IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.6, Issue.10, Oct. 2018 E-ISSN: 2347-2693

Ontology based News Extraction System using Vanilla Recurrent Neural Network

Shine K George[1*], Jagathy Raj V. P[2]

[1]Dept. of Computer Applications, Cochin University of Science and Technology, Cochin, Kerala, India
[2]School of Management Studies, Cochin University of Science and Technology, Cochin, Kerala, India

**Corresponding Author shineucc@gmail.com, Tel.: +91 9447189662*

Available online at: www.ijcseonline.org

Accepted: 19/Oct/2018, Published: 31/Oct/2018

Abstract— News channels established a 24-hour news habit which gets updated virtually in every second. Archiving becomes a challenging process since the news production is huge. Viewers are interested in news stories as it delivers useful and detailed information in short form. The news story created based on the history and the latest news updates The journalists access news archives to get details about the news happened related to the new happenings. Searching archives, fetching and linking related news is a tedious job for a reporter. In this work, a system is suggested which uses ontology and vanilla recurrent neural network to create news automatically for a query. The framework is evaluated using BLEU method and correlated with human evaluation. Ontology completeness decides the quality of the news generated.

Keywords— Ontology, Deep learning, recurrent neural network, news generation, Personalization

I. Introduction

Enormous news contents are generated in every second which keeps on growing exponentially. Fast pace of production is due to more news organizations are come into existence and new types of news media hit the market. Fetching the right content from a huge library is a difficult job for journalists. Ontology helps to resolve this situation [1],[29],[30].

News desk journalists, editors and news producers are accountable for making news script and compiling it into story. The process starts with preparing a news script. Obviously reporters have to search the news library to get relevant information. Since the news related to a single event itself is huge, there is every chance that news correspondent may get hundreds of news for his/her single search query. To make a story, journalists may need to search archive with different keywords to collect information required to make the news story. It is a hectic task for the journalist considering the time limit.

On the contrary, if the news reporter gets the news generated by the machine taking into considering all related events archived based on his/her search query terms provides a much more personalized experience for the journalist.
The system introduced in this work is based on simple recurrent neural network and ontology. The content list of this paper is as follows. Literature review is summarized in Section II. The suggested system is explained in section III. In section IV, assessment and experimental results are carefully studied and finally, section V concludes the work mentioned in this paper

II. Related Work

Traditional news extraction frameworks are recommendation applications focuses on content and user [2], [3]. Several hybrid approaches are also formed by combining traditional approach [4], [5], [6], [7], [8], [9]. One of the drawbacks of traditional approach is the inability to capture semantic meaning of news stories and search queries. It has been overcome by using semantic web tools [10], [11], [12].

News generation and summarization systems are developed based on natural language processing techniques in current years [13,14,15,16].

Neural networks are used extensively in automatic creation of text due to its immense capabilities and increased computational power [17,18]. The quality of text created is subject to the neural model used. Recently, these models are used for generating news, news head- lines, news comments etc. [19,20,21, 22,23].

III. PROPOSED SYSTEM

Recurrent Neural Network (RNN) is an extension of simple neural network which stores the previous state and uses sequential information. It has a context layer which maintains previous information. One of the simplest form of RNN is vanilla-RNN. RNNs works well for a short sequential information and does not remember long sequences. Long short term memory (LSTM) is developed to solve this issue which is a variant of RNN. The important concept of LSTM is the cell state [24,25].

