https://orcid.org/0000-0002-6497-5056
https://orcid.org/0000-0003-3631-1268
https://orcid.org/0000-0003-3601-8555

УДК 004.83

С.Д. ПОГОРІЛИЙ[*], А.А. КРАМОВ[*], Ф.М. ЯЦЕНКО[*]

МЕТОД АНАЛІЗУ КОГЕРЕНТНОСТІ УКРАЇНОМОВНИХ ТЕКСТІВ ІЗ ВИКОРИСТАННЯМ РЕКУРЕНТНОЇ НЕЙРОННОЇ МЕРЕЖІ

[*]Київський національний університет імені Тараса Шевченка, м. Київ, Україна

Анотація. *У роботі обґрунтовано актуальність вирішення задачі оцінки когерентності (цілісності) текстів та здійснено аналіз відповідних методів комп'ютерної лінгвістики. Автоматизована оцінка когерентності тексту відноситься до задач обробки природної мови, тому її варто розглядати як AI-повну задачу. Для здійснення оцінки когерентності тексту використовуються методи машинного навчання і комп'ютерної лінгвістики: метод опорних векторів, нейронні мережі різної архітектури тощо. Доцільним є використання нейронних мереж у зв'язку з відсутністю залежності від експертних знань. Варто зазначити необхідність урахування семантичної складової тексту. Семантичне формалізоване представлення одиниць тексту (слів чи речень) здійснюється за допомогою попередньо навчених моделей відповідно до предметної області інформації. Відповідні нейронні мережі варто проектувати за допомогою згорткових чи рекурентних шарів, які дозволяють здійснювати обробку вхідних даних нефіксованого розміру (слів та речень). Детально розглянуто принцип роботи методу розподіленого представлення речень із використанням рекурентної нейронної мережі. Перевагою застосування рекурентних шарів є наявність зворотного зв'язку у нейронах: вихідне значення з попереднього кроку потрапляє на вхід нейрона. Такий підхід відображає процес сприйняття тексту читачем, адже аналіз поточної інформації можливий за рахунок наявності попередньо отриманих знань. Створено рекурентну нейронну мережу та здійснено її навчання на множині україномовних наукових статей. Для покращення точності методу здійснено попередню обробку текстів статей, що містили некоректні послідовності символів у зв'язку з автоматизованою екстракцією їх вмісту з PDF-файлів. Проведено аналіз точності методу за допомогою експериментального вирішення задач розрізнення документів та вставки. Отримані результати можуть свідчити про можливість застосування методу на основі рекурентної нейронної мережі для оцінки когерентності україномовних текстів.*

Ключові слова: *когерентність тексту, розподілене представлення речень, рекурентна нейронна мережа, нейрони зі зворотним зв'язком, модель Word2Vec.*

Аннотация. *В работе обоснована актуальность решения задачи оценки когерентности (целостности) текста и осуществлен анализ соответствующих методов компьютерной лингвистики. Автоматизированная оценка когерентности текста относится к задачам обработки естественного языка, поэтому её стоит рассматривать как AI-полную задачу. Для осуществления оценки когерентности текста используются различные методы машинного обучения и компьютерной лингвистики. Целесообразно использовать нейронные сети в связи с отсутствием зависимости от экспертных знаний. Стоит отметить необходимость учитывать семантическую составляющую текста. Семантическое формализованное представление единиц текста осуществляется с помощью предварительно обученных моделей в соответствии с предметной областью. Соответствующие нейронные сети стоит проектировать с помощью сверточных или рекурентных слоев, которые позволяют обрабатывать входные данные нефиксированного размера (слов или предложений). Детально рассмотрен принцип работы метода распределенного представления предложений с помощью рекуррентной нейронной сети. Преимуществом использования рекуррентных слоев является наличие обратной связи в нейронах: исходное значение с предыдущего шага попадает на вход нейрона. Такой подход отображает процесс восприятия текста читателем, ведь анализ текущей информации возможен с помощью наличия предварительно полученных знаний. Создано рекуррентную нейронную сеть и осуществлено её обучение на множестве украиноязычных научных статей. Для улучшения точности метода осуществлена предварительная обработка текстов статей, которые содержали некорректные последовательности символов в связи с автоматизированной экстракцией их содержания с PDF-файлов. Проведен анализ точности метода с помощью экспериментального решения задач различения документов и*

