Ideography Leads Us to the Field of Cognition: A Radical-Guided Associative Model for Chinese Text Classification

Hanqing Tao[1], **Shiwei Tong**[1], **Kun Zhang**[3], **Tong Xu**[1,2], **Qi Liu**[1,2], **Enhong Chen**[1,2]*, **Min Hou**[1,2]

[1] Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China
[2] School of Data Science, University of Science and Technology of China
[3] School of Computer Science and Information Engineering, Hefei University of Technology
{hqtao, tongsw, minho}@mail.ustc.edu.cn, {tongxu, qiliuql, cheneh}@ustc.edu.cn, {zhang1028kun}@gmail.com

Abstract

Cognitive psychology research shows that humans have the instinct for abstract thinking, where association plays an essential role in language comprehension. Especially for Chinese, its ideographic writing system allows radicals to trigger semantic association without the need of phonetics. In fact, subconsciously using the associative information guided by radicals is a key for readers to ensure the robustness of semantic understanding. Fortunately, many basic and extended concepts related to radicals are systematically included in Chinese language dictionaries, which leaves a handy but unexplored way for improving Chinese text representation and classification. To this end, we draw inspirations from cognitive principles between ideography and human associative behavior to propose a novel **R**adical-guided **A**ssociative **M**odel (**RAM**) for Chinese text classification. RAM comprises two coupled spaces, namely *Literal Space* and *Associative Space*, which imitates the real process in people's mind when understanding a Chinese text. To be specific, we first devise a serialized modeling structure in *Literal Space* to thoroughly capture the sequential information of Chinese text. Then, based on the authoritative information provided by Chinese language dictionaries, we design an *association* module and put forward a strategy called *Radical-Word Association* to use ideographic radicals as the medium to associate prior concept words in *Associative Space*. Afterwards, we design an *attention* module to imitate people's matching and decision between *Literal Space* and *Associative Space*, which can balance the importance of each associative words under specific contexts. Finally, extensive experiments on two real-world datasets prove the effectiveness and rationality of RAM, with good cognitive insights for future language modeling.

Introduction

Due to the uniqueness of ideography and great potential for future applications, the research of Chinese text classification has been appealing in recent years (Liao, Sun, and Gu 2019; Tao et al. 2019). However, traditional Chinese text classification methods usually ignore the essence of Chinese ideographic characteristics (DeFrancis 1986), that is, associative information guided by ideographic radicals.

Most of the time, when people receive a certain text, they will not only grasp it according to the literal features of the text, but also expand a series of association in their minds based on those features (Kuvska et al. 2016). In fact, the language symbols in a text that we can directly obtain or perceive are literal features. For Chinese, its writing system derived from pictographs makes its literal features ideographic (Unger 2004). Moreover, as the semantic component used to compose *Phono-semantic Compound Characters* (Tung 2012) which take up over 80% of all Chinese characters, each radical has a pictorial glyph origin. This vivid feature has been inherited for thousands of years, often allowing readers to understand the meaning of Chinese characters without knowing their pronunciation, which forms a unique cognitive process in the mode of conveying semantics compared with English and other phonetic languages (Tzeng et al. 1979). As shown by the example in Figure 1, ideographic radicals prompt us to associate prior concepts with corresponding Chinese characters: "*climate*" for "snow", "*foot*" for "road", "*water*" for "slippery", etc., which helps us grasp relevant attributes of characters and approximate the core idea of classification label "Traffic".

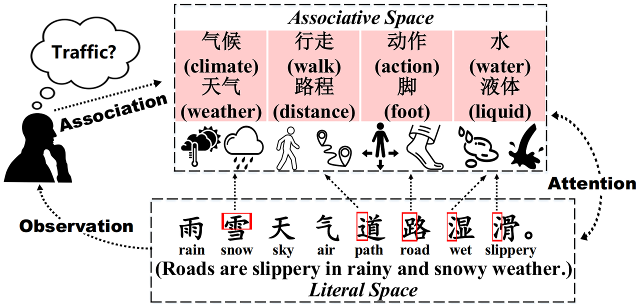

Figure 1: The cognitive process when understanding a Chinese text and inferring its possible label, where radical of each Phono-semantic Compound Character is circled in red.

Meanwhile, association in psychology refers to the psychological connection between concepts, events or mental states, usually derived from specific experiences (Klein 2018). It allows people to use prior concepts outside a given text to assist comprehension during reading. Actually, associative behavior is a fundamental and effective principle in psycho-linguistics for explaining examples of cognition and knowledge learning through accumulated experience (Dick-

[*] Corresponding author

inson 2012). More importantly, language research is inseparable from cognitive science (Marslen-Wilson et al. 1980; Elman 1990), and more and more researchers have regarded language learning as a cognitive phenomenon (Isac and Reiss 2013; Huth et al. 2016; Ellis 2019). However, based on these interdisciplinary theories of psychology and cognitive science, how to leverage association mechanism to import desired human prior concepts into Natural Language Processing (NLP) is an urgent problem for current deep learning, which also faces great challenges.

However, traditional text modeling methods often ignore the participation of human cognitive behavior and association in the process of text comprehension, just stick to the analysis of the literal space in isolation to deal with the linguistic symbols (Montague 1974). This perspective is very limited now, especially for short texts whose literal features are very sparse. Therefore, introducing some external information reasonably to enrich text representation is more in line with human cognition. Fortunately, as a treasure-house of human knowledge, language dictionaries (e.g., Xinhua Dictionary and Oxford English Dictionary) are very efficient for inferring basic information of radicals (roots), characters, words and common concepts to help understand texts in our daily life (Nielsen 2008). But little attention has been paid to the importance and utilization of language dictionaries in Chinese NLP tasks, which leaves plenty of space for further improving text representation and classification.

