Using Fractional Latent Topic to Enhance Recurrent Neural Network in Text Similarity Modeling

Yang Song(✉), Wenxin Hu, and Liang He

Department of Computer Science and Technology,
East China Normal University, Shanghai 200241, China
ysong@ica.stc.sh.cn, wxhu@cc.ecnu.edu.cn, lhe@cs.ecnu.edu.cn

Abstract. Recurrent neural networks (RNNs) have been widely used in text similarity modeling for text semantic representation learning. However, referring to the classical topic models, a text contains many different latent topics, and the complete semantic information of the text is described by all the latent topics. Previous RNN based models usually learn the text representation with the separated words in the text instead of topics, which will bring noises and loss hierarchical structure information for text representation. In this paper, we proposed a novel fractional latent topic based RNN (FraLT-RNN) model, which focuses on the text representation in topic-level and largely preserve the whole semantic information of a text. To be specific, we first adopt the fractional calculus to generate latent topics for a text with the hidden states learned by a RNN model. Then, we propose a topic-wise attention gating mechanism and embed it into our model to generate the topic-level attentive vector for each topic. Finally, we reward the topic perspective with the topic-level attention for text representation. Experiments on four benchmark datasets, namely TREC-QA and WikiQA for answer selection, MSRP for paraphrase identification, and MultiNLI for textual entailment, show the great advantages of our proposed model.

Keywords: Latent topic · Fractional calculus · Recurrent neural network

1 Introduction

Text similarity modeling is a crucial issue in many neural language processing (NLP) tasks, such as paraphrase identification [4,10], question answering [25,35], and textual entailment [18,23]. Take the paraphrase identification task as an example, text similarity is utilized to assess whether the two pieces of texts are semantically equivalent.

Recently, the recurrent neural networks (RNNs) have gained popularity in text similarity modeling, due to its good performance and less human interventions. Specifically, a hidden vector is learned for each word in the text via a hidden state in

G. Li et al. (Eds.): DASFAA 2019, LNCS 11447, pp. 173–190, 2019.
https://doi.org/10.1007/978-3-030-18579-4_11

RNN, and the whole text is represented by the aggregation of all the hidden vectors. Then, the similarity of a pair of texts is calculated with their representations and a similarity function. Most RNN based models, including those embedded with attention mechanisms, focus on using sequential hidden vector in word-level to generate the text representation, while the hierarchical structures of the text, such as features in phrase-level and sentence-level, are neglected. However, a piece of text has complicated structures, it is essential to understand and represent the text both sequentially and hierarchically [15].

Referring to the well-known topic models, such as latent dirichlet allocation (LDA) [3], it can be found that words in the text generate various topics, and the distribution of the topics demonstrates the semantic representation of the text. Noting that a topic is generated by a group of words, [4,35] use the aligned textual information to model text similarity, which first locates the textual snippets that have the same semantic meanings in the text pair, and then highlights the weight of those textual snippets for text representation. However, this method usually focuses on the co-occurrent words between the text pair, instead of the high-level topics, which will bring noise during text similarity modeling. Take the following text pair as an example for illustration.

> *T1: A child gets a fever, but he has no symptoms of influenza.*
> *T2: A kid with the symptoms of mild influenza and low fever can be cured by the Oseltamivir.*

Obviously, there are two topics in T1, namely "the child has a fever" as topic 1, and "the child has no influenza" as topic 2. It can be seen that topic 1 is relevant to T2, while topic 2 is not. Therefore, T1 and T2 should have a lower similarity. However, textual alignment approaches pay more attention on the aligned words, such as "fever" and "influenza", and will generate a higher similarity score for T1 and T2 and lead to mismatching problem. Hence, modeling the text similarity in topic-level is meaningful and promising, which should arouse much attention.

To the best of our knowledge, how to generate topics based on the words interactions and use interactions among topics for text similarity modeling are still not well studied in RNN. In this paper, we propose a **fra**ctional **l**atent **t**opic based RNN (**FraLT-RNN**) model, where the hierarchical features, namely features in word-level and topic-level, as well as the word sequential patterns, are incorporated into RNN for text representation by means of the fractional latent topics. To be specific, we first adopt the fractional latent topic generator to learn latent topics based on the hidden states learned by a RNN structure. In particular, the fractional latent topic generator is derived from the fractional calculus, which computes the function's integral in fractional order, instead of integer order, and has been successfully introduced into image processing for generation and denoising, due to its excellent characteristics in memory and heredity. Then, we design a topic-wise attention mechanism to generate an topic-level attentive vector for each latent topic, which measures the perspective of the latent topic and enhances the interactions between a text pair. Finally, the latent topics are rewarded by the attentive vector for text representation and similarity calculation. We evaluate

our FraLT-RNN model on four benchmark collections, namely Trec-QA and WikiQA for question answering, MSRP for paraphrase identification, and MultiNLI for textual entailment. The experimental results show great advantages of the proposed FraLT-RNN on text similarity modeling. It is notable that we achieve the new state-of-the-art performance on TREC-QA, WikiQA, and MSRP. Furthermore, our model is comparable to if not better than the recent neural network based approaches on MultiNLI.

The contribution of this paper is summarized as follows:

- We propose a new fractional latent topic based RNN model, where the text is represented in topic-level for better semantic capturing and understanding.
- This is the first attempt to introduce the fractional calculus into neural language processing for latent topics generation.
- We conduct elaborate analyses of the experimental results on three text similarity tasks, which provides a better understanding of the effectiveness of our model.

