LDA-Based Scoring of Sequences Generated by RNN for Automatic Tanka Composition

Tomonari Masada[1 [0000−0002−8358−3699]] and Atsuhiro Takasu[2]

[1] Nagasaki University, 1-14 Bunkyo-machi, Nagasaki-shi, Nagasaki, Japan
masada@nagasaki-u.ac.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
takasu@nii.ac.jp

Abstract. This paper proposes a method of scoring sequences generated by recurrent neural network (RNN) for automatic Tanka composition. Our method gives sequences a score based on topic assignments provided by latent Dirichlet allocation (LDA). When many word tokens in a sequence are assigned to the same topic, we give the sequence a high score. While a scoring of sequences can also be achieved by using RNN output probabilities, the sequences having large probabilities are likely to share much the same subsequences and thus are doomed to be deprived of diversity. The experimental results, where we scored Japanese Tanka poems generated by RNN, show that the top-ranked sequences selected by our method were likely to contain a wider variety of subsequences than those selected by RNN output probabilities.

Keywords: Topic modeling · Sequence generation · Recurrent neural network · Automatic poetry composition

1 Introduction

Recurrent neural network (RNN) is a class of artificial neural network that has been providing remarkable achievements in many important applications. *Sequence generation* is one among such applications [10][6]. By adopting a special architecture like LSTM [8] or GRU [3][4], RNN can learn dependencies among word tokens appearing at distant positions. When we use sequence generation in realistic situations, we may sift the sequences generated by RNN to obtain only useful ones. Such sifting may give each generated sequence a score representing its usefulness relative to the application under consideration. While an improvement of the generated sequences can also be achieved by modifying the architecture of RNN [5], we here consider a scoring method that can be tuned and applied separately from RNN. In particular, this paper proposes a method achieving a diversity of subsequences appearing in highly scored sequences.

We may score the generated sequences by using their output probabilities in RNN. However, this method is likely to give high scores to the sequences containing subsequences popular among the sequences used for training RNN. Consequently, the sequences of high score are doomed to look alike and to show

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_33

only a limited diversity. In contrast, our scoring method uses latent Dirichlet allocation (LDA) [2]. Topic models like LDA can extract diverse topics from training documents. By using the per-topic word probabilities LDA provides, we can assign high scores to the sequences containing many words having a strong relevance to some particular topic. Our scoring method is expected to select the sequences individually being relevant to some particular topic and together being relevant to diverse topics. We performed an evaluation experiment by generating Japanese Tanka poems with RNN. After training RNN under different settings, we chose the best setting in terms of validation perplexity. We then generated random sequences by using RNN under the best setting. The generated sequences were scored by using their output probabilities in RNN or by using our LDA-based method. The results show that our LDA-based method could select more diverse sequences in the sense that a wider variety of subsequences were obtained as the parts of the top-ranked Tanka poems.

2 Method

2.1 Preprocessing of Tanka Poems

Sequence generation is one among the important applications of RNN [10][6]. This paper considers generation of Japanese Tanka poems. We assume that all Tanka poems for training RNN are given in Hiragana characters with no voicing marks. Tanka poems have a 5-7-5-7-7 syllabic structure and thus consist of five subsequences, which are called *parts* in this paper. Here we give an example of Tanka poem taken from *The Tale of Genji*: "mo no o mo hu ni / ta ti ma hu he ku mo / a ra nu mi no / so te u ti hu ri si / ko ko ro si ri ki ya." In this paper, we use *Kunrei-shiki* romanization for Hiraganas. While the first part of the standard syllabic structure consists of five syllables, that of this example consists of six. In this manner, a small deviation from the standard 5-7-5-7-7 structure is often observed. Our preprocessing addresses this kind of deviation.

First, we put a spacing character '_' between each neighboring pair of parts and also at the tail of the poem. Moreover, the first Hiragana character of each of the five parts is "uppercased," i.e., marked as distinct from the same character appearing at the other positions. The above example is then converted to: "MO no o mo hu ni _ TA ti ma hu he ku mo _ A ra nu mi no _ SO te u ti hu ri si _ KO ko ro si ri ki ya _." Second, we represent each Tanka poem as a sequence of *non-overlapping* character bigrams, which are regarded as vocabulary words composing each Tanka poem. However, only to the parts containing an even number of Hiragana characters, we apply an additional modification. The special bigram '(_ _)' is put at the tail of such parts in place of the spacing character '_'. Consequently, the non-overlapping bigram sequence corresponding to the above example is obtained as: "(MO no) (o mo) (hu ni) (_ _) (TA ti) (ma hu) (he ku) (mo _) (A ra) (nu mi) (no _) (SO te) (u ti) (hu ri) (si _) (KO ko) (ro si) (ri ki) (ya _)." Finally, we put a token of the special words 'BOS' and 'EOS' at the head and the tail of each sequence, respectively. The sequences preprocessed in this manner were used for training RNN and also for training LDA.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_33