ISSN 1028-9763. Математичні машини і системи, 2019, № 4

вставки. Полученные результаты могут свидетельствовать о возможности применения метода на основе рекуррентной нейронной сети для оценки когерентности украиноязычных текстов.
Ключевые слова: *когерентность текста, распределенное представление предложений, рекуррентная нейронная сеть, нейрон с обратной связью, модель Word2Vec.*

Abstract. *The urgency of solving the problem of text coherence estimation has been justified in the paper. The comparative analysis of the corresponding methods of computer linguistic has been performed. The automated estimation of text coherence falls into a category of natural language processing; therefore, it should be considered as an AI-complete task. In order to perform the estimation of text coherence machine learning methods and computer linguistics means are used. It is advisable to use neural networks because such an approach does not require expert knowledge. It should be noted that the semantic component of a text should be taken into account. The semantic formalized representation of text units (words or sentences) is performed using a previously trained model according to a subject area of information. The corresponding neural networks can be designed using convolutional and recurrent layers that allow the processing of input data with unfixed size (words and sentences). The principle of the distributed sentence representation method using a recurrent neural network has been considered in detail. The key advantage of recurrent layers is the availability of feedback connections within neurons: the initial value from the previous step goes to the input of the neuron. Such an approach shows the process of reading a text by a reader because the analysis of current information should be based on previously retrieved knowledge. A recurrent neural network has been created. The training process of the network on the set of Ukrainian scientific articles has been performed. In order to increase the accuracy of the method, the previous processing of articles has been made. The articles contained incorrect symbol sequences due to the automated extraction of its content from PDF files. The analysis of method accuracy has been implemented basing on experimental examination of the method on document discrimination task and insertion task. The results obtained can indicate the method based on recurrent neural networks can be used for the coherence estimation of Ukrainian texts.*
Keywords: *text coherence, distributed sentence representation, recurrent neural network, neurons with a feedback connection, Word2Vec model.*

DOI: 10.34121/1028-9763-2019-4-9-16

1. Вступ

Постійна динаміка зростання потужностей обчислювальних систем вказує на актуальність вирішення AI-повних проблем – задач, складність яких еквівалентна головному завданню штучного інтелекту (створити системи прийняття рішень аналогічно людському мозку). До класу AI-повних задач варто віднести обробку природної мови (natural language processing – NLP). Галузь NLP поєднує напрями штучного інтелекту і комп'ютерної лінгвістики, а саме, вивчає проблеми автоматизованого аналізу та синтезу природної мови: машинний переклад, інформаційний пошук, генерування та розпізнавання мовлення тощо. Вирішення задач такого типу здійснюється за допомогою використання методології машинного навчання чи різних моделей прийняття рішень. Однією з базових задач обробки природної мови є здійснення автоматизованої оцінки когерентності тексту. Під когерентністю (цілісністю) тексту розуміють міру зв’язності речень, пов’язаних між собою у логічний і синтаксичний спосіб. Когерентний текст адаптований для зручного читання, тому дозволяє краще сприймати інформацію, яку намагається донести автор до читача. Задача оцінки когерентності тексту є актуальною для відбору необхідних даних серед множини текстової інформації, зважаючи на постійну динаміку її приросту. Аналіз цілісності текстової інформації дозволяє водночас авторам підвищити якість статей відповідно до читацької аудиторії, а також є допоміжним етапом визначення релевантних матеріалів пошуковими системами. Незважаючи на існування терміна «когерентність тексту», немає загальноприйнятого способу визначення його значення. Різні науковці пропонують свої підходи для обрахування когерентності через визначення складних функцій чи представлення тексту у вигляді деревовидних структур і графів [1–4]; інші підходи передбачають використання методів ма-

шинного навчання [5–7]. Наявність актуальних робіт щодо визначення когерентності тексту свідчить про важливість дослідження методів вирішення цієї задачі.

Більшість методів автоматизованої оцінки когерентності пропонується для розрахунку цілісності англомовних текстів. У роботі [5] було досліджено використання графів семантичної схожості для відстеження впливу семантичної узгодженості речень на загальну оцінку цілісності україномовних текстів. Незважаючи на активний розвиток досліджень у напрямі обробки природних мов, дослідження когерентності для україномовних текстів поки знаходиться на початковому етапі.