2 Related Work

Recently, the deep neural networks have been widely used in text similarity modeling [26,32], especially the recurrent neural networks (RNN) due to their capacity in modeling the sentence with variable length. [7,42] applied the long short-term memory (LSTM) [12] based RNN model to obtain the semantic relevance between text pairs for the community based question selection.

To capture the salient information for better sentence representations, the attention mechanism was introduced into the neural networks [25,31]. [41] proposed an attentive interactive neural network, which focused on the interactions between text segments for answer selection. In addition, the interactions in sentence-level or word-level are incorporated for the attentive weight generation within the RNN framework. In [29], the attentive weights for an answer sentence relied on the interactions with the question sentence. In [25], the word-by-word interactions were utilized for the attentive sentence representations.

Most of the previous work focused on representing the text in word-level, while the hierarchical structure of the text is neglected. Topic models, such as PLSA [13] and LDA [3], showed that words in the text generate various latent topics, and the perspectives of the latent topics demonstrate the semantic meaning of the text. Furthermore, topic models had shown great advantages in text understanding. In this paper, we will attempt to incorporate latent topics into RNN for text similarity modeling.

3 Fractional Latent Topic Based Recurrent Neural Network

In this section, we will introduce our fractional latent topic based RNN (FraLT-RNN) model for text similarity modeling in detail. For a better understanding,

we first give a brief introduction of the traditional RNN models as well as some notations used in this paper.

Given a pair of text as $X = \{x_1, x_2, ..., x_m\}$ and $Y = \{y_1, y_2, ..., y_n\}$, we let $\mathbf{x}_i$ and $\mathbf{y}_j$ denote the embedding representations of the word x_i and y_j respectively. The traditional RNN approaches model each text separately. Take the text X as an example, we suppose the sequential hidden state as $\mathbf{h}_1^x, \mathbf{h}_2^x, ..., \mathbf{h}_m^x$ and each hidden state corresponds to a word in text X. Then, the hidden state can be calculated by:

$$\mathbf{h}_i^x = f(\mathbf{x}_i, \mathbf{h}_{i-1}^x), \tag{1}$$

where f can be defined with the long short-term memory (LSTM) model or the gated recurrent unit (GRU), and $\mathbf{h}_i^x$ contains the context information from the first word to the current one [16]. After that, the text is represented by the attention [29] method over the hidden states as

$$\mathbf{H}_X = \sum_{i=1}^{m} \alpha_i \mathbf{h}_i^x, \tag{2}$$

where α_i is the attention of $\mathbf{h}_i^x$. Finally, the similarity score is computed according to the two text representations ($\mathbf{H}_X$ and $\mathbf{H}_Y$) and a similarity function, such as cosine similarity.

3.1 Framework of FraLT-RNN

Previous work on RNN based text similarity modeling mainly learned the text representation in word-level, namely using the word representation for sentence representation and similarity modeling, while the hierarchical structures of a text pair are neglected. Topics of a text are generated by groups of words, and the distribution of the topics has been used for text similarity modeling. Referring to the topic models, it can be found that the semantic meaning of a text is more susceptible to the perspective of the topics it involves compared with the word-level features. In this paper, we facilitate the hierarchical structures of a text in similarity modeling and proposed a fractional latent topic based RNN (FraLT-RNNN) model, which automatically learn latent topic and the topic attentive vector for text representation and similarity modeling.

The framework of our proposed FraLT-RNN model is shown in Fig. 1. Compared with the traditional RNN based models which learn the representation of a text with separated and independent words, our model takes a group of words in the text as a whole to generate latent topic which can better capture the hierarchical information and the topic-level interactions in the text pair. Specifically, we first generate the fractional latent topics for text $\mathbf{Y}$ by mean of a fractional latent topic generator which involves the hidden states learned by RNN and a fractional calculus. With this step, the latent topics in the text will be captured and encoded. Then, a topic-wise attention gating mechanism is presented and embedded into our model, which controls the information flow between the text pair. As shown in Fig. 1, besides the topic representation, the word-level attention based text representations ($\mathbf{H}_X$ and $\mathbf{H}_Y$) are also the inputs of the gating to

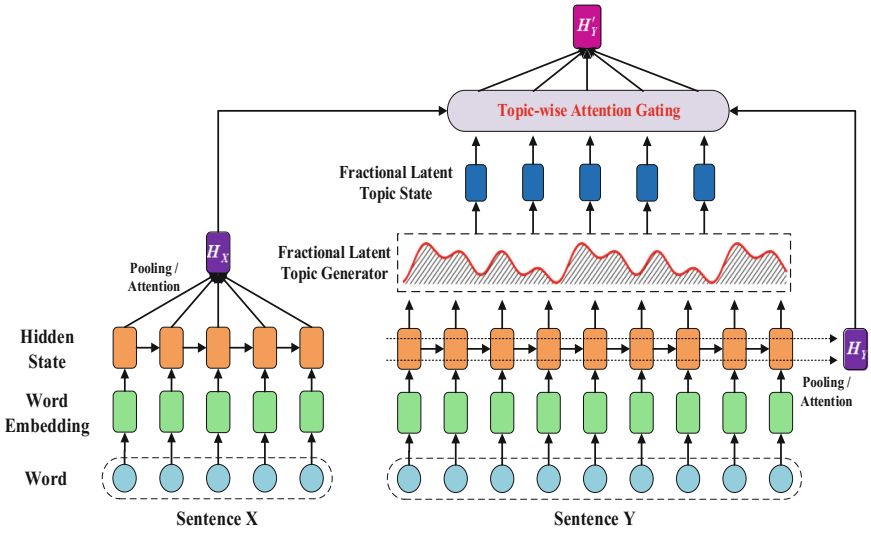


Fig. 1. A framework of fractional latent topic based RNN

help determine the perspectives of the topics. We will give a detailed description of each step in the following sections.