The hidden layer depends on current input at time n and n-1 step (previous).

$$a^{n}=\tanh(\text{weight}_{ax}C^{(n)}-\text{Weight}_{aa}a^{(n-1)}+b_{a}) \qquad (1)$$

The softmax function is used as the last layer to determine the probability of an output.

$$\overline{y}^{n} = \text{softmax}(W_{ya}a^{n}) \qquad (2)$$

LSTM cell uses three gates which are input gate, output gate and forget gate. Input gate (5) updates the previous state values. Add gate (4) includes the new candidate values to the state. Forget gate (3) determines which all values should not pass through and to be forgotten.

$$ft_{n} = \sigma\ (\text{weight}_{ft} \cdot [a_{n-1}, x_{n}] + b_{ft}) \qquad (3)$$
$$ad_{n} = \tanh\ (\text{weight}_{ad} \cdot [a_{n-1}, x_{n}] + b_{ad}) \qquad (4)$$
$$in_{n} = \sigma\ (\text{weight}_{in} \cdot [a_{n-1}, x_{n}] + b_{in}) \qquad (5)$$

Where input sequence is denoted by $x = (x_1, . . ., x_n)$, RNN calculates the hidden vector sequence $a = (a_1, . . ., a_n)$ and output vector sequence $y = (y_1, . . ., y_n)$ by iterating the equations (1) and (2) from $n = 1$ to n. The symbol weight denotes weight matrices and b represents bias vector. σ denotes sigmoid function.

The suggested architecture is represented in figure 1. The ontology used in this framework is Open Calais ontology which is developed by Thomson Reuters and follows the International Press Communications Council (IPTC). Vanilla Char RNN-LSTM algorithm is used to train the model to generate news from ontology keywords.

The Framework has two processes namely training and testing. News annotations are used for model training. News is generated by using ontology keywords/tags as input to neural model.

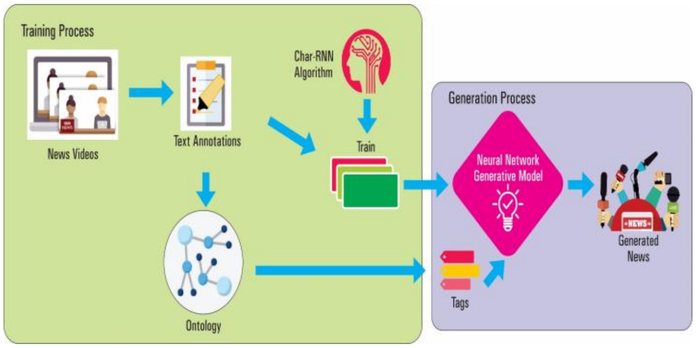


Figure 1: News Generation System

Two hidden LSTM layers and one fully connected layer is used in this model. In each hidden layer 256 hidden states are there. The model was trained for 1000 epochs. Learning rate 0.9 and dropout ratio is 0.3. The Algorithm 1 describes the procedure.

```
Algorithm 1 Vanilla RNN-LSTM model
    Input: Input news annotations x
    Output: Trained Vanilla CharRNN-LSTM Model.
Training Steps
  Start Procedure
  for several epochs of training do
    for each character $c_i$ in x do
        Run encoding on $c_i$
        Run one step of NN optimization & Compute gradi- ents of the loss
        Update the parameters according to this gradient
  end for end for
  End Procedure

Testing
Input: Input Ontology Tags x
Output: News generated
Testing Steps
for each tag $ta_i$ in x do
      Generate the next n characters of news using the trained model

  end for
```

IV. EVALUATION AND RESULTS

The approach in this work consists of manual and automatic evaluation. Automatic evaluation techniques are not full proof in checking the news produced by system because of the complexity of natural languages. So we are also considering human evaluation in this case. Finally, conclusion is made by correlating human score with automatic evaluation metric value.

Figure 2 is a graphical representation of average training loss versus number of epochs.