Метою цієї роботи є аналіз методів оцінки цілісності тексту за допомогою нейронних мереж та експериментальна перевірка методу на основі рекурентних мереж на корпусі української мови.

2. Методи вирішення задачі оцінки когерентності текстів

Для оцінки когерентності текстів були запропоновані моделі, основані на аналізі граматичних і семантичних властивостей елементів тексту. Метод Entity Grid [1] полягає у виділенні ключових сутностей у реченні та подальшій оцінці частоти зміни їх ролей. RST-аналіз [2] є модифікацією методу Entity Grid: здійснюється аналіз зміни дискурсивних ролей сутностей. У методі Entity Graph [3] відслідковується зміна ролей сутності тексту у графічний спосіб. Ці методи потребують відповідних експертних знань, якість яких має суттєвий вплив на вихідний результат. Для уникнення цієї залежності застосовують інші підходи, що передбачають отримання необхідних закономірностей за допомогою навчання нейронної мережі. Нейронні мережі та інші методи машинного навчання широко використовуються в галузі комп'ютерної лінгвістики. Незважаючи на залежність якості навчання мережі від навчальних даних, застосування вибірки з достатньою кількістю прикладів дозволяє отримати узагальнюючу модель оцінки цілісності текстів. Для автоматизованої оцінки когерентності тексту використовуються методи з різноманітною архітектурою нейронних мереж: згорткова мережа [6], рекурентна і рекурсивна мережі [7]. Використання таких типів мереж обумовлене можливістю здійснювати обробку вхідних даних нефіксованого розміру. Застосування згорткових мереж дозволяє аналізувати семантичні канали вхідних речень окремо з подальшим об'єднанням отриманих результатів. Процес оцінки когерентності тексту поділяється на 2 етапи: формування векторного представлення речень та власне розрахунок міри когерентності. Для здійснення векторного представлення речень використовується навчена модель Word2Vec, що здійснює перетворення речення у матричну форму. Подальше трансформування матричної форми нефіксованого розміру до векторного представлення здійснюється за допомогою шарів згортки і субдискретизації, що за допомогою карт ознак дозволяють виділити різні властивості вхідних даних. Далі виконується застосування повнозв'язних шарів і функції softmax для формування кінцевого результату – оцінки цілісності вхідних речень.

Під час роботи згорткової мережі виконується пряме проходження сигналу від входу до виходу. Таким чином, кожний окремий прохід не залежить від попереднього. Проводячи аналогію із процесом читання, можна вважати, що кожне наступне слово є незалежним від попереднього. Однак нейрони головного мозку сприймають і аналізують кожне наступне слово тексту на основі вже прочитаного. Прочитавши частину речення, можна зрозуміти значення слова в даному контексті або, навіть, його передбачити. Для вирішення цієї проблеми використовують нейрони зі зворотним зв'язком, які мають додаткові входи зі значеннями, отриманими на попередньому кроці роботи. Такі нейрони «пам'ятають» попередні значення і можуть бути використані довільну кількість разів за один прохід сигналу. Ця властивість має такі переваги:

• можливість подавати на вхід мережі дані з нефіксованим розміром, що дозволяє здійснювати обробку речень різної довжини;

• зменшення кількості вільних параметрів нейронної мережі, адже можливе багаторазове використання одного нейрона.

Нейронні мережі, що використовують наведені вище нейрони, називаються рекурентними [8] і широко застосовуються для вирішення задач, пов'язаних з обробкою текстів. Таким чином, можна зробити висновок про доцільність застосування методу розподіленого представлення речень (англ. distributed sentence representation) з використанням рекурентної нейронної мережі для оцінки когерентності тексту.

3. Метод розподіленого представлення речень за допомогою рекурентної нейронної мережі

Початковим етапом роботи методу є здійснення попередньої обробки вхідного тексту: токенізації (розподілення тексту на окремі речення; з кожного речення формується набір слів). Далі виконується семантичне представлення слова у вигляді вектора розмірністю K за допомогою навченої моделі Word2Vec чи GloVe. Таким чином, кожне речення s може бути представлене як набір векторів:

$$s = \left\{ \mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{n_s} \right\}, \tag{1}$$

де n_s – кількість слів речення. Наступним кроком є здійснення векторного представлення речення. Цей етап є ключовим у роботі методу, адже для його реалізації використовується рекурентний шар мережі. На рис. 1 зображений приклад обробки вхідного речення за допомогою нейронів рекурентного шару.

