УДК 004.83

д.т.н., проф. **Погорілий С.Д.** (КНУ)
Крамов А.А. (КНУ)
Білецький П.В. (КНУ)

DOI: https://doi.org/10.17721/2519-481X/2019/65-08

МЕТОД ОЦІНКИ КОГЕРЕНТНОСТІ УКРАЇНОМОВНИХ ТЕКСТІВ З ВИКОРИСТАННЯМ ЗГОРТКОВОЇ НЕЙРОННОЇ МЕРЕЖІ

Задача оцінки когерентності тексту є однією із актуальних задач комп'ютерної лінгвістики. Аналіз цілісності текстової інформації використовується для написання та відбору документів, що дозволяють передати читачу ідею автора у зрозумілий спосіб. Важливість цієї задачі підтверджує наявність актуальних робіт, присвячених її вирішенню. Автоматизовані методи оцінки цілісності тексту основані на методології машинного навчання, що полягають у формалізованому представленні тексту та подальшому виявленні закономірностей для формування вихідного результату. Метою роботи є аналітичний огляд різних методів оцінки цілісності тексту; обґрунтування вибору методу та здійснення його адаптації відповідно до особливостей української мови; виконання експериментальної перевірки ефективності роботи пропонованого методу для україномовного корпусу.

В роботі здійснено порівняльний аналіз методів оцінки когерентності англомовних текстів на основі методології машинного навчання. За результатами проведеного аналізу обґрунтовано доцільність застосування методів з використанням попередньо навчених універсальних моделей формалізованого представлення елементів тексту. До таких методів відносяться моделі на основі нейронних мереж різної архітектури: рекурентні та згорткові мережі. Такі типи мереж використовуються для обробки текстів, адже дозволяють здійснювати обробку вхідних даних нефіксованого розміру – речень чи слів. Незважаючи на властивість рекурентних мереж враховувати попередні дані, що певним чином відтворює процес сприйняття інформації читачем, для проведення експериментального дослідження обрано згорткову нейронну мережу. Такий вибір обумовлений здатністю згорткових мереж відслідковувати зв'язки між сутностями незалежно від відстані між ними. В роботі детально описано принцип роботи методу на основі згорткової нейронної мережі, розглянуто її архітектуру. Для перевірки ефективності роботи розглянутого методу на множині україномовних текстів створено застосування з використанням згорткової нейронної мережі. Формалізоване представлення елементів тексту здійснено за допомогою попереднього навчання моделі семантичного представлення слів на корпусі україномовних анотацій наукових статей. Виконано навчання сформованої мережі з використанням навченої моделі. Проведено експериментальну перевірку ефективності роботи методу на множині наукових статей для вирішення задач розрізнення документів і вставки. На основі отриманих результатів можна зробити висновок про доцільність використання згорткової нейронної мережі для оцінки когерентності україномовних текстів.

Ключові слова: когерентність тексту, згорткова нейронна мережа, семантична узгодженість речень, задача розрізнення документів, задача вставки.

Вступ. З постійним зростанням потужності обчислювальних систем з'явилася можливість автоматизованого вирішення задач, що вирішувались людиною власноруч. До таких задач належить лінгвістичний аналіз, що полягає у дослідженні різнорівневих аспектів усного та писемного мовлення людини. Автоматизований лінгвістичний аналіз здійснюється засобами комп'ютерної лінгвістики. Підкласом лінгвістичного аналізу є задача дослідження семантичної складової текстів. Семантичний аналіз дозволяє з'ясувати смислове значення тексту та його складових, знаходячи зв'язки між реченнями та словами. В галузі комп'ютерної лінгвістики цей аналіз розглядається як задача здійснення формалізованого представлення складових частин тексту та подальше дослідження їх взаємозв'язку.