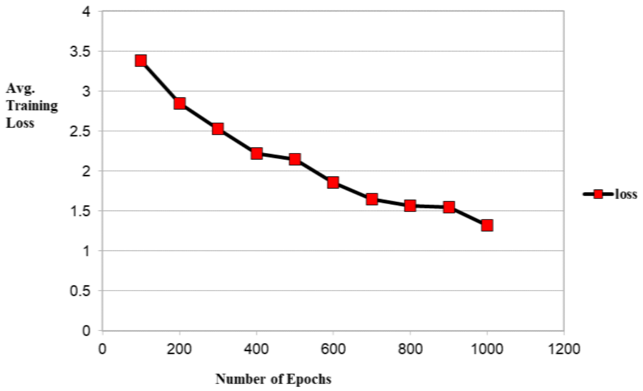


Figure 2: Avg. training loss Vs No. of Epochs

4.1 Dataset

The dataset used is BBC news dataset [26]. They belong to five different categories like business, entertainment, politics, sports and technology and is collected in the year 2004–2005. Two third of the data is inputted to train the model and remaining news is used for testing purpose.

4.2 BLEU metric

Precision can be measured using BLEU method. Precision is calculated by making human text as reference and diversity is measured by using generated text as reference. BLEU scales from 0 to 1 and 1 is the highest value. All n-gram matches between system and reference news were computed in this work [27,28]. Table 1 presents the BLEU score values. Figure 3 illustrates the scores with respect to training epochs.

Table 1: Average BLEU score Vs Number of Epochs

Epoch	BLEU-1	BLEU-2	BLEU-3	BLEU-4
100	0.0913	0.0365	0.0128	0.0064
200	0.1209	0.0484	0.0169	0.0085
300	0.1873	0.0749	0.0262	0.0131
400	0.1998	0.0799	0.0280	0.0140
500	0.2049	0.0820	0.0287	0.0143
600	0.2111	0.0844	0.0296	0.0148
700	0.2234	0.0894	0.0313	0.0156
800	0.2286	0.0914	0.0320	0.0160
900	0.2291	0.0916	0.0321	0.0160
1000	0.2307	0.0923	0.0323	0.0161

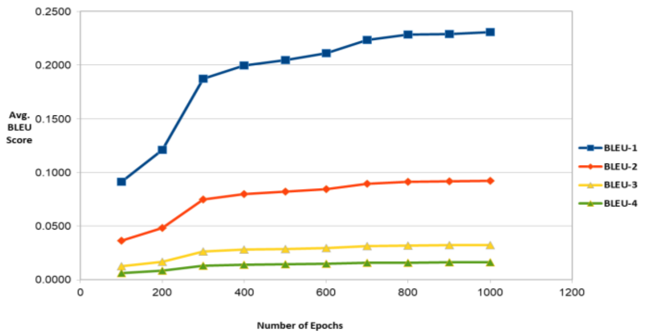


Figure 3: BLEU score

4.3 Human Evaluation

Human experts and non-experts with distinct backgrounds are selected as evaluators. The evaluation is based on the basis of fluency, adequacy and quality. They are also told to judge whether the news generated is by human or computer. A questionnaire was prepared and score level is from 0 to 5 except the judgement of the source of news (human or computer). A yes or no question is included in the questionnaire to judge news source. Table 2 summarizes human evaluation score

Table 2. Human ratings

Model	Quality (5)	Fluency (5)	Adequac y (5)	Machine Generated
Char RNN-LSTM	2	2	1	8/10

4.4 Discussions

The tags from Open Calais ontology is fed as the seed data for news generation. The intuition here is that if the system has enough data to learn from, it will produce news which has some correspondence to the reference news. Feeding the tag words at an interval results in generation of news which has some information from past news related with the tags. Hence the final generated news will contain information from past related news annotations with which each of the input tags are related to.

The easiest test for human experts was to predict whether the generated news is manmade or machine generated. Most of them considered the generated news as machine text. Training vanilla recurrent neural network and generating meaning full text is a difficult job since it learns and generate text by character by character. Both human and BLEU values are correlating in this work.

V. Conclusion and Future Scope

We presented a news generation system using ontology and neural networks. The outcomes are promising for further research in this area. The possibility of using models based on different deep learning algorithms can also be investigated. The dataset used in this work consists of different topics. The inclusion of related news in the dataset may help the model to recognize the context.