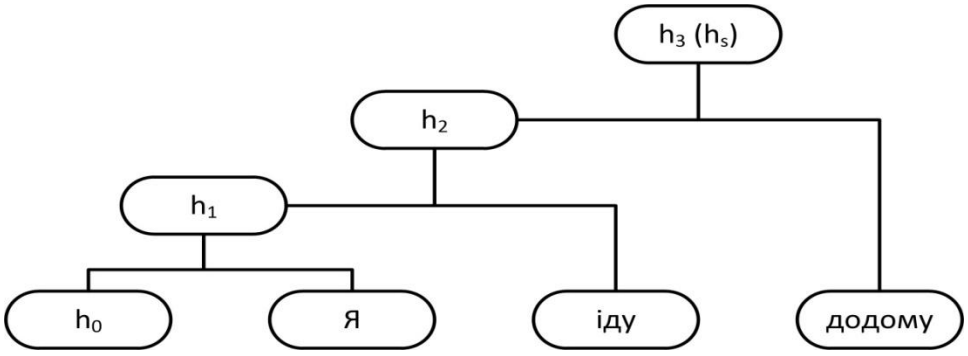

Рисунок 1 – Векторне представлення речення за допомогою рекурентного шару мережі

Значення вектора $\mathbf{h}_t$ у відповідний момент часу обраховується у такий спосіб:

$$\mathbf{h}_t = f(V_{\text{Recurrent}} \cdot \mathbf{h}_{t-1} + W_{\text{Recurrent}} \cdot \mathbf{w}^t + \mathbf{b}_{\text{Recurrent}}), \tag{2}$$

де $V_{\text{Recurrent}}$ і $W_{\text{Recurrent}}$ – матриці вільних параметрів розмірністю $K \times K$, $\mathbf{b}_{\text{Recurrent}}$ – вектор зсуву, $f = \tanh$ – нелінійна функція активації.

Наступним кроком є об'єднання речення у групи – «вікна» речень фіксованої довжини. Результатом об'єднання речень є конкатенація їх векторів у вектор $\mathbf{h}_c$ розмірністю $(L \times K) \times 1$, де L – кількість речень у «вікні». Значення кількості речень є фіксованим: $L = 3$. Далі вектор $\mathbf{h}_c$ подається на вхід повнозв'язних шарів, вихідним результатом яких є значення міри когерентності «вікна» y_c. Оцінка когерентності тексту D розраховується як добуток мір цілісності всіх «вікон», сформованих у тексті:

$$S_D = \prod_{c \in D} y_c \,. \tag{3}$$

Чим більше значення S_D, тим вища оцінка когерентності тексту.

4. Структура рекурентної нейронної мережі

Для реалізації зазначеного в попередньому розділі методу було створено рекурентну нейронну мережу. Кількість вхідних шарів рівна розміру «вікна». Враховуючи, що нейронна мережа повинна працювати з реченнями довільної довжини, було використано додатковий шар маскування, що дозволяє керувати розмірністю вхідних даних відповідно до заданої маски. Таким чином, у межах групи, що подається на вхід нейронної мережі, відбувається пошук речення з максимальною довжиною (кількість слів у реченні). Далі відбувається доповнення всіх речень у групі значеннями маски для регулювання довжини відповідно до максимального значення. Під час проходження сигналу через нейронну мережу шар маскування фільтрує необхідні значення, відтворюючи початкову форму речення.

