистичних моделей були розроблені таким чином на базі навчальних корпусів даних. Вони поєднують різні евристики, використовуючи ключові слова, положення та довжину речень, частоту слів або заголовки [2].

Наш метод базується на поданні тексту як графа у вигляді рекурентної штучної нейронної мережі. Підходи до класифікації на основі графів враховують власну структуру тексту замість того, щоб обробляти текст як прості набори лексичних термінів. Таким чином, вони здатні охоплювати і виражати інформацію якісніше при визначенні важливих понять і ключових текстових одиниць [3].

Вибрані фрагменти тексту використовуються у графовій конструкції як вершини, вони

* corresponding author: ostap.solomon@gmail.com

можуть бути словами, фразами, реченнями або цілими абзацами. На сьогодні впроваджено багато успішних систем, що враховують компроміс між багатством вмісту та граматикою тексту. Згідно з цим підходом найважливішими текстовими одиницями є ті, що найбільш пов'язані між собою у графі. Вони і використовуються для побудови остаточного підсумку тексту [4]. Для виявлення зв'язків між текстовими одиницями (ребрами графу) існує кілька підходів:

- перекривання слів;
- подібність за косинус-відстанню;
- подібність за чутливістю до запитів.

Також деякі автори пропонують комбінації попередніх методів із контрольованими навчальними функціями [5].

Ці алгоритми використовують різні методи пошуку інформації для визначення найважливіших текстових одиниць (вершин графу) для побудови узагальнення [7]. Алгоритм TextRank, розроблений у [6], і алгоритм LexRank, розроблений у [7], засновані на оцінці лексичних текстових одиниць (речень або слів) із використанням алгоритмів PageRank [8]. Інші класифікаційні алгоритми на основі графів, такі як HITS [9] або позиційна функція [10], також можуть бути застосовані для вирішення поставленої проблеми.

Постановка задачі

Презентація тексту у вигляді графу фундаментально відрізняється від звичайної обробки тексту як послідовності лексичних термінів, що дає змогу отримувати більше інформації про структуру тексту та його значення.

Особливістю та основною відмінністю навчання без учителя є те, що йому не потрібні великі підготовані масиви тренувальних даних із заздалегідь виділеними ключовими фразами.

Основна ціль роботи – проаналізувати ці відмінності та особливості, а також розглянути метод автоматичного узагальнення (підсумовування) тексту для глибшого розуміння та подальшого використання в дисертаційній роботі.

Структурне подання методу

Для завдання автоматичного підбиття підсумків алгоритм моделює будь-який документ як граф, що використовує текстові одиниці як вузли [11]. Функція для обчислення подібності текстових одиниць потрібна для побудови зв'язків між ними. Ця функція використовується для розрахунку ваг ребер графу. Чим більша подібність між текстовими одиницями, тим важливішою буде вага ребра між ними в графі [13].

Алгоритм визначає, наскільки подібні дві текстові одиниці на основі вмісту, який обидва поділяють. Це перекриття розраховується як число загальних лексичних знаків між ними, поділених на довжину кожного.

Функція, представлена в оригінальному алгоритмі, може бути формалізована таким чином.

Нехай дано два речення S_i, S_i, представлені у формі n слів, що в S_i та в S_j відповідно подані як

$$S_i = w_1^i, w_2^i, \ldots, w_n^i .$$

Функція подібності для S_i та S_j визначена так:

$$\mathrm{Sim}(S_i, S_j) = \frac{|\{w_k | w_k \in S_i w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)} .$$

Результатом цього процесу є щільний граф, що представляє документ. З цього графу алгоритм може обчислити важливість кожної вершини. Найбільш значимі текстові одиниці відбираються і подаються в тому ж порядку, що і з'являються в узагальненому документі ключових фраз.

Незалежно від типу та характеристик елементів, доданих до графу, оцінка важливості текстової одиниці складається з таких основних етапів:

1. Визначення текстових одиниць, які найкраще визначають поставлене завдання (домен) і додавання їх у вершини графу.

2. Визначення зв'язків між цими текстовими одиницями і використання їх для побудови ребер між вершинами на графі. Ребра можуть бути орієнтовані або неорієнтовані, збалансовані або незбалансовані.

3. Ітерація алгоритму оцінки до конвергенції.

4. Сортування вершин на основі їх кінцевої оцінки та прийняття рішень відповідно до значень, прикріплених до кожної вершини та ребра.