2.2 Tanka Poem Generation by RNN

We downloaded 179,225 Tanka poems from the web site of International Research Center for Japanese Studies.[1] 3,631 different non-overlapping bigrams were found to appear in this set. Therefore, the vocabulary size was 3,631. Among the 179,225 Tanka poems, 143,550 were used for training both RNN and LDA, and 35,675 were used for validation, i.e., for tuning free parameters. We implemented RNN with PyTorch[2] by using LSTM or GRU modules. RMSprop [11] was used for optimization with the learning rate 0.002. The mini-batch size was 200. The number of hidden layers was three. The dropout probability was 0.5. Based on an evaluation in terms of validation set perplexity, the hidden layer size was set to 600. Since the validation perplexity of GRU-RNN was slightly better than that of LSTM-RNN, GRU-RNN was used for generating Tanka poems.

2.3 LDA-based Sequence Scoring

This paper proposes a new method of scoring the sequences generated by RNN. We use latent Dirichlet allocation (LDA) [2], the best-known topic model, for scoring. LDA is a Bayesian probabilistic model of documents and can model the difference in semantic contents of documents as the difference in mixing proportions of topics. Each topic is in turn modeled as a probability distribution defined over vocabulary words. We denote the number of documents, the vocabulary size, and the number of topics by D, V, and K, respectively. By performing an inference for LDA via variational Bayesian inference [2], collapsed Gibbs sampling (CGS) [7], etc., over training set, we can estimate the two groups of parameters: θ_{dk} and ϕ_{kv}, for $d = 1, \ldots, D$, $v = 1, \ldots, V$, and $k = 1, \ldots, K$. The parameter θ_{dk} is the probability of the topic k in the document d. Intuitively, θ_{dk} quantifies the importance of each topic in each document. The parameter ϕ_{kv} is the probability of the word v in the topic k. Intuitively, ϕ_{kv} quantifies the relevance of each vocabulary word to each topic. For example, in autumn, people talk about fallen leaves more often than about blooming flowers. Such topic relevancy of each vocabulary word is represented by ϕ_{kv}.

In our experiment, we regarded each Tanka poem as a document. The inference for LDA was performed by CGS, where we used the same set of Tanka poems as that used for training RNN. Therefore, $D = 143,550$ and $V = 3,631$ as given in Subsection 2.2. K was set to 50, because other values gave no significant improvement. The Dirichlet hyperparameters of LDA were tuned by a grid search [1] based on validation set perplexity. Table 1 gives an example of the 20 top-ranked words in terms of ϕ_{kv} for three among $K = 50$ topics. Each row corresponds to a different topic. The three topics represent blooming flowers, autumn moon, and singing birds, respectively from top to bottom. For example, in the topic corresponding to the top row, the words "ha na" (flowers), "ha ru" (spring), "ni ho" (the first two Hiragana characters of the word "ni ho hi," which means fragrance), and "u me" (plum blossom) have large probabilities.

[1] http://tois.nichibun.ac.jp/database/html2/waka/menu.html
[2] http://pytorch.org/

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_33

Table 1. An Example of Topic Words obtained by CGS for LDA

(HA na) (ha na) (HA ru) (no _) (hi su) (U ku) (ni _) (YA ma) (hu _) (sa to) (ru _) (NI ho) (U me) (ha ru) (ta ti) (HU ru) (SA ki) (U no)
(tu ki) (no _) (NA ka) (ni _) (TU ki) (A ka) (KU mo) (te _) (ka ke) (wo _) (so ra) (ki yo) (ha _) (KA ke) (ka ri) (hi no) (yu ku) (KO yo)
(su _) (to ki) (HO to) (ko ye) (NA ki) (NA ku) (KO ye) (HI to) (ho to) (ni _) (ni na) (su ka) (ku _) (MA tu) (HA tu) (ku ra) (MA ta) (ya ma)

