Multimodal neural pronunciation modeling for spoken languages with logographic origin

Minh Nguyen
National University of Singapore
elenguy@nus.edu.sg

Gia H. Ngo
National University of Singapore
ngohgia@u.nus.edu

Nancy F. Chen
Institute for Infocomm Research
Singapore
nancychen@alum.mit.edu

Abstract

Graphemes of most languages encode pronunciation, though some are more explicit than others. Languages like Spanish have a straightforward mapping between its graphemes and phonemes, while this mapping is more convoluted for languages like English. Spoken languages such as Cantonese present even more challenges in pronunciation modeling: (1) they do not have a standard written form, (2) the closest graphemic origins are logographic Han characters, of which only a subset of these logographic characters implicitly encodes pronunciation. In this work, we propose a multimodal approach to predict the pronunciation of Cantonese logographic characters, using neural networks with a geometric representation of logographs and pronunciation of cognates in historically related languages. The proposed framework improves performance by 18.1% and 25.0% respective to unimodal and multimodal baselines.

1 Introduction

In phonographic languages, there is a direct correspondence between graphemes and phonemes (Defrancis, 1996), though this correspondence is not always one-to-one. For example, in English, the word *table* corresponds to the pronunciation ["teɪ.bl], in which each alphabetic character corresponds to one phoneme, and the character *e* is mapped to silence. However, in logographic languages, the correspondence between graphemes and phonemes is more ambiguous (Defrancis, 1996), as only some sub-units in a grapheme are indicative of its phonemes. Korean[1], Vietnamese[2] and Chinese languages (e.g. Cantonese) are examples of logographic languages, all belonging to the Han logographic family. Similar to pronunciation modeling in phonographic languages, in which words are broken down into characters and modeling is done at the character level, pronunciation modeling in logographic languages requires decomposing logographs into sub-units and extracting only sub-units carrying pronunciation hints. As the correspondence of Han logograph to phoneme is intricately complex with many sub-rules or exceptions (Hashimoto, 1978), it is challenging to computationally model these correspondences using white box approaches (e.g. graphical model). Instead, we exploit neural networks, as they (1) can flexibly model the implicit similarity of grapheme-phoneme relationships across languages with Han origin, (2) can automatically learn the most relevant knowledge representation with minimal feature engineering (LeCun et al., 2015), such as extracting pronunciation hints from logographic representations.

Due to historical contact, there is much lexical overlap across Han logographic languages, as they borrowed words from one another (Rokuro, 1969; Miyake, 1997; Loveday, 1996; Sohn, 2001; Alves, 1999). As a result, cognates in different languages are written using identical graphemes but pronounced differently. For example, [she] in Mandarin and [sip] in Cantonese are cognates; their pronunciations are different yet they are written using the same logograph (懾), which represents "admire". Though Han logographic languages are mutually unintelligible (Tang and Van Heuven, 2009; Handel, 2015), the correspondence of Han logographic graphemes to phonemes across languages is often similar in systematic ways (Cai et al., 2011; Frellesvig and Whitman, 2008; Miyake, 1997). The shared characteristics in pronunciation of cognates could be leveraged in deciphering the pronunciation of Han logographs. In this work, we proposed a neural pronuncia-

[1] A large portion of Korean vocabulary are Sino-Korean written in Hanja (Korean logographs) (Sohn, 2001)

[2] Traditional Vietnamese vocabulary comprises of Sino-Vietnamese words written by Chinese logographs and locally-invented Nom logographs (Alves, 1999).

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2916–2922
Brussels, Belgium, October 31 - November 4, 2018.

tion model that exploits both embeddings of logographs and cognates' phonemes. The proposed model significantly improves pronunciation prediction of logographs in Cantonese.

2 Related Work

The basic units in writing (graphemes) of Han logographic languages are logographs. A word contains one or more logographs and a logograph consists of one or more radicals. The pronunciation of a logograph corresponds to a syllable which has three phonemes: onset, nucleus and coda.

Grapheme-to-phoneme (G2P) approaches such as (Xu et al., 2004; Chen et al., 2016) predicted a Han logograph's pronunciation from its local context in a phrase. This was similar to predicting a Latin word's pronunciation from its surrounding words, essentially treated individual logographs as the basic units of the model and did not delve further into the logographic sub-units (the radicals).