In response to the problems and limitations mentioned above, we propose a novel **R**adical-guided **A**ssociative **M**odel (**RAM**) for Chinese text classification, which can take both literal features and human prior concepts into consideration with the help of language dictionaries. Specifically, 1) we first introduce a *Literal Space* and devise a serialized structure to model the sequential information of Chinese text; 2) Then, we propose an *association* module and a strategy of using radicals as the medium for *Radical-Word Association*, so as to model associative contents in *Associative Space*; 3) After that, we design an *attention* module by imitating the cognitive process in people's mind to model the matching and decision between *Literal Space* and *Associative Space*; 4) Finally, we conduct extensive experiments, where the experimental results not only demonstrate the effectiveness and rationality of RAM, but also provide good cognitive insights for future language modeling.

Related Work

Text Classification & Deep Learning

Text classification is a fundamental natural language processing (NLP) task, which plays an indispensable role in various scenarios, such as document retrieval, news filtering, public opinion analysis (Hotho et al. 2005; D'Andrea et al. 2019). Recent years have witnessed the success of deep learning in this field, no matter in terms of the construction of deep classification model (Aggarwal et al. 2012; Kim 2014; Wang et al. 2018) or word embedding approaches including CBOW, Skipgram, GloVe and so on (Mikolov et al. 2013; Pennington et al. 2014; Li et al. 2019a). Given the sequential property of human language, Recurrent Neural Network (RNN) (Elman 1990), its improved version Long Short-Term Memory (LSTM) (Hochreiter et al. 1997) and Bidirectional LSTM (BiLSTM) have been proposed for capturing the long-range information of the context (Graves et al. 2013), which has a profound effect on the subsequent study of text modeling. Currently, there has been another wave in the field of natural language processing, that is the emerging model of pre-training (Peters et al. 2018). Among them, the most successful model might be BERT (Devlin et al. 2018), which combines Transformer's powerful representation ability with some language-related pre-training goals to address many NLP tasks while exhibiting impressive performance (Yu and Jiang 2019; Zhang et al. 2019c).

Human Cognitive Modeling

No matter in the early days or now, imitating human cognitive principles has always been the original intention of deep learning (Bezdek 1992; Sardi et al. 2020). Initially, the fully-connected edges designed in artificial neural networks ideally mimic the numerous dendrites of nerve cells. Then, to improve the nonlinear expression ability of neural networks, the activation functions (e.g., sigmoid and ReLU) were proposed by imitating the activation threshold of biofilm action potential (Hodgkin et al. 1952; Krizhevsky et al. 2012). More importantly, the attention mechanism (Vaswani et al. 2017; Fukui et al. 2019) was proposed to mimic the fact of eye allocation when people are reading a text or observing an image (Yang et al. 2016; Yin et al. 2018), which indeed exhibits superior performance and psychological interpretability at the same time (Zhang et al. 2019a,b).

Chinese-specific Methods

In recent years, the human brain investigations about the differences between Chinese and phonetic languages have prompted researchers to explore the uniqueness of Chinese lansigns (Tan et al. 2000; Hung et al. 2014). Scholars have found that the low-level features of Chinese characters such as radical (Sun et al. 2014), pinyin (Wang et al. 2019), stroke (Cao et al. 2018) and glyph (Wu et al. 2019) also have certain semantics. By introducing them into word or sentence representation learning, the performance can indeed be improved. At the same time, for the study of Chinese downstream tasks, a proper text modeling method can highlight the characteristics of Chinese, which is an important factor to improve the performance (Zhou et al. 2016; Peng, Cambria, and Zou 2017). Lately, Tao et al. (2019) have achieved impressive results by directly introducing radicals to participate in Chinese text representation and classification.

To seek common ground while reserving differences, our work draws inspirations from cognitive principles between ideographic radicals and associative behavior, and take advantages of deep learning to provide a novel insight into the task of Chinese text modeling and classification.

Radical-guided Associative Model

In this section, we will elaborate on the technical details of our **R**adical-guided **A**ssociative **M**odel (**RAM**) for addressing the problem of Chinese text classification.

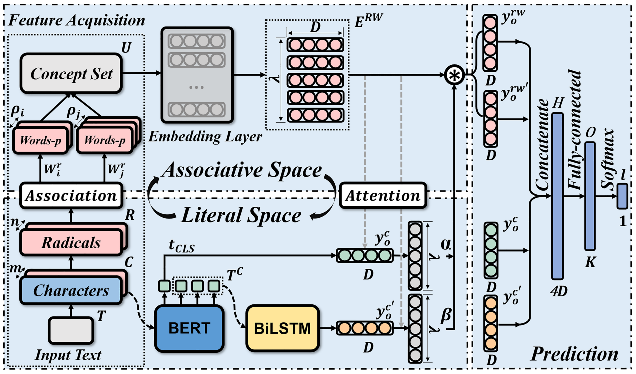

Figure 2: The overall architecture of our Radical-guided Associative Model (RAM).

Problem Definition

Given an arbitrary unlabeled text $T = \{x_1, x_2, ..., x_m\}$ and a pre-defined label set S containing K different labels, the goal of our task is to train and obtain a classification function $\mathcal{F}$ with the ability to assign a proper label $l \in S$ for T:

$$\mathcal{F}(T) \to l, \tag{1}$$

where $x_i \in T$ $(0 \leq i \leq m)$ stands for the feature vector of the i-th token in T after text preprocessing.

Overall Architecture of RAM

As shown in Figure 2, RAM mainly comprises two coupled spaces (i.e., *Literal Space* and *Associative Space*), together with one *Feature Acquisition* process and two modules, namely *association* and *attention*. In addition, classification is implemented in the *Prediction* component. The technical details of each part will be elaborated as follows.

Feature Acquisition. In line with the cognitive behavior of people observing a text and acquiring features shown in Figure 1, we first propose the *Feature Acquisition* process to extract the features required for our model, i.e., characters, radicals and words. In fact, there are six kinds of Chinese characters according to "six writings"[1], but only the radicals of *Phono-semantic Compound Characters* contain informative semantics (Tung 2012). Therefore, in this paper, we mainly pay attention to *Phono-semantic Compound Characters* and their radicals. Then, we will use these radicals to obtain corresponding associative words with the help of three Chinese language dictionaries.

