Neural-based Chinese Idiom Recommendation for Enhancing Elegance in Essay Writing

Yuanchao Liu, Bo Pang, Bingquan Liu

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

{ycliu, bpang, liubq}@hit.edu.cn

Abstract

Although the proper use of idioms can enhance the elegance of writing, the active use of various expressions is a challenge because remembering idioms is difficult. In this study, we address the problem of idiom recommendation by leveraging a neural machine translation framework, in which we suppose that idioms are written in one pseudo target language. Two types of real-life datasets are collected to support this study. Experimental results show that the proposed approach achieves promising performance compared with other baseline methods.

1 Introduction

Nearly every language has some ancient idioms, aphorisms, and sayings from history (Muzny et al., 2013; Moussallem et al., 2018). Chinese idioms, also known as "ready phrases" and usually consist of only four characters, can reveal complex meaning and enhance the conciseness and elegance of writing if properly used. For example, in the text segment "*一夜春雷雨，朋友圈的微商如**雨后春笋**般冒了出来。*" (*During the thunderstorm overnight, microbusinessmen in the circle of friends sprang up like bamboo shoots after rain.*), the author elegantly describes the rapid emergence of things in large numbers by properly using the popular idiom "***雨后春笋***" (*When it rains in spring, many bamboo shoots grow simultaneously*). Therefore, automatically recommending idioms that are pertinent to the input context is an appealing task because remembering idioms is difficult for most people.

To this end, one typical and straightforward approach is to regard idiom recommendation as a standard classification problem and assign a piece of context to one idiom label by training corresponding classifiers. Whereas by doing so, the meaningful text information in the idiom itself tends to be ignored. Intuitively, combining textual information in the context and idiom in the training stage may be helpful. However, texts in the idiom are usually written in ancient classical Chinese for conciseness; thus, they are highly different from those in the context and difficult to directly utilize for classifying unseen contexts. In most cases, such as in the aforementioned example, few common words or characters are shared between the idiom and the surrounding context.

In this study, we provide a new perspective for idiom recommendation by formulating it as a translation problem, in which the idioms are assumed to be written with a pseudo target language because they are usually written in ancient Chinese and have special and limited vocabularies. We propose a machine translation-based approach that operates in three stages. First, an attention-based neural network is used to encode the context sequence (source language). Second, the coded context attention vector is decoded into one intermediate sequence (target language). Third, the final recommended idioms are selected through sequence mapping.

The remainder of this paper is organized as follows. The related work is surveyed in Section 2. Sections 3 and 4 present the proposed approach and experimental results, respectively. Finally, conclusions and future directions are drawn in Section 5.

2 Related works

Our task can be viewed as a content-based recommendation, and the closely related work includes scientific article citation (He et al., 2010), news (Lu et al., 2014), and quotation (Tan et al., 2015) recommendations. He et al. (2010) used a context-aware approach and measured the relevance between context and candidate items for scientific citation recommendation. Tan et al. (2015) proposed a supervised ranking framework to recommend quotes for writing. The difference

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5522–5526
Florence, Italy, July 28 - August 2, 2019.

between idiom recommendation and the above ones is that idioms were usually formed in ancient times and commonly written in classical Chinese, thereby exhibiting few common surface features with context.

Sequence-to-sequence models (Sutskever et al., 2014; Cho et al., 2014) have recently received great success in various tasks, such as machine translation (Bahdanau et al., 2015), image caption generation (Xu et al., 2015), and text summarization (Chopra et al., 2016). Cho et al. (2014) showed that the performance of a basic encoder–decoder rapidly deteriorates as the length of input context increases. Correspondingly, attention mechanism (Bahdanau et al., 2015; Luong et al., 2015) has been proposed to address such types of problem.

3 Methodology

We formulate idiom recommendation as a context-to-idiom machine translation problem by using the encoder–decoder framework. Figure 1 shows the architecture of our approach. This scheme works by taking an idiom-bearing sentence and yielding the idiom as output. The framework consists of five layers from the embedding (bottom) to the prediction (top) layer. The encoder and decoder separately receive the words in the source context sentence and characters in the target idiom as inputs. The implementation of each layer is presented as follows.