Однією з задач семантичного аналізу є визначення ступеня пов'язаності речень тексту між собою – когерентності тексту. Під когерентністю текстів розуміють наявність семантичного, стилістичного та граматичного зв'язків між елементами тексту, що визначають його цілісність [1]. Оцінка когерентності дозволяє визначати структурну узгодженість тексту, виявляти логічний взаємозв'язок семантично схожих слів, який тематично пов'язує речення

між собою. Показник когерентності також певною мірою характеризує ступінь зв'язності речень – когезію тексту.

Зважаючи на невелику кількість робіт, що досліджують методи оцінки когерентності україномовних текстів, актуальною є задача автоматизованого розрахунку ступеню цілісності текстів різної тематики україномовного корпусу. Метою цієї роботи є дослідження існуючих методів оцінки когерентності текстів та здійснення експериментальної перевірки ефективності використання методу на основі згорткових нейронних мереж для україномовних текстів.

Аналіз методів оцінки когерентності тексту. Для оцінки ступеню когерентності текстів використовуються різні моделі представлень текстів та засоби машинного навчання. Методи *Entity Grid* та *Entity Graph* основані на дослідженні закономірностей зміни ролей сутностей (іменних груп, власних назв) у тексті. Розглядаються три різні варіанти сутності: об'єкт, суб'єкт та відсутність ролі. Метод Entity Grid [2] полягає в наступному: формується множина векторів ознак на основі модифікації ролі кожної сутності в різних частинах тексту. Далі за допомогою методів машинного навчання, а саме методу опорних векторів, здійснюється навчання моделі класифікатора. Вихідним значенням моделі є бінарне значення, що інтерпретується у наступний спосіб: когерентний текст чи ні. Відмінність методу Entity Graph [3] полягає у побудові двочасткового орієнтованого графу з вершин сутностей і речень на основі їх ролей; далі здійснюється трансформація отриманого графу до орієнтованого проекційного графу. Цілісність тексту розраховується як усереднене значення напівстепені виходу графу. На відміну від методу Entity Grid, вихідне значення методу Entity Graph є дійсним числом.

Побудова *графу семантичної схожості* здійснює аналіз семантичної узгодженості речень тексту за допомогою попередньо навчених моделей векторного представлення слів і речень [4]. Використання графічного підходу дозволяє відслідковувати процес розрахунку оцінки цілісності тексту, тобто з'ясувати причину отримання вихідного результату. Це допомагає корегувати якість тексту відповідно до вихідного значення методу.

Недоліком розглянутих методів є залежність від попередньої обробки тексту – пошук сутностей, розмітка їх ролей, векторне представлення слів та речень. Такий підхід ускладнює використання цих методів для різних мов. Для уникнення цього обмеження використовуються методи на основі нейронних мереж із застосуванням універсальних моделей попередньої обробки тексту та його складових. Архітектура нейронних мереж відповідних методів містить рекурентні і згорткові шари, що обумовлено обробкою вхідних даних з нефіксованою розмірністю (речень з різною кількістю слів). *Метод, оснований на використанні рекурентної мережі* [5, 6], полягає у послідовному здійсненні векторного представлення вхідних речень та подальшому виконанні регресійної моделі (послідовності повнозв'язних шарів) для отримання вихідної оцінки. Застосування власне рекурентних шарів обумовлене наявністю зворотного зв'язку у нейронах шарів цього типу. Зворотній зв'язок прослідковується також під час аналізу тексту читачем: поточна інформація сприймається на основі попередньо отриманих знань. Недоліком методу на основі рекурентної нейронної мережі є залежність ефективності пошуку взаємозв'язку семантично близьких сутностей залежно від відстані в тексті: чим більше слів між сутностями, тим складніше виявити взаємозв'язок між ними.

Метод на основі згорткової нейронної мережі [7–10] дозволяє відслідковувати зв'язки між сутностями незалежно від їхніх позицій в тексті. Крім того, використання багатоканальної структури дозволяє в подальшому розширити кількість та тип вхідних даних. Розглянемо більш детально принцип роботи цього методу та архітектуру мережі.