1) Character Type Masking. As intuitively depicted in Figure 3, given an input Chinese text T containing m characters, we first segment it into a character sequence $C = \{c_1, c_2, ..., c_m\}$ according to the string operation, where C actually stands for the character-level feature of T. Then, by referring to Chinese Character Type Dictionary (Liang et al. 2019), we are able to label each character with a type tag so as to realize the *Character Type Masking* process:

$$Mask(c_i) = \begin{cases} 1 & c_i = C_p, \\ 0 & c_i = Others, \end{cases} \tag{2}$$

where C_p represents *Phono-semantic Compound Characters*. $Mask(\cdot)$ denotes the masking function, and c_i $(0 \leq i \leq m)$ is the i-th character in C.

2) Radical Distilling. After getting the mask code of each character, we could carry out the following *Radical Distilling* process. That is to say, in order to extract the radicals that have significant ideographic effects from text T (i.e., radicals of *Phono-semantic Compound Characters*) and thus help to convey semantics, we need to remove other useless contents. So, we multiply each character's mask code with itself to determine which characters could retain for querying radicals from Chinese dictionary:

$$R = Radical\_Query(C \odot Mask(C)), \tag{3}$$

where $\odot$ is an element-wise product operation, and $Radical\_Query$ operation allows us to map each Chinese character into a single radical with the help of Xinhua Dictionary (Han 2009). Additionally, we filter out all the repeated radicals in R to avoid redundant processing. As a result, $R = \{r_1, r_2, ..., r_n\}$ is the distilled radicals of character sequence C, where $n \in [0, m]$.

3) Radical-Word Association. Instead of using radicals directly as an additional feature (Tao et al. 2019), we regard the distilled radicals as the medium for associating highly relevant associative words that indicate attributes and extensional meaning. Formally, we call this strategy *Radical-Word Association*, which corresponds to the *association* module in Figure 2. As a result, associative words connected with *Phono-semantic Compound Characters* are denoted as *Words-p* (red). By referring to Radical Concept Dictionary (Hong and Huang 2012), each distilled radical $r_i \in R = \{r_1, r_2, ..., r_n\}$ will correspond to a list of associative words:

$$W_i^r = Concept\_Query(r_i) = \{w_1^r, w_2^r, ..., w_{\rho_i}^r\}. \tag{4}$$

Here, $\rho_i \geq 1$ denotes the number of associative words for r_i, which will vary from radical to radical. Therefore, all the radicals R extracted from text T could form a set of associative words $U = \{w_1, w_2, ..., w_\lambda\}$:

$$U = \bigcup_{i=1}^{n} W_i^r, \tag{5}$$

where U actually stands for the imported external word-level feature for T and $\lambda = \sum_{i=1}^{n} \rho_i$ denotes the total words number of U. Since different radicals may correspond to the same associative words, the set operation here allows repeated associative words to be merged into one.

Literal Space Modeling. Given an input Chinese text T containing m characters, RAM will literally project it into a character sequence $C = \{c_1, c_2, ..., c_m\}$ for subsequent processing (each punctuation will also be regarded as a character). Then, we devise a deep modeling structure by harnessing the power of pre-trained BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018), which has embraced abundant "*accumulated experience*" (statistical language information) based on very large training materials (Dickinson 2012), to obtain the sentence representation $t_{CLS} \in \mathcal{R}^{1 \times D}$ and character representation $T^C = \{t_1, t_2, ..., t_m\}$ of T as follows:

$$t_{CLS},\ T^C = BERT([CLS], C), \tag{6}$$

[1] https://en.wikipedia.org/wiki/Chinese_characters

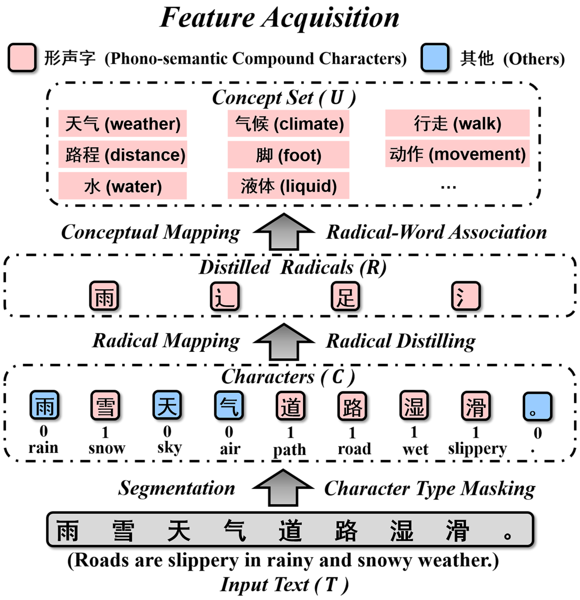

Figure 3: An intuitive illustration of the *Feature Acquisition* process for Chinese text.

where the first token $[CLS]$ added in front of every sequence C is always a special classification token, and the final hidden state t_{CLS} corresponding to this token is used as the aggregate sequence representation for classification tasks. Because t_{CLS} acts as the output of BERT for later classification, we also use y_o^c to denote it for convenience. Meanwhile, T^C represents the hidden vectors of corresponding m characters contained in text T. Then, we treat the hidden states of BERT output as the initialization vectors and then send them together as a sequence into BiLSTM to further learn the context dependencies, which is depicted in Figure 4.

Formally, we take the rest output of BERT, i.e., $T^C = \{t_1, t_2, ..., t_m\}$, as the sophisticated representations for each character c_i $(1 \leq i \leq m)$ in C. Afterwards, we apply BiLSTM to further imitate the *conceptual change* (Council et al. 2000) under the specific context of T^C, which is consistent with the process of people adjusting to a new text based on their accumulated experience. Thus, given the vector embedding sequence of BERT output T^C, the hidden vectors of BiLSTM are calculated by receiving T^C as input:

$$\begin{aligned} \overrightarrow{h_i} &= LSTM(\overrightarrow{h}_{i-1}, s_i), \\ \overleftarrow{h_i} &= LSTM(\overleftarrow{h}_{i+1}, s_i), \\ y_i &= concatenate(\overrightarrow{h_i}, \overleftarrow{h_i}), \end{aligned} \qquad (7)$$

where $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ denote the forward hidden vector and backward hidden vector respectively at the i-th time step s_i $(1 \leq i \leq m)$ in the BiLSTM unit. While y_i is the concatenation of $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$. As a result, the final output of BiLSTM (i.e., y_m) will integrate the forward and backward contextual information. For convenience, we also use $y_o^{c'}$ to denote it for subsequent calculation.

