Deep Learning Approach in the Context of Information Retrieval for Solving both Automatic Natural Language Generation and Automatic Text Generation Problems

Yury Krapivin
Brest State Technical University
Brest, Belarus, 224000
Email: ybox@list.ru

***Abstract*—The article presents the solution of practical application of the deep learning approach in the context of information retrieval based on the usage of LSTM and GPT-2 language models for solving both automatic natural language generation and automatic text generation problems.**

***Keywords*—natural language, information retrieval, automatic natural language synthesis, automatic text synthesis, machine learning**

I. Introduction

The rapid development of information technology with each passing year increases the importance of obtaining prompt and high-quality information both for solving simple daily tasks (for example, related to the choice of a product or service, receiving region and world news) and for business analysis and making management decisions. The use for this purpose a global data network - the Internet, as well as a variety of related services and applying artificial intelligence technology applications has become an integral part of everyday life. Services and applications are turning from consumers to data generators insensibly and increasingly.

II. Natural language generation and text generation technologies

“Imperva Bad Bot Report 2021” – a research of “Imperva”, cybersecurity leader, whose mission is to protect data and all paths to it, suggests the growing scale and widespread impact of bots in daily life. It divide such applications in two groups:

1) Bad bots – interact with applications in the same way a legitimate user would, making them harder to detect and prevent. They enable high-speed abuse, misuse, and attacks on websites, mobile apps, and APIs. They allow bot operators, attackers, unsavory competitors, and fraudsters to perform a wide array of malicious activities. Such activities include web scraping, competitive data mining, personal and financial data harvesting, brute-force login, digital ad fraud, spam, transaction fraud, and more. [1]
2) Good bots – support the various business and operational goals of their owners—from personal users to large multinationals and can be categorized by the following four groups (“Fig. 1”):
   - Feed fetcher – Bots that ferry website content to mobile and web applications, which they then display to users.
   - Search engine bots – Bots that collect information for search engine algorithms, which is then used to make ranking decisions.
   - Commercial crawlers – Spiders used for authorized data extractions, usually on behalf of digital marketing tools.
   - Monitoring bots – Bots that monitor website availability and the proper functioning of various online features.

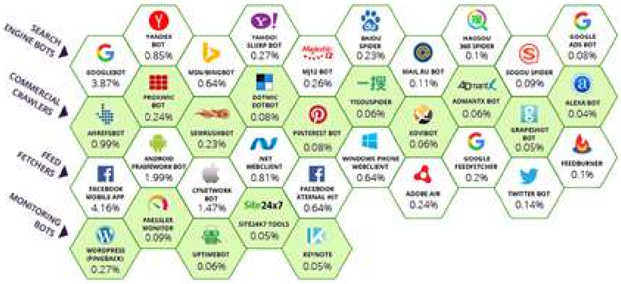


Figure 1. The four groups of good bots.

Imperva Research Labs saw the highest percentage of bad bot traffic (25.6%) since the inception of the report in 2014, while traffic from humans fell by 5.7%. More than 40% of all web traffic requests originated from a bot last year.

It is worth to note that activity of the "smart" applications also reveals in the creation of natural language data: vivid examples are dialogue systems and services that provide interaction with the user, as well as systems for automatic content creation.

Robot journalism is an actively developing area of research refers to the generation of news stories by algorithms based on data without human-journalistic intervention and these news stories are then published automatically on news websites. It relies on natural language generation (NLG) technology, which is a subfield of artificial intelligence. The main objective of NLG technology is to design text generation systems that create readable explanatory stories based on data. Defining the grammatical and syntax rules of a language within an NLG system allows the automatic creation of various documents, reports, explanations, and summaries by the algorithms on the basis of the input data [2]. Currently, NLG systems help news media organizations to generate news stories and, in this context, the emergence of robot journalism is a result of the convergence of NLG technology and the news media sector. Narrative Science, Automated Insights, Yseop, and Arria are some of the large technology firms that are progressing NLG technology [3] and providing robot journalism services to news media organizations such as Associated Press, Forbes, and Yahoo. [4]

Another natural language processing technology related to tasks where the target is a single or several sentences or even the set of words – text generation. It also finds its applications in developing of question answering, dialog, machine translation and information retrieval systems. In the context of the last one, the results of solving these problems used for search-engine optimization to increase the rate of a web-resource in the search-engine output, as well as speedup and simplify the process of web-resource developing.