Метод оцінки ступеню когерентності текстів за допомогою згорткової нейронної мережі. Вхідними даними методу є україномовний текст, вихідними – оцінка когерентності методу. Спочатку виконується попередня обробка тексту, а саме:

- токенізація – розбиття тексту на речення і слова;
- лематизація слів – приведення іменників до називного відмінку та дієслів до форми інфінітиву;
- здійснення векторного представлення слів.

Векторне представлення слів здійснюється за допомогою попередньо навченої моделі Word2Vec [11]; розмірність вектору $d = 300$. Формалізація речення відбувається у наступний спосіб:

$$S = \left\{ \mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{|S|} \right\}, \quad (1)$$

де $|S|$ – кількість слів в реченні; $\mathbf{w}_i \in \mathrm{R}^d$ – векторне представлення слова речення, $i = 1..|S|$. Отже, дані подаються на вхід нейронної мережі у формі матриці. Здійснення розрахунку цілісності тексту відбувається на основі аналізу семантичної узгодженості речень між собою. На вхід мережі подається угрупування речень («вікно»); кількість елементів групи фіксована: $L = 3$. Кількість елементів угрупування може змінюватися, однак збільшення розміру «вікна» істотно не впливає на результат. Для аналізу семантичної узгодженості речень «вікна» за допомогою повнозв'язних шарів мережі потрібно виконати векторне представлення речень. Таким чином, наступним кроком є здійснення перетворення матричної форми речення до векторної для подальшого аналізу взаємозв'язку речень тексту. Для цього використовується операція згортки, що застосовує 2 послідовних шари: згортковий і субдискретизації (див. рис. 1).