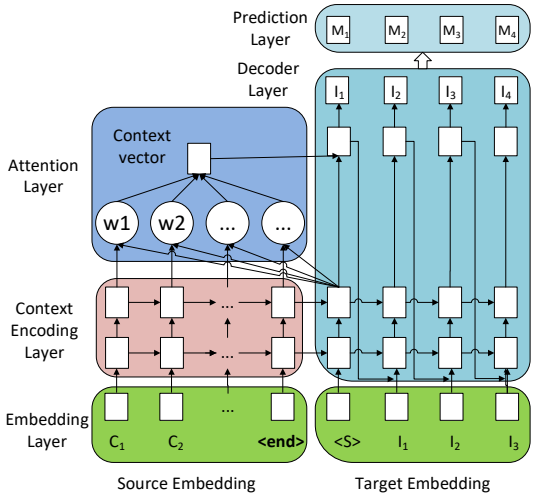


Figure 1: Graphical illustration of the proposed model.

3.1 Embedding Layer

The model determines the source and target embeddings to retrieve the corresponding word representations. A vocabulary is initially selected for context and idiom separately. For context, only the frequent words (fre. ≥ 2) are treated as unique to reduce the effect of noise that is usually caused by low-frequency words. For target idioms, all the unique Chinese characters shown in the idioms are used to create the vocabulary because there is a relatively limited character set for the idioms.

3.2 Context Encoding Layer

The word embeddings retrieved from the embedding layer are fed into the encoder for the source language C (context) and decoder for the target language I (idiom). We use a bidirectional long short-term memory (BiLSTM) network (Graves et al., 2013) to capture the left and right contexts of each word in the input.

$$[\vec{h}^{\,C}_i, \vec{c}^{\,C}_i] = \overrightarrow{LSTM_C}(t_i, \vec{h}^{\,C}_{i-1}, \vec{c}^{\,C}_{i-1}), \quad (1)$$

$$[\overleftarrow{h}^{\,C}_i, \overleftarrow{c}^{\,C}_i] = \overleftarrow{LSTM_C}(t_i, \overleftarrow{h}^{\,C}_{i+1}, \overleftarrow{c}^{\,C}_{i+1}), \quad (2)$$

where $h \in \mathbb{R}^{d\times 1}$ and $c \in \mathbb{R}^{d\times 1}$ are the hidden and cell states of the LSTM, respectively; $\rightarrow$ ($\leftarrow$) indicates the forward (backward) pass; and t_i is the input context word vector at time step i. Then, the output for each input is the concatenation of the two vectors from both directions. The bottom half of the decoding layer for the idiom also takes the same measures, whereas the Chinese character is used for each time step. The last source state from the encoder is passed to the decoder when the decoding process is initiated.

3.3 Attention Layer

Various words in the long context are generally of different importance. For example, the context words "*冒*" (*sprang up*) and "*出来*" (*show up*) in the aforementioned example are intuitively strong indicators for recommending the idiom "*雨后春笋*" (*When it rains in spring, many bamboo shoots grow simultaneously*). Thus, increased attention should be given to such words. Consequently, a feasible solution is to introduce attention mechanism. Thus, various attention weights are given to different input words.

We use a global attentional model (Luong et al., 2015) to obtain the attention vector. This model consists of the following stages:

1. The current target hidden state is compared with all the source states to calculate the attention weights α_{ts}, as shown as follows:

$$\alpha_{ts} = \frac{\exp(score(\tilde{h}_t, \bar{h}_s))}{\sum_{s'=1}^{S} \exp(score(\tilde{h}_t, \bar{h}_{s'}))}, \quad (3)$$

where the function $score$ is used to produce attention weights. In the training stage, we extend the target hidden state as $\tilde{h}_t = V[h_t; h_m]$, where $V \in \mathbb{R}^{d\times 2d}$, h_t is the target hidden state, and h_m is the average of the embedding of all the words in the modern plain text meaning of the idiom. Then, we compare the extended target hidden state $\tilde{h}_t$ with each of the source hidden states $\bar{h}_s$ to compute $score$ (i.e., $score(\tilde{h}_t, \bar{h}_s) = \tilde{h}_t^{\mathrm{T}} W \bar{h}_s$, where $\mathrm{W} \in \mathbb{R}^{d\times 2d}$).