In today's digitalized world, search engines have become one of the most powerful tools on the Internet and an essential part of our daily live. By consolidating and organizing the wealth of information available online, search engines like Google [5], Yahoo [6] or Bing [7] help billions of online users find the content they need at a rapid pace. In 2019, almost 30 percent of global web traffic was generated via online search usage, showing the vital role these platforms play in directing and navigating user flows to different websites. For example, according to [8] the global marketing share percentage, in terms of the use of search engines heavily favours Google, with over 92% ("Fig. 2").

It is interesting to note that Google's large market share is still on the increase. In April 2017 the market share for Google was 77.43%.[9]

The analysis of search queries has revealed a stable downward trend of the number of keywords usage. One

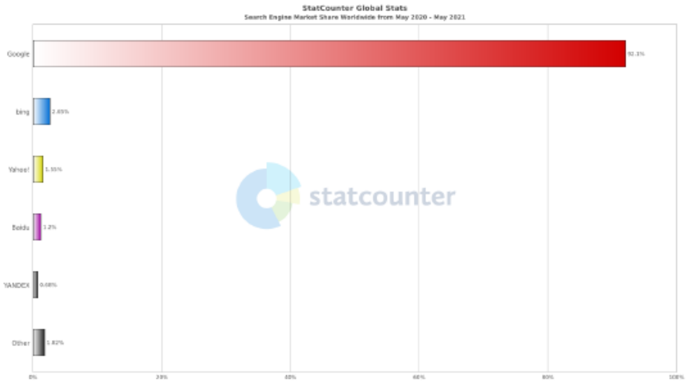


Figure 2. The global marketing share of search engines.

of the examples is the average number of typed search terms during online search in the United States as of January 2016. During that month, 20.14 percent of all U.S. online search queries contained tree keywords ("Fig. 3") and the situation changes heavily. As of

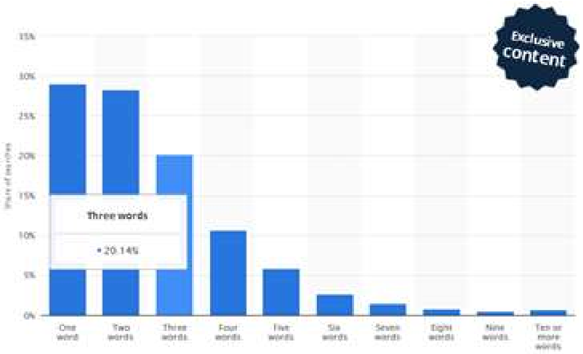


Figure 3. Average number of search terms for online search queries in the United States as of January 2016.

January 2020, 40 percent of all U.S. online search queries contained two keywords. Three word search terms accounted for 22.74 percent of searches. Queries up to three words accounted for over 80 percent of online searches in the United States. [10] As search engines rank search results in order of relevancy, meaning that the most valuable links for users' queries are displayed prominently on the results page, high rankings have become one of the top digital priorities for companies worldwide. [11] That is why tools like UberSuggest's [12] or SemRush [13] are quite popular among web-masters and usually used for analysis of the web traffic, user queries statistics etc. Some of those routine but important functions: candidate words and word sequences for semantic core generation[1], as well as the scratch of the future web-pages content suggestion can be automated in the context of the automatic text and automatic language generation tasks.