While we are unaware of any work that derives features for pronunciation prediction from logographs, there are recent work in deriving representation of logographs for various semantic tasks. Some methods (Shi et al., 2015; Ke and Hagiwara, 2017; Nguyen et al., 2017; Zhuang et al., 2017) decomposed logographs into sub-units using expert-defined rules and then extracted the relevant semantic features. Other methods use convolutional neural network to extract features from the images of logographs (Dai and Cai, 2017; Liu et al., 2017; Toyama et al., 2017). Other works combined multiple level of information for feature extraction, using both logograph and sub-units obtained from logograph decomposition (Dong et al., 2016; Han et al., 2017; Peng et al., 2017; Yu et al., 2017; Yin et al., 2016).

In this work, we explicitly looked at the relationship between a logograph's constituent radicals and its pronunciation. Among Han logographs, 81% of frequently used logographs are semantic-phonetic compounds (Li and Kang, 1993) which consist of radicals that might contain phonetic or semantic hints (Hsiao and Shillcock, 2006). The pronunciation of a logograph could conceivably be predicted from the phonetic radicals. Furthermore, the relative position of radicals in the logograph might also offer clues about it pronunciation. Table 1 shows an example of such intricate relationships between a logograph's pronunciation and its constituent radicals. All Han logographs in the table have a common phonetic radical (in red), which offers an inkling of the pronunciation of these logographs. For instance, logographs that have the phonetic radical on the left (剖 and 部) share a similar pronunciation in Korean (in blue) while logographs that have the phonetic radical on the right (陪, 賠, and 蓓) share a similar pronunciation in Mandarin, Cantonese and Vietnamese. Note that for each logograph, their pronunciations across the different languages share similarities: when the phonetic radical is on the left, the nucleus ends in a back vowel like *u* or *o*, whereas when the phonetic radical is on the right, the nucleus ends in a front vowel like *i*.

Position of 音	音		音		
Logograph	剖	部	陪	賠	蓓
Mandarin	pou	bu	**pei**	**pei**	bei
Cantonese	fau	bou	**pui**	**pui**	bui
Korean	**pwu**	**pwu**	**pay**	**pay**	**pay**
Vietnamese	phau	bo	**boi**	**boi**	bui

Table 1: The position of radicals affects pronunciations. All logographs share a common radical in red. Similar pronunciations for 剖 and 部 are bolded in blue. Similar pronunciations for 陪, 賠, and 蓓 are bolded in green. The pronunciation of a logograph in Mandarin, Cantonese, Korean and Vietnamese are represented by Pinyin, Jyutping, Yale, and Vietnamese alphabet symbols respectively.

The example in Table 1 explains the motivation for our proposed approach to predict a logograph's pronunciation by modelling both the constituent radicals and their geometric positions. Furthermore, the proposed approach can generalize to unseen logographs if the co-occurrence patterns of their constituent radicals have been learnt.

3 Model

We first describe a geometric decomposition of logographs and then different neural pronunciation models for logographs. Finally, we present a multimodal neural model that incorporates both logographic input and the cognates' phonemes in predicting pronunciation of logographs.

Representation of Han logographs

The majority of logographs (characters) in Han logographic language family comprise of a radical that indicates its nominal semantic category and a phonetic radical that gives an inkling of the pronunciation (Defrancis, 1996). Thus, patterns of co-occurrence of radicals across logographs might

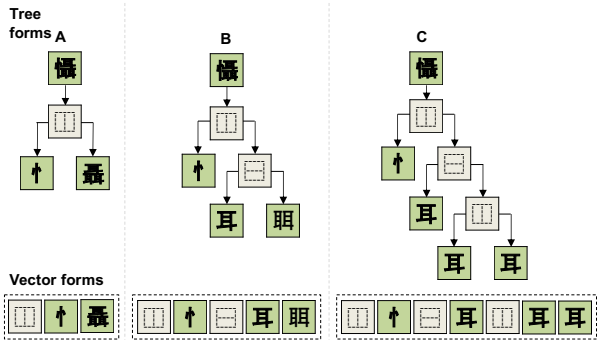

*Figure 1: Geometric representation of the logograph "admire". **A**, **B** and **C** are equivalent decomposition of the same logograph but with different levels of granularity. The geometric representation comprises of both the radicals and geometric operators, which can be used to reconstruct the original logograph.*

be exploited to find the phonetic radicals, which in turn can suggest the corresponding pronunciation of a logograph. Using this intuition, we model the pronunciation of logographs at the radical level.

