Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), Honolulu, Hawaii, January 27 - February 1, 2019

University of Science and Technology of China (USTC)

A Radical-aware Attention-based Model for Chinese Text Classification

Hanqing Tao[1], Shiwei Tong[1], Hongke Zhao[1], Tong Xu[1, 2]*, Binbin Jin[1], Qi Liu[1, 2]

[1]Anhui Province Key Lab. 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

Introduction

Motivation: Chinese is a language derived from Oracle Bone Inscriptions (pictographs), which is essentially different from English or other phonetic languages. Radical is a semantic unit of Chinese with some graphic characteristics, which might help us to recognize semantics.

Fact : Most existing studies on text classification are professionally conducted for English, which may lose effectiveness on Chinese materials due to the huge difference between Chinese and English.

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

How to build a text classification model adapted to Chinese language?

Glyph Origin	Radical (Chinese Characters)	English
	亻 (仆, 伴)	Man (servant, partner)
	目 (看, 瞳)	Eye (look, pupil)
	扌 (打, 挖)	Hand (hit, dig)
	日 (晴, 暗)	Sun (sunny, dark)
	雨 (雾, 霜)	Rain (fog, frost)
	山 (峰, 崖)	Mountain (peak, cliff)

A radical is often related to certain concepts

Challenges:

- Radical is useful, but how to introduce it to existing works is difficult.
- The properties of Chinese is hard to model.
- How to combine radicals with other features is also challenging.

Table 1: Characters with the same radical "insect".

Chinese Characters	Radical	English
蝇	虫	fly
蚊	虫	mosquito
蜂	虫	bee
虱	虫	louse
蚁	虫	ant

Table 2: Words with the same character "cattle".

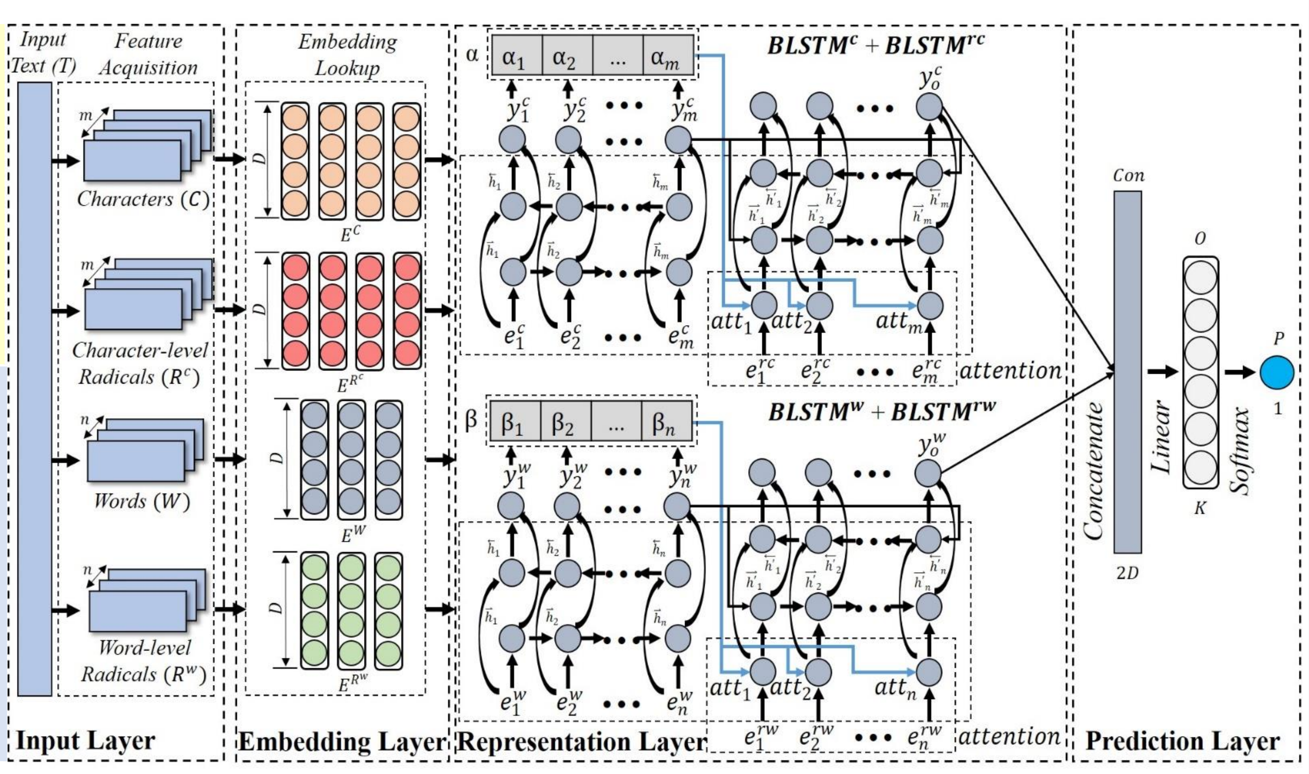
Chinese Words	Chinese Characters	English
公牛	公 (male) + 牛 (cattle)	bull
母牛	母 (female) + 牛 (cattle)	cow
牛奶	牛 (cattle) + 奶 (milk)	milk
牛肉	牛 (cattle) + 肉 (meat)	beef
牛角	牛 (cattle) + 角 (horn)	horn

Radical-aware Attention-based Four-Granularity model

Different from previous work, our goal is to take advantage of radicals and leverage four different granularities of features to comprehensively model Chinese texts. Furtherly, we systematically integrate these features into the task of Chinese text classification, so that to deal with the huge difference between Chinese and English.