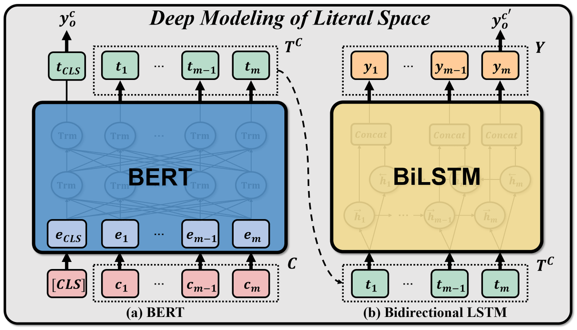

Figure 4: Diagrammatic sketch of *Literal Space* modeling.

Associative Space Modeling. As mentioned above, the ideographic characteristics of Chinese characters are deeply rooted and ubiquitous (Tan et al. 2000), which is a crucial factor for readers to associate relevant concepts with radicals. Now that we have obtained associative words through the *association* module described in *Feature Acquisition* process, we should further represent those words and highlight the information that we need.

1) Associative Word Embedding. In order to represent the associative words in concept set $U = \{w_1, w_2, ..., w_\lambda\}$ for subsequent calculation, we need to map each word into a low-dimension real-value vector. Here, we apply an external well pre-trained embedding model based on distributional assumption (Mikolov et al. 2013; Le and Mikolov 2014) and an *Embedding Layer* to get the embedding vectors for words obtained by *Radical-word Association*:

$$E^{RW} = Embedding(U) = \{e_1^{rw}, e_2^{rw}, ..., e_\lambda^{rw}\}, \qquad (8)$$

where λ denotes the total associative words number of U.

2) Attention Mechanism. The attention mechanism in deep learning is essentially similar to the selective visual attention mechanism of human beings. In fact, as for reading comprehension, people usually tend to first read through the sentence to form a preliminary cognition in their minds, and then back to select and match the proper concepts based on the overall context of the sentence (Taatgen et al. 2007). Inspired by this cognitive process, we design an *attention* module which can focus our model on relatively important associative words in U back with the consideration of learned contextual representation explained before, i.e., y_o^c and $y_o^{c'}$.

Formally, we regard y_o^c and $y_o^{c'}$ as *querys*, E^{RW} as *key* and *value* at the same time to implement attention mechanism. That is, given the associative word representations obtained in *Associative Space*, i.e., $E^{RW} = \{e_1^{rw}, e_2^{rw}, ..., e_\lambda^{rw}\}$, we use the contextual representations obtained in *Literal Space*, i.e., y_o^c and $y_o^{c'}$, to attend to each associative word $w_i \in U$ and get the attention weight for each $e_\epsilon^{rw} \in E^{RW}$ and $e_\theta^{rw} \in E^{RW}$ $(1 \leq \epsilon \leq \lambda, 1 \leq \theta \leq \lambda)$:

$$\begin{aligned} \alpha' &= [\alpha_1', ..., \alpha_\epsilon', ..., \alpha_\lambda'], \ \alpha_\epsilon' = f(y_o^c, e_\epsilon^{rw}), \\ \beta' &= [\beta_1', ..., \beta_\theta', ..., \beta_\lambda'], \ \beta_\theta' = f(y_o^{c'}, e_\theta^{rw}), \end{aligned} \qquad (9)$$

where $\alpha' \in \mathcal{R}^{1 \times \lambda}$ and $\beta' \in \mathcal{R}^{1 \times \lambda}$ are two vectors for E^{RW} respectively, representing the attention weight from two contextual aspects of y_o^c and $y_o^{c'}$. Besides, α_ϵ' and β_θ' denote the

ϵ-th or the θ-th weight of an associative word respectively, and $f(\cdot,\cdot)$ denotes the distance function which is stated as an element-wise dot product operation in this paper. Then, we need to normalize $\alpha^{'}$ and $\beta^{'}$ with the $softmax$ function:

$$\begin{aligned}\alpha_i &= \frac{exp\ (\alpha_i^{'})}{\sum_{\epsilon=1}^{\lambda} exp\ (\alpha_\epsilon^{'})},\ where \sum_{i=1}^{\lambda} \alpha_i = 1,\\ \beta_j &= \frac{exp\ (\beta_j^{'})}{\sum_{\theta=1}^{\lambda} exp\ (\beta_\theta^{'})},\ where \sum_{j=1}^{\lambda} \beta_j = 1.\end{aligned} \tag{10}$$

Afterwards, the two-aspect attentive representations y_o^{rw} and $y_o^{rw'}$ for associative words could be obtained through attentive weighted sum as:

$$y_o^{rw} = \sum_{\epsilon=1}^{\lambda} \alpha_\epsilon e_\epsilon^{rw},\ y_o^{rw'} = \sum_{\theta=1}^{\lambda} \beta_\theta e_\theta^{rw}, \tag{11}$$

where α_ϵ is the ϵ-th dimensional value of $\alpha \in \mathcal{R}^{1\times\lambda}$, and β_θ is the θ-th dimensional value of $\beta \in \mathcal{R}^{1\times\lambda}$ $(1 \le \epsilon \le \lambda,\ 1 \le \theta \le \lambda)$. Consequently, the attentive representations y_o^{rw} and $y_o^{rw'}$ have precisely fused the information of *Literal Space* and *Associative Space* together. Like the information processing in human brain, the attention mechanism herein is actually a bridge between literal and associative spaces.

