UDC 004.8
DOI: 10.25140/2411-5363-2017-4(10)-110-115

Volodymyr Fomenko, Heorhii Loutskii, Pavlo Rehida, Artem Volokyta

THEMATIC TEXTS GENERATION ISSUES BASED ON RECURRENT NEURAL NETWORKS AND WORD2VEC

Urgency of the research. *The problem of natural language generation is becoming more actual in recent days due to the growing demand for automated generation of object descriptions, article excerpts, news summaries, passages in microblogging services, response messages used by chat bots, etc. Thus, the problem is to generate a text given the context. This paper deals with the problem of generating text specifically in Russian since each language group requires an individual approach.*

Target setting. *There is no method to generate thematic texts automatically, especially in Russian language, that gives well-interpreted and suitable results.*

Actual scientific researches and issues analysis. *In the past few years, more articles have been devoted to the topic of generating thematic texts, due to the emergence of new methods for sequences generation using recurrent neural networks. However, approaches related specifically to thematic texts generations, in Russian are insufficiently explored.*

Uninvestigated parts of general matters defining. *This article focuses on a study and analysis of the proposed approach for generating Russian-language thematic texts. It is specialized in one language group and specific approach in terms of model selection.*

The research objective. *Create model trained on a group of short passages that identifies a context of a text and as output generates a well-interpreted natural text in Russian.*

The statement of basic materials. *The analysis of the joint use of the RNN and word2vec models is conducted. Approaches for the transformation of the input text, analysis of sentences structure, prediction of subsequent parts of speech, prediction of following words and the general model structures are proposed. The results of the models are appeared to be well interpreted and meaningful.*

Conclusions. *The iinterpretability, structure and parameters of the models that showed the best results for the generation were analyzed. The approach proved to be good for generating thematic texts. The results and analysis of the subsequent steps are given.*

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. At the same time, the problem of generating thematic texts with the help of recurrent neural networks [1] is still little understood, moreover, in the context of the Russian language. Because of that, a new approach to the generation of Russian-language thematic texts with the use of recurrent neural networks in combination with word2vec has been offered.

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. In particular, in [2] the generation of English texts on a general topic on a basis of recurrent neural networks is studied, in [3] the application of recurrent neural networks for the generation of an English-language image descriptions is investigated, and in [4] the application of recurrent neural networks for constructing a model capable of an English-speaking dialogue system with the user is developed.

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]. An alternative is to predict the next word given N previous words. Here, N usually varies from 5 to 20 [5]. The approach where the next character of the text is being predicted has a big advantage in terms of a small number of classes of elements: the size of the alphabet and separating symbols. The other approach that learns words sequences has significantly more variants, depending on the size of the vocabulary of training data.

In this article we focus on developing model that deals with word sequences to extract more data from every word by using word2vec transformations.

Basic definitions. A recurrent neural network (RNN) [1, 6] is a type of artificial neural network that involves directed cycles in memory. The input to such networks is a sequential signal. Each element of the sequence is successively transmitted to the same neurons, which return their prediction to themselves together with its next element until the sequence ends.

LSTM [7] is a type of Recurrent Neural Network that has a complex dynamics and makes it easy to "remember" information for an extended number of timestamps. The "long-term" memory is stored in the memory cell vector. Despite the fact that many LSTM architectures differ in their connection structure and activation functions, all LSTM architectures have explicit memory cells for storing information for long periods. LSTM can decide to overwrite the memory location, load it, or save it for the next time step. The architecture demonstrated itself better than RNN in a number of tasks [8, 9, 10].

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:

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

where o_t is the output vector of the model.

The goal of the generative model is to maximize the total logarithm of the probability of the training sequence. Even considering the fact that the latent elements of the network are deterministic, our choice of network prediction will be stochastic, because the states of its output elements define the conditional distribution $P(x_{t+1}|x_{\le t}) = \text{softmax}(o_t)$.