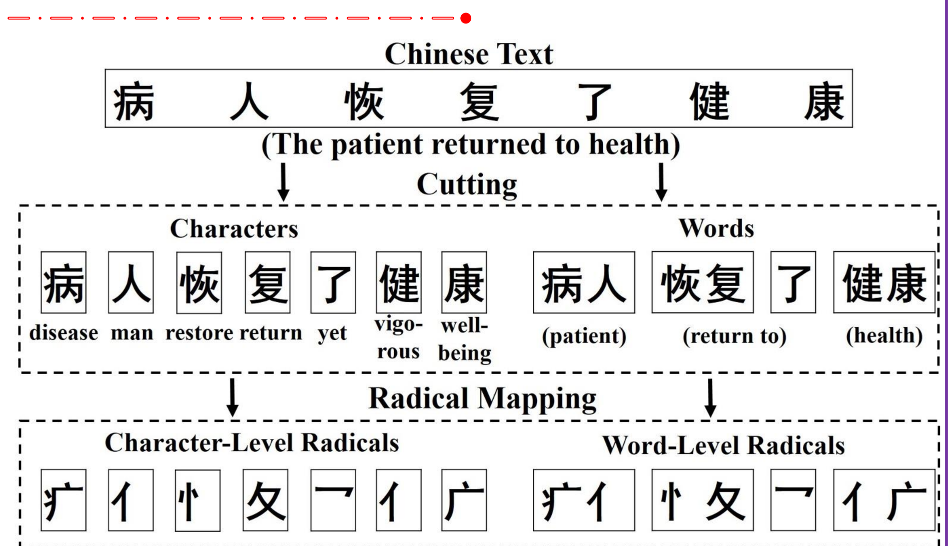
The Radical-aware Attention-based Four-Granularity (RAFG) Model

- *Input Layer*
- *Embedding Layer*
- *Representation Layer*
- *Prediction Layer*



Feature Acquisition

- *Characters (C)*
- *Words (W)*
- *Character-level Radicals (R^c)*
- *Word-level Radicals (R^w)*



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

Implementation

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

Hidden Representation Calculation:

Given a specific feature embedding sequence of a sentence $s = \{x_1, x_2, \dots, x_N\}$, the hidden vector of a BLSTM is calculated as follows:

$$\overrightarrow{h_t} = LSTM(\overrightarrow{h}_{t-1}, x_t),$$
$$\overleftarrow{h_t} = LSTM(\overleftarrow{h}_{t+1}, x_t),$$
$$y_t = [\overrightarrow{h_t}, \overleftarrow{h_t}],$$

Attention Mechanism:

1) To capture the interrelations between radicals and their corresponding characters or words;
2) The radical information can be further modified by the attention weight sum of character context and word context.

$$\alpha' = [\alpha'_1, ..., \alpha'_\epsilon, ..., \alpha'_m],\ \alpha'_\epsilon = f(y^c_\epsilon, e^{rc}_i), 1 \leqslant \epsilon \leqslant m,$$
$$\beta' = [\beta'_1, ..., \beta'_\theta, ..., \beta'_n],\ \beta'_\theta = f(y^w_\theta, e^{rw}_j), 1 \leqslant \theta \leqslant n,$$
$$\alpha_i = \frac{exp\,(\alpha'_\epsilon)}{\sum_{\epsilon=1}^{m} exp\,(\alpha'_\epsilon)},\ where \sum_{i=1}^{m} \alpha_i = 1,$$
$$\beta_j = \frac{exp\,(\beta'_\theta)}{\sum_{\theta=1}^{n} exp\,(\beta'_\theta)},\ where \sum_{j=1}^{n} \beta_j = 1,$$
$$\widetilde{e}^{rc}_i = \sum_{\epsilon=1}^{m} \alpha_\epsilon y^c_\epsilon,\ \widetilde{e}^{rw}_j = \sum_{\theta=1}^{n} \beta_\theta y^w_\theta,$$

Prediction:

$$Con = y^c_o \oplus y^w_o, \quad O = sigmoid\,(Con \times W), \quad P = argmax\,(softmax\,(O)).$$

Objective Function:

$$Loss = -\sum_{T \in Corpus} \sum_{i=1}^{K} p_i(T) \log p_i(T)$$

Experiments

1. Dataset: We conduct experiments on two real-world datasets.

Table 3: Statistics of *dataset#1* and *dataset#2*.