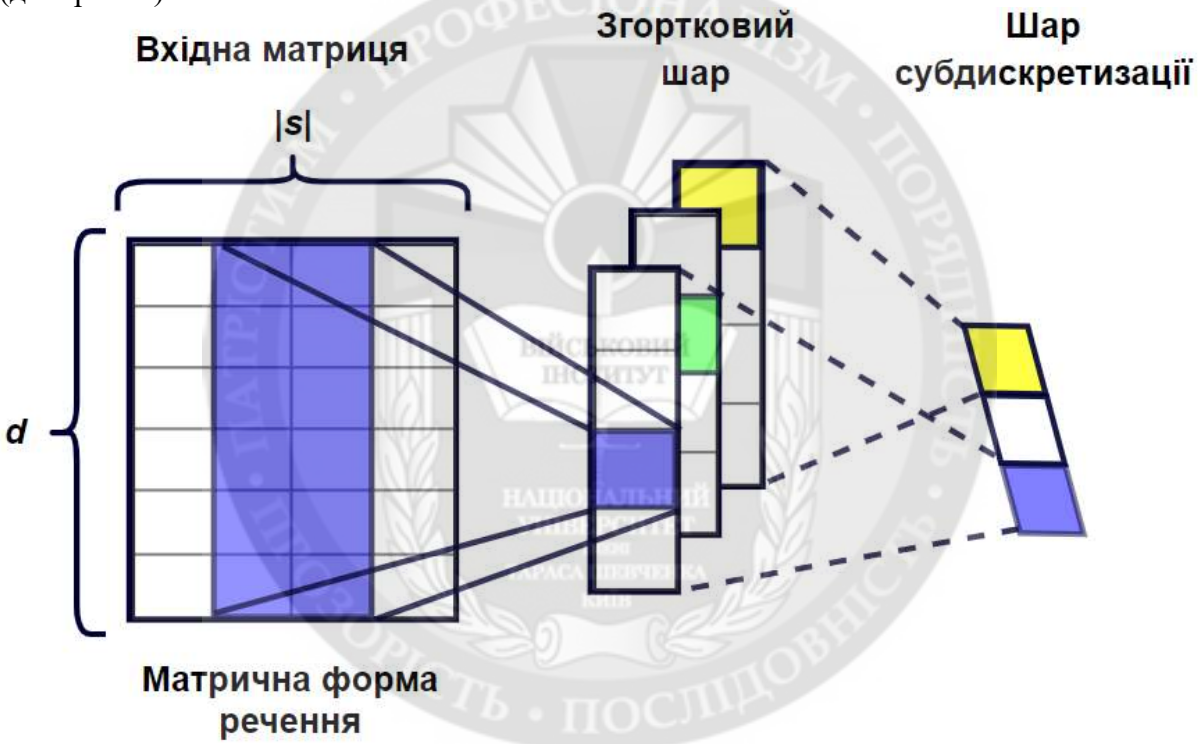

Рисунок 1 – Застосування операції згортки до матриці речення для формування вектору ознак

Згортковий шар здійснює виділення ознак з набору речень, використовуючи фільтр – матрицю ваг $F \in \mathrm{R}^{d \times m}$, де m – кількість слів, що обробляються фільтром одночасно. В ході виконання дослідження визначено доцільне (відносно точності моделі та часу навчання) значення параметра $m : m = 3$. Фільтр рухається вздовж стовпців матриці S з одиничним кроком; результатом виконання кожного кроку є вектор $c \in \mathrm{R}^{|S|-m+1}$, кожен елемент якого обраховується у наступний спосіб:

$$c_i = (S * F)_i = \sum_{k,j} \left(S_{[:i-m+1:i]} \otimes F \right)_{kj}, \quad (2)$$

де $\otimes$ – операція попарного добутку елементів двох векторів.

Для виділення множини ознак застосовується набір таких фільтрів. Результатом роботи фільтрів є множина карт характеристик (див. «Згортковий шар» на рис. 1). До карт характеристик застосовується нелінійна функція-випрямляч $\max(0, x)$ (rectified linear unit, ReLU). Для виконання агрегації отриманих результатів використовується шар субдискретизації. Вихідна матриця характеристик C обробляється у наступний спосіб: з кожної колонки обирається найбільше значення. Таке перетворення можна представити у вигляді відображення (див. «Шар субдискретизації» на рис. 1):

$$\text{pool}(C): \mathrm{R}^{|S|-m+1} \to \mathrm{R}. \tag{3}$$

До отриманих векторів речень застосовується операція конкатенації, тобто перетворення набору векторів до одного. Результуючий вектор подається на вхід повнозв'язних шарів; вихідним результатом є оцінка цілісності групи речень, яка отримується шляхом застосування функції softmax до отриманого результату. Для кожної групи тексту обраховується ступінь її цілісності. Оцінка когерентності тексту в цілому здійснюється у наступний спосіб:

$$S_D = \prod_{q \in D} p(y_q = 1), \tag{4}$$

де $p(y_q = 1)$ – оцінка когерентності угрупування; $q \in D$ – множина угрупувань документу D. Для пари документів $\langle D_1, D_2 \rangle$ вірним є наступне твердження: якщо $S_{D_1} > S_{D_2}$, то документ D_1 більш когерентний, ніж документ D_2. На рис. 2 наведена архітектура створеної згорткової нейронної мережі.

