УДК 004.75

ПІБ авторів

НАЙМЕНУВАННЯ СТАТТІ

Анотація. Авторська анотація є коротким викладом основної ідеї роботи. Обсяг анотації обмежений в 400 знаків.
Abstract in English. Up to 400 characters with spaces.

Ключові слова: не більше 5 слів.

СТРУКТУРА СТАТТІ

Актуальність теми дослідження.
Постановка проблеми.
Аналіз останніх досліджень і публікацій.
Виділення недосліджених частин загальної проблеми.
Постановка завдання.
Викладення основного матеріалу.
Висновки.
Довідка про авторів (укр. та англ.).
Розширена анотація (укр. або англ., на мові відмінній від публікації).

Вимоги до оформлення

Основні вимоги до оформлення статті: шрифт: Times New Roman, текст статті набирається 14 розміром шрифту. При формуванні десяткових чисел використовувати тільки крапку – 0.95.

Рисунки та таблиці.

Основні вимоги до підпису рисунку: шрифт Times New Roman, розмір шрифту 14, підпис розміщується по центру під рисунком.

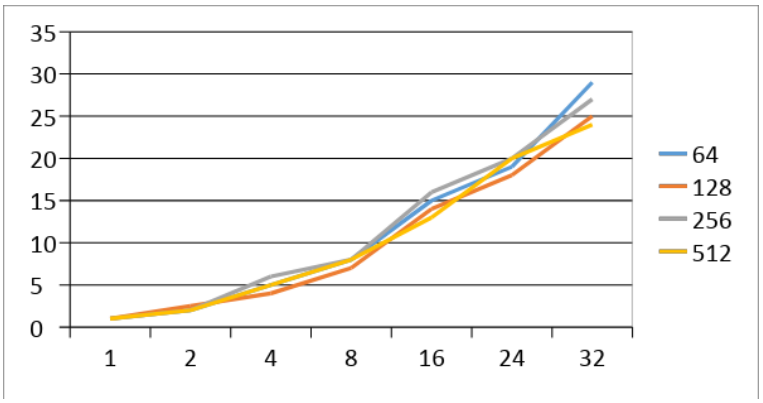

Рис. 1. Підпис до рисунку.

Приклад оформлення таблиці.

Таблиця 1

Заголовок таблиці

Заголовок	Заголовок	Заголовок	Заголовок
текст	текст	текст	текст
текст	текст	текст	текст
текст	текст	текст	текст

Математичні формули.

Скористуйтесь Equation у Word 2010 або вище. (Insert -> Equation, або комбінація гарячих клавіш Alt+=). Приклад розміщення формули:

$$y = a + b \quad (1)$$

Нумерація ставиться праворуч у дужках.

Список використаних джерел.

Список використаних джерел оформляти згідно з вимогами ДСТУ ГОСТ 8302-2015.

ПРИКЛАД СТАТТІ.

UDC 004.8

Володимир Фоменко, Георгій Луцький, Павло Регіда, Артем Волокита

ПИТАННЯ ГЕНЕРАЦІЇ ТЕМАТИЧНИХ ТЕКСТІВ НА ОСНОВІ РЕКУРЕНТНИХ НЕЙРОННИХ МЕРЕЖ ТА WORD2VEC

Volodymyr Fomenko, Heorhii Loutskii, Pavlo Rehida, Artem Volokyta

THEMATIC TEXTS GENERATION ISSUES BASED ON RECURRENT NEURAL NETWORKS AND WORD2VEC

У статті розглядається питання генерації псевдовипадкових текстів на задану тематику. Для генерації використовуються рекурентні нейронні мережі (LSTM) з попередньою обробкою слів за допомогою моделі word2vec. Тема тексту задається за допомогою набору ключових слів. Моделі тренуються на наборі російськомовних статей.

Ключові слова: генерація тексту, рекурентні нейронні мережі, довга короткочасна пам’ять, word2vec.

Рис.: 3. Табл.: 1. Бібл.: 13.

The paper deals with the issues of generating pseudo-random texts on a given topic. For generation the recurrent neural networks (LSTM) with preliminary pre-processing of words using the word2vec model are used. The text topic is assigned using a set of keywords. The models are trained on a dataset of Russian-language articles.