Модель алгоритму оцінки важливості текстової одиниці

Основна ідея, реалізована на основі графової моделі оцінки, – "голосування" або "ре-

комендація". Коли одна вершина зв'язана з іншою, вона голосує за цю вершину. Чим більша кількість голосів, відданих за вершину, тим більше значення самої вершини. Крім того, важливість вершини, яка здійснює голосування, визначає, наскільки є важливим саме голосування, і ця інформація також враховується за допомогою моделі оцінки. Отже, оцінка, пов'язана з вершиною, визначається на підставі голосів, відданих за неї, і оцінок вершин, що власне голосують.

Формально, нехай $G = (V, E)$ є напрямленим графом із набором вершин V і набором ребер E, де E – це підмножина $V \times V$. Для заданої вершини V_i нехай $\mathrm{In}(V_i)$ буде набором вершин, які вказують до неї (попередники), і нехай $\mathrm{Out}(V_i)$ буде сукупністю вершин V_i, що вказують від неї (наступники). Оцінка вершини V_i визначається як

$$S(V_i) = (1-d) + d \sum_{j \in \mathrm{In}(V_i)} \frac{1}{|\mathrm{Out}(V_j)|} S(V_j),$$

де d – коефіцієнт згасання, який можна встановити між 0 і 1. Він грає роль ймовірності стрибка від заданої вершини на іншу випадкову вершину в графі [14].

Починаючи від довільних значень, призначених для кожного вузла у графі, обчислення повторюється до досягнення збіжності нижче певного порога. Конвергенція досягається, коли коефіцієнт помилки для будь-якої вершини графа падає нижче заданого порога. Частота помилок вершини V_i визначається як різниця між "реальними" оцінками вершини $S(V_i)$ і балом, обчисленим при ітерації, k, $S^k(V_i)$. Оскільки реальна оцінка невідома апріорі, ця частота помилок апроксимується з різницею між оцінками, обчисленими за двома послідовними ітераціями:

$$S^{(k+1)}(V_i) - S^k(V_i).$$

Після запуску алгоритму оцінка зв'язується з кожною вершиною, яка представляє "важливість" вершини всередині графа. Остаточні значення, отримані після закінчення роботи алгоритму, не впливають на вибір вихідного значення. Кількість ітерацій до конвергенції може бути різною [12].

Алгоритм вилучення ключових текстових одиниць

Алгоритм вилучення ключових текстових одиниць працює повністю без учителя і виконується таким чином.

Першим кроком попередньої обробки, необхідним для вмикання програми синтаксичних фільтрів, є маркування і анотація тексту з частинами мовлення. Щоб уникнути надмірного збільшення розміру графа, ми розглядаємо лише окремі текстові одиниці як кандидати на додавання до графа, усі мультислівні ключові текстові одиниці реконструюються у фазі подальшої обробки.

Наступним кроком усі граматичні елементи, що проходять синтаксичний фільтр, додаються до графа, а між лексичними одиницями, які збігаються у контексті слів, додається ребро. Після побудови неорієнтованого незрівноваженого графа, оцінка, пов'язана з кожною вершиною, встановлюється в початкове значення 1. Модель, описана в попередньому розділі, виконується на цьому графі впродовж декількох ітерацій, доки не зійдеться (зазвичай за 20–30 ітерацій, з порогом 0,0001).

Після отримання остаточної оцінки для кожної вершини графа вершини сортуються в зворотному порядку їх оцінок, а верхні T вершин у рейтингу зберігаються для подальшої обробки. Хоча T може бути будь-яким фіксованим значенням, зазвичай від 5 до 20, цей алгоритм використовує більш гнучкий підхід, який визначає кількість ключових текстових одиниць залежно від розміру тексту.

Під час подальшої обробки всі текстові одиниці, вибрані як потенційні ключові слова за допомогою алгоритму, позначаються в тексті, а послідовності сусідніх ключових слів колапсують у багатослівне ключове слово. Наприклад, у тексті "рубі код для обробки моделі" якщо обидва слова "рубі" та "код" вибрані як потенційні ключові слова, то вони згортаються в одне ключове слово "рубі" оскільки є сусідніми.

Навчання без учителя

Навчання з учителем має деякі позитивні властивості, зокрема здатність створювати інтерпретовані правила про те, які якості харак-

теризують ключову фразу. Проте вони також вимагають великої кількості навчальних даних. У цьому випадку потрібно багато документів із відомими ключовими фразами-узагальненнями.