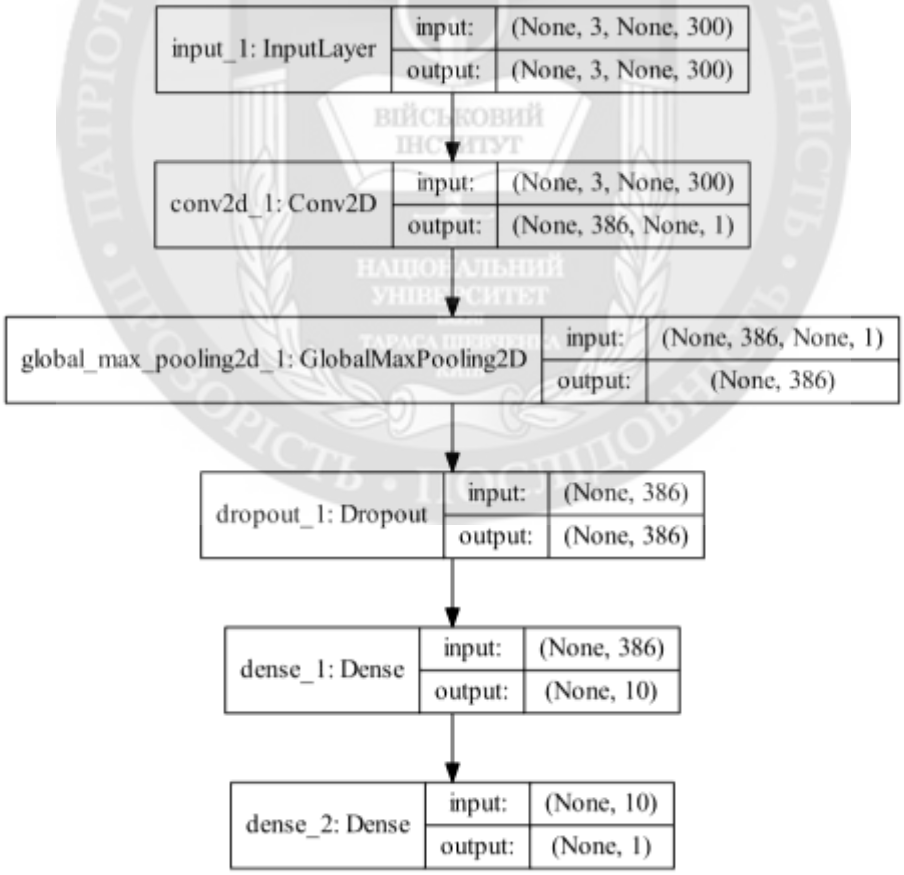

Рисунок 2 – Архітектура згорткової нейронної мережі для оцінки цілісності угрупування речень

Підготовка даних та навчання мережі. Навчання моделі семантичного векторного представлення слів Word2Vec виконано на множині україномовних анотацій наукових статей різних галузей. Навчальну та тестові вибірки сформовано з множини повних версій наукових статей, написаних українською мовою. Повні версії статей було отримано шляхом їх автоматизованої екстракції з веб-сайтів наукових журналів [12] та подальшої обробки відповідних PDF-файлів, тому тексти статей містили різні послідовності символів, які не мали смислового навантаження. Автоматизованим програмним способом виправлено частину помилок в оригінальних файлах, зокрема, таких як злиття слів; видалено частину таблиць, що до цього ідентифікувалися як речення. Варто звернути увагу на процес формування маркованого набору навчальних даних, а саме на приклади некогерентних угрупувань. Спочатку для отримання некогерентних груп речень виконувалось перемішування речень в межах тексту. Однак такий підхід формування прикладів виявився неефективним під час розрахунку метрик оцінки точності роботи методу. Заміна такого підходу на перемішування речень в межах кожної групи дозволила значно підвищити ефективність навчання нейронної мережі, а також точність методу. Для уникнення перенавчання мережі додатково використано шар Dropout. Кількість нейронів в повнозв'язному шарі та карт характеристик було підібрано експериментальним чином.

Для створення застосування використано мову програмування Python 3.6. Формування архітектури мережі та виконання навчання реалізовано з використанням засобів бібліотеки Keras [13].

Експериментальна перевірка роботи методу. Для оцінки точності методу обраховано метрики для двох задач: вставки та розрізнення документів. Спільним початковим кроком вирішення задач є оцінка когерентності вихідного тексту.

В задачі розрізнення наступним кроком є оцінка когерентності тексту, в якому порядок речень змінено випадковим чином. Якщо міра когерентності початкового тексту більше, ніж у модифікованого, текст вважається розпізнаним правильно.

Наступним кроком для задачі вставки є отримання оцінки когерентності тексту, в якому випадково вибране речення послідовно вставляється в усі можливі позиції в тексті. Якщо міра когерентності початкового тексту вища, ніж когерентність для всіх модифікацій, текст вважається розпізнаним правильно.