3.2 Fractional Latent Topic

Traditional topic models, such as PLSA [13] and LDA [3], usually use probabilities to analysis the latent topic distribution of the text, and have shown great advantages on text understanding and decoding. However, those models take the text as a bag of words and assume that words are independent with other, while the contextual information is neglected during latent topic generation. On the other hand, RNN models, such as LSTM and GRU, are excellent in processing and generating word sequence via taking the term dependency into consideration. Therefore, directly heap up the RNN model and traditional topic model for latent topic generation will miss contextual information such as sequential structure and term dependency in the text pair. To cope with this problem, we propose a fractional latent topic generator, which learns the latent topic of a text by means of the fractional calculus. Furthermore, the latent topic learned by the fractional latent topic generator is defined as the **fractional latent topic**. In the rest of this subsection, we will provide an insight into the fractional calculus, and then introduce the approach of the fractional latent topic generation.

Fractional Calculus. Fractional calculus, including fractional integral and fractional differential, which has been successfully used in image generation and denoising [2,22]. Different from the ordinary integral and differential which conduct computing in integer order, the factional calculus refines the computational

step into fractional order and considers the value of time-delayed states (namely the states before the current state) for the integral or differential implementation. Therefore, the fractional calculus has excellent characteristics of memory and heredity in processing sequential data. In this paper, we focus on the fractional integral, which is introduced in [17]. Suppose $f(x)$ is a continues function, then the Riemann-Louville definition of the fractional integral in α order is formulated as:

$$I^{\alpha} f(x) = \frac{1}{\Gamma(\alpha)} \int_0^x (x-t)^{\alpha-1} f(t) dt. \tag{3}$$

where I^{α} stands for the fractional integral operator, and $\Gamma(\cdot)$ is the gamma function which is defined by the Formula 4.

$$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt \tag{4}$$

Taking the term $\int_0^x (x-t)^{\alpha-1} f(t) dt$ in Formula 3 into consideration, the fractional integral of $f(x)$ in order α can be seen as the convolution operation between $f(x)$ and $x^{\alpha-1}$, which involves all states of function $f(\cdot)$ in the time interval $[0, x]$. If we take the $f(\cdot)$ as a function of memory, 0 is the beginning of the memory and x is the end of the memory, then with this step, the fractional integral incorporates all states of $f(\cdot)$ in the memory period $[0, x]$ based on their convolutional interactions, and generates a overall perspective of the memory $f(\cdot)$. This also explains why the fractional integral operator has the capacity to process sequential data and has good performance in memory.

Fractional Latent Topic Generation. Since a text is usually composed by a word sequence, the latent semantic information of a text is closely related to the words semantic meanings and the contextual information such as term dependency and sequential structure [1,14]. In this paper, therefore, we assume that the latent topic of a text can be reasoned and inferred from the occurred words' semantic representation and contextual information. It is notable that word semantic representation can be easily learned by the hidden state of a RNN model. Regarding to the contextual information, we first adopt a contextual window for text snippet sampling by sliding over the input word sequence, which has been reported to be effective in contextual feature extraction [27]. Then, we use a fractional latent topic generator derived from fractional calculus to aggregate contextual features and generate latent topics based on the text snippets. Figure 2 shows the procedure of the fractional latent topic generation.

Text Snippet Sampling. In RNN, hidden states are used to learn words representations. Since a text can be seen as a word sequence, the corresponding hidden state sequence can be regarded as a mirroring or projection of the given text. Therefore, we adopt a contextual window which slides from the beginning to the end of the hidden state sequence for text snippet sampling. As is shown in Fig. 2, the red box stands for the contextual window, the length of the contextual

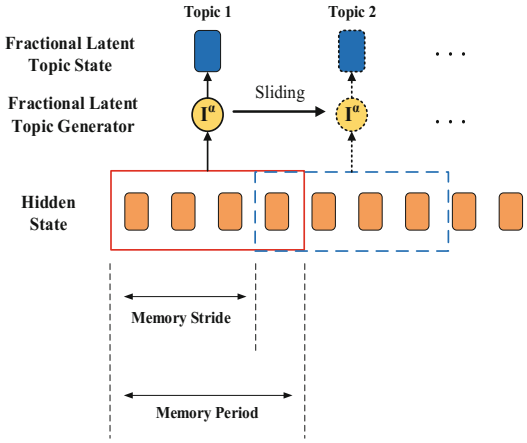


Fig. 2. Procedure of the Fractional Latent Topic Generation. Where the red box stands for the contextual window, and the blue box indicates the location where the contextual window will sliding to in the next time. (Color figure online)

window is defined as the memory period and denoted by p, and the step length that the contextual window moving forward is defined as the memory stride. After the contextual window sliding over the whole text with a constant stride, we obtain all text snippets for the latent topic generation.