III. Practical application of the deep learning approach in the context of information retrieval

In regard to methods suitable to solve those tasks, along with summarization-based approaches such as extractive summarization [15], fractal summarization that is used to generate a brief skeleton of summary at the first stage, and the details of the summary on different levels of the document are generated on demands of users [16], linguistic and semantic analysis [17], the analysis revealed the machine learning methods can be used. It is worth to mention that methods of supervised learning are not effective enough due to the next limitations: they need large amount of annotated data for learning a proper task, which is often not easy available, and they fail to generalize for tasks other than they have been trained for. With the development of neural networks and deep learning approaches, models of vector representations of words [18], recurrent neural networks, Long-Short-Term-Memory (LSTM) or Gated Recurrent Unit (GRU) architectures and their variations and combinations [19-21] began to be used most often.

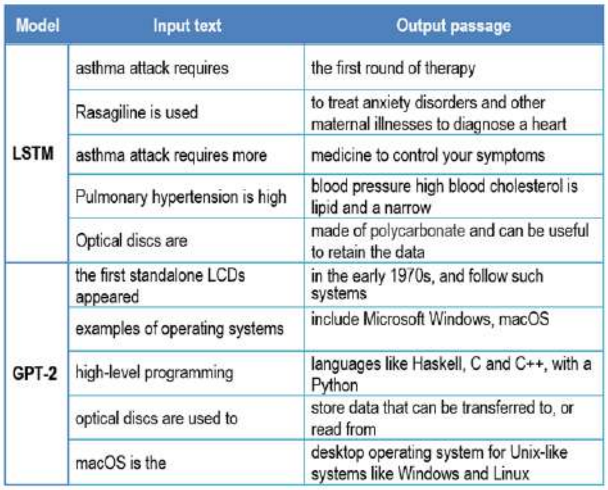
Taking into account typical length of the search query the choice of LSTM as the model for generation passages up to fourteen words length (that is a bit less than an average sentence length, for example for the well-known LOB Corpus [22] it is about nineteen words) implemented with the help of Tensorflow source-platform [23] and Keras framework [24] (main parameters are: ReLu, SoftMax activation functions, Adam optimizer) seems obvious. The dataset of about 200 thousand words length was created on the basis of a well-known MEDIQA-QA dataset for answer-ranking [25], encouraging research in medical question answering systems, and consisting of consumer health questions and passages selected from reliable online sources, as well as 120 thousand words length text corpus referred to IT-domain was collected from Wikipedia. Standard procedures of preprocessing related to cleaning, tokenization, and splitting into training and test data were performed.

Recurrent models typically factor computation along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time, they generate a sequence of hidden states h_{t}, as a function of the previous hidden state $h_{\mathrm{t}-1}$ and the input for position t. This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples. The appearance of Generative Pre-trained Transformer (GPT) models by OpenAI gave ability to overcome limitations. The first one – GPT has the architecture which facilitated transfer learning and can perform various natural language processing tasks with very little fine-tuning. It also showed the power of generative-pre-training and opened up avenues for other models, which could unleash this potential better with lager datasets and more parameters. The second one – GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages. GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text. The diversity of the dataset causes this simple goal to contain naturally occurring demonstrations of many tasks across diverse domains. GPT-2 is a direct scale-up of GPT, with more than 10X the parameters and trained on more than 10X the amount of data. [26] GPT-2 124M was fine tuned and used to generate the text passages up to the same length as the previous models as well as much more longer to create the web-page content template suitable for manual or semi-automatic postprocessing according to the objectives of the web-resource.