Навчання без учителя є набагато ближчим до навчання біологічного інтелекту. У виділенні ключових фраз навчання без учителя позбавляє потреби в навчальній інформації. Навчання без учителя підходить до проблеми з іншого боку. Замість того щоб намагатися вивчити явні функції, які характеризують ключові фрази, алгоритм використовує структуру самого тексту для визначення ключових фраз, які видаються "центральними" у тексті [6].

Цей принцип ґрунтується на понятті "престиж" або "рекомендація" із соціальних мереж. Таким чином, він взагалі не спирається на будь-які попередні навчальні дані та може працювати на будь-якому довільному фрагменті тексту. Алгоритм виводить висновок просто на основі внутрішньої архітектури та властивостей тексту. Таким чином, алгоритм легко переноситься на нові домени, мови та платформи.

Висновки

Досліджено елементи теорії графів і процеси автоматичного нейромережевого узагальнення текстових даних та встановлено, що графовий алгоритм подання тексту здатний охоплювати і виражати інформацію значно якісніше, ніж лінійні алгоритми. Цей метод відрізняється тим, що враховує власну структуру тексту замість того, щоб обробляти текст як прості ряди лексичних термінів.

Саме перетворення тексту на багатовимірний орієнтований граф відкриває потенціал значно абстрактнішої обробки. З теоретичної точки зору, маючи достатні потужності, у перспективі можна збільшувати мережу як углиб для більш детального опрацювання тексту, так і вшир для опрацювання більших об'ємів даних. Практичне застосування своєю чергою охоплює велику область безперервного опрацювання не тільки соціальних мереж і новин, блогів, статей чи повідомлень, а й сферу освіти, генетики та медицини.

References

[1] D. Das and A.F.T. Martins, "A survey on automatic text summarization", Carnegie Mellon University, Language Technologies Institute, Tech. Rep., 2007.

[2] C.Y. Lin and E.H. Hovy, "Identifying topics by position", in *Proc. 5th Conf. Appl. Natural Language Proc.*, Washington D.C., March 1997.

[3] Y. Ouyang *et al.*, "Learning similarity functions in graph-based document summarization", in *Lecture Notes in Computer Science*, vol. 5459, W. Li and D.M. Aliod, eds. Springer, 2009, pp. 189–200. doi: 10.1007/978-3-642-00831-3_18

[4] R. Barzilay and K. McKeown, "Sentence fusion for multi-document news summarization", *Computational Linguistics*, vol. 31, no. 3, pp. 297–328, 2005. doi: 10.1162/089120105774321091

[5] G. Salton *et al.*, "Automatic text structuring and summarization", *Inform. Proces. Manag.*, vol. 33, no. 2, pp. 193–207, 1997. doi: 10.1016/S0306-4573(96)00062-3

[6] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts", in *Proc. EMNLP 2004*, Barcelona, Spain, July 2004, pp. 404–411.

[7] G. Erkan and D.R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization", *J. Artif. Intell. Res.*, vol. 22, no. 1, pp. 457–479, 2004.

[8] L. Page *et al.*, "The PageRank citation ranking: Bringing order to the web", in *Proc. 7th Int. World Wide Web Conf.*, Brisbane, Australia, 1998, pp. 161–172.

[9] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment", *JACM*, vol. 46, no. 5, pp. 604–632, 1999. doi: 10.1145/324133.324140

[10] P.J.J. Herings *et al.*, "Measuring the power of nodes in digraphs", Maastricht University, Maastricht Research School of Economics of Technology and Organization, TI 2001–096/1, 2001.

[11] C.D. Manning *et al.*, *Introduction to Information Retrieval.* Cambridge, UK: Cambridge University Press, 2008.

[12] F. Geche *et al.*, "Boolean neuro-functions and synthesis of the recognition device in the neuro-base", *Bulletin of the National University "Lviv Polytechnic". Ser. Comp. Sci. Inform. Technol.*, no. 598, pp. 44–50, 2007.

[13] V. Kotsovsky *et al.*, "Estimates of the magnitude of integer weighted coefficients of two-threshold neuronal elements", *Bulletin of the National University "Lviv Polytechnic". Ser. Comp. Sci. Inform. Technol.*, no. 694, pp. 292–296, 2011.