Точність вирішення задач обраховується як відношення кількості коректно визначених прикладів до їх загальної кількості. В таблиці 1 наведено результати вирішення розглянутих вище задач.

Таблиця 1

Результати роботи методу залежно від параметрів нейронної мережі

Параметри мережі та варіанти формування некогерентних прикладів вибірки	Точність вирішення задачі вставки, %	Точність вирішення задачі розрізнення, %
Перемішування тексту в цілому	0.79	100.00
Перемішування в межах «вікна»	12.66	97.67
Перемішування в межах «вікна» з вдосконаленими параметрами нейронної мережі	13.64	99.44

Точність вирішення задачі розрізнення документів досягає значення більше 97%, що обумовлено здійсненням формування навчальної вибірки нейронної мережі у спосіб, аналогічний до алгоритму обрахунку відповідної метрики задачі. Задача вставки складніша порівняно із задачею розрізнення документів. Завдання такого типу можна зустріти у різноманітних тестах оцінювання знань іноземної мови, де потрібно знайти оригінальну позицію речення в тексті. На точність вирішення задачі вставки впливає стиль написання вхідного тексту (наявність орфографічних та граматичних помилок, непослідовність викладення матеріа-

лу). Здійснення зміни порядку речень в межах «вікна» та вдосконалення параметрів мережі дозволили підвищити точність вирішення задачі вставки. З отриманих результатів можна зробити висновок про доцільність використання згорткових нейронних мережі для оцінки когерентності україномовних текстів.

Висновки. В роботі здійснено порівняльний аналіз автоматизованих методів оцінки цілісності текстів, обґрунтовано доцільність застосування методу на основі згорткової нейронної мережі та виконано експериментальне дослідження ефективності роботи навченої мережі для вирішення задач розрізнення документів і вставки. На основі проведеного аналізу та отриманих результатів експериментальної перевірки методу можна зробити наступні висновки:

- доцільним є використання методів оцінки цілісності тексту на основі попередньо навчених універсальних семантичних моделей; основною перевагою такого підходу є відсутність залежності методу від експертних знань для певної мови;
- застосування згорткових нейронних мереж дозволяє здійснювати обробку вхідних даних нефіксованого розміру – речень тексту; перевагою такого типу мережі порівняно з рекурентною є можливість відслідковувати взаємозв'язки елементів тексту незалежно від відстані між ними;
- результати вирішення задач розрізнення документів та вставки свідчать про можливість використання навченої мережі для оцінки когерентності україномовних текстів;
- точність методу може бути підвищена за рахунок формування навчальної вибірки з текстів, призначених для завдань вставки речення в оригінальну позицію тексту; такий підхід дозволить водночас збільшити точність виконання задачі вставки, а також зберегти поточні показники ефективності вирішення задачі розрізнення документів.

ЛІТЕРАТУРА:

1. Grosz B. J., Weinstein S., Joshi A. K. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*. 1995. Vol. 21. No 2. P. 203–225.
2. Barzilay R., Lapata M. Modeling local coherence: An entity-based approach. *Computational Linguistics*. 2008. Vol. 34, No 1. P. 1–34.
3. Guinaudeau C., Strube M. Graph-based local coherence modeling. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 2013. Vol. 1. P. 93–103.
4. Погорілий С.Д., Крамов А.А. Метод розрахунку когерентності українського тексту. *Реєстрація, зберігання і обробка даних*. 2018. № 4. С. 64–75.
5. Li J., Hovy E. A model of coherence based on distributed sentence representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* October 25-29, 2014. Doha, Qatar. P. 2039–2048.
6. Mesnil G., He X., Deng L., Bengio Y. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. *INTERSPEECH 2013*. August 25-29, 2013. Lyon, France. P. 3771–3775.
7. 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*. November 6-10, 2017. Singapore, Singapore. P. 2027–2030.
8. Kim Y. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. October 2014. Doha, Qatar. P. 1746–1751.
9. Severyn A., Moschitti A. Learning to rank short text pairs with convolutional deep neural networks. *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. August 9-13, 2015. Santiago, Chile. P. 373–382.
10. Severyn A., Moschitti A. Modeling Relational Information in Question-Answer Pairs with Convolutional Neural Networks. *arXiv preprint arXiv:1604.01178*. 2016.
11. Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems*. December 5-10, 2013. Lake Tahoe, Nevada. P. 3111–3119.