Fractional Latent Topic Generator. In the text snippet sampling step, we obtain various text snippets. The fractional latent topic generator, which is derived by fractional calculus, aims to learn a latent topic for each text snippet. It is notable that the fractional calculus is defined on a continues function, while the hidden state in RNN is a discrete variable. Therefore, we are required to transform the fractional calculus to its discrete format. Formally, suppose a text snippet contains the hidden states as $\mathbf{h}^y_{i-p+1}$, ..., $\mathbf{h}^y_{i-1}$, $\mathbf{h}^y_i$, then the hidden topic t_i in this memory period is calculated by the fractional latent topic generator as

$$t_i = \frac{1}{\Gamma(\alpha)} \sum_{j=i-p+1}^{i} (i+1-j)^{\alpha-1} \mathbf{h}^y_j, \tag{5}$$

where p is the length of the memory period, and the fractional calculus order $\alpha \in (0, 1]$ restricts the weights of hidden states. Different text snippets will generate different fractional latent topics, for example "Topic 1" and "Topic 2" in Fig. 2. With this step, we can obtain all fractional latent topics of the given text on the sampled text snippets.

3.3 Topic-Wise Attention

Intuitively, different topics tend to have different perspectives. Noting that the similarity of a text pair can be measured by the perspectives relevance of the

topics, we are motivated to decrease the perspective distance between topics in a relevant text pair, and increase the distance for an irrelevant pair, after generating the fractional latent topics for each text. In particular, we present a topic-wise attention gating mechanism for our model, which automatically rewards the perspectives of the topics with an attentive vector.

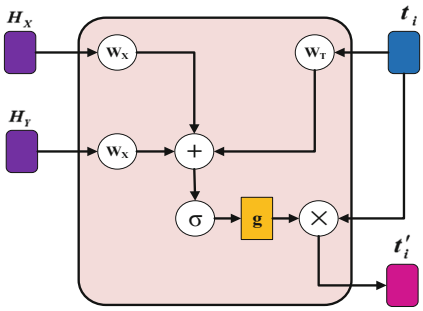


Fig. 3. Topic-wise attention gating mechanism

The structure of our topic-wise attention gating are shown in Fig. 3. Different from the traditional attention based RNN models, our gating mechanism makes full use of the information in two texts rather than a single one for the topic perspective rewarding. Specifically, to reward the perspective of a topic, our topic-wise attention mechanism mainly performs the following two steps, namely relevance measurement and perspective rewarding.

Relevance Measurement. The relevance measurement step measures the relevance between the fractional latent topic $\mathbf{t}_i$ of text $\mathbf{Y}$ and the text $\mathbf{X}$, which serves as a good criteria to decide how much information in the fractional latent topic should be rewarded. In particular, the whole semantic meaning of the text $\mathbf{Y}$ should also be taken into consideration during the relevance calculation, since the fractional latent topic $\mathbf{t}_i$ is generated on a segment of text $\mathbf{Y}$, instead of the whole $\mathbf{Y}$. More concretely, the relevance is formulated as:

$$\mathbf{g} = \sigma(\mathbf{W}_X\mathbf{H}_X + \mathbf{W}_Y\mathbf{H}_Y + \mathbf{W}_t\mathbf{t}_i + \mathbf{b}), \tag{6}$$

where $\mathbf{W}_X$, $\mathbf{W}_Y$ and $\mathbf{W}_t$ are weight matrices, $\mathbf{b}$ is a bias vector, and $\sigma(\cdot)$ is an element-wise sigmoid function. It is worth noting that the obtained $\mathbf{g}$ is a vector, which reflects the relevance in each hidden dimension.

Perspective Rewarding. In topic perspective rewarding step, the original fractional latent topic $\mathbf{t}_i$ is refined by the rewarding vector $\mathbf{g}$ obtained by Formula 6. With this step, the perspective tendency of a topic is modified to be more distinguishable for better text similarity modeling. The rewarded fractional latent topic is formulated as:

$$\mathbf{t}'_i = \mathbf{g} \odot \mathbf{t}_i, \tag{7}$$

where $\odot$ denotes the element-wise multiplication, and $\mathbf{t}'_i$ is the new rewarded hidden topic.

With the above two steps, the topic perspective gap in relevant text pair will be narrowed down, while it is opposite for the irrelevant one. It will have a big influence on the whole text modeling with the text is represented by

$$\mathbf{H}'_Y = \sum_{i=1}^{\tau} \mathbf{t}'_i. \tag{8}$$

4 Empirical Study

4.1 Datasets and Evaluation Metrics

To evaluate the effectiveness of our proposed model, we conduct experiments on three well-known text similarity tasks, namely question answering, paraphrase identification, and textual entailment.

Question Answering. Given a question and a list of candidate answers, the question answering task is to rank the candidates according to their similarities with the question. Two widely used datasets, namely TREC-QA and WikiQA, are adopted in our experiments. TREC-QA was created by Wang et al. [33] based on the QA track (8–13) data of Text REtrieval Conference. WikiQA [38] is an open domain QA dataset in which all answers were collected from the Wikipedia. Both TREC-QA and WikiQA have the train, development and test sets, and each sample is labeled as 1 or 0 to indicate whether the candidate answer is right or wrong for a given question. The statistics of the datasets are presented in Table 1. The performance of answer selection is usually measured by the mean average precision (MAP) and mean reciprocal rank (MRR) [25].