[14] I. Cmoc *et al.*, "Realization of the neural element based on previous calculations", *Bulletin of the National University "Lviv Polytechnic". Ser. Comp. Sci. Inform. Technol.*, no. 710, pp. 11–18, 2011.

О.В. Иванишин, А.Е. Батюк

МЕТОД АВТОМАТИЧЕСКОГО ЭКСТРАКТИВНОГО ОБОБЩЕНИЯ ТЕКСТА НА ОСНОВЕ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ

Проблематика. В статье рассматривается решение задания автоматического экстрактивного обобщения текста на базе рекуррентной искусственной нейронной сети с использованием графовой интерпретации текста и алгоритма оценки важности текстовой единицы. Абстрактный подход намного сложнее, чем экстрактивный, поскольку он требует создания собственного вектора мысли сети, который не обязательно должен включать слова из входного текста, а также должен быть грамматически правильно построен. Алгоритм оценки важности текстовой единицы заключается в использовании принципа рейтинга рекомендаций, который балансирует веса графа в зависимости от популярности текстовых единиц. Принцип обучения без учителя является более близким к обучению биологического интеллекта и не требует помеченных подготовленных данных.

Цель исследования. Целью работы является анализ метода автоматического экстрактивного обобщения текста на основе рекуррентных искусственных нейронных сетей с использованием парадигмы обучения без учителя.

Методика реализации. Предлагается алгоритм достижения более глубокой абстрактной обработки текста при помощи интерпретации текста в виде графа. Использованы элементы теории графов, теории и методы проектирования алгоритмов. Алгоритм оценки важности текстовой единицы использует принцип рейтинга рекомендаций.

Результаты исследования. В относительном сравнении продуктивность нейронной сети на базе ориентированного графа почти на 5 % превышает неориентированную версию. При помощи графовой интерпретации продуктивность сети на 15 % выше, чем обычное лексическое семантическое *n*-грамм представление.

Выводы. Метод отличается тем, что учитывает собственную структуру текста вместо того, чтобы обрабатывать текст как простые ряди лексических семантических терминов. Именно преобразование текста в многомерный ориентированный граф открывает потенциал значительно более абстрактной обработки. Практическое применение в свою очередь охватывает большую область непрерывной обработки не только социальных сетей и новостей, блогов, статей или сообщений, но и сферу образования, генетики и медицины.

Ключевые слова: рекуррентная искусственная нейронная сеть; экстрактивное обобщение текста; глубокое обучение; обработка природного языка; обучение без учителя.

O.V. Ivanyshyn, A.E. Batyuk

METHOD OF AUTOMATIC EXTRACTIVE TEXT SUMMARIZATION ON THE BASIS OF RECURRENT NEURAL NETWORKS

Background. The article deals with the solution of the problem of automatic extractive text summarization on the basis of recurrent artificial neural network, using graph interpretation of the text and a text unit importance estimator. Abstractive approach is much more complex than extractive as it requires network to generate personal thought vector which is not obliged to contain words from input text as well as it should be built grammatically correct. The text unit importance estimator uses recommendation rating principle which balances the graph weights depending on the popularity of text units. The principle of unsupervised learning is much closer to real biological learning process and doesn't require labeled preprocessed dataset.

Objective. The aim of the paper is the method of automatic extractive text summarization based on recurrent artificial neural networks using unsupervised learning.

Methods. An algorithm for the achievement of deeper abstract text processing using the interpretation of the text in the form of a graph is proposed. The algorithm uses elements of graph theory and methods of algorithms' design. The text unit importance estimator uses recommendation rating principle.

Results. In relative comparison, the performance of the directed graph based on neural network is almost 5 % higher than undirected graph network version. Using graph interpretation algorithm, the network performance is 15 % higher than the usual simple lexical *n*-gram representation.

Conclusions. This method is characterised in that it takes into account its own structure of the text, instead of processing the text as simple rows of lexical and semantic terms. It is the transformation of the text into a multidimensional oriented graph that opens the potential for much more abstract text processing. Practical application, in its turn, covers a large area of continuous processing of not only social networks and news, blogs, articles or communications, but also the fields of education, genetics and medicine.

Keywords: recurrent neural network; extractive text summarization; deep learning; natural language processing; unsupervised learning.

Рекомендована Радою
факультету прикладної математики
КПІ ім. Ігоря Сікорського

Надійшла до редакції
25 квітня 2018 року

Прийнята до публікації
6 вересня 2018 року