We investigated two representations of radicals in a logograph. In the first approach, a logograph is represented as a bag of its unordered constituent radicals (BoR), encoded as a vector of radical counts. The second approach is to use a decomposition of radicals in the logograph that retains the original geometric organization of the radicals. The geometric decomposition (GeoD) approach preserves important cues about the word's pronunciation in the relative position of the radicals. For example, differentiating the left radical from the right radical in a left-right semantic-phonetic compound allows more effective extraction of pronunciation hints. In addition, radicals that should be interpreted together are closer spatially in the GeoD representation, making the knowledge representation easier to learn. Note that the GeoD representation is lossless as the original logograph can be reconstructed perfectly (details in Appendix A). Figure 1 shows the geometric decomposition of the Han logograph "admire" at three levels of granularity.

Neural pronunciation prediction models

Figure 2 and Figure 3 show two neural pronunciation prediction models of logographs. In Figure 2, each logograph is treated as an ordered "bag of radicals" (BoR). For example, assume the vocabulary of radicals in the whole dataset is [𠂆, 氵, 耳, 灬], the word 慑 ("admire" - see Figure 1) is represented by a vector of counts $[1, 0, 3, 0]$, corresponding to one radical 𠂆 and three radicals 耳.

The BoR is input to a multilayer perceptron (MLP) with three layers of size 750, 500, 250. L2 regularization of 1e-4 is applied to the hidden layers. The three dropout layers have dropout probabilities of 0.5, 0.5, and 0.2, respectively. As the output variables are categorical, cross-entropy loss was used.

We investigated two structures for predicting output phonemes (i.e. onset, nucleus, coda). In the first structure, output phonemes were predicted independently using the last hidden layer. The second structure made a sequential prediction (1) the coda was first predicted using the last hidden layer (2) the nucleus was predicted using both the final hidden layer and the predicted coda, and (3) the onset was predicted using the last hidden layer together with the predicted coda and nucleus. The second structure was motivated by a stronger dependency between the nuclues and coda. For example, the nucleus and coda are often grouped together as a single unit (rime/final) in the syllabic structure of most languages (Kessler and Treiman, 2002). In our experiments, the sequential structure yielded lower error rates so it is used in all neural network models.

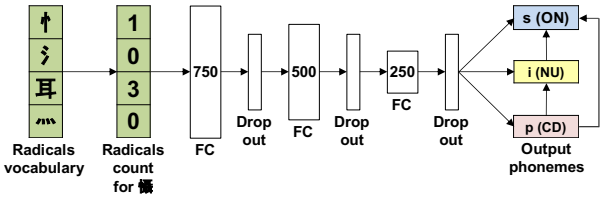

Figure 2: Pronunciation model of logographs using multilayer perceptron (MLP). FC: Fully connected.

In Figure 3, each logograph is represented by its geometric decomposition (GeoD). For example, the logograph 慑 is represented by a sequence of radicals and geometric operators shown in Figure 1C. The neural prediction model consists of two LSTM layers with 256 memory cells each. Input and recurrent dropout (Gal and Ghahramani, 2016) of 0.2 and 0.5 are applied to the LSTM layers to prevent overfitting.

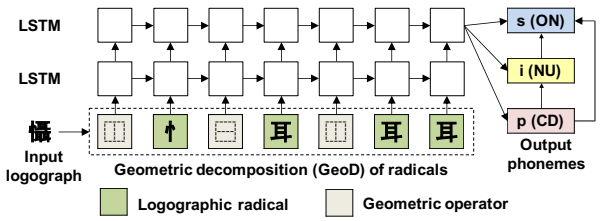

Figure 3: Neural pronunciation model with geometric decomposition of logographs.

Multimodal neural pronunciation model of logographs

In this section, we want to model the pronunciations of a logograph in the target language Cantonese using multimodal information from both the logograph and phonemes of the cognates, as shown in Figure 4. Given a vocabulary of phonemes in the source languages related to Cantonese (Mandarin, Korean, Vietnamese), the cognates' phonemes are encoded as an indicator vector, with an element equals 1 if the corresponding phoneme in the vocabulary appears in a cognate's pronunciation, and 0 otherwise.