Paraphrase Identification. The paraphrase identification task can be treated as a binary classification problem, and the goal is to judge whether two texts are paraphrases or not according to their similarity. We utilize the Microsoft Research Paraphrase corpus (MSRP) [5] for experiment, which is constructed from a large corpus of temporally and topically clustered news articles. The MSRP dataset contains 4,076 sentence pairs in the training set, and 1,725 ones in the test set. Each text pair is labeled with 1 or 0 to indicate whether the two text are paraphrases or not. To evaluate the performance, two widely used metrics, namely accuracy (Acc) and F1 score are adopted [39].

Textual Entailment. For a sentence pair, one of a sentence can be seen as the premise and the other as the hypothesis. The textual entailment task is to judge whether the hypothesis can be inferred by the premise according to their similarity. We use the Multi-Genre Natural Language Inference (MultiNLI) corpus for experiment, which is a crowd-sourced collection of sentence pairs annotated

Table 1. Statistics of the Datasets. "Avg QL", "Avg AL", "Avg Para1L", "Avg Para2L", "Avg Sent1L" and "Avg Sent2L" denote the average length of questions, answers, the first paragraphs, the second paragraphs, the first sentences and the second sentences respectively.

Answer selection	Dataset		# of Questions	Avg QL	Avg AL
	TREC-QA	train	1162	7.57	23.21
		dev	65	8.00	24.9
		test	68	8.63	25.61
	WikiQA	train	873	7.16	25.29
		dev	126	7.23	24.59
		est	243	7.26	24.59
Paraphrase Identification	Dataset		# of Paragraph Pair	Avg Para1L	Avg Para2L
	MSRP	train	4077	18.99	18.93
		test	1725	18.82	18.80
Textual Entailment	Dataset		# of Sentence Pair	Avg Sent1L	Avg Sent2L
	MultiNLI	train	392,702	19.91	10.12
		matched	10,000	19.40	10.08
		mismatched	10,000	19.90	10.98

with textual entailment information. In MultiNLI, the relationship between a sentence pair is classified into three categories, namely neutral, contradiction, and entailment. During our experiments, we assign the value of -1, 0, 1 to the label of neutral, contradiction, and entailment respectively. Furthermore, this corpus has served as the basis for the shared task of the RepEval 2017 Workshop[1] at EMNLP in Copenhagen. and the evaluation metric used in this corpus is accuracy (Acc) [36].

4.2 Training

We use the bidirectional LSTM (BLSTM) [9] model as the function in Formula 1 to obtain the original hidden states, which can effectively mitigate the gradient vanish problem. Then, we utilize the Manhattan distance similarity function with $l1$ norm and restrict it to a range of [0, 1] for text similarity calculation [20]:

$$s(X, Y) = exp(-||\mathbf{H}_{\mathbf{X}}^{'} - \mathbf{H}_{\mathbf{Y}}^{'}||_1) \tag{9}$$

where, $\mathbf{H}_{\mathbf{X}}^{'}$ and $\mathbf{H}_{\mathbf{Y}}^{'}$ are text representations learned by the proposed FraLT-RNN model. The predicted probability of a text pair labeled as 1 or 0 is defined according to the relevance score: $\hat{p}(c = 1|X, Y) = s(X, Y)$ and $\hat{p}(c = 0|X, Y) = 1 - s(X, Y)$.

[1] https://repeval2017.github.io/shared/.

For each text pair, the loss function is defined by the cross-entropy of the predicted and true label distributions for training:

$$L(X, Y; c) = -\sum_{j=0}^{C-1} p(c = j|X, Y) \log \hat{p}(c = j|X, Y) \tag{10}$$

where C is the number of classes, and $p(c = 1|X, Y)$ is the gold probability of label c, which equals to 1 with ground truth and otherwise is 0.

4.3 Parameter Settings

We implement the proposed FraLT-RNN model by using TensorFlow. The optimization is relatively straightforward with standard back-propagation [24]. We apply stochastic gradient descent method Adagrad [6] with mini-batches (64 in size), which can be easily parallelized on single machine with multi-cores. The rectifier linear unit $ReLU = max(0, x)$ is adopted as the activation function, which is a common choice in the deep learning literature [19]. For regularization, we use dropout [11] strategy for our model, and the dropout rate is selected from $[0.0, 0.1, 0.2, 0.5]$. Regarding to the word embeddings, we adopt the pre-trained 100-dimensional GloVe word vectors[2], which are trained based on the global word co-occurrence [21]. Moreover, the fractional integral order α is valued from 0.1 to 1 with the stride of 0.1, and the memory period is selected in the set $[2, 3, 4, 5, 6, 7, 8, 9, 10]$.

5 Experimental Results and Analyses

5.1 Effectiveness of FraLT-RNN

To investigate the effect of our FraLT-RNN model, the BLSTM based RNN model which does not involve any topic information, and the recently proposed word-level attention mechanism [29] are utilized for comparisons. Table 2 shows the performance of various models for question answering, paraphrase identification, and textual entailment tasks. It is observed that we achieve significant improvements over classical BLSTM and attention based BLSTM models on all datasets, by incorporating fractional latent topic into text representation. It is also notable that the classical BLSTM model relies more on the attention mechanism to capture the salient information for text similarity modeling. However, the attention method mainly focuses on measuring the weight of each hidden state, while does not pay specifical attention to the surrounding context of the words in a text pair. Moreover, the attentive weight is produced after obtaining all the hidden states, which neglects the internal interactions and hierarchical structure of the text during hidden state generation. In contrast, our proposed FraLT-RNN model can explicitly capture the internal relations and the hierarchical features between two texts, by incorporating the fractional latent topics

[2] http://nlp.stanford.edu/data/glove.6B.zip.

and topic-wise attentions for text representation. Therefore, our FraLT-RNN model integrated can yield better performance than the traditional attention mechanism.