The examples of results are represented below ("Fig. 4"). The achieved text sequences can be further processed with the set of filters or stages that can be organized in a pipeline to identify named entities, syntagmata, taxanomic categories with the help of functionality provided by available standalone natural language and statistic processing libraries or linguistic processors. For example, at the first stage – the extraction of named entities – the names of programming languages were achieved: “Haskell”, “C”, “C++”, “Python” from a text fragment “languages like Haskell, C and C++, with a Python”, as well as, the names of the operating systems “Windows”, “Linux”, “macOS” – achieved from the input sequences “include Microsoft Windows, macOS” and “desktop operating system for Unix-like systems like Windows and Linux”. In particular, for identifying synonymy relations the multilingual lexical database MModWN can be used, due to the ability to take in consideration lexical units membership of the most informative lexico-grammatical, syntactical and semantic classes and achieve proper synonymic sets in different languages [9] or hybrid knowledge bases of ostis-systems and models for representing various types of knowledge within the framework of such a knowledge base. [27] The last one option might be useful if it is important to ensure possibility to use within the ostis-systems of various types of knowledge.

In regard to the web-page content template of the web-page, its structure depends on the type of the web-resource (a homepage, a magazine website, an e-commerce website, a landing page etc.) and can be

[1] The semantic core is a list of all words on the project’s topic, which serves as a basis for creating a common structure and individual pages of the site. An appropriate compilation of the semantic core will help the site to get into the top search engine results. [14]



Model	Input text	Output passage
LSTM	asthma attack requires	the first round of therapy
	Rasagiline is used	to treat anxiety disorders and other maternal illnesses to diagnose a heart
	asthma attack requires more	medicine to control your symptoms
	Pulmonary hypertension is high	blood pressure high blood cholesterol is lipid and a narrow
	Optical discs are	made of polycarbonate and can be useful to retain the data
GPT-2	the first standalone LCDs appeared	in the early 1970s, and follow such systems
	examples of operating systems	include Microsoft Windows, macOS
	high-level programming	languages like Haskell, C and C++, with a Python
	optical discs are used to	store data that can be transferred to, or read from
	macOS is the	desktop operating system for Unix-like systems like Windows and Linux

Figure 4. The examples of text passages achieved via LSTM and GPT-2 language models.

organized in unit-structure style, where the content for every item generates separately taking into account the sub-topic and related dataset, which is used to fine tune or prepare the language model discussed above.

IV. Conclusion

The analysis of the machine learning methods to solve the problem of automatic natural language generation as well as automatic text generation problems have been surveyed. A practical application of the deep learning approach in the context of information retrieval based on the usage of LSTM and GPT-2 language models has been performed. Proposed solution is language independent and can be reinforced with functionality of the natural language analysis provided by linguistic processor or standalone natural language and statistic processing libraries together with multilingual lexical databases or hybrid knowledge bases of ostis-systems.

References

[1] Bad Bot Report - Imperva. Available at: https://www.imperva.com/blog/bad-bot-report-2020-bad-bots-strike-back/. (accessed 2021, Jun).

[2] Reiter, E., Dale, R. Building applied natural language generation systemsn, *Natural Language Engineering*, 1997, vol. 3, no. 1, pp. 57–59.

[3] Van der Kaa, H. A. J. Krahmer, E. J. Journalist versus news consumer: The perceived credibility of machine written news. *Proc. of the Computation + Journalism Symposium New York, NY, USA.*, ACL, 2014.

[4] Firat F., Robot journalism. *The International Encyclopedia of Journalism Studies, 2019*, DOI: 10.1002/9781118841570.iejs0243.

[5] Google. Available at: https://www.google.com/. (accessed 2021, Jun).

[6] Yahoo. Available at: https://www.yahoo.com. (accessed 2021, Jun).

[7] Bing. Available at: https://www.bing.com. (accessed 2021, Jun).

[8] Search Engine Market Share Worldwide - StatCounter Global Stats. Available at: https://gs.statcounter.com/search-engine-market-share#monthly-202005-202105-bar. (accessed 2021, Jun).

[9] Krapivin, Y. Information Retrieval and Machine Translation in Solving the Task of Automatic Recognition of Adopted Fragments of Text Documents, *Proceedings of Open Semantic Technology on Intelligent Systems, International Conference, 21-23 Febr. 2019*: Belarusian State University of Informatics and Radioelectonics; ed. : V. Golenkov. – Minsk, 2019. – P. 289–292.