The keywords extracted by Ontology from news in the dataset are less in number. Since the feeding input data for generating news by neural model is ontology keywords, sufficient number ontology tags are necessary for improving the quality of news generated.

REFERENCES

[1] D. C. Wimalasuriya and D. Dou, “Ontology-based information extraction: An introduction and a survey of current approaches,” Journal of Information Science, vol. 36, no. 3, pp. 306–323, 2010.

[2] H. Dai and Mobasher B, “Integrating Semantic Knowledge with Web Usage Mining for Personalization,” Web Mining: Applications and Techniques, pp. 276–306, 2004.

[3] Balabanovic, Marko and Shoham, and Yoav, “Fab: Content-based, collabora- tive recommendation, Communications of the ACM,” vol. 40, no. 3, pp. 66–72, 1997.

[4] A. S. Das, M. Datar, A. Garg, and S. Rajaram, “Google news personalization,” Proceedings of the 16th international conference on World Wide Web - WWW 07, pp. 271–280, May 2007.

[5] Merialdo, Bernard and Lee, Kyung Tak and Luparello, Dario and Roudaire, and Jeremie, “Automatic Construction of Personalised TV News Programs,” Proceedings of the Seventh ACM International Conference on Multimedia (Part 1), ACM Multimedia, Orlando, Florida, USA, pp. 323–331.

[6] A.-H. Tan, C. Teo, and Heng Mui Keng, “Learning user profiles for personalized information dissemination,” IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227), pp. 183–188, 1998.

[7] Cotter, Paul and Smyth, Barry: PTV Intelligent Personalized TV Guides, in Proceedings of the 17th National Conference on Artificial Intelligence, AAAI 2000, Austin, Texas, pp. 957–964, 2000.

[8] Konstan, Joseph A and Miller, Bradley N and Maltz, David and Her- locker, Jonathan L and Gordon, Lee R and Riedl, and John, “ACM,” GroupLens: Applying Collaborative Filtering to UseNet News, in Commun, vol. 40, no. 3, pp. 77–87, Mar. 1997.

[9] Liu Jiahui, Dolan Peter and Pedersen Elin Rønby, “Personalized News Recommendation Based on Click Behavior,” Proceedings of the 15th Inter- national Conference on Intelligent User Interfaces, IUI ’10, ACM, Hong Kong, China, pp. 7–10, Feb. 2010.

[10] S. Jokela, M. Turpeinen, T. Kurki, E. Savia, and R. Sulonen, “The role of structured content in a personalized news service,” Proceedings of the 34th Annual Hawaii International Conference on System Sciences, pp. 1–10, 2001.

[11] L. Ardissono, L. Console, and I. Torre, “An Adaptive System for the Personalised Access to News, in AI Commun,” AI*IA 99: Advances in Artificial Intelligence Lecture Notes in Computer Science, vol. 14, no. 3, pp. 129–147, 2001.

[12] IJntema Wouter, Goossen Frank, Frasincar Flavius and Hogenboom Fred- erik, “-Based News Recommendation,” Proceedings of the 2010 EDBT/ICDT Workshops, EDBT ’10, Lausanne, Switzerland, pp. 22–26, Mar. 2010.

[13] L. Leppänen, M. Munezero, M. Granroth-Wilding, and H. Toivonen, “Data-Driven News Generation for Automated Journalism ,” Proceedings of the 10th International Conference on Natural Language Generation, Association for Computational Linguistics, Santiago de Compostela, Spain, pp. 188–197, 2017.

[14] Zadbuke(unpublished), “Automatic Summarization of News Articles using TextRank,” International Journal of Advanced Research in Computer Science and Software Engineering , vol. 6, no. 3, pp. 124–127, Mar. 2016.

[15] Riya Jhalani, Yogesh Kumar, and Meena, “An Abstractive Approach For Text Summarization,” International Journal of Advanced Computational Engi- neering and Networking, ISSN: 2320-2106, vol. 5, no. 1, Jan. 2017.