Prediction

In order to systematically integrate and fully learn the information of the obtained four different representations for Chinese text T: two-aspect contextual representations learned through literal features, i.e., y_o^c and $y_o^{c'}$; attentive representations derived from jointly modeling of literal features and associative words, i.e., y_o^{rw} and $y_o^{rw'}$, we first conduct a concatenation operation:

$$H = [y_o^c;\ y_o^{c'};\ y_o^{rw};\ y_o^{rw'}], \tag{12}$$

where $H \in \mathcal{R}^{1\times 4D}$ is the vector concatenated through dimension with an advantage of retaining all the information (Zhang et al. 2019a). Afterwards, we leverage the fully connected neural network to learn the hidden interactions and enhancements among these four representations:

$$O = \sigma(W^{(l)} \times H + b^{(l)}),\ \sigma(x) = \frac{1}{1+e^{-x}}, \tag{13}$$

where $W^{(l)}$ and $b^{(l)}$ respectively denotes the weight matrix and bias vector fitted by the fully-connected linear neural network, and $O \in \mathcal{R}^{1\times K}$ is its output. Note that K represents the size of label set S which has been stated in *Problem Definition*. Finally, the predicted label l could be classified through the $softmax$ function and $argmax$ operation:

$$l = argmax(softmax(O)). \tag{14}$$

Training Strategy

Loss Function. As the multi-class classification task exhibits an output of distribution with different probabilities on various classes, we need to judge the most significant class and distinguish it from others. According to Zhou et al. (2016) and Tao et al. (2019), cross-entropy is a good way to effectively scale the significance of probability distribution. So, we apply it as our loss function for training RAM:

$$\mathcal{L} = -\sum_{T\in\mathcal{D}} \sum_{i=1}^{K} p_i(T) \log p_i(T), \tag{15}$$

where $p_i(T)$ denotes the calculated probability of assigning label $l_i \in S$ $(1 \le i \le K)$ for text T, and $\mathcal{D}$ is the dataset which T belongs to.

Model Initialization. Before training, a proper initialization is beneficial for optimizing our model. The specific setting for hyperparameters are illustrated as follows.

In conjunction with the architecture of RAM, we apply the pre-trained Chinese BERT model with 24-layer, 1024-hidden, 16-heads and 330M parameters[2]. In addition, we obtained a well pre-trained Chinese word embedding model[3] with a dimension of 256 for representing associative words in U. The embedding dimension D is also set as 256, while the hidden size of BiLSTM is set as 1,024. To prevent our model from overfitting, we add dropout mechanism in front of the embedding layer and fully-connected layer with a drop rate of 0.5. As for the scale size λ of concept set U, it is a variable which will be adapted to Radical Concept Dictionary dynamically. Besides, we randomly initialize the parameters of BiLSTM and the fully connected neural network. Finally, we apply *Adagrad* optimizer with a learning rate of 0.01. During implementation, we use *MXNet* to build our model and train it with two 2.30GHz Intel(R) Xeon(R) Gold 5218 CPUs and a Tesla V100-SXM2-32GB GPU.

Experiments

Dataset Description

To fit the problems studied in this paper, we selected two real-world datasets to evaluate our model: the Chinese News Title Dataset (CNT) and Fudan Chinese Text Dataset (FCT). To ensure reproducibility, the segmentation ratio of the training set and testing set of these two datasets are consistent with the public ones.

- **CNT** (Zhou et al. 2016) is a public dataset which covers a wide range of 32 different categories of Chinese news. After preprocessing and filtering the useless text whose length is lower than 2, it contains 47,693 texts for taining and 15,901 for testing, which is a quite ideal choice for validating the robustness of different methods.
- **FCT**[4] is an official dataset provided by Fudan University with 20 categories covering abundant academic texts for validation. To guarantee the quality of implementation, we carefully preprocessed this dataset by correcting and removing unreadable samples. As a result, it contains 8,220 texts for training and 8,115 for testing.

Dictionary Preparation

In order to guarantee the reliability of our model, we apply three formal Chinese Dictionary datasets to support the process of *Character Type Masking*, *Radical Mapping* and *Conceptual Mapping* process. Accordingly, Chinese Character

[2] https://github.com/ymcui/Chinese-BERT-wwm

[3] https://spaces.ac.cn/archives/4304

[4] https://www.kesci.com/home/dataset/5d3a9c86cf76a600360edd04

Methods	CNT				FCT			
	Accuracy	Recall	F1-score	ΔF1 (%)	Accuracy	Recall	F1-score	ΔF1 (%)
(1) TextCNN (char)	0.6123	0.6127	0.6059	-39.71	0.7481	0.4041	0.4095	-104.73
(2) TextCNN (word)	0.7706	0.7707	0.7695	-10.00	0.9012	0.6270	0.6643	-26.20
(3) TextRNN (char)	0.6992	0.6993	0.6995	-21.01	0.8361	0.4925	0.5174	-62.04
(4) TextRNN (word)	0.8023	0.8025	0.8025	-5.47	0.8704	0.5149	0.5372	-56.05
(5) BERT (char, fine-tuned)	0.8124	0.8120	0.8117	-4.29	0.9096	0.7635	0.7910	-5.98
(6) C-LSTM (char+word)	0.8186	0.8187	0.8183	-3.45	0.9204	0.6856	0.7218	-16.15
(7) C-BLSTM (char+word)	0.8230	0.8231	0.8225	-2.91	0.9204	0.6847	0.7216	-16.18
(8) RAFG (char+word+radical)	0.8324	0.8325	0.8325	-1.68	0.9241	0.7140	0.7408	-13.17
(9) RAM (char+word+radical)	**0.8464**	**0.8461**	**0.8465**	-	**0.9423**	**0.8058**	**0.8383**	-

Table 1: Experimental results of comparison methods on CNT dataset and FCT dataset.

Type Dictionary[5] contains all the information about *Phono-semantic Compound Characters*; Xinhua Dictionary[6] contains the necessary radical information for mapping each character to a radical; Radical Concepts Dictionary[7] includes detailed conceptual information for all Chinese radicals, with over 1,000 concept words in total and 6 concept words for each radical on average.