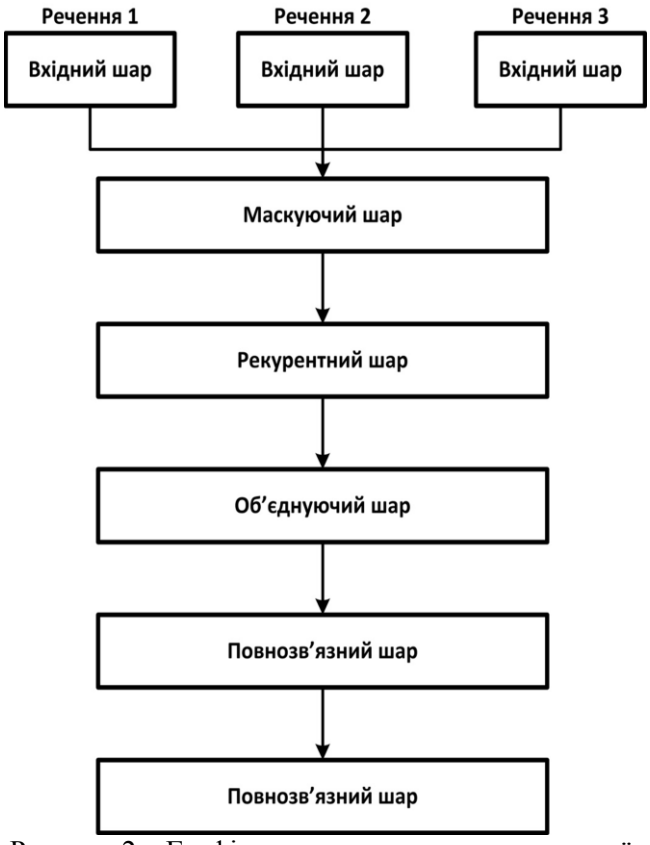

Рисунок 2 – Графічне представлення рекурентної нейронної мережі для оцінки когерентності текстів

Застосований у роботі віконний підхід означає, що нейронна мережа приймає та аналізує декілька речень одночасно. Тому було використано декілька вхідних шарів для передачі вхідних значень до шару маскування. Далі сигнал проходить до спільного рекурентного шару. Здійснюється конкатенація результату роботи рекурентного шару; результуючий вектор подається на вхід повнозв'язного шару. Графічне представлення нейронної мережі наведено на рис. 2.

5. Підготовка даних та навчання нейронної мережі

Попередня обробка тексту передбачає здійснення лематизації і векторного представлення тексту. Токенізацію україномовних текстів та лематизацію слів виконано за допомогою утиліти LanguageTool API NLP UK [9], яка має програмний інтерфейс для мови Python. Векторне представлення слів здійснено за допомогою навченої моделі Word2Vec [10].

Важливим етапом попередньої обробки текстів є усунення небажаних символів у тексті. Експериментальне дослідження методу здійснювалося на множині україномовних наукових статей. Текстова інформація статей була отримана шляхом екстракції даних із відповідних PDF-документів [11], тому тексти містили некоректні послідовності символів. Певні частини речень трактувалися як кінець речення, наприклад, точка у скороченнях ініціалів, позначеннях дат (років, століть), дробових числах, уточненнях, посиланнях на таблиці і рисунки тощо. Наведені послідовності символів некоректно інтерпретувалися моделлю векторного представлення слів, а також не були пов'язані з іншими реченнями, то-

му було вирішено їх вилучати. Для цього було створено відповідний програмний модуль та виправлено частину цих недоліків. Крім того, вирішено вилучати некоректно розпізнані таблиці та колонтитули статей, що відображалися як єдиний рядок або один символ без смислового навантаження. Схожі проблеми виникали під час обробки відформатованих списків.

Для навчання мережі було створено дві множини: навчальну і перевірочну. Навчання здійснювалося на 4091 файлі; відношення навчальної вибірки до перевірочної складало 70/30. Для уникнення перенавчання було застосовано метод раннього зупину, що дозволяє відстежити перенавчання мережі та зберегти найоптимальніший варіант ваг мережі. Динаміку навчання мережі представлено на рис. 3 і рис. 4, а саме зображено графіки значення точності та функції витрат залежно від епохи, відповідно. Крива функції витрат для перевірочної множини вказує, що з десятої епохи розпочався процес перенавчання.

Для створення програмних модулів використано мову програмування Python 3.6. Рекурентну нейронну мережу реалізовано за допомогою бібліотеки Keras – прикладного програмного інтерфейсу для бібліотек TensorFlow, CNTK і Theano. Для пришвидшення навчання було застосовано версію TensorFlow для графічних процесорів – tensorflow-gpu, що використовує бібліотеку NVIDIA CUDA Deep Neural Network (cuDNN).