The geometric decomposition (GeoD) of the logograph is fed to two LSTM layers. The output at the last time step is concatenated together with the multilingual phonemic vector and used as input for a multi-layer perceptron (MLP). The MLP and LSTM setups are the same as those in Figure 2 and Figure 3 respectively. Deep supervision (Szegedy et al., 2015) was applied by using the output of the LSTM to make auxiliary prediction of the output phonemes. Note that the auxiliary prediction should be identical to the main prediction. While predicting the same target, the main prediction used both cognate phonemes and the logograph while the auxiliary prediction used only the logograph. This was to ensure features extracted from the logographs are useful for pronunciation prediction and are complementary to the features extracted from the multilingual phonemes.

4 Experiments

We investigate whether Cantonese phonemes could be predicted using Han logographs and the cognates' phonemes from Mandarin, Korean, and Vietnamese. The prediction output are Cantonese onsets, nuclei and codas. The experimental design is motivated by the nature of Han-logographic languages. A Chinese logograph (character) is phonologically equivalent to a syllable in English while the constituent radicals are analogous to alphabet letters (with far less phonetic information). While in most languages, a syllable's pronunciation is influenced by neighboring syllables, most Han-logographic languages are monosyllabic and a logograph's pronunciation is rarely affected by neighboring logographs. Therefore, pronunciation prediction at the logograph (character) level for Han logographs is more appropriate. We use string error rate (SER) and token error rate (TER) as evaluation metrics. A wrongly predicted phoneme (onset, nucleus or coda) is counted as one token error. A syllable containing token error(s) is counted as one string error. All the neural networks were trained using Adam (Kingma and Ba, 2014).

Data

The dataset is extracted from the UniHan database,[3] which is a pronunciation database of logographs from Han logographic languages and maintained by the Unicode consortium. For each entry in the dataset, a logograph corresponds to phonemes in Cantonese, Mandarin, Korean and Vietnamese, represented by Jyutping,[4] Pinyin,[5] Yale,[6] and Vietnamese alphabet symbols respectively.[7] We randomly partition the dataset into two sets, with 80% for training and the other 20% for testing. Overall, there are 16,011 entries in the training set and 4,002 entries in the test set. 1000 entries of the training set are used as the development set for hyper-parameters fine-tuning.

In the test set, only 16% of logographs have pronunciations in all non-target languages, while 6% of logographs have no non-target language pronunciation. The availability of pronunciations in non-target languages differs from logograph to logograph. For example, some logographs have Mandarin and Korean pronunciations, while others only have Mandarin pronunciations.

Predicting pronunciation using logograph input

We compared the neural networks against a decision tree baseline. The decision tree baseline was implemented using scikit-learn (Pedregosa et al., 2011). The input of the decision tree (DT) model is the BoR representation of the logograph, while the input of neural networks can be either BoR or GeoD. The MLP network in Figure 2 uses BoR, while the LSTM in Figure 3 uses GeoD as input. All models output phonemes in Cantonese.

From Table 2, the neural network (MLP) outperforms decision tree when using BoR input. Both the SER and TER of the MLP model are lower than those of the decision tree. The LSTM model using GeoD leads to the lowest SER and TER, suggesting the benefits of relative positional

[3] https://www.unicode.org/charts/unihan.html
[4] https://en.wikipedia.org/wiki/Jyutping
[5] https://en.wikipedia.org/wiki/Pinyin
[6] https://en.wikipedia.org/wiki/Yale_romanization_of_Korean
[7] https://en.wikipedia.org/wiki/Vietnamese_alphabet