[10] U.S. online search query size 2020 – Statista. Available at: https://www.statista.com/statistics/269740/number-of-search-terms-in-internet-research-in-the-us/#statisticContainer. (accessed 2021, Jun).

[11] Online search usage - Statistics & Facts – Statista. Available at: https://www.statista.com/topics/1710/search-engine-usage/?#topicHeader_wrapper. (accessed 2021, Jun).

[12] Ubersuggest's Free Keyword Tool, Generate More Suggestions. Available at: https://neilpatel.com/ubersuggest. (accessed 2021, Jun).

[13] Serpstat — Growth hacking tool for SEO, PPC and content marketing. Available at: https://serpstat.com/. (accessed 2021, Jun).

[14] Six ways to collect a semantic kernel for your site: tools overview. Available at: https://serpstat.com/blog/how-to-collect-a-semantic-core-for-a-site/. (accessed 2021, Jun).

[15] Voronkov N.V. Metody, algoritmy i modeli sistem avtomaticheskogo referirovanija tekstovyh dokumentov.dis. kand. teh. nauk, Minsk, 2007.165 p.

[16] Yang, C., Wang, F. Fractal Summarization for Mobile Device to Access Large Documents on the Web // WWW '03: *Proceedings of the 12th international conference on World Wide Web*, 20 May. – 2003. – P. 215–224.

[17] Lipnitsky S.F., Stepura L.V., J. Algoritmy sozdaniya giperteksta na osnove situativno-sintagmaticheskoi seti [Algorithms of hypertext creation based on situational-syntagmatic network]. Izvestiya Natsionalnoi akademii nauk Belarusi. Seriya fiziko-tehnicheskih nauk, 2010, no. 3, pp.90-95.

[18] Kim D. [et al.] Lexical feature embedding for classifying dialogue acts on Korean conversations. *Proc. of 42th Winter Conference on Korean Institute of Information Scientists and Engineers*, 2015, pp. 575–577.

[19] Ravuri S., Stolcke A., Recurrent neural network and LSTM models for lexical utterance classification. *Proc. of 16th Annual Conference of the International Speech Communication Association*, 2015, pp. 135–139.

[20] Dey R., Salemt F., Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks. *IEEE 60th International Midwest Symposium on Circuits and Systems*, IEEE, 2017, pp. 1597–1600.

[21] Hochreiter, S., Schmidhuber, J. Long Short-term Memory. *Neural computation*, 1997, vol. 9, no. 8, pp. 1735–1741.

[22] Johansson, S. Manual of information to accompany the Lancaster-Oslo-Bergen Corpus of British English, for use with digital computers. Available at: http://clu.uni.no/icame/manuals/LOB/INDEX.HTM/. (accessed 2021, Jun).

[23] TensorFlow. Available at : https://www.tensorflow.org/. (accessed 2021, Jun).

[24] Keras. Available at: https://keras.io/. (accessed 2021, Jun).

[25] Question-Driven Summarization of Answers to Consumer Health Questions. Available at: https://arxiv.org/abs/2005.09067. (accessed 2021, Jun).

[26] Better Language Models and Their Implications. Available at: https://openai.com/blog/better-language-models/. (accessed 2021, Jun).

[27] Davydenko, I. "Semantic models, method and tools of knowledge bases coordinated development based on reusable components," *Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed., BSUIR. Minsk , BSUIR, 2018, pp.99–118.

Подход на основе глубокого обучения в контексте информационного поиска для решения задач как автоматического синтеза естественного языка так и автоматического синтеза текста

Крапивин Ю.Б.

В статье представлено решение практического применения подхода глубокого обучения в контексте поиска информации на основе использования языковых моделей LSTM и GPT-2 для решения задач как автоматического синтеза естественного языка, так и автоматического синтеза текста.

Ключевые слова: естественный язык, информационный поиск, автоматический синтез естественного языка, автоматический синтез текста, машинное обучение.

Received 04.07.2021