Word2vec [11] is a technology comprised of models used to convert words to word embeddings. These models are two-layer neural networks that processes text. Word2vec input is a text corpus and its output is a set of vectors: feature vectors for words in that corpus. The purpose and usefulness of Word2vec is to group the vectors of similar words together in vector-space. Word2vec creates vectors that are distributed numerical representations of word features, features such as the context of individual words.

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.

The input of the model is given by a sequence of N last words or punctuation marks. Then the sequence is directed to the input of two models. The first model preprocesses the input sequence and converts the words into the appropriate parts of the speech and punctuation marks into the corresponding codes. The processed sequence is forwarded to the Recurrent Model 1, which outputs a part of the speech of the word to be predicted.

The second model also preprocesses input sequence, converting words and punctuation marks into the corresponding vectors using the word2vec model. Then, the processed sequence is forwarded to the Recurrent Model 2, which outputs the vector representation of the following word.

Model number 3 aggregates the output data of both models and on their basis, as well as using a dictionary of words with the corresponding vector representations of word2vec model, and predicts the next word in the text. At the end, the predicted word is appended to the end of the text.

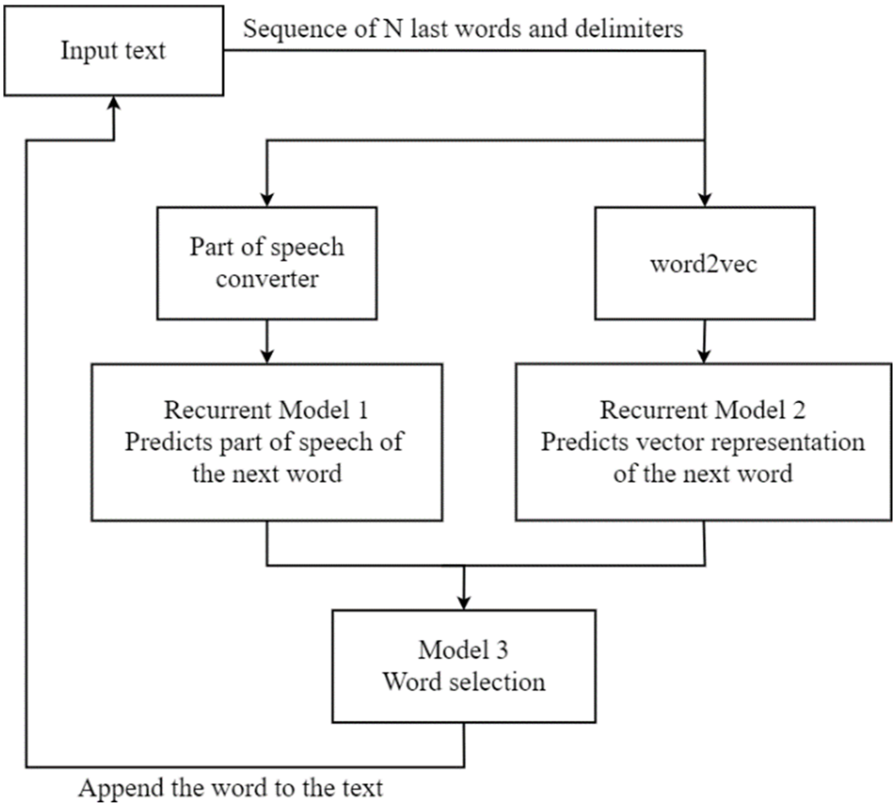
Input text
Sequence of N last words and delimiters
Part of speech converter
word2vec
Recurrent Model 1 Predicts part of speech of the next word
Recurrent Model 2 Predicts vector representation of the next word
Model 3 Word selection
Append the word to the text

Fig. 1. General model structure

Recurrent models structure. Recurrent Model 1 is responsible for words' parts of speech prediction and consists of two layers – LSTM and Dense one. As an input, it takes a sequence of N parts of speech and then predicts the part of speech of the following word.

