Moto: Enhancing Embedding with Multiple Joint Factors for Chinese Text Classification

Xunzhu Tang*
National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology
Wuhan, China
tangxz@hust.edu.cn

Rujie Zhu
Department of Electrical and Computer Engineering
University of Central Florida
Orlando, FL, USA
rujie.zhu@ucf.edu

Tiezhu Sun
Momenta
Suzhou, China
suntiezhu@momenta.ai

Shi Wang*
Institute of Computing Technology, Chinese Academy
beijing, China
wangshi@ict.ac.cn

***Abstract*—Recently, language representation techniques have achieved great performances in text classification. However, most existing representation models are specifically designed for English materials, which may fail in Chinese because of the huge difference between these two languages. Actually, few existing methods for Chinese text classification process texts at a single level. However, as a special kind of hieroglyphics, radicals of Chinese characters are good semantic carriers. In addition, Pinyin codes carry the semantic of tones, and Wubi reflects the stroke structure information, *etc*. Unfortunately, previous researches neglected to find an effective way to distill the useful parts of these four factors and to fuse them. In our works, we propose a novel model called Moto: Enhancing Embedding with Multiple Joint Factors. Specifically, we design an attention mechanism to distill the useful parts by fusing the four-level information above more effectively. We conduct extensive experiments on four popular tasks. The empirical results show that our Moto achieves SOTA 0.8316 (F_1-score, 2.11% improvement) on Chinese news titles, 96.38 (1.24% improvement) on Fudan Corpus and 0.9633 (3.26% improvement) on THUCNews.**

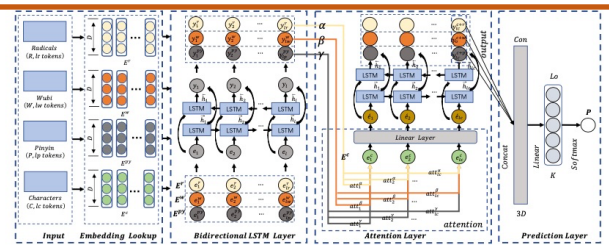

Fig. 2: Network architecture of **Moto**, including four-granularity representations of Chinese: Radicals, Wubi, Pinyin, and Characters.

TABLE I: Experimental results of different methods on Chinese news titles, Fudan Corpus, Douban movie review, and THUCNews.

Methods	Chinese news titles dataset #1	Chinese news titles dataset #2	Fudan Corpus	THUCNews
	F1(P,R)	F1(P,R)	F1(P,R)	F1(P,R)
SVM+BOW(C)	0.7421 (0.7440, 0.7420)	0.7252 (0.7268, 0.7255)	0.8434 (0.8373, 0.8495)	0.8713 (0.8811, 0.8618)
SVM+BOW(R)	0.4697 (0.4652, 0.4809)	0.4691 (0.4636, 0.4813)	0.8187 (0.8216, 0.8158)	0.8641 (0.8637, 0.8646)
SVM+BOW(W)	0.6021 (0.6041, 0.6002)	0.4852 (0.4783, 0.4923)	0.8303 (0.8229, 0.8378)	0.8638 (0.8597, 0.8679)
SVM+BOW(Py)	0.7290 (0.7309, 0.7271)	0.6702 (0.6874, 0.6539)	0.8359 (0.8367, 0.8352)	0.8703 (0.8778, 0.8629)
Four LSTMs (C + R + W + Py)	0.8072 (0.8078, 0.8074)	0.7904 (0.7912, 0.7910)	0.8826 (0.8841, 0.8811)	0.9018 (0.9022, 0.9014)
Four BiLSTMs (C + R + W + Py)	0.8098 (0.8103, 0.8103)	0.7915 (0.7925, 0.7921)	0.8899 (0.8990, 0.8809)	0.9122 (0.9191, 0.9054)
RAFG	0.8181 (0.8181, 0.8187)	0.7999 (0.7993, 0.8010)	0.9172 (0.9201, 0.9144)	0.9002 (0.9033, 0.8972)
cw2vec(stroke-level)	– (–, –)	– (–, –)	0.9520 (0.9528, 0.9511)	0.9329 (0.9433, 0.9227)
C-LSTMs (C)	0.8108 (0.8102, 0.8114)	0.7931 (0.7944, 0.7929)	0.8801 (0.8828, 0.8774)	0.9033 (0.9054, 0.9012)
C-LSTMs (C + R + W + Py)	0.8163 (0.8177, 0.8149)	0.7956 (0.7951, 0.7972)	0.8823 (0.8775, 0.8871)	0.9036 (0.9068, 0.9004)
C-BiLSTMs (C)	0.8140 (0.8153, 0.8127)	0.7757 (0.7754, 0.7922)	0.9213 (0.9309, 0.9118)	0.9236 (0.9290, 0.9183)
C-BiLSTMs (C + R + W + Py)	0.8211 (0.8246, 0.8177)	0.7939 (0.7957, 0.7922)	0.9264 (0.9384, 0.9147)	0.9294 (0.9332, 0.9257)
Moto(BiLSTM)	**0.8316** (0.8346, 0.8287)	**0.8168** (0.8192, 0.8144)	**0.9638** (0.9671, 0.9605)	**0.9633** (0.9679, 0.9588)