Key words: text generation, recurrent neural networks, long short-term memory, word2vec.

Fig.: 3. Tabl.: 1. Bibl.: 13.

Target setting. Due to the growing demand for automated generation of object descriptions, article excerpts, news summaries, etc., generation of thematic texts has become an actual topic in the recent years.

Actual scientific researches and issues analysis. In connection to the invention of new methods and approaches in the field of artificial intelligence, the topic of text generation has become more studied in recent years.

Uninvestigated parts of general matters defining. Despite a considerable number of works devoted to the application of recurrent neural networks for the text

generation, the problem of thematic text generation remains little investigated. Moreover, in connection with the fact that models behave differently for each language group, it is necessary to conduct a separate study and a separate selection of parameters for the each language. Therefore, this work focuses on the generation of thematic texts in Russian.

The research objective. The purpose of this paper is to investigate the application of the recurrent neural networks in combination with word2vec to generate thematic texts specifically for the Russian language. As a solution, the article will focus on creating a model that generates Russian-language text on a given topic using the above-mentioned structures and analyzing its interpretability and parameters.

The statement of basic materials. The standard formulation of the task of pseudo-random text generation occurs in two forms. In the first form, the goal is to predict the next character of the text given N previous characters, where N usually varies from 50 to 1000 [5]. To predict the next element in a sequence, specifically, the next word in the sentence, the Generative LSTM is used. Having the sequence of input vectors $(x_1,...,x_T)$, the model uses the sequence of its output vectors $(o_1,...,o_T)$, to have a sequence of predictable distributions $P(x_{t+1}|x_{\le t}) = \text{softmax}(o_t)$, where the distribution of softmax function is given by:

$$\text{P}(\text{softmax}(\text{o}_t) = j) = \frac{\exp(o_t^{(j)})}{\sum_k \exp(o_t^{(k)})}, \quad (1)$$

where o_t is the output vector of the model.

General model structure. The structure of the model was chosen to predict the sequence of sentences most accurately, while taking into account the correspondence between the parts of speech, punctuation marks and the topic of the text. Fig. 1 illustrates its main components.

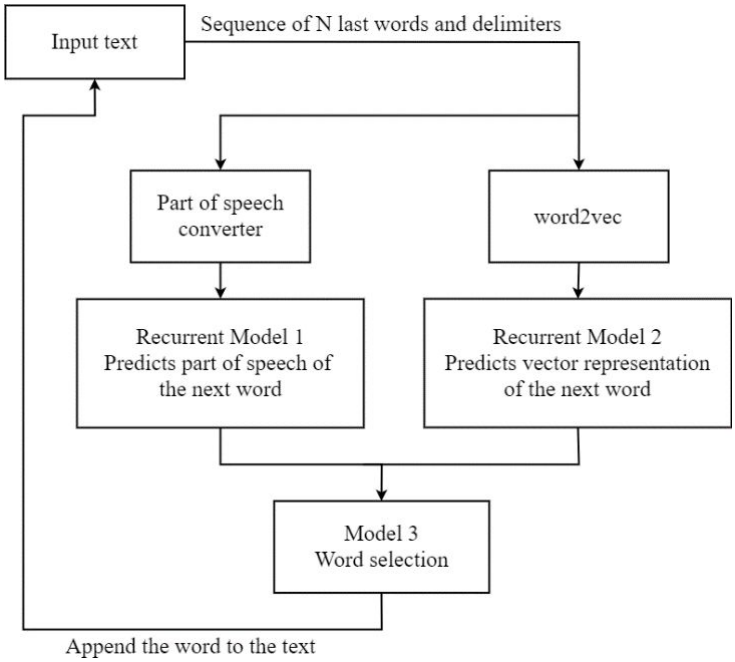

Fig. 1. General model structure

Experiments. The training stage consisted of splitting the articles into sequences of N words while marking $N+1$ word as the target variable. The experiments were done with N varying from 5 to 20 and the final results presented here were held with the value of 13. Final dataset consisted of 300,000 samples where the word prediction model reached the loss of 0.0195. The part of speech model reached the loss of 1.31417.

