Glyph-aware Embedding of Chinese Characters

Falcon Z. Dai
Toyota Technological Institute at Chicago
dai@ttic.edu

Zheng Cai
The University of Chicago
jontsai@uchicago.edu

paper: https://arxiv.org/abs/1709.00028 code: https://github.com/falcondai/chinese-char-lm

Highlights

- **a novel character embedding model that explicitly incorporates visual appearance of Chinese characters.**
- **a quantitative study of the contribution of sub-glyph visual information in Chinese NLP tasks.**
- **new state-of-the-art results on a Chinese segmentation benchmark task.**

Introduction & Hypothesis

Unlike English script which is ***alphabetic*** with a small alphabet, Chinese script is ***logographic*** with a large set of characters which are meaningful individually and in combination.
One of the distinctive advantages of character-level or subword-level modeling is their high coverage, i.e., few or no out-of-vocabulary (**OOV**) tokens, with a small set of tokens.
With Chinese corpora, there is a strong case for modeling at character-level as since the ***segmentation*** of characters into words is usually unavailable, Written Chinese, Japanese and Korean usually do not contain word segmentation as Western languages do.

original: 这是一篇有趣的文章

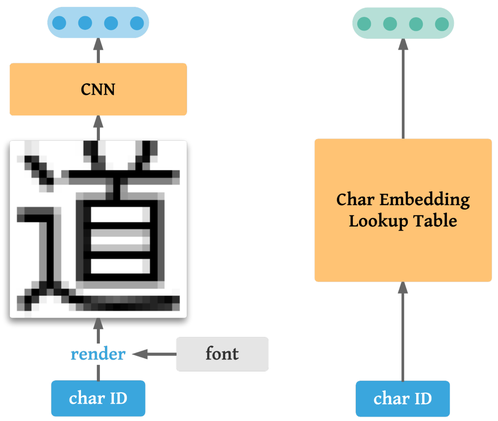
segmented: 这 是 一篇 有趣的 文章

It is well-known that many Chinese characters' written form, their *glyphs*, share common sub-structures and some of these sub-structure are informative of the semantics and phonetics of the characters.

We hypothesize that the semantic information of sub-glyph structures can help improve the character embeddings and thus improve performance in Chinese NLP tasks.

CNN Embedder

We feed the glyph as an input to a feed-forward neural network (FNN) model, an *embedder*, that outputs an *embedding vector* which, in both the segmentation task and the language modeling task, is then consumed by a recurrent neural network to make predictions.
A traditional ID embedder is defined as a trainable embedding lookup table, while a glyph-aware embedder is defined as a CNN output of a rendered glyph of a character.
For the CNN embedder, we used a two layer ReLU-gated CNN: 32 (7, 7) filters with (2, 2) stride in the first layer, 16 (5, 5) filters (2, 2) stride in the second layer, and a fully-connected layer at the end.



Downstream Tasks

Language Modeling

We model language model in character level as:

$$p(c_1, \cdots, c_n) = p(c_1) \prod_{i=2}^{n} p(c_i | c_1, \cdots, c_{i-1})$$

We compare traditional glyph-unaware character level language model (ID embedder) with the proposed glyph-aware embedding (CNN embedder).

Embedder	Test Perplexity
ID embedder	**47.53**
Linear embedder	71.51
CNN embedder	55.51
ID embedder + linear embedder	54.69
ID embedder + CNN embedder	**47.75**

Segmentation

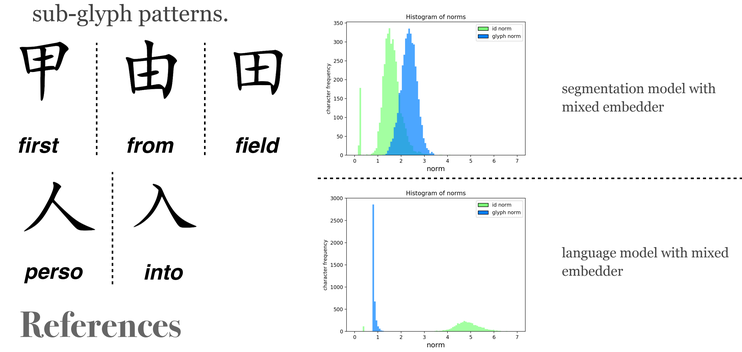
We applied the proposed CNN embedder in classic segmentation task. We use Peking University dataset (PKU) and Microsoft Research dataset (MSR) from the Second International Chinese Word Segmentation Bakeoff to compare the proposed CNN embedder with the ID embedder. We use Bidirectional LSTM(BiLSTM) to do classification

	RNN segmentors	embedder	Precision	Recall	F1
PKU	GRU	ID	87.41	84.14	85.75
		CNN	90.03	**89.54**	**89.78**
		ID + CNN	**90.46**	88.80	89.62
	BiLSTM	ID	96.06	94.66	95.36
		CNN	94.73	94.88	94.81
		ID + CNN	**96.91**	**95.41**	**96.15**
	NWS (Cai and Zhao 2016)		95.5	94.9	95.16

	RNN segmentors	embedder	Precision	Recall	F1
MSR	GRU	ID	86.97	85.25	86.10
		CNN	**89.93**	86.79	**88.33**
		ID + CNN	88.81	**87.19**	88.00
	BiLSTM	ID	97.34	**97.25**	97.29
		CNN	97.07	96.98	97.03
		ID + CNN	**97.82**	97.04	**97.43**
	NWS (Cai and Zhao 2016)		96.1	96.7	96.4

Analysis & Discussion

- Visual similarity does *not* always imply semantic/syntactic/phonetic similarity between characters. This might explain the lower performance of CNN embedder (vs ID embedder).
- Each character only has exactly one sample (we limit vocabulary size to 4K). The CNN embedder might be overfitting and unable to learn sub-glyph patterns.



References

Cai, Deng, and Hai Zhao. "Neural word segmentation learning for Chinese." arXiv preprint arXiv:1606.04300 (2016).
The Second International Chinese Word Segmentation Bakeoff took place over the summer of 2005 and the results were presented at the 4th SIGHAN Workshop, held at IJCNLP'05, October 14-15.