2. Then, the context vector c_t is calculated as the weighted average of the source states.

$$c_t = \sum_s \alpha_{ts} \bar{h}_s \quad (4)$$

3. Finally, the attention vector α_t is derived by combining the context vector with the current target hidden state h_t.

$$\alpha_t = tanh(W_c[c_t; h_t]) \quad (5)$$

3.4 Decoder Layer

Given the attention vector α_t and all the previously predicted target idiom characters $\{I_1, \cdots, I_{t-1}\}$, the decoder layer defines a probability over the translation by decomposing the joint probability into the ordered conditionals to predict the next character I_t.

$$\mathrm{p(I)} = \prod_{t=1}^{T} p(I_t | \{I_1, \ldots I_{t-1}\}, \alpha_t) \quad (6)$$

We use BiLSTM to model each conditional probability (Bahdanau et al., 2015). In the decoder layer, we create a candidate character table for different locations in the idiom to decrease the decoding space. For example, when generating the first character in the preceding example, "雨" (*rain*) is eligible because it is in the table of Position 1, which consists of all the unique characters shown in the first position of all idioms. Thus, many other ineligible characters that are not in this table will be naturally ignored.

3.5 Prediction Layer

Many standard idioms are present in our work compared with the traditional machine translation. Therefore, the translated character sequences in this layer are further mapped into the standard idioms in the idiom set (i.e., I_* to M_* in Figure 1). To achieve this goal, we use edit distance (Navarro, 2001) to find the most similar idiom from the standard idiom set as the prediction result.

4 Experiments

4.1 Experimental settings

Datasets. We carry out experiments on two datasets, which are referred as BN and WB respectively. The datasets are collected from *Weibo* and *Baidu News* as two data sources to get the short context by inputting the idiom as the query. Table 1 provides the details of the datasets.

Table 1. Details of the datasets[1].

Dataset	# of total pairs	# of snippets per Idiom	# of Idioms
WB	167,844	≈176	956
BN	163,817	≈171	956

Baselines and Evaluation Metrics. We conduct experiments using the following baselines: (1) Elastic Net, (2) KNN (K-Nearest Neighbor), (3) Multinomial naive Bayes, (4) LinearSVC. We use the scikit-learn (Version 0.19) implementation[2] of the above models (using the default settings) for the experiments. We also experiment with several neural network based classification approaches, namely, (5) TextCNN (Convolutional neural network) (Kim et al., 2014) and (6) Bi-LSTM-RNN (Graves et al., 2013), and (7) HierAtteNet (Hierarchical attention network) (Yang et al., 2017). All the review texts are segmented into Chinese words using Jieba[3].

We mainly use recall as the primary recommendation metrics in accordance with the study of He et al. (2010). We remove the original idioms from the testing documents. The recall is defined as the percentage of original idioms that appear in the recommended ones. Moreover, we also use smoothed BLEU[4], which is widely used in MT performance evaluation, to examine the intermediate results of our approach.

Training Details. We use a minibatch stochastic gradient descent (SGD) algorithm and Adadelta (Zeiler, 2012) to train each model. A total of 12 training epochs is conducted, and a simple learning rate schedule begins with a learning rate of 1.0, followed by six epochs. Then, the learning rate is divided every epoch. Each SGD update direction is computed using a minibatch of 128 snippets. We set the dropout to 0.2, target max length to 4, and source max length to 50. The

[1] The datasets are available at http://u.163.com/syyAdG6P, pass code: YdgIfzHn
[2] http://scikit-learn.org/
[3] https://pypi.python.org/pypi/jieba/
[4] http://www.nltk.org/

pretrained Chinese word and Chinese idiom character embeddings are trained by word2vec (Mikolov et al., 2013) toolkit, and unseen words are assigned with unique random vectors. Both languages have a set of embedding weights because they actually come from the same mother language, although considerable differences exist in their vocabulary sets.

4.2 Results and Analysis

In the first experiment, we compare the performance of our approach with baseline methods. We separate our datasets into 8:1:1 as the training, validation, and test sets. Table 2 summarizes the performance comparison on WB and BN datasets.

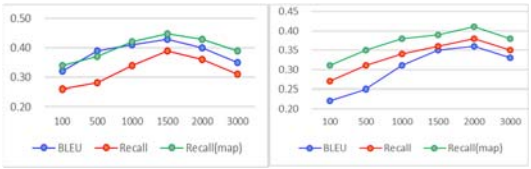
Table 2. Comparison with baseline methods.