Comparison Methods

- **TextRNN (char/word)** (Mikolov et al. 2011) refers to the plain recurrent neural network which processes tokens sequentially. To compare the functionality of Chinese feature granularity in different scenarios, we set character-level and word-level feature as the input respectively.
- **TextCNN (char/word)** (Kim 2014) is a convolutional neural network-based model for text classification. With the same aim of comparision like TextRNN, the input is also set as two kinds, i.e., character-level and word-level. We apply *jieba*[8] as the segmentation tool to obtain the word-level feature.
- **C-LSTMs / C-BLSTMs** (Zhou et al. 2016) are two Chinese-specific text classification models applying two independent L-STMs to concatenate word and character features together. Since both characters and words are important features for Chinese text, they make up for the disadvantages of using one kind of feature unilaterally. And C-BLSTMs is the bidirectional version of C-LSTMs.
- **BERT** (Devlin et al. 2018) stands for the current state-of-the-art pre-training model for natural language processing, which is usually applied in English materials and performs well. We here take it as an important baseline and fine-tune it to validate the rationality and effectiveness of the design of our RAM model.
- **RAFG** (Tao et al. 2019) is another Chinese-specific text classification method. This SOTA baseline is a four-granularity model, which proposes two extra kinds of radicals (character-level and word-level radicals) together with corresponding Chinese characters and words to help Chinese text classification. In fact, we might know that radical is a special low-level feature which does not possess the property of "context", so the way of RAFG directly integrating radicals with Chinese characters and words is somewhat imperfect and taking every radical into consideration is a little irrational. To utilize radicals more properly and avoid the hidden adverse effects of wrong Chinese word segmentation (CWS), we systematically design the *Radical-Word Association* strategy to regard radicals as a kind of medium for associating prior concept words, which could filter out uninformative radicals through *Radical Distilling* process and provide a more rational way for utilizing radicals.

Experimental Results

The comparison results on two datasets are shown in Table 1. The results are quite revealing in several ways, from which we can see that our RAM model is able to substantially achieve the best results on both datasets, no matter in terms of Accuracy, Recall or F1-score. This proves that RAM has gained a better comprehension of Chinese texts, hence the performance is boosted. However, there are still some thought-provoking findings in this table.

Firstly, by comparing Chinese-specific methods (6-9) with those non-Chinese-specific ones (1-5), we could notice that the feature granularity of Chinese text is a crucial factor for classification performance: utilizing character or word feature unilaterally is worse than combing them together, which proves that they can make up for each other and meanwhile Chinese word segmentation may cause loss of information unavoidably. But secondly, we could infer from Table 1 that although BERT (5) only takes Chinese characters as input, it can maintain a stable performance on both datasets (more stable ΔF1), which confirms that after large-scale corpus pre-training, character-level features can also obtain better robustness when faced with different corpora. These findings are quite consistent with the study in (Li et al. 2019b). Thirdly, looking back on our modeling of the three features in Chinese (character, radical and word), we can find that we only use the character features of Chinese text literally, and meanwhile the word features are associated via the medium of radical. This process perfectly avoids the adverse effects of Chinese word segmentation errors, which plays a non-negligible role in promoting the performance of RAM. Fourthly, the results are clear that RAM has a comprehensive improvement in performance compared with the most advanced BERT, which shows that our modeling strategy based on cognitive principles can better grasp the purport of Chinese text. Last but not least, through the comparison between RAFG and RAM, we could learn that a more rational method of utilizing radicals is beneficial for better understanding hence harnessing the messages conveyed by radicals, especially in terms of cognitive modeling.

[5] http://zidian.kxue.com/

[6] http://zidian.aies.cn/

[7] http://xh.5156edu.com/page/z2443m7618j19616.html

[8] https://github.com/fxsjy/jieba

Methods	CNT				FCT			
	Accuracy	Recall	F1-score	ΔF1 (%)	Accuracy	Recall	F1-score	ΔF1 (%)
(1) RAM-association	0.8433	0.8436	0.8426	-0.46	0.9420	0.7953	0.8334	-0.59
(2) RAM-attention	0.8445	0.8450	0.8442	-0.27	0.9405	0.7937	0.8350	-0.40
(3) RAM	**0.8464**	**0.8461**	**0.8465**	-	**0.9423**	**0.8058**	**0.8383**	-

Table 2: Ablation results of RAM: (1) RAM without the association module (whole associative space modeling); (2) RAM without the attention module (attention mechanism for sorting associative words in the light of given context).

Source Text	Distilled Radicals	Top 2 Associative Words for each Distilled Radical (sorted by attention weights)	Ground Truth (Dataset)	Predicted (Dataset)
Chinese: 拯救夏日胃口的肉菜：茄汁里脊 **English**: Salvation of Summer Appetite: Tenderloin in Eggplant Sauce.	扌、攵、艹、氵	**Chinese**: (动作、行为)、(做、行为)、(植物、农业)、(液体、水) **English**: (action, behavior)、(do, behavior)、(plant, agriculture)、(liquid, water)	Food (CNT)	Food (CNT)
Chinese: 皮草早春混搭新时髦 **English**: The new fashion of mixing and matching fur in early spring.	艹、氵、扌、斤、日、髟	**Chinese**: (草木、材料)、(状态、液体)、(动作、行为)、(性质、工具)、(时间、太阳) 、(毛发、胡须) **English**: (vegetation, material)、 (condition, liquid)、(action, behavior)、(property, tool)、(time, sun)、 (hair, beard)	Dress (CNT)	Dress (CNT)
Chinese: 利用城市污泥防治水土流失 **English**: Applying urban sludge to prevent and control soil erosion.	刂、土、氵、阝	**Chinese**: (工具、动作)、(土地、建筑)、(流体、水)、(山地、地形) **English**: (tool, action)、(soil, building) 、(fluid, water) 、(mountainous region, topography)	Environment (FCT)	Environment (FCT)
Chinese: 短跑运动员专项力量练习的设计与选择 **English**: Design and Selection of Special Strength Practice for Sprinters.	矢、足、辶、力、页、纟、讠、扌	**Chinese**: (长度、度量)、(脚、动作)、(行走、路程)、(力量、行为)、(颈部、数量)、(纺织、行为)、(交流、语言)、(手、动作) **English**: (length, measurement)、(foot, action)、(walk, distance)、(strength, behavior)、(neck, number)、(weave, behavior)、(communication, language)、(hand, action)	Sports (FCT)	Sports (FCT)

Figure 5: A case study for some Chinese source texts, where the Phono-semantic Compound Characters are all painted in red.