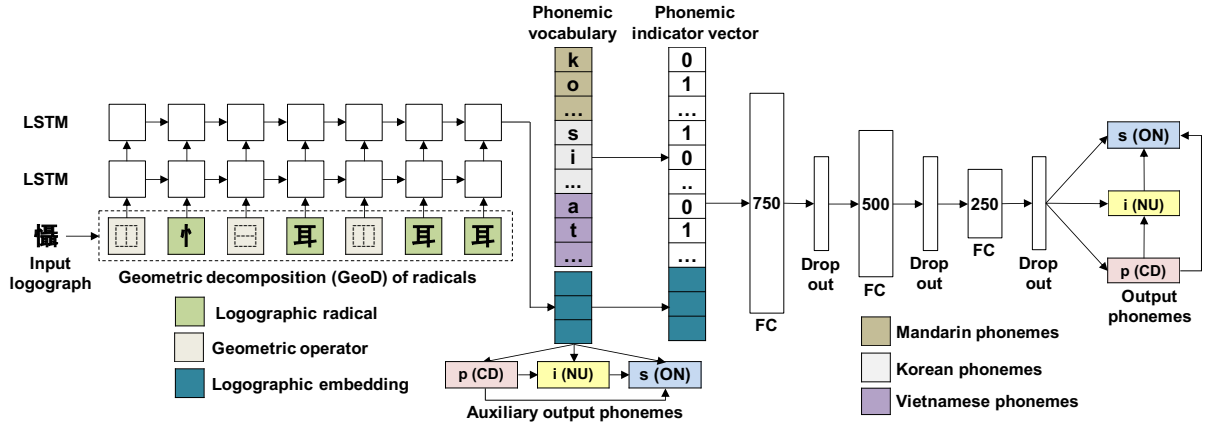

Figure 4: Multimodal neural pronunciation prediction model using logographs' geometric representation and cognates' phonemes.

information of radicals in predicting pronunciation. The trends of onset, nucleus and coda error rates are similar to those of TER and SER. However, as the gap of of error rate between MLP (BoR) and LSTM (GeoD) for TER and SER are quite small, using BoR instead of GeoD can be a good computation-accuracy trade-off.

Method	SER	TER	On.	Nu.	Cd.
DT (BoR)	63.8	39.8	50.7	45.7	22.9
MLP (BoR)	59.2	33.6	44.5	38.6	17.8
LSTM (GeoD)	**58.4**	**32.6**	**43.3**	**37.4**	**17.1**

Table 2: Prediction error rates of Cantonese phonemes by decision tree (DT), MLP and LSTM using only logographic input. Best results are in bold.

Predicting pronunciation using multimodal input

The input of the models are logographs and cognate phonemes from Mandarin, Korean and Vietnamese. Table 3 shows that the proposed multimodal neural network exploits multimodal and geometric information effectively. The relative improvement reaches 18.2% and 33.3% for SER and TER respectively. The last rows in Table 2 and Table 3 show that by combining Korean, Mandarin and Vietnamese phonemes input with GeoD, the prediction performance improves by 54.1% relative in TER and by 65.5% relative in SER. Moreover, using solely logograph input resulted in higher onset error (43.3%) than nucleus error (37.4%) while using both logographs and multilingual phonemes improves the onset error (23.5%) to be lower than nucleus error (24.6%). This suggests that logographs and phonemes of cognates provide complementary information about the pronunciation of a logograph, which in this case, most notably at the onset position. While logographs usually carry hints about phonemes at the nucleus and coda position but not at the onset position, multilingual phonemes input might carry hints about pronunciation at all three positions.

Method	SER	TER	On.	Nu.	Cd.
DT (BoR, ph)	44.0	24.8	29.8	29.9	14.7
MLP (BoR, ph)	38.5	19.6	23.4	24.8	10.5
LSTM (GeoD, ph)	**37.2**	**18.6**	**22.6**	**23.4**	**9.8**

Table 3: Prediction error rates of Cantonese phonemes by multimodal models; BoR: Bag of Radicals; GeoD: Geometric Decomposition; ph:phonemes. Best results are in bold.

5 Discussion

We have empirically shown that the systematic yet tenuous correspondence between pronunciations of cognates in Han logographic languages can be exploited for pronunciation modeling using neural networks. Moreover, combining logograph with cognate pronunciations further improves pronunciation prediction. These results could be potentially applied to speech processing tasks such as speech synthesis, where the construction of pronunciation dictionaries are expert labor-intensive, especially for under-resourced spoken languages.

For future work, recursive neural network (Tai et al., 2015) can be used as it is better suited for the hierarchical logographic decomposition. Besides, incorporating more detailed relationship between radicals (e.g. (Zhuang et al., 2017)) can help improve the model. The proposed approaches can also be applied to other languages such as Min Nan or Hakka, which are spoken languages that are even less well-documented than Cantonese.

References

Mark J Alves. 1999. What's so Chinese about Vietnamese. In *Papers from the ninth annual meeting of the Southeast Asian Linguistics Society*, pages 221–242.

Zhenguang G Cai, Martin J Pickering, Hao Yan, and Holly P Branigan. 2011. Lexical and syntactic representations in closely related languages: Evidence from Cantonese–Mandarin bilinguals. *Journal of Memory and Language*, 65(4):431–445.

Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2016. Acoustic data-driven pronunciation lexicon generation for logographic languages. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5350–5354. IEEE.

Falcon Z Dai and Zheng Cai. 2017. Glyph-aware Embedding of Chinese Characters. *EMNLP 2017*, page 64.

John Defrancis. 1996. Graphemic indeterminacy in writing systems. *Word*, 47(3):365–377.

Chuanhai Dong, Jiajun Zhang, Chengqing Zong, Masanori Hattori, and Hui Di. 2016. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications*, pages 239–250. Springer.

Bjarke Frellesvig and John Whitman. 2008. The Japanese-Korean vowel correspondences. *Japanese/Korean Linguistics*, 13:15–28.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

He Han, Yang Xiaokun, Wu Lei, Yan Hua, Gao Zhimin, Feng Yi, and Townsend George. 2017. Dual long short-term memory networks for sub-character representation learning. *arXiv preprint arXiv:1712.08841*.

Zev Handel. 2015. The classification of Chinese: sinitic (the Chinese language family). In *The Oxford handbook of Chinese linguistics*, pages 34–44. Oxford University Press.

Mantaro J Hashimoto. 1978. Current developments in Sino—Vietnamese studies. *Journal of Chinese Linguistics*, pages 1–26.

Janet Hui-wen Hsiao and Richard Shillcock. 2006. Analysis of a chinese phonetic compound database: Implications for orthographic processing. *Journal of psycholinguistic research*, 35(5):405–426.

Yuanzhi Ke and Masafumi Hagiwara. 2017. Radical-level Ideograph Encoder for RNN-based Sentiment Analysis of Chinese and Japanese. *arXiv preprint arXiv:1708.03312*.

Brett Kessler and Rebecca Treiman. 2002. Syllable structure and the distribution of phonemes in english syllables.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. ArXiv: 1412.6980.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436.

Y Li and JS Kang. 1993. Analysis of phonetics of the ideophonetic characters in Modern Chinese. *Information analysis of usage of characters in modern Chinese*, pages 84–98.

Frederick Liu, Han Lu, Chieh Lo, and Graham Neubig. 2017. Learning character-level compositionality with visual features. *arXiv preprint arXiv:1704.04859*.

Leo J Loveday. 1996. *Language contact in Japan: A sociolinguistic history*. Clarendon Press.

Marc Hideo Miyake. 1997. Pre-Sino-Korean and Pre-Sino-Japanese: reexamining an old Problem from a modern perspective. *Japanese/Korean Linguistics*, 6:179–211.

Viet Nguyen, Julian Brooke, and Timothy Baldwin. 2017. Sub-character Neural Language Modelling in Japanese. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 148–153.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Haiyun Peng, Erik Cambria, and Xiaomei Zou. 2017. Radical-based hierarchical embeddings for Chinese sentiment analysis at sentence level. In *The 30th International FLAIRS conference. Marco Island*.

Kono Rokuro. 1969. The Chinese writing and its influence on the Scripts of the Neighbouring Peoples with special reference to Korea and Japan. *Memoirs of the Research Department of the Toyo Bunko (The Oriental Library) No*, 27:117–123.

Xinlei Shi, Junjie Zhai, Xudong Yang, Zehua Xie, and Chao Liu. 2015. Radical embedding: Delving deeper to Chinese radicals. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 594–598.

Ho-Min Sohn. 2001. *The Korean Language*. Cambridge University Press.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Chaoju Tang and Vincent J Van Heuven. 2009. Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, 119(5):709–732.

Yota Toyama, Makoto Miwa, and Yutaka Sasaki. 2017. Utilizing Visual Forms of Japanese Characters for Neural Review Classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 378–382.

Jun Xu, Guohong Fu, and Haizhou Li. 2004. Grapheme-to-phoneme conversion for chinese text-to-speech. In *Eighth International Conference on Spoken Language Processing*.

Rongchao Yin, Quan Wang, Peng Li, Rui Li, and Bin Wang. 2016. Multi-granularity Chinese word embedding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 981–986.

Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. 2017. Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 286–291.

Hang Zhuang, Chao Wang, Changlong Li, Qingfeng Wang, and Xuehai Zhou. 2017. Natural Language Processing Service Based on Stroke-Level Convolutional Networks for Chinese Text Classification. In *Web Services (ICWS), 2017 IEEE International Conference on*, pages 404–411. IEEE.