Dataset		Count	Len (Avg. / Max)	Class
Dataset#1	Train	47,952	17.8 / 56	32
	Test	15,986	17.7 / 56	
Dataset#2	Train	36,431	16.7 / 46	32
	Test	12,267	16.7 / 43	

2. Experimental Results of Different Methods

- Word-level radicals are worthy of attention.
- Our model (RAFG) gains a higher performance than any other comparison methods

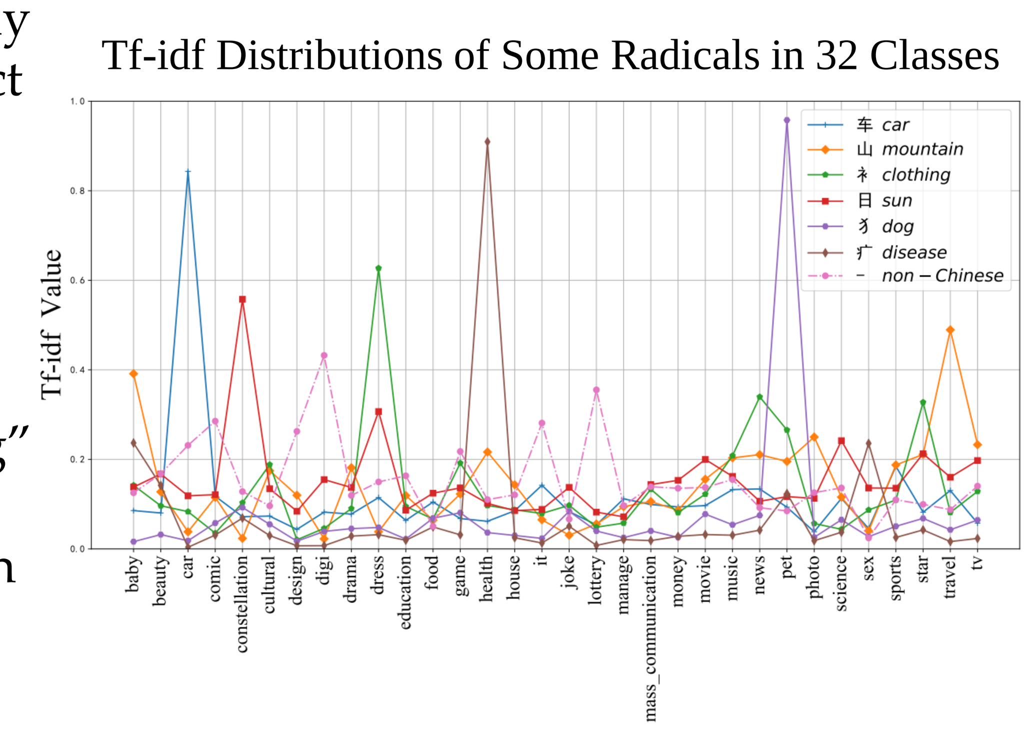
Table 4: Experimental results of different methods on $dataset\#1$ and $dataset\#2$.

Methods	*Dataset#1* F_1 (P, R)	*Dataset#2* F_1 (P, R)
SVM + BOW (W)	0.7552 (0.7639, 0.7514)	0.7341 (0.7459, 0.7303)
SVM + BOW (C)	0.7421 (0.7440, 0.7420)	0.7252 (0.7268, 0.7255)
SVM + BOW (R^w)	0.6834 (0.6913, 0.6800)	0.6762 (0.6858, 0.6729)
SVM + BOW (R^c)	0.4697 (0.4652, 0.4809)	0.4691 (0.4636, 0.4813)
LSTM (E^C)	0.7077 (0.7108, 0.7077)	0.6871 (0.6926, 0.6887)
LSTM (E^W)	0.8029 (0.8034, 0.8031)	0.7875 (0.7893, 0.7885)
Four LSTMs ($E^W + E^C + E^{R^w} + E^{R^c}$)	0.8072 (0.8078, 0.8074)	0.7904 (0.7912, 0.7910)
Four BLSTMs ($E^W + E^C + E^{R^w} + E^{R^c}$)	0.8098 (0.8103, 0.8103)	0.7915 (0.7925, 0.7921)
C-LSTMs ($E^W + E^C$)	0.8112 (0.8118, 0.8115)	0.7931 (0.7944, 0.7929)
C-BLSTMs ($E^W + E^C$)	0.8128 (0.8135, 0.8131)	0.7956 (0.7951, 0.7972)
Ours (RAFG)	**0.8181** (0.8181, 0.8187)	**0.7999** (0.7993, 0.8010)

3. Discussion

- The peaks and valleys exactly reflect the classification effect of radicals.
- Radicals can help recognize semantics and classify Chinese texts.
- For example, the original meaning of radical "clothing" is closed to the concept of class "dress", where the high tf-idf value is a convincing indication.

Tf-idf Distributions of Some Radicals in 32 Classes



4. Conclusion

- Our work presents a novel insight on how to leverage radicals.
- Simply introducing radicals to Chinese text classification cannot improve the performance well.
- Making rational use of radicals is necessary.
- Attention mechanism in RAFG can enhance the effect of radicals.
- Extensive experiments demonstrate the superiority of our model and the effectiveness of radicals in the task of Chinese text classification.