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[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

AAAI-21

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. Through extensive experiments on two real-world datasets, our model not only shows its effectiveness and rationality, but also provides good cognitive insights for future language modeling.

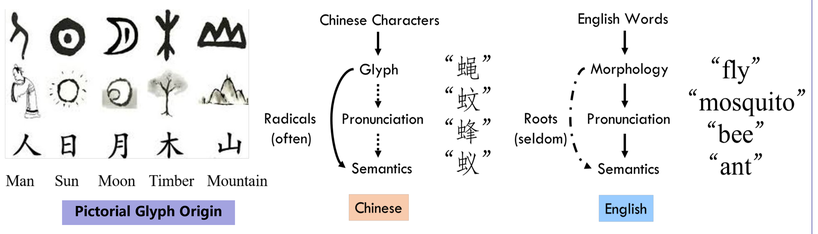
Background

Motivation:

- Language research should consider the involvement of **association**;
- Vivid pictorial glyph origin makes ideography **deeply rooted** and **ubiquitous** in Chinese, which could trigger semantic association more easily than alphabetical languages;
- As the semantic component used to compose **Phono-semantic Compound Characters** (形声字, PCC) which take up over 80% of all Chinese characters, each radical of them could serve as a **medium** for associating relevant prior concepts.

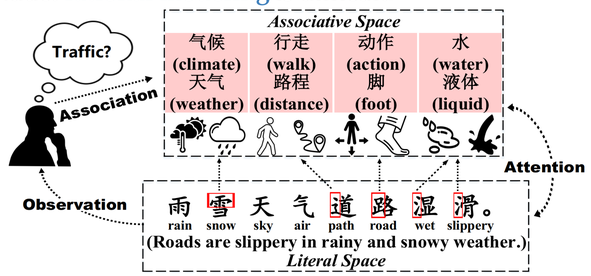
Fact :

- 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;
- Introducing some **external information** reasonably to enrich text representation is more in line with human cognition.



How to build a text classification model adapted to the uniqueness of Chinese language?

Challenges: Associative behavior is vital for the **robustness** of semantic understanding, but how to introduce it to existing works and conduct **rational modeling** is difficult.



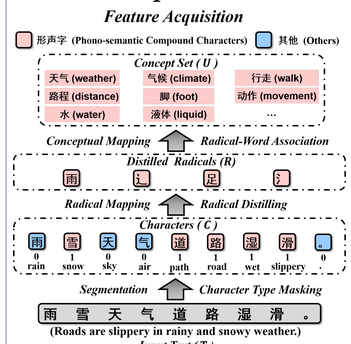
Problem Definition: To select **the most appropriate assignment** to an untagged text from a predefined set of tags.

Methodology

Implementation

- Given: A predefined set of tags $\boldsymbol{S}$;
- Input space: An untagged text $\boldsymbol{T}$;
- Output space: The most appropriate assignment $\boldsymbol{l} \in \boldsymbol{S}$;
- Task: To learn a classification function $\boldsymbol{F}$: $\boldsymbol{F}(\boldsymbol{T}) \rightarrow \boldsymbol{l}$.

Feature Acquisition:



Characters: $C = \{c_1, c_2, ..., c_m\}$

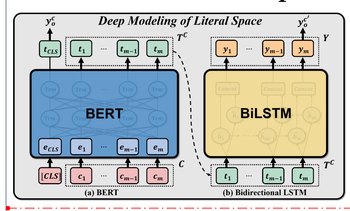
PCC Masking : $Mask(c_i) = \begin{cases} 1 & c_i = C_p, \\ 0 & c_i = Others, \end{cases}$

Radical Distilling: $R = Radical\_Query(C \odot Mask(C)),$

Radical-Word Association: $W_i^r = Concept\_Query(r_i) = \{w_1^r, w_2^r, ..., w_{\rho_i}^r\}$

Concept Set: $U = \bigcup_{i=1}^{n} W_i^r$

Literal & Associative Space Modeling & Attention Mechanism:



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

$$\overrightarrow{h_i} = LSTM(\overrightarrow{h}_{i-1}, s_i),$$

$$\overleftarrow{h_i} = LSTM(\overleftarrow{h}_{i+1}, s_i),$$

$$y_i = concatenate(\overrightarrow{h_i}, \overleftarrow{h_i}),$$

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

$$\alpha' = [\alpha'_1, ..., \alpha'_\epsilon, ..., \alpha'_\lambda],\ \alpha'_\epsilon = f(y_o^c, e_\epsilon^{rw}),$$

$$\beta' = [\beta'_1, ..., \beta'_\theta, ..., \beta'_\lambda],\ \beta'_\theta = f(y_o^{c'}, e_\theta^{rw}),$$

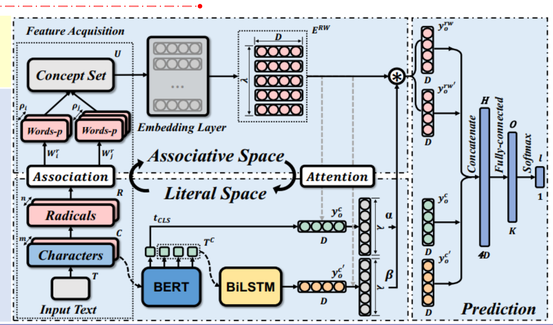
$$\alpha_i = \frac{exp(\alpha'_i)}{\sum_{i=1}^{\lambda} exp(\alpha'_i)},\ where \sum_{i=1}^{\lambda} \alpha_i = 1,$$

$$\beta_j = \frac{exp(\beta'_j)}{\sum_{\theta=1}^{\lambda} exp(\beta'_\theta)},\ where \sum_{j=1}^{\lambda} \beta_j = 1.$$

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

Framework

- *Feature Acquisition*
- *Literal Space*
- *Associative Space*
- *Prediction*



Prediction:

$$H = [y_o^c;\ y_o^{c'};\ y_o^{rw};\ y_o^{rw'}],\ O = \sigma(W^{(l)} \times H + b^{(l)}),\ \sigma(x) = \frac{1}{1+e^{-x}},\ l = argmax(softmax(O)).$$

Objective Function:

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

Experiments

1. **Dataset:** We apply three Chinese dictionary datasets and conduct experiments on two real-world datasets.

- Character Type Dictionary —— *Character Type Masking* process;
- Xinhua Dictionary —— *Radical Mapping* process;
- Radical Concepts Dictionary —— *Conceptual Mapping* process;

- Chinese News Title Dataset (CNT) with 32 gold classification labels;
- Fudan Chinese Text Dataset (FCT) with 20 gold classification labels.

2. Experimental Results of Different Methods

- Our model (RAM) can gain a **better performance** and **robustness** than any other baseline methods;
- 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**.

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

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**	-

3. Ablation Study

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).

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**	-

- The degraded performance indeed verifies the **necessity** of each module;
- Cognitive modules can **help grasp semantics** of Chinese texts better;
- These results validate the importance of **accumulated experience** and highlights the essential role of **association** mechanism in language comprehension.

Conclusion

- Our model is solidly based on the **cognitive principles** between ideography and human associative behavior, which presents **a novel insight** into Chinese text classification and future language modeling;
- Making rational use of radicals is critical for modeling the **essence** of ideography;
- Our work has gone some way towards enhancing our understanding of **ideographic Chinese** and **human cognition**;
- More research on this topic needs to be undertaken before the association between radicals and cognitive concepts is more clearly understood;
- Imitating human cognitive principles is the **future** of AI.

Speaker: Hanqing Tao
Email: hqtao@mail.ustc.edu.cn
Lab Homepage: http://bigdata.ustc.edu.cn/