Ablation Study & Case Study

As mentioned earlier, RAM is solidly based on the cognitive principles between ideography and human associative behavior (Ellis 2019). To validate the design of RAM and determine how each module affects the final results, we conduct an ablation study by removing each module respectively, which is summarized in Table 2. According to the results, we observe an obvious decline in the performance of RAM on both datasets no matter the *attention* module or the *associative* module is removed, which indeed verifies the necessity of each module. Meanwhile, we can see that removing the *association* module leads to more performance degradation, which further validates the importance of accumulated experience and highlights the essential role of association mechanism in language comprehension.

To provide some intuitionistic examples for explaining why our model gains a better performance than any other baseline methods, we conduct a case study similar with (Qin et al. 2018, 2020) to see what is happening in the working flow of RAM, where the specific cases could be found in Figure 5. Taking the first example to say, we notice that associative words and literal features can enhance each other, i.e., associative words "*plant*" and "*agriculture*" associated by RAM are important clues for inferring the concept of "Eggplant", while other associative words (e.g., "*action*" suggests the attribute of "salvation", and "*liquid*" indicates the property of "sauce") could be regarded as complementary contents for source text thus helping us grasp less prominent but global semantics. Then, for the second example, associative words "*vegetation*", "*material*" and "*hair*" globally reflect the trait of "fur", while "*condition*", "*property*" and "*time*" together help us recognize the semantics of "fashion" hence lead us to the idea of "Dress". Moreover, as for the remaining two examples, we could also know that the abstract concepts provided by associative words are mostly informative indicators for determining the ground truths. Although there might be some associative words which are not directly related to the semantics of ground truths (e.g., "*liquid*" for "Dress" and "*weave*" for "Sports"), those words actually reflect the original meaning of corresponding radicals, which will be balanced under the *attention* module and maybe helpful in another context. In fact, when we humans are associating related concepts to help text comprehension in our minds, we tend to think of all possible meanings. This is similar to the unconscious iceberg effect (Freud 2005; Rogers 2014), i.e., although some associative contents seem to be irrelevant to the classification ground truth of current text, the sufficient associative information is actually a key hidden factor to grasp original meanings of characters and ensure the understanding robustness. In summary, all the above findings could finally enable us to confirm the rationality and effectiveness of our model.

Conclusion

In this paper, we conducted an explorative but focused study on Chinese text classification from a cognitive viewpoint of human beings, and proposed a novel **R**adical-guided **A**ssociative **M**odel (**RAM**) for this task. RAM comprises two coupled spaces called *Literal Space* and *Associative Space*, which ideally imitates the real process in people's mind when understanding a Chinese text. Through extensive experiments, our study has gone some way towards enhancing our understanding of Chinese and human cognition.

Acknowledgments

This research was partially supported by grants from the National Key Research and Development Program of China (No. 2018YFB1004300 and No. 2016YFB1000904) and the National Natural Science Foundation of China (No.s 61727809 and 61922073).

References

Aggarwal, C. C.; and Zhai, C. 2012. A survey of text classification algorithms. In *Mining text data*, 163–222. Springer.

Bezdek, J. C. 1992. On the relationship between neural networks, pattern recognition and intelligence. *Int. J. Approx. Reason.* 6: 85–107.

Cao, S.; Lu, W.; Zhou, J.; and Li, X. 2018. cw2vec: Learning Chinese Word Embeddings with Stroke n-gram Information. In *AAAI*.

Council, N. R.; et al. 2000. *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press.

D'Andrea, E.; Ducange, P.; Bechini, A.; Renda, A.; and Marcelloni, F. 2019. Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications* 116: 209–226.

DeFrancis, J. 1986. *The Chinese language: Fact and fantasy*. University of Hawaii Press.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Dickinson, A. 2012. Associative learning and animal cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1603): 2733–2742.

Ellis, N. C. 2019. Essentials of a theory of language cognition. *The Modern Language Journal* 103: 39–60.

Elman, J. L. 1990. Finding structure in time. *Cognitive science* 14(2): 179–211.

Freud, S. 2005. *The unconscious*, volume 8. Penguin UK.

Fukui, H.; Hirakawa, T.; Yamashita, T.; and Fujiyoshi, H. 2019. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10705–10714.

Graves, A.; Mohamed, A.-r.; and Hinton, G. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, 6645–6649. IEEE.

Han, Z. 2009. Xinhua Zidian (Xinhua Dictionary).

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Hodgkin, A. L.; and Huxley, A. F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology* 117(4): 500.

Hong, J.-F.; and Huang, C.-R. 2012. A hanzi radical ontology based approach towards teaching chinese characters. In *Workshop on Chinese Lexical Semantics*, 745–755. Springer.

Hotho, A.; Nürnberger, A.; and Paaß, G. 2005. A brief survey of text mining. In *Ldv Forum*, volume 20, 19–62. Citeseer.

Hung, Y.-h.; Hung, D. L.; Tzeng, O. J.-L.; and Wu, D. H. 2014. Tracking the temporal dynamics of the processing of phonetic and semantic radicals in Chinese character recognition by MEG. *Journal of Neurolinguistics* 29: 42–65.

Huth, A. G.; De Heer, W. A.; Griffiths, T. L.; Theunissen, F. E.; and Gallant, J. L. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532(7600): 453.

Isac, D.; and Reiss, C. 2013. *I-language: An introduction to linguistics as cognitive science*. Oxford University Press.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. Association for Computational Linguistics. doi:10.3115/v1/D14-1181. URL http://www.aclweb.org/anthology/D14-1181.