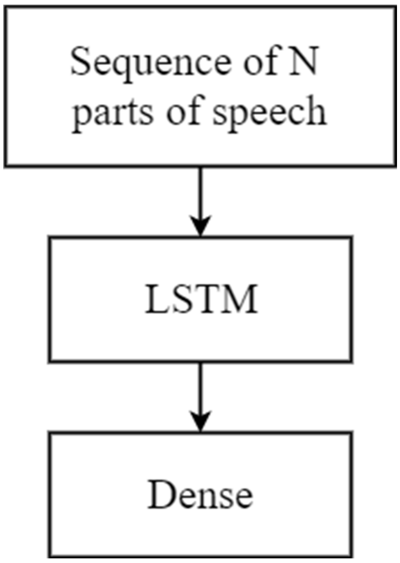

Fig. 2. Structure of the part of speech prediction model

Recurrent Model 2 is responsible for words prediction and consists of 12 layers and is the main part of the general model responsible for words generation.

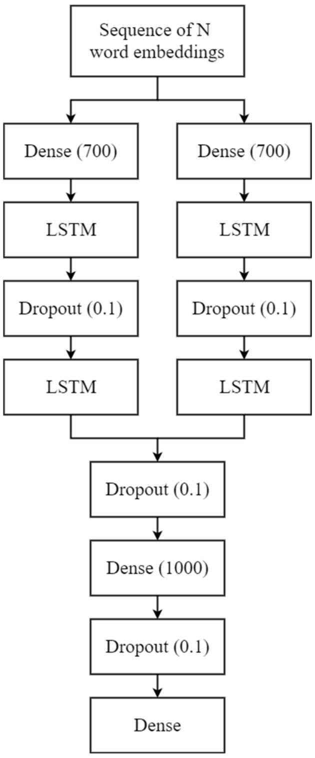

Fig. 3. Structure of the word prediction model

The input in format of N word embeddings firstly goes to two parallel networks, both of which consists of Dense layer, followed by LSTM along with Dropout layers and, lastly, the LSTM layer.

Finally, outputs of two networks are stacked together, pushed to the network of 4 Dense and Dropout layers, and the last Dense layer outputs the word embedding of the following word.

Data. To convert words to the word2vec format, the pre-trained fast Text model [12] was used. The model was trained on an open archive of Russian Wikipedia articles for the 2016 year [13]. To simplify the work of the model, the text content was preliminarily filtered, leaving only Russian-language words and punctuation marks ",", "-", ".", "!", "?". The brackets with the inner content were also removed. To filter out rare words, such as proper names, only those words that occurred at least 20 times were left in the set. All sentences that contained previously defined rare words were removed from the text. Finally, about articles with a total number of 300,000 words were kept in the training set.

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.

The testing stage evaluation was based on observations of generated text. Text generation process consisted of giving the network initial context and iterating the prediction phase until at least 30 words were produced. To give the network the context, first $N/2$ elements of the sequence were manually set. Table illustrates the produced results.

Table

Examples of generated texts

Given context	Produced results
1	2
экономика страна импорт экспорт налог	на судоходных реках или каналах основные статьи импорта , нефть , автомобили выделяются производительностью . в итоге помимо примечательности возращения быстроменяющихся соответственных производительностей, выяснилось место соответственного спецучреждения.

End table

1	2
азия китай китайский японский тайвань корея корейский	он учитывал и возможности придворного спектакля в итальянском духе . в последствии сингапурец помимо двадцатишестилетняя время также получает более высокие позиции в списке.
электрический клавиатура программа программирование компьютер	объекты виртуальной реальности должны вести себя аналогично постредактирования самонастраивающихся систем представления . тexдокументации 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.

5. Graves, A. (2013). *Generating sequences with recurrent neural networks.* arXiv preprint arXiv:1308.0850.

6. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to sequence learning with neural networks.* In Advances in neural information processing systems (pp. 3104-3112).

7. Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory.* Neural computation, 9(8), 1735-1780.

8. Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). *Speech recognition with deep recurrent neural networks.* In Acoustics, speech and signal processing (icassp), 2013 ieee international conference on (pp. 6645-6649). IEEE.

9. Graves, A., & Schmidhuber, J. (2009). *Offline handwriting recognition with multidimensional recurrent neural networks.* In Advances in neural information processing systems (pp. 545-552).

10. Eck, D., & Schmidhuber, J. (2002). *A first look at music composition using lstm recurrent neural networks.* Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, 103.

11. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781

12. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). *Enriching word vectors with subword information.* arXiv preprint arXiv:1607.04606.

13. Wikimedia downloads. Retrieved from http://dumps.wikimedia.org.

УДК 004.8

Володимир Фоменко, Георгій Луцький,
Павло Регіда, Артем 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. Ця 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 не вирішених раніше частин загальної проблеми. *Стаття присвячена вивченню та аналізу запропонованого підходу для генерації тематичних текстів, зокрема написаних російською мовою. 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.*

Виклад 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.*
Рис.: 3. Табл.: 1. Бібл.: 13.

УДК 004.8

Владимир Фоменко, Георгий Луцкий, Павел Регида, Артем Волокита

ВОПРОСЫ ГЕНЕРАЦИИ ТЕМАТИЧЕСКИХ ТЕКСТОВ НА ОСНОВЕ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ И WORD2VEC

В статье рассматривается вопрос генерации псевдослучайных текстов на заданную тематику. Для генерации текстов используются рекуррентные нейронные сети (LSTM) с предварительной обработкой слов с помощью модели word2vec. Тема текста задается с помощью набора ключевых слов. Модели тренируются на наборе русскоязычных статей.

Ключевые слова: *генерация текста; рекуррентные нейронные сети; долгая краткосрочная память; word2vec.*
Рис.: 3. Табл.: 1. Библ.: 13.

Fomenko Volodymyr – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute” (37 Pobedy Av., 03056 Kyiv, Ukraine).
Фоменко Володимир Андрійович – студент, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» (просп. Перемоги, 37, м. Київ, 03056, Україна).
Фоменко Владимир Андреевич – студент, Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского» (просп. Победы, 37, г. Киев, 03056, Украина).
E-mail: vlfomenk@gmail.com
Loutskii Heorhii – Doctor of Technical Sciences, Professor, Professor of Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute” (37 Pobedy Av., 03056 Kyiv, Ukraine).
Луцький Георгій Михайлович – доктор технічних наук, професор, професор кафедри обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» (просп. Перемоги, 37, м. Київ, 03056, Україна).
Луцкий Георгий Михайлович – доктор технических наук, профессор, профессор кафедры вычислительной техники, Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского» (просп. Победы, 37, г. Киев, 03056, Украина).
E-mail: georgijluckij80@gmail.com
Регіда Павло Геннадійович – аспірант, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» (просп. Перемоги, 37, м. Київ, 03056, Україна).
Регида Павел Геннадиевич – аспирант, Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского» (просп. Победы, 37, г. Киев, 03056, Украина).
Rehida Pavlo – PhD student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute” (37 Pobedy Av., 03056 Kyiv, Ukraine).
E-mail: pavel.regida@gmail.com
Volokyta Artem – PhD in Technical Sciences, Associate Professor, Associate Professor of Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute” (37 Pobedy Av., 03056 Kyiv, Ukraine).
Волокита Артем Миколайович – кандидат технічних наук, доцент, доцент кафедри обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» (просп. Перемоги, 37, м. Київ, 03056, Україна).
Волокита Артем Николаевич – кандидат технических наук, доцент, доцент кафедры вычислительной техники, Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского» (просп. Победы, 37, г. Киев, 03056, Украина).
E-mail: artem.volokita@kpi.ua

Fomenko, V., Loutskii, H., Rehida, P., Volokyta, A. (2017). Thematic texts generation issues based on recurrent neural networks and word2vec. *Technical sciences and technologies*, no. 4 (10), pp. 110-115.