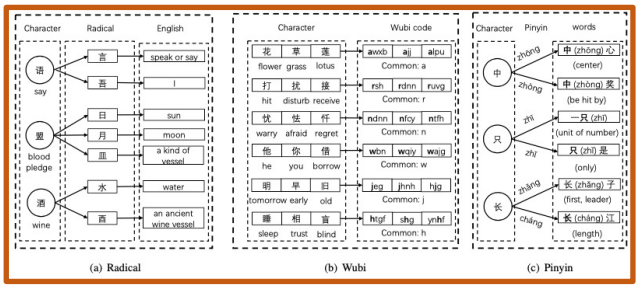

Fig. 1: We employ three kind of representations to enhance the character embedding. Figure (a) indicates that radicals can show more details of characters; Figure (b) shows that Wubi code can capture the structure information of characters; And figure (c) expresses that the Pinyin (with pronounce) is important to Chinese characters.

Radical or radical-like components serving as the basic units for building Chinese characters has been explored in [1], [9]. Commonly, radicals has the following two features. The first one is that one radical normally has one or two types. '言' (speak) is itself, but it becomes '讠' in character '语'. The second is that radicals have specific meanings. In Figure 1(a), the character '语' (*talk or speak*) have radicals : '讠' (*the same as* 言, *speak*) and '吾' (*I*). Obviously, radical provides extra photographic features of characters.

Wubi is another effective representation of Chinese characters, which includes more comprehensive structure information compared to radical. Each element in a Wubi code represents a type of structure (or stroke) in characters. In Figure 1(b), '花' (flower), '草' (grass), and '莲' (lotus) are all related to plants, and their Wubi codes 'awxb', 'ajj', and 'alpu' have one common letter 'a', which is corresponding to radical '艹'. Therefore, Wubi is an efficient approach to capture structure features of Chinese characters.

Pinyin is a English-like expression approach of Chinese characters. Besides, Pinyin is highly relevant to semantics - one character may have multiple pronunciations corresponding to different semantic meanings [3], which is called polyphone

When comparing four LSTMs (C + R + W + Py), Four BiLSTMs (C + R + W + Py), RAFG, and cw2vec, we can find that RAFG which takes attention mechanism achieves the best performance, whose average F_1-value is 0.8589, higher than Four LSTMs (0.8455)) and Four BiLSTMs (0.8509). Moreover, cw2vec achieves the best performance in Fudan Corpus and THUCNews. Additionally, for C-LSTMs (C), C-LSTMs (C + R + W + Py), C-BiLSTMs(C), and C-BiLSTMs(C + R + W + P), the results indicate that methods with bidirectional version achieve better performance. At the same time, four-granularity model is better than single character-level model. Figure 4 plots that the comparison in F_1-value among C-BiLSTMs, RAFG, and our model Moto. We can see that Moto achieves the best performance in the most classes in dataset #1 and dataset #2.

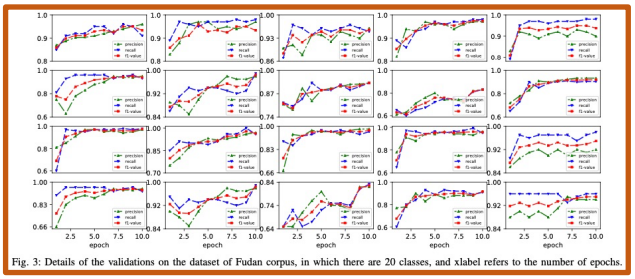

Fig. 3: Details of the validations on the dataset of Fudan corpus, in which there are 20 classes, and xlabel refers to the number of epochs.

We provide the comparison results with SVM+BOW employing characters, radicals, Wubi codes, and Pinyin codes as features respectively. Table I shows that SVM + BOW (C) achieves the best average F_1-value 0.7955, 2.5% higher than SVM + BOW (Py) in four Chinese text classification tasks. At the same time, Wubi gets average F_1-value 0.6954, as well radical gets 0.6554. The results indicate that all these four aspects are carriers of semantics in Chinese, and character plays the most important role in them.

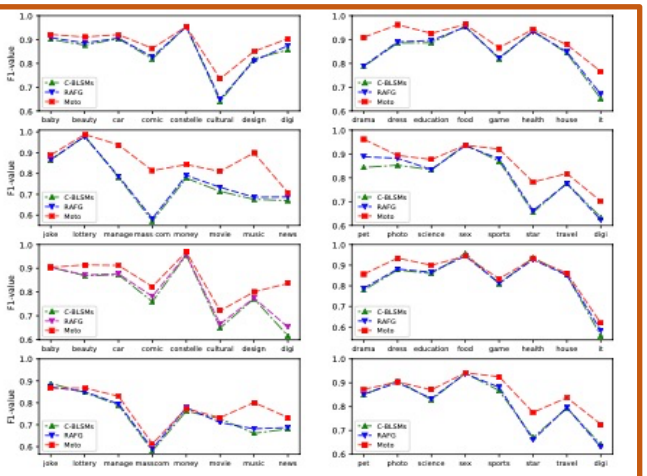

Fig. 4: Detailed comparison on the dataset of Chinese news titles, Sub-figures in former two rows describe the dataset#1, and sub-figures in the later two rows are related to dataset#2.