Klein, S. B. 2018. *Learning: Principles and applications*. Sage Publications.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Kuvska, M.; Trnka, R.; Kubena, A. A.; and Ruvzivcka, J. G. 2016. Free Associations Mirroring Self- and World-Related Concepts: Implications for Personal Construct Theory, Psycholinguistics and Philosophical Psychology. *Frontiers in Psychology* 7.

Le, Q.; and Mikolov, T. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, 1188–1196.

Li, B.; Drozd, A.; Guo, Y.; Liu, T.; Matsuoka, S.; and Du, X. 2019a. Scaling Word2Vec on Big Corpus. *Data Science and Engineering* 1–19.

Li, X.; Meng, Y.; Sun, X.; Han, Q.; Yuan, A.; and Li, J. 2019b. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3242–3252.

Liang, S.; Tang, X.; Hu, R.; Wu, J.; and Liu, Z. 2019. Measurement and Application of Chinese Component Semantic Ability Based on Distributed Representation. In *CCL*.

Liao, J.; Sun, F.; and Gu, J. 2019. Combining Concept Graph with Improved Neural Networks for Chinese Short Text Classification. Technical report, EasyChair.

Marslen-Wilson, W.; and Tyler, L. K. 1980. The temporal structure of spoken language understanding. *Cognition* 8(1): 1–71.

Mikolov, T.; Kombrink, S.; Burget, L.; Černockỳ, J.; and Khudanpur, S. 2011. Extensions of recurrent neural network language model. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5528–5531. IEEE.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Montague, R. 1974. Formal Philosophy: Selected Papers, RH Thomason, Ed.

Nielsen, S. 2008. The effect of lexicographical information costs on dictionary making. *Lexikos* 18.

Peng, H.; Cambria, E.; and Zou, X. 2017. Radical-based hierarchical embeddings for chinese sentiment analysis at sentence level. In *The 30th International FLAIRS conference. Marco Island*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* .

Qin, C.; Zhu, H.; Xu, T.; Zhu, C.; Jiang, L.; Chen, E.; and Xiong, H. 2018. Enhancing person-job fit for talent recruitment: An ability-aware neural network approach. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 25–34.

Qin, C.; Zhu, H.; Xu, T.; Zhu, C.; Ma, C.; Chen, E.; and Xiong, H. 2020. An enhanced neural network approach to person-job fit in talent recruitment. *ACM Transactions on Information Systems (TOIS)* 38(2): 1–33.

Rogers, A. 2014. *The base of the iceberg: Informal learning and its impact on formal and non-formal learning*. Verlag Barbara Budrich.

Sardi, S.; Vardi, R.; Meir, Y.; Tugendhaft, Y.; Hodassman, S.; Goldental, A.; and Kanter, I. 2020. Brain experiments imply adaptation mechanisms which outperform common AI learning algorithms. *Scientific Reports* 10(1): 1–10.

Sun, Y.; Lin, L.; Yang, N.; Ji, Z.; and Wang, X. 2014. Radical-enhanced chinese character embedding. In *International Conference on Neural Information Processing*, 279–286. Springer.

Taatgen, N. A.; Van Rijn, H.; and Anderson, J. 2007. An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning. *Psychological Review* 114(3): 577.

Tan, L. H.; Spinks, J. A.; Gao, J.-H.; Liu, H.-L.; Perfetti, C. A.; Xiong, J.; Stofer, K. A.; Pu, Y.; Liu, Y.; and Fox, P. T. 2000. Brain activation in the processing of Chinese characters and words: a functional MRI study. *Human brain mapping* 10(1): 16–27.

Tao, H.; Tong, S.; Zhao, H.; Xu, T.; Jin, B.; and Liu, Q. 2019. A Radical-Aware Attention-Based Model for Chinese Text Classification. In *AAAI*.

Tung, T. 2012. *The Six Scripts Or the Principles of Chinese Writing by Tai Tung: A Translation by LC Hopkins, with a Memoir of the Translator by W. Perceval Yetts*. Cambridge University Press.

Tzeng, O. J.; Hung, D. L.; Cotton, B.; and Wang, W. S. 1979. Visual lateralisation effect in reading Chinese characters. *Nature* 282(5738): 499.

Unger, J. M. 2004. *Ideogram: Chinese characters and the myth of disembodied meaning*. University of Hawaii Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, S.; Huang, M.; and Deng, Z. 2018. Densely Connected CNN with Multi-scale Feature Attention for Text Classification. In *IJCAI*, 4468–4474.

Wang, Y.; Ananiadou, S.; and Tsujii, J. 2019. Improving clinical named entity recognition in Chinese using the graphical and phonetic feature. *BMC Medical Informatics and Decision Making* 19.

Wu, W.; Meng, Y.; Han, Q.; Li, M.; Li, X.; Mei, J.; Nie, P.; Sun, X.; and Li, J. 2019. Glyce: Glyph-vectors for Chinese Character Representations. *arXiv preprint arXiv:1901.10125* .

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489.

Yin, Y.; Huang, Z.; Chen, E.; Liu, Q.; Zhang, F.; Xie, X.; and Hu, G. 2018. Transcribing Content from Structural Images with Spotlight Mechanism. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* .

Yu, J.; and Jiang, J. 2019. Adapting BERT for Target-Oriented Multimodal Sentiment Classification. In *IJCAI*.

Zhang, K.; Lv, G.; Wang, L.; Wu, L.; Chen, E.; Wu, F.; and Xie, X. 2019a. Drr-net: Dynamic re-read network for sentence semantic matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7442–7449.

Zhang, K.; Zhang, H.; Liu, Q.; Zhao, H.; Zhu, H.; and Chen, E. 2019b. Interactive attention transfer network for cross-domain sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5773–5780.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019c. ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129* .

Zhou, Y.; Xu, B.; Xu, J.; Yang, L.; and Li, C. 2016. Compositional recurrent neural networks for Chinese short text classification. In *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*, 137–144. IEEE.