Our sequence scoring uses the ϕ_{kv}'s, i.e., the per-topic word probabilities, learned by CGS for LDA. Based on the ϕ_{kv}'s learned from the training set, we can estimate the topic probabilities of unseen documents by *fold-in* [1]. In our case, bigram sequences generated by RNN are unseen documents. When the fold-in procedure estimates θ_{dk} for some k as far larger than $\theta_{dk'}$ for $k' \neq k$, we can say that the document d is exclusively related to the topic k. In this manner, LDA can be used to know if a given Tanka poem is exclusively related to some particular topic. By using the fold-in estimation of θ_{dk} for a Tanka poem generated by RNN, we compute the entropy $-\sum_{k=1}^{K} \theta_{dk} \log \theta_{dk}$, which is called *topic entropy* of the Tanka poem. Smaller topic entropies are regarded as better, because smaller ones correspond to the situations where the Tanka poems relate to some particular topic more exclusively. In other words, we would like to select the poems showing a topic consistency. Since LDA can extract a wide variety of topics, our scoring method is expected to select the sequences individually showing a topic consistency and together showing a topic diversity.

3 Evaluation

The evaluation experiment compared our scoring method to the method based on RNN output probabilities. The output probability in RNN can be obtained as follows. We generate a random sequence with RNN by starting from the special word 'BOS' and then randomly drawing words one by one until we draw the special word 'EOS.' The output probability of the generated sequence is the product of the output probabilities of all tokens, where the probability of each token is the output probability at each moment during the sequence generation. Our LDA-based scoring was compared to this probability-based scoring.

We first investigate the difference of the top-ranked Tanka poems obtained by the two compared scoring methods. Table 2 presents an example of the top five Tanka poems selected by our method in the left column and those selected based on RNN output probabilities in the right column. To obtain these top-ranked poems, we first generated 100,000 Tanka poems with the GRU-RNN. Since the number of poems containing grammatically incorrect parts was large, a grammar check was required. However, we could not find any good grammar check tool for archaic Japanese. Therefore, as an approximation, we regarded the poems containing at least one part appearing in no training Tanka poem as grammatically incorrect. After removing grammatically incorrect poems, we

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_33

Table 2. Top five Tanka poems selected by compared methods

(rank)	Topic entropy	Output probability
1	si ku re yu ku ka tu ra ki ya ma no i ro hu ka ki mo mi ti no i ro ni si ku re hu ri ke ri	o ho a ra ki no mo ri no si ta ku sa ka mi na tu ki mo ri no si ta ku sa ku ti ni ke ru ka na
2	ti ha ya hu ru yu hu hi no ya ma no ka mi na hi no mi mu ro no ya ma no mo mi ti wo so mi ru	hi sa ka ta no ku mo wi ni mi yu ru tu ki ka ke no tu ki ka ke wa ta ru a ma no ka ku ya ma
3	ka he ru ya ma hu mo to no mi ti ha ki ri ko me te hu ka ku mo mu su hu wo ti no ya ma ka se	ti ha ya hu ru ka mo no ka ha na mi ta ti ka he ri ki ru hi to mo na ki a hu sa ka no se ki
4	o ho ka ta mo wa su ru ru ko to mo ka yo he to mo o mo hu ko ko ro ni o mo hu ha ka ri so	ta ka sa ko no wo no he no sa ku ra sa ki ni ke ri mi ne no ma tu ya ma yu ki hu ri ni ke ri
5	u ti na hi ku ko ro mo te sa mu ki o ku ya ma ni u tu ro hi ni ke ri a ki no ha tu ka se	ta ka sa ko no wo no he no sa ku ra na ka mu re ha a ri a ke no tu ki ni a ki ka se so hu ku

assigned to the remaining ones a score based on each method. Table 2 presents the resulting top five Tanka poems for each method.

Table 2 shows that when we use RNN output probabilities (right column), it is difficult to achieve topic consistency. The fourth poem contains the words "sa ku ra" (cherry blossom) and "yu ki" (snow). The fifth one contains the words "sa ku ra" (cherry blossom) and "a ki ka se" (autumn wind). In this manner, the poems top-ranked based on RNN output probabilities sometimes contain the words expressing different seasons. This is prohibitive for Tanka composition. In contrast, the first poem selected by our method contains the words "si ku re" (drizzling rain) and "mo mi ti" (autumnal tints). Because drizzling rain is a shower observed in late autumn or in early winter, the word "mo mi ti" fits well within this context. In the third poem selected by our method, the words "ya ma" and "hu mo to" are observed. The former means mountain, and the latter means the foot of the mountain. This poem also shows a topic consistency. However, a slight weakness can be observed in the poems selected by our method. The same word is likely to be used twice or more. While refrains are often observed in Tanka poems, some future work may introduce an improvement here.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_33