Table 2. Comparison with Various RNN Models. "BLSTM" stands for BLSTM with no more optimization, and "A-BLSTM" stands for traditional word attention based BLSTM. "$*$" and "$+$" imply significant improvements over "BLSTM" and "A-BLSTM" respectively.

Model	TREC-QA		WikiQA		MSRP		MultiNLI	
	MAP	MRR	MAP	MRR	Acc(%)	F1	Matched (Acc %)	Mismatched (Acc %)
BLSTM	0.6487	0.6991	0.6581	0.6691	73.6	81.8	67.5	67.1
A-BLSTM	0.7369*	0.8208*	0.7258*	0.7394*	75.4*	82.7*	71.1*	70.8*
FraLT-RNN	**0.8359**$^{*+}$	**0.8962**$^{*+}$	**0.7401**$^{*+}$	**0.7519**$^{*+}$	**81.2**$^{*+}$	**87.5**$^{*+}$	**81.9**$^{*+}$	**81.3**$^{*+}$

5.2 Comparison with Recent Progress

In addition to the classical BLSTM model, we compare our model with the recent progress in question answering, paraphrase identification and textual entailment.

Table 3. Performance comparisons on TREC-QA

System	MAP	MRR
Wang, Liu, and Zhao 2016 [31]	0.7369	0.8208
Wang and Ittcheriah 2015 [34]	0.7460	0.8200
Santos et al. 2016 [25]	0.7530	0.8511
Wang, Mi, and Ittycheriah 2016 [35]	0.7714	0.8447
Chen et al. 2018 [4]	0.8227	0.8886
FraLT-RNN	**0.8359**	**0.8962**

Results on Question Answering. Table 3 and Table 4 summarize the results on TREC-QA and WikiQA respectively. [25,31,40] are the recent attention based models that focus on the word-level attentive text representations. It is observed that our proposed model achieves the new state-of-the-art performance on both TREC-QA and WikiQA. Specifically, we outperform the best results on TREC-QA and WikiQA with absolute improvements of 0.132 and 0.0043 in terms of MAP, and 0.0076 and 0.0069 in terms of MRR. Regarding to the word alignment models [4,34,35], which take the aligned words and the neighboring texts into consideration for text representation, our model is also much more effective and does not rely on the laboursome feature engineering. This is mainly owing to the hierarchical structure embedded in our model, namely the latent topics, which are neglected in the above models.

Table 4. Performance comparisons on WikiQA

System	MAP	MRR
Santos et al. 2016 [25]	0.6886	0.6957
Yin et al. 2015 [40]	0.6921	0.7108
Wang, Mi, and Ittycheriah 2016 [35]	0.7058	0.7226
Wang, Liu and Zhao 2016 [31]	0.7341	0.7418
Chen et al. 2018 [4]	0.7358	0.7450
FraLT-RNN	**0.7401**	**0.7519**

Table 5. Performance comparisons on MSRP

System	Acc	F1
Hu et al. 2014 [15]	69.9	80.9
Socher et al. 2011 [28]	76.8	83.6
Yin and Schutze 2015 [39]	78.1	84.4
Chen et al. 2018 [4]	77.3	84.0
He, Gimpel, and Lin 2015 [10]	78.6	84.7
FraLT-RNN	**81.2**	**87.5**

Results on Paraphrase Identification. The results from recent work on MSRP are summarized in Table 5. [39] presented a convolutional neural network based deep learning architecture, which modeled interaction features at multiple levels of granularity. However, their model relied much on the pretraining step. In [10], a similar model was proposed, which also used a CNN model for feature extraction at a multiplicity of perspectives. We observe that our model also achieve the best results among the existing work. Furthermore, our model automatically generates fractional latent topics with a fractional latent topic generator, which requires no more parameter besides the fractional order α and has a lower computation complexity.

Table 6. Performance comparisons on MultiNLI

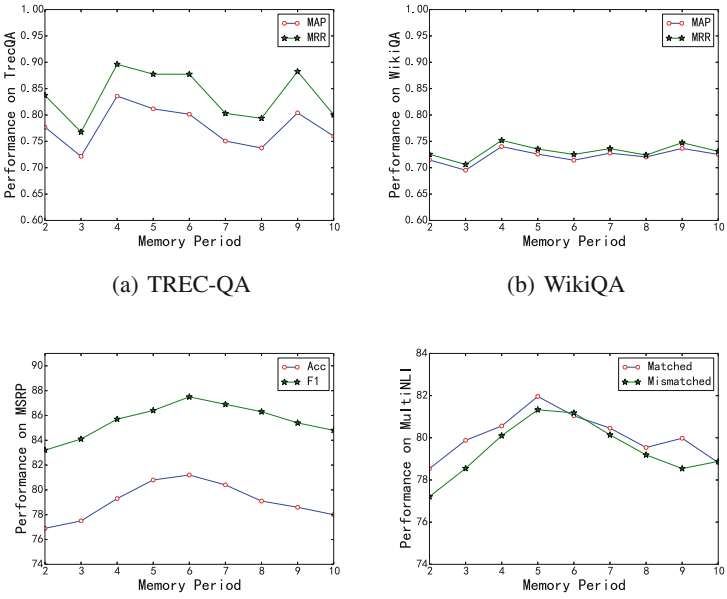
System	Matched	Mismatched
Williams, Nangia, and Bowman 2017 [37]	72.3	72.1
Gong, Luo, and Zhang 2018 [8]	78.8	77.8
Tay, Tuan, and Hui 2017 [30]	78.7	77.9
Kim, Kang, and Kwak 2018 [18]	79.1	78.4
Radford et al. 2018 [23]	**82.1**	**81.4**
FraLT-RNN	81.9	81.3