*12.*Pogorilyy S., Kramov A. Automated extraction of structured information from a variety of web pages. In *11th International Conference of Programming UkrPROG, UkrPROG 2018*. May 22-24, 2018. Kyiv, Ukraine. P. 149–158.

13.Keras: The Python Deep Learning library. URL: https://keras.io (дата звернення: 26.09.2019).

REFERENCES:

1. Grosz, B., Weinstein, S. and Joshi, A.K. (1995). "Centering: A framework for modeling the local coherence of discourse". *Computational linguistics*, 21(2), pp. 203–225.

2. Barzilay, R. and Lapata, M. (2008). "Modeling Local Coherence: An Entity-Based Approach". *Computational Linguistics*, 34(1), pp. 1–34.

3. Guinaudeau, C. and Strube, M. (2013). "Graph-based local coherence modeling". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. pp. 93–103.

4. Pogorilyy, S.D. and Kramov, A.A. (2018). "Metod rozrakhunku kogherentnosti ukrajinsjkogho tekstu" [Method of the coherence evaluation of Ukrainian text], *Data Recording, Storage & Processing*, 20(4), pp. 64–75.

5. Li, J. and Hovy, E. (2014). "A model of coherence based on distributed sentence representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 2039–2048.

6. Mesnil, G., He, X., Deng, L. and Bengio, Y. (2013). "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding". In: *INTERSPEECH 2013*. pp. 3771–3775.

7. Cui, B., Li, Y., Zhang, Y. and Zhang, Z. (2017). "Text Coherence Analysis Based on Deep Neural Network". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. pp. 2027–2030.

8. Kim, Y. (2014). "Convolutional neural networks for sentence classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1746–1751.

9. Severyn, A. and Moschitti, A. (2015). "Learning to rank short text pairs with convolutional deep neural networks". In: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. pp. 373–382.

10.Severyn, A. and Moschitti, A. (2019). *Modeling Relational Information in Question-Answer Pairs with Convolutional Neural Networks*. [online] Arxiv.org. Available at: https://arxiv.org/pdf/1604.01178.pdf [Accessed 26 Sep. 2019].

11.Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems*. pp. 3111–3119.

12.Pogorilyy, S. and Kramov, A. (2018). "Automated extraction of structured information from a variety of web pages". In: *11th International Conference of Programming UkrPROG, UkrPROG 2018*. CEUR Workshop Proceedings, pp.149-158.

13.Keras.io. (2019). *Home - Keras Documentation*. [online] Available at: https://keras.io [Accessed 26 Sep. 2019].

д.т.н., проф. Погорелый С.Д., Крамов А.А., Билецкий П.В.

МЕТОД ОЦЕНКИ КОГЕРЕНТНОСТИ УКРАИНОЯЗЫЧНЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ

Задача оценки когерентности текста является одной из актуальных задач компьютерной лингвистики. Анализ целостности текстовой информации используется для написания и отбора документов, что позволяют понятным образом передать идею автора читателю. Важность этой задачи подтверждает наличие актуальных работ, посвященных её решению. Автоматизированные методы оценки целостности текста основаны на методологии машинного обучения, что заключаются в формализованном представлении текста и дальнейшем определении закономерностей для формирования выходного результата. Целью работы есть аналитический обзор разных методов оценки целостности текста; обоснование выбора метода и осуществление его адаптации соответственно с особенностями украинского языка; исполнение экспериментальной проверки эффективности работы предложенного метода для украиноязычного корпуса.