Table 3. Five most frequent parts observed in the top 200 Tanka poems

Topic entropy	Output probability
hi sa ka ta no (9)	a ri a ke no tu ki no (13)
ko ro mo te sa mu ki (8)	hi sa ka ta no (12)
a ri a ke no tu ki ni (8)	ta ka sa ko no (12)
ho to to ki su (7)	ho to to ki su (11)
a ri a ke no tu ki (6)	a si hi ki no (10)

We next investigate the diversity of selected Tanka poems. We picked up the 200 top-ranked poems given by each method and then split each poem into five parts to obtain $1,000$ parts in total. Since the resulting set of 1,000 parts included duplicates, we grouped those parts by their identity and counted duplicates. Table 3 presents the five most frequent parts for each method. When we used RNN output probabilities (right column), "a ri a ke no tu ki no" appeared 13 times, "hi sa ka ta no" 12 times, and so on, among the 1,000 parts coming from the 200 top-ranked poems. In contrast, when we used our LDA-based scoring (left column), "hi sa ka ta no" appeared nine times, "ko ro mo te sa mu ki" eight times, and so on. That is, there were less duplicates for our method. Moreover, we also observed that while only 678 parts among 1,000 were unique when RNN output probabilities were used, 806 were unique when our method was used. It can be said that our method explored a larger diversity.

4 Previous Study

While there already exist many proposals of sequence generation using RNN, LDA is a key component in our method. Therefore, we focus on the proposals using topic modeling. Yan et al. [12] utilize LDA for Chinese poetry composition. However, LDA is only used for obtaining word similarities, not for exploring topical diversity. The combination of topic modeling and RNN can be found in the proposals not related to automatic poetry composition. Dieng et al. [5] propose a combination of RNN and topic modeling. The model, called TopicRNN, modifies the output word probabilities of RNN by using long-range semantic information of documents captured by an LDA-like mechanism. However, when we generate random sequences with TopicRNN, we need to choose one document among the existing documents as a seed. This means that we can only generate sequences similar to the document chosen as a seed. While TopicRNN has this limitation, it provides a valuable guide for future work. Our method detaches sequence selection from sequence generation. It may be better to directly generate the sequences having some desirable property regarding their topical contents.

5 Conclusions

This paper proposed a method for scoring sequences generated by RNN. The proposed method was compared to the scoring using RNN output probabilities.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_33

The experiment showed that our method could select more diverse Tanka poems. In this paper, we only consider the method for obtaining better sequences by screening generated sequences. However, the same thing can also be achieved by modifying the architecture of RNN. As discussed in Section 4, Dieng et al. [5] incorporate an idea from topic modeling into the architecture of RNN. It is an interesting research direction to propose an architecture of RNN that can directly generate sequences diverse in topics. With respect to the evaluation, it is a possible research direction to apply evaluations using BLEU [9] or even human subjective evaluations for ensuring the reliability.

Acknowledgments

This work was supported by Grant-in-Aid for Scientific Research (B) 15H02789.

References

1. Asuncion, A., Welling, M., Smyth, P., Teh, Y. W.: On smoothing and inference for topic models. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09), pp. 27–34 (2009)
2. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint, arXiv:1409.1259 (2014)
4. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint, arXiv:1412.3555 (2014)
5. Dieng, A. B., Wang, C., Gao, J., Paisley, J.: TopicRNN: A recurrent neural network with long-range semantic dependency. arXiv preprint, arXiv:1611.01702 (2016)
6. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint, arXiv:1308.0850 (2013)
7. Griffiths, T. L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. U. S. A. **101**, Suppl. 1, 5228–35 (2004)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02), pp. 311–318 (2002)
10. Sutskever, I., Martens, J., Hinton, G.: Generating text with recurrent neural networks. In: Proceedings of the 28th International Conference on Machine Learning (ICML '11), pp. 1017–1024 (2011)
11. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. In: COURSERA: Neural Networks for Machine Learning **4** (2012)
12. Yan, R., Jiang, H., Lapata, M., Lin, S.-D., Lv, X., Li, X.: i, Poet: Automatic Chinese poetry composition through a generative summarization framework under constrained optimization. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI '13), pp. 2197–2203 (2013)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_33