Results on Textual Entailment. Table 6 shows the comparison with recent work on matched and mismatched problems of MultiNLI collection. It can be seen that our model outperforms most of the recent work, and can be comparable to if no better than the state-of-the-art model [23]. [23] makes use of task-aware input transformations to achieve effective transfer for natural language understanding. However, their model requires to learn a pre-trained model at first, and then tunes parameters in the pre-trained model for a better performance. During the model tuning step, numerous parameters need to be learned. In contrast,

our FraLT-RNN model does not need a pre-trained model and only requires one parameter, namely the fractional integral order, besides the RNN parameters during model learning.

5.3 Influence of the Parameters in Fractional Latent Topic Generation

For the proposed FraLT-RNN model, there are two components in fractional latent topic generation, namely the text snippet sampling and the fractional latent topic generator. The memory period and the fractional calculus order is the main parameter in the text snippet sampling and the fractional latent topic generator respectively. We conduct experiments on the four datasets to investigate the influence of the two parameters.

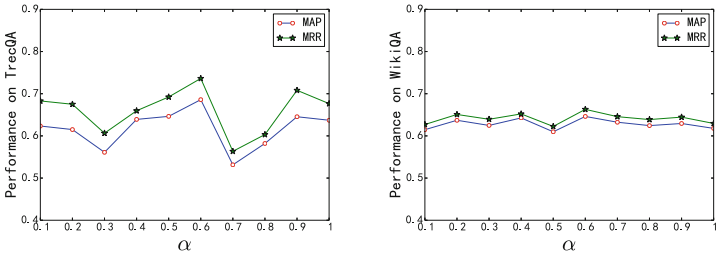


(a) TREC-QA (b) WikiQA

(c) MSRP (d) MultiNLI

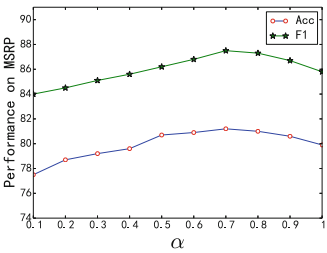
Fig. 4. Influence of the memory period.

Figure 4 show the influence of the memory period on the FraLT-RNN. The memory period decides the range of the term dependency, i.e. a large memory period implies a long-term dependency. It can be seen that TREC-QA is more sensitive on the memory period compared with the other three datasets. To ensure a better and steady performance of FraLT-RNN, it is recommended to select the memory period value of the contextual window in the set $[4, 5, 6, 7]$. Regarding to the fractional calculus order α, Fig. 5 illustrates its influence on

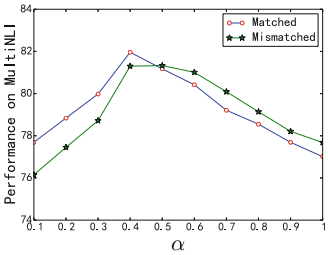


(a) TREC-QA

(b) WikiQA



(c) MSRP



(d) MultiNLI

Fig. 5. Influence of the fractional calculus order α.

the proposed FraLT-RNN model. It is interesting to find that TREC-QA is also more sensitive on the fractional calculus order compared with the other three datasets, which is consistent with the memory period parament. Furthermore, to obtain a more reliable result, it is recommended to assign the value of fractional calculus order α in the interval $[0.4, 0.8]$.

6 Conclusion and Future Work

In this paper, we proposed a fractional latent topic based RNN model, which incorporate the hierarchical structures of the text during text representation. In particular, we provide a novel latent topic generation approach, which is implemented by means of the fractional calculus. To the best of our knowledge, this is the first attempt to apply the fractional calculus in natural language processing. Experiments on four benchmark datasets, namely TREC-QA and WikiQA for question answering, MSRP for paraphrase identification, and MultiNLI for textual entailment show the great advantages of our proposed model. It is notable that we achieve the new state-of-the-art results on TREC-QA, WikiQA, and MSRP. It is also interesting to find that the TREC-QA is more sensitive to the parameters of fractional latent topic generator. In the future, we will investigate how to apply the fractional calculus into more natural language processing tasks and deep learning models to cope with the time series and hierarchical structure problems.

Acknowledgement. This research is funded by the Science and Technology Commission of Shanghai Municipality (No. 18511105502), Shanghai Municipal Commission of Economy and Informatization (No. 170513) and Xiaoi Research. The computation is performed in the Supercomputer Center of ECNU. The second author is the corresponding author.