[16] J. Gong, W. Ren, and P. Zhang, “An automatic generation method of sports news based on knowledge rules,” 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017. [17] Larseidnes, “Auto-Generating Clickbait With Recurrent Neural Networks,” Lars Eidnes' blog, 19-Apr-2018. [Online]. Available: https://larseidnes.com/2015/10/13/auto-generating-clickbait-with-recurrent-neural-networks/. [Accessed: 26-Sep-2018].

[18] The Unreasonable Effectiveness of Recurrent Neural Networks. [Online]. Available: http://karpathy.github.io/2015/05/21/rnn-effectiveness/. [Accessed: 03-Oct-2018].

[19] Ayana, Shen Shi-Qi, Lin Yan-Kai, Tu Cun-Chao, Zhao Yu, Liu Zhiyuan and Sun Mao-Song, “Recent advances on neural headline generation,” Journal of Computer Science and Technology, vol. 32, no. 4, pp. 768–784, Jul. 2017.

[20] D. Zhou, L. Guo, and Y. He, “Neural Storyline Extraction Model for Storyline Generation from News Articles,” Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1727–1736, 2018.

[21] H. T. Zheng, W. Wang, W. Chen and A. K. Sangaiah, “Automatic Gener- ation of News Comments Based on Gated Attention Neural Networks,” in IEEE Access, vol. 6, pp. 702–710, 2018.

[22] Park Keunchan, Lee Jisoo and Choi Jaeho, “Deep Neural Networks for News Recommendations,” Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM, New York, NY, USA, pp. 2255–2258, 2017.

[23] Chen Kuan-Yu (unpublished), “Chen Kuan-Yu ,” in IEEE/ACM, Transactions on Audio, Speech, and Language Processing, vol. 23, no. 8, pp. 1322–1334, Aug. 2015.

[24] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] Martin Sundermeyer, Ralf Schlüter and Hermann Ney, “LSTM neural net- works for language modeling,” 13th annual conference of the international speech communication association, Portland, Oregon, USA, pp. 194–197, 2012.

[26] Greene D and Cunningham P, “Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering,” in Proc. 23rd International Conference on Machine learning, ICML ’06, ACM, New York, NY, USA, pp. 377–384, 2006.

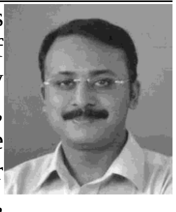
[27] Z. Shi, X. Chen, X. Qiu, and X. Huang, “Toward Diverse Text Generation with Inverse Reinforcement Learning,” Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp. 4361–4367, 2018.

[28] K. E. A. Papineni, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, pp. 311–318, 2002.

[29]Apurva Dube, Pradnya Gotmare, “Semantics Based Document Clustering”, International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.4, pp.25-30, August 2017

[30]M.Chahbar, A.Elhore, Y.Askane, “PERO2: Machine Teaching based on a Normalized Ontological Knowledge Base”, International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.5, pp.63-74, October 2017

Authors Profile



SHINE K GEORGE- Shine K George is currently pursuing Ph.D. in Department of Computer Applications, at Cochin University of Science and Technology, Kochi, Kerala, India. He is also working as associate professor in Department of Computer Applications, Union Christian College, Aluva, Kerala, India. His main research work focuses on ontology, knowl- edge management, machine learning and deep learning. He has 14 years of teaching experience and 7 years of Research Experience.

JAGATHY RAJ V. P- Dr. Jagathy Raj V. P., currently working as professor in Operations and Systems Management at School of Management Studies, Cochin Uni- versity of Science and Technology, Kochi, is a Ph.D holder in industrial engineering and management from Indian Institute of Technology (IIT), Kharagpur,India. Dr. Jagathy Raj did his B.Tech in Electrical Engineering from University of Kerala, India and M.Tech (electronics with communication as specialization) and MBA (Systems and Operations Management) from Cochin University of Science and Technology, Kerala, India.