В работе осуществлен сравнительный анализ методов оценки когерентности англоязычных текстов на основе методологии машинного обучения. По результатам проведенного анали-

за обосновано целесообразность применения методов с использованием предварительно обученных универсальных моделей формализованного представления элементов текста. К этому методу относятся модели на основе нейронных сетей разной архитектуры: рекуррентные и сверточные сети. Такие типы сетей используются для обработки текстов, потому что позволяют осуществлять обработку входных данных с нефиксированным размером – предложений и слов. Несмотря на свойство рекуррентных сетей учитывать предыдущие данные, что определенным образом воспроизводит процесс восприятия информации читателем, для проведения экспериментального исследования выбрано сверточную нейронную сеть. Такой выбор обусловлен способностью сверточных нейронных сетей отслеживать связи между сущностями независимо от расстояния между ними. В работе подробно описан принцип работы метода на основе сверточной нейронний сети, рассмотрена её архитектура. Для проверки эффективности работы рассмотренного метода на множестве украиноязычных текстов создано приложение с использованием сверточной нейронной сети. Формализованное представление элементов текста осуществлено с помощью предварительного обучения модели семантического представления слов на корпусе украиноязычных аннотаций научных статей. Выполнено обучение сформированной сети с использованием обученной модели. Проведена экспериментальная проверка эффективности работы метода на множестве научных статей для решения задач различения документов и вставки. На основе полученных результатов можно сделать вывод о целесообразности использования сверточной нейронной сети для оценки когерентности украиноязычных текстов.

***Ключевые слова:** когерентность текста, сверточная нейронная сеть, семантическая согласованность предложений, задача различения документов, задача вставки.*

Prof. Pogorilyy S.D., Kramov A.A., Biletskyi P.V.

METHOD FOR COHERECE EVALUATION OF UKRAINIAN TEXTS USING CONVOLUTIONAL NEURAL NETWORK

The estimation of text coherence is one of the most actual tasks of computer linguistics. Analysis of text coherence is widely used for writing and selection of documents. It allows clearly conveying the idea of an author to a reader. The importance of this task can be confirmed by the availability of actual works that are dedicated to solving it. Different automated methods for the estimation of text coherence are based on the methodology of machine learning. Corresponding methods are based on of formal text representation and following detection of regularities for the generation of an output result. The purpose of this work is to perform the analytic review of different automated methods for the estimation of text coherence; to justify method selection and adapt it due to the features of the Ukrainian language; to perform the experimental verification of the effectiveness of the suggested method for a Ukrainian corpus.

In this paper, the comparative analysis of the methods for the estimation of coherence of English texts basing on a machine learning methodology has been performed. The expediency of application of methods that are based on trained universal models for the formalized representation of text components has been justified. The following models using neural networks with different architecture can be considered: recurrent and convolutional networks. These types of networks are widely used for text processing because they allow processing input data with an unfixed structure like sentences or words. Despite the ability of recurrent neural networks to take into account previous data (this behavior is similar to text perception by the reader), the convolutional neural network for conducting experimental research has been chosen. Such choice has been made due to the ability of convolutional neural networks to detect relations between entities regardless of the distance between them. In this paper, the principle of the method basing on the convolutional neural network and the corresponding architecture has been described. Program application for the verification of the suggested method effectiveness has been created. Formalized representation of text elements has been performed using a previously trained model for the semantic representation of words; the training process of this model has been implemented on the corpus of Ukrainian scientific abstracts. The training of the formed networks using pre-trained model has been performed. Experimental verification of method effectiveness for solving of document discrimination task and insert task has been made on the set of scientific articles. The results obtained may indicate that the method using convolutional neural networks can be used for further estimation of coherence of Ukrainian texts.

***Keywords:** text coherence, convolutional neural network, semantic consistency of sentences, document discrimination task, insert task.*