References

1. Al-Anzi, F.S., AbuZeina, D.: Toward an enhanced Arabic text classification using cosine similarity and latent semantic indexing. J. King Saud University-Comput. Inf. Sci. **29**(2), 189–195 (2017)
2. Bai, J., Feng, X.C.: Fractional-order anisotropic diffusion for image denoising. IEEE Trans. Image Process. **16**(10), 2492–2502 (2007)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
4. Chen, Q., Hu, Q., Huang, J.X., He, L.: CA-RNN: using context-aligned recurrent neural networks for modeling sentence similarity. In: AAAI (2018)
5. Dolan, B., Quirk, C., Brockett, C.: Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In: Proceedings of the 20th International Conference on Computational Linguistics, p. 350. Association for Computational Linguistics (2004)
6. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(Jul), 2121–2159 (2011)
7. Fang, H., Wu, F., Zhao, Z., Duan, X., Zhuang, Y., Ester, M.: Community-based question answering via heterogeneous social network learning. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
8. Gong, Y., Luo, H., Zhang, J.: Natural language inference over interaction space. arXiv preprint arXiv:1709.04348 (2017)
9. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
10. He, H., Gimpel, K., Lin, J.: Multi-perspective sentence similarity modeling with convolutional neural networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1576–1586 (2015)
11. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
13. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42**(1–2), 177–196 (2001)
14. Hofmann, T.: Probabilistic latent semantic indexing. In: ACM SIGIR Forum, vol. 51, pp. 211–218. ACM (2017)
15. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional neural network architectures for matching natural language sentences. In: Advances in Neural Information Processing Systems, pp. 2042–2050 (2014)
16. Jozefowicz, R., Zaremba, W., Sutskever, I.: An empirical exploration of recurrent network architectures. In: International Conference on Machine Learning, pp. 2342–2350 (2015)

17. Kilbas, A.A.A., Srivastava, H.M., Trujillo, J.J.: Theory and Applications of Fractional Differential Equations, vol. 204. Elsevier Science Limited, Amsterdam (2006)
18. Kim, S., Hong, J.H., Kang, I., Kwak, N.: Semantic sentence matching with densely-connected recurrent and co-attentive information. arXiv preprint arXiv:1805.11360 (2018)
19. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
20. Mueller, J., Thyagarajan, A.: Siamese recurrent architectures for learning sentence similarity. In: AAAI, vol. 16, pp. 2786–2792 (2016)
21. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
22. Pu, Y., Wang, W., Zhou, J., Wang, Y., Jia, H.: Fractional differential approach to detecting textural features of digital image and its fractional differential filter implementation. Sci. China Series F Inf. Sci. **51**(9), 1319–1339 (2008)
23. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (2018)
24. Rumelhart, D.E., Hinton, G.E., Williams, R.J., et al.: Learning representations by back-propagating errors. Cogn. Mod. **5**(3), 1 (1988)
25. dos Santos, C.N., Tan, M., Xiang, B., Zhou, B.: Attentive pooling networks. CoRR, abs/1602.03609 **2**(3), 4 (2016)
26. Severyn, A., Moschitti, A.: Learning to rank short text pairs with convolutional deep neural networks. In: SIGIR, pp. 373–382 (2015)
27. Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: A latent semantic model with convolutional-pooling structure for information retrieval. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 101–110. ACM (2014)
28. Socher, R., Huang, E.H., Pennin, J., Manning, C.D., Ng, A.Y.: Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In: Advances in Neural Information Processing Systems, pp. 801–809 (2011)
29. Tan, M., Santos, C.D., Xiang, B., Zhou, B.: LSTM-based deep learning models for non-factoid answer selection. arXiv preprint arXiv:1511.04108 (2015)
30. Tay, Y., Tuan, L.A., Hui, S.C.: A compare-propagate architecture with alignment factorization for natural language inference. arXiv preprint arXiv:1801.00102 (2017)
31. Wang, B., Liu, K., Zhao, J.: Inner attention based recurrent neural networks for answer selection. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1288–1297 (2016)
32. Wang, D., Nyberg, E.: A long short-term memory model for answer sentence selection in question answering. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), vol. 2, pp. 707–712 (2015)
33. Wang, M., Smith, N.A., Mitamura, T.: What is the jeopardy model? A quasi-synchronous grammar for QA. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (2007)
34. Wang, Z., Ittycheriah, A.: FAQ-based question answering via word alignment. arXiv preprint arXiv:1507.02628 (2015)

35. Wang, Z., Mi, H., Ittycheriah, A.: Sentence similarity learning by lexical decomposition and composition. arXiv preprint arXiv:1602.07019 (2016)
36. Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1112–1122. Association for Computational Linguistics (2018). http://aclweb.org/anthology/N18-1101
37. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426 (2017)
38. Yang, Y., Yih, W.T., Meek, C.: WikiQA: a challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2013–2018 (2015)
39. Yin, W., Schütze, H.: Convolutional neural network for paraphrase identification. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 901–911 (2015)
40. Yin, W., Schütze, H., Xiang, B., Zhou, B.: ABCNN: attention-based convolutional neural network for modeling sentence pairs. arXiv preprint arXiv:1512.05193 (2015)
41. Zhang, X., Li, S., Sha, L., Wang, H.: Attentive interactive neural networks for answer selection in community question answering. In: AAAI, pp. 3525–3531 (2017)
42. Zhao, Z., Lu, H., Zheng, V.W., Cai, D., He, X., Zhuang, Y.: Community-based question answering via asymmetric multi-faceted ranking network learning. In: AAAI, pp. 3532–3539 (2017)