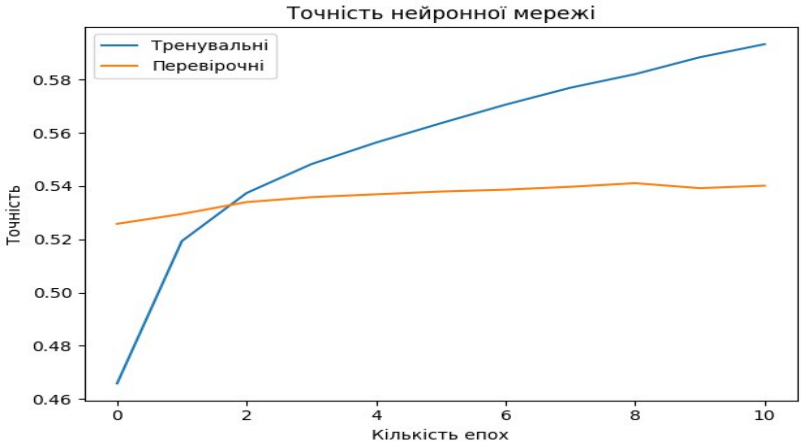

Рисунок 3 – Залежність точності нейронної мережі від порядкового номеру епохи навчання

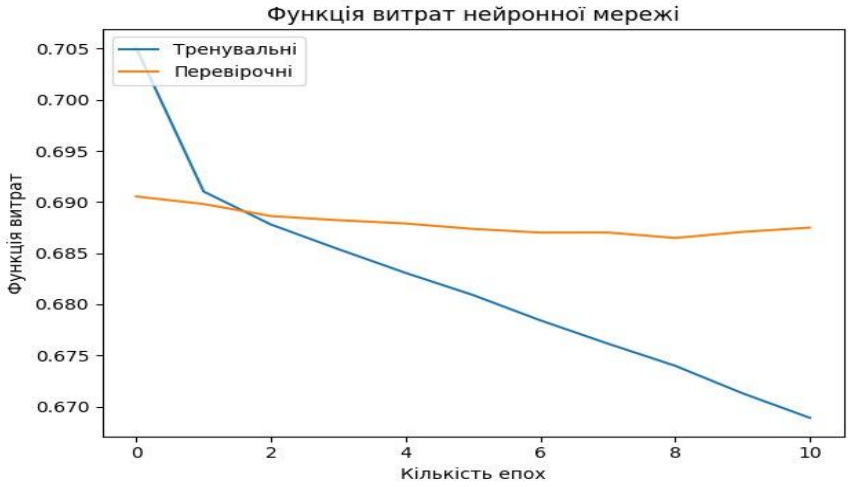

Рисунок 4 – Залежність значення функції витрат від порядкового номеру епохи навчання

6. Експериментальне дослідження точності методу

Після навчання рекурентної нейронної мережі було показано здатність методу вирішувати задачі розрізнення документів та вставки.

Задача розрізнення документів (англ. document discrimination task): для кожного тексту обраховується значення когерентності. Далі випадковим чином здійснюється зміна порядку розташування речень у тексті. Для отриманого тексту обраховується значення когерентності і порівнюється з відповідним значенням оригінального тексту. Якщо значення когерентності для оригінального тексту виявилось вище, ніж для зміненого, вважається, що метод успішно виконав задачу розрізнення для вибраного тексту. Оцінка точності методу розраховується, як відношення успішно оброблених документів до їх загальної кількості.

Задача вставки така: для кожного тексту випадковим чином вибирається речення. Це речення вилучається з тексту, а потім вставляється на кожну можливу позицію в тексті. Для кожного з отриманих варіантів тексту обраховується значення когерентності. Текст вважається успішно обробленим, якщо оригінальний порядок речень має найвище значення цілісності порівняно зі зміненими версіями. Оцінка точності методу розраховується аналогічно до задачі розрізнення документів.

Для розрахунку наведених вище метрик використано 1226 файлів. Результати обчислень наведено в табл. 1.