Table 1

Examples of generated texts

Given context	Produced results
электрический клавиатура программа программирование компьютер	объекты виртуальной реальности должны вести себя аналогично постредактирования самонастраивающихся систем представления . техдокументации компонуются относительно машинального положения в дальнейшем используема .

Conclusions. The paper has demonstrated the ability of Long Short-Term Memory recurrent neural networks along with word2vec network to generate thematic meaningful Russian-language texts. It can be seen that the use of such combination produces qualitative results. A model that produces interpretable results has been developed and its parameters has been studied.

There are several directions for future work. One is to change the model structure, increasing the number of hidden units and adding more layers. Another is to increase the size of training dataset to give the model more context. These changes will definitely improve the results. It also would be interesting to test the model on different languages.

References

1. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010, September). *Recurrent neural network based language model.* In Interspeech (Vol. 2, p. 3).
2. Sutskever, I., Martens, J., & Hinton, G. E. (2011). *Generating text with recurrent neural networks*. In Proceedings of the 28th International Conference on Machine Learning (ICML-11) (pp. 1017-1024).
3. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). *Show and tell: A neural image caption generator*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
4. Shang, L., Lu, Z., & Li, H. (2015). *Neural responding machine for short-text conversation*. arXiv preprint arXiv:1503.02364.

ДОВІДКА ПРО АВТОРІВ

Волокита Артем Миколайович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Volokyta Artem – associate professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: artem.volokita@kpi.ua

Регіда Павло Генадійович – аспірант, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Rehida Pavlo – PhD student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: pavel.regida@gmail.com

РОЗШИРЕНА АНОТАЦІ

В.А.Фоменко, Г.М.Луцький, П.Г.Регіда, А.М.Волокита

ПИТАННЯ ГЕНЕРАЦІЇ ТЕМАТИЧНИХ ТЕКСТІВ НА ОСНОВІ РЕКУРЕНТНИХ НЕЙРОННИХ МЕРЕЖ ТА WORD2VEC

Актуальність теми дослідження. Проблема генерації текстів стає більш актуальною в останні дні у зв'язку зі зростаючим попитом на автоматичне створення описів об'єктів, уривків статей, підсумків новин, повідомлень у службах мікроблогів, відповідей чат-ботів тощо. Таким чином, проблемою є створення текстів, що відповідають заданій тематиці. Дана робота присвячена проблемі генерації текстів саме російською мовою, оскільки кожна мовна група вимагає індивідуального підходу.

Постановка проблеми. Відсутність добре інтерпретованого методу для автоматичного створення російськомовних тематичних текстів за допомогою рекурентних нейронних мереж.

Аналіз останніх досліджень і публікацій. Протягом останніх років з’являється все більше статей присвячених генерації тематичних текстів, зокрема, завдяки появі нових методів генерації послідовностей з використанням рекурентних нейронних мереж. Проте підходи специфічні для генерації тематичних текстів, особливо російською мовою, все ще недостатньо вивчені.

Виділення недосліджених частин загальної проблеми. Дана стаття присвячена вивченню та аналізу запропонованого підходу для генерації тематичних текстів, зокрема на російській мові. Дослідження сфокусовано на вивченні застосування рекурентних нейронних мереж та word2vec.

Постановка завдання. Завданням є створити модель, натреновану на групі уривків російськомовних статей, що навчиться визначати контекст тексту, і як результат видавати добре інтерпретований текст за тією ж самою тематикою.

Викладення основного матеріалу. Проведено аналіз спільного використання моделей RNN та word2vec. Описано підходи для обробки вхідного тексту, аналізу структури речень, прогнозування наступних частин мови, прогнозування наступних слів та структури відповідних моделей. Результати виявились добре інтерпретованими та змістовними.

Висновки. Проаналізовано зміст, структуру та параметри моделей, які показали найкращі результати для генерації текстів. Підхід показав себе добре для створення тематичних текстів. Наведені результати експериментів та аналіз наступних кроків.

Ключові слова: генерація тексту, рекурентні нейронні мережі, довга короткочасна пам’ять, word2vec.