Method	WB	BN
Elastic Net (loss=hinge)	0.239	0.378
KNN (n-neibous=10)	0.182	0.225
Multinomial Naive Bayes	0.164	0.314
LinearSVC	0.221	0.339
Bi-LSTM-RNN	0.294	0.395
TextCNN	0.325	0.386
HATT	0.362	0.412
Proposed method	0.412	0.448

Evidently, the proposed method notably outperforms all the other baseline methods on both datasets due to the following reasons. First, user-generated content is inherently noisy. The classification performance may be adversely affected by the considerable classes because of the hundreds of idioms present. Conversely, the proposed method focuses on the salient words in the context, thereby alleviating the adverse effect of noisy words to some extent. Second, the proposed encoder–decoder framework provides substantial advantages in this task: in comparison with many classification approaches that regard the entire idiom as a classification label, our approach considers the relationship between the context and the character inside the idiom by using attention-based neural machine translation architecture because some characters in the idiom have a close relationship with the context.

Notably, neither the attention-based NMT nor other approaches effectively perform in recommendation across the two datasets. The recall values of BN and WB are 44.8% and 41.2%, respectively, thereby indicating that nearly half of the context cannot obtain the original idiom recommended. One possible reason is that the quality of the corpora considerably influences the result. Sometimes, selecting the suitable idioms according to the context may be relatively difficult for experienced people, not to mention for the models.

In the second experiment, we intend to examine the performance with different number of iterations. Subpanels (a) and (b) of Figure 2 depict the BLEU and recall of BN and WB datasets, respectively, when the iteration number varies from 100 to 3000. The result shows that the recommendation performance can greatly improve by increasing the number of iterations, thereby obtaining excellent results for iterations of approximately 1000 to 1500. However, after considerable iterations (greater than 2000), decreasing trends are observed for the model performance. This result is due to overfitting of the training data with numerous iterations. Moreover, when mapping is added, an increase is observed in the recall, this indicates that the transformation in prediction layer is necessary to recommend the idiom from the standard set.



(a) BN (b) WB

Figure 2: Metrics as a function of the number of iterations of our model on both datasets.

5 Conclusion

In this study, we address the appealing problem of idiom recommendation on the basis of the surrounding context and formulate it as a translation task. The evaluation results over two datasets demonstrate the effectiveness of the proposed approach. In the future, several ways of extending our model (e.g., exploring more attention mechanisms, such as location attention) are suggested to encode the context, because some particular locations in the context may be more important for different idioms. Moreover, substantial research will be conducted to propose other approaches for target language generation, which is one of the intermediary steps in our approach for the final idiom recommendation.

6 Acknowledgments

This study was supported by the National Natural Science Foundation of China (61672192 and 61572151).

References

Bahdanau, D., Cho, K., & Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. *ICLR 2015.*

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. 2014. On the properties of neural machine translation: Encoder–Decoder approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111

Graves A, Mohamed A, Hinton G. 2013. Speech Recognition with Deep Recurrent Neural Networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pages 6645–6649.

He, Q., Pei, J., Kifer, D., Mitra, P., and Giles, L. 2010. Context aware citation recommendation. *In Proceedings of the 19th international conference on World wide web.* pages 421–430.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014*, pages 1746–1751.

Lu, M., Qin, Z., Cao, Y., Liu, Z., & Wang, M. 2014. Scalable news recommendation using multi-dimensional similarity and jaccard-kmeans clustering. *Journal of Systems & Software*, 95(9), pages 242-251.

Luong M T , Pham H , Manning C D. Effective Approaches to Attention-based Neural Machine Translation. *EMNLP 2015*.

Mikolov T , Chen K , Corrado G , et al. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR. arXiv:1301.3781v1.*

Moussallem et al., 2018. LIDIOMS: A Multilingual Linked Idioms Data Set. arXiv:1802.08148

Muzny G. and Zettlemoyer L. 2013. Automatic Idiom Identification in Wiktionary. *EMNLP 2013*. pages 1417–1421

Navarro, Gonzalo. 2001. A guided tour to approximate string matching. *ACM Computing Surveys*. 33 (1): 31–88.

Sumit Chopra, Michael Auli, Alexander M. Rush. 2016. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks, *NAACL 2016*. pages 93-98.

Sutskever I, Vinyals O, Le Q V. 2014. Sequence to Sequence Learning with Neural Networks. *NIPS 2014*. pages 3104-3112.

Tan, J., Wan, X., Xiao, J. 2015. Learning to Recommend Quotes for Writing. *AAAI 2015.* pages 2453-2459

Xu K., Ba J., Kiros R. . 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *ICML2015*.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E.. 2017. Hierarchical Attention Networks for Document Classification. *NAACL 2017*, pages 1480-1489.

Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. arXiv:1212.5701.