Наведені в таблиці результати свідчать про можливість застосування методу на основі рекурентної нейронної мережі для оцінки когерентності україномовних текстів. Відмінність оцінок задач розрізнення документів і вставки полягає у виборі підходу до навчання нейронної мережі, аналогічному задачі розрізнення; крім того, задача вставки розглядає більше варіантів порівнянь з оригіналом тексту. На точність методів також впливають формат вхідних текстових даних та якість їх попередньої обробки. В роботі використовувалися тексти наукових статей, отримані в результаті їх програмної екстракції з PDF-документів. Науковий стиль написання, що включає наявність формул і таблиць, а також похибка обробки PDF-документів значно ускладнюють задачу і певною мірою впливають на результати роботи методу.

Таблиця 1 – Вирішення задач розрізнення та вставки за допомогою методу на основі рекурентної нейронної мережі

Задача	Тип текстів	Кількість успішно оброблених текстів	Загальна кількість текстів	Точність, %
Розрізнення документів	навчальні і тестові	3537	4091	87
	тестові	980	1226	80
Вставки	навчальні і тестові	559	4091	14
	тестові	104	1226	9

7. Висновки

Проведено порівняльний аналіз основних методів оцінки когерентності текстів та здійснено експериментальну перевірку методу, що ґрунтується на основі рекурентних нейронних мереж, на множині україномовних наукових статей. Перевагою використання методів машинного навчання на основі нейронних мереж є відсутність залежності від експертної попередньої обробки вхідних даних. Зважаючи на наявність зворотного зв’язку в архітектурі рекурентної мережі, доцільно використовувати цей тип мережі. Нейрони зі зворотним

зв’язком певною мірою відтворюють процес сприйняття читачем текстової інформації: аналіз попередніх даних дозволяє краще зрозуміти контекст вживання поточного слова чи передбачити наступне.

Здійснено експериментальну перевірку ефективності роботи методу на основі рекурентної нейронної мережі для множини україномовних текстів. З отриманих результатів можна зробити висновок щодо можливості використання цього методу для здійснення оцінки когерентності текстів. Точність роботи методу може бути покращена за допомогою вдосконалення існуючих інструментів попередньої обробки текстової інформації та збільшення кількості вільних параметрів мережі.

СПИСОК ДЖЕРЕЛ

1. Barzilay R., Lapata M. Modeling local coherence: An entity-based approach. *Computational Linguistics*. 2008. Vol. 34, N 1. P. 1–34.
2. Feng V.W., Lin Z., Hirst G. The impact of deep hierarchical discourse structures in the evaluation of text coherence. *Proc. of the 25th International Conference on Computational Linguistics (COLING 2014)*. 2014. P. 940–949.
3. Guinaudeau C., Strube M. Graph-based local coherence modeling. *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*. 2013. Vol. 1. P. 93–103.
4. Погорілий С.Д. Програмне конструювання: підручник серії «Автоматизація наукових досліджень» / ред. академіка АПН України О.В. Третяка. Київ: ВПЦ, Київський університет, 2007. 438 с.
5. Погорілий С.Д., Крамов А.А. Метод розрахунку когерентності українського тексту. *Реєстрація, зберігання і обробка даних*. 2018. № 4. С. 64–75.
6. Cui B., Li Y., Zhang Y., Zhang Z. Text Coherence Analysis Based on Deep Neural Network. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, 6–10 November 2017). Singapore, 2017. P. 2027–2030.
7. Li J., Hovy E. A model of coherence based on distributed sentence representation. *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar, 25–29 October 2014). Doha, Qatar, 2014. P. 2039–2048.
8. Хайкин С. Нейронные сети: полный курс. 2-е изд. Киев, 2016. 1104 с.
9. LanguageTool API NLP UK. URL: https://github.com/brown-uk/nlp_uk (дата звернення: 13.10.2019).
10. Моделі: lang-uk. URL: http://lang.org.ua/uk/models (дата звернення: 13.10.2019).
11. Pogorilyy S., Kramov A. Automated extraction of structured information from a variety of web pages. *Proc. of the 11th International Conference of Programming UkrPROG 2018* (Kyiv, Ukraine, 22–24 May 2018). Kyiv, Ukraine, 2018. P. 149–158.

Стаття надійшла до редакції 23.10.2019