Don't Just Scratch the Surface: Enhancing Word Representations for Korean with Hanja

Kang Min Yoo*, **Taeuk Kim*** and **Sang-goo Lee**
Department of Computer Science and Engineering
Seoul National University, Seoul, Korea
{kangminyoo,taeuk,sglee}@europa.snu.ac.kr

Abstract

We propose a simple yet effective approach for improving Korean word representations using additional linguistic annotation (i.e. Hanja). We employ cross-lingual transfer learning in training word representations by leveraging the fact that Hanja is closely related to Chinese. We evaluate the intrinsic quality of representations learned through our approach using the word analogy and similarity tests. In addition, we demonstrate their effectiveness on several downstream tasks, including a novel Korean news headline generation task.

1 Introduction

There is a strong connection between the Korean and Chinese languages due to cultural and historical reasons (Lee and Ramsey, 2011). Specifically, a set of logograms with very similar forms to the Chinese characters, called **Hanja**[1], served in the past as the only medium for written Korean until **Hangul**, the Korean alphabet, and **Jamo** came into existence in 1443 (Sohn, 2001). Considering this etymological background, a substantial portion of Korean words are classified as Sino-Korean, a set of Korean words that had originated from Chinese and can be expressed in both Hanja and Hangul (Taylor, 1997), with the latter becoming commonplace in modern Korean.

Based on these facts, we assume the introduction of Hanja-level information will aid in grounding better representation for the meaning of Korean words (e.g. see Fig. 1). To validate our hypothesis, we propose a simple yet effective approach to training Korean word representations, named **Hanja-level SISG** (Hanja-level Subword Information Skip-Gram), for capturing the semantics of Hanja and subword structures of Korean and introducing them into the vector space. Note that it is also quite intuitive for native Koreans to resolve the ambiguity of (Sino-)Korean words with the aid of Hanja. We conjecture this heuristic is attributed to the fact that Hanja characters are logograms, each of which contains more lexical meaning when compared against its counterpart, the Hangul character, which is a phonogram. Accordingly, in this work, we focus on exploiting the rich extra information from Hanja as a means of constructing better Korean word embeddings. Furthermore, our approach enables us to empirically investigate the potential of character-level cross-lingual knowledge transfer with a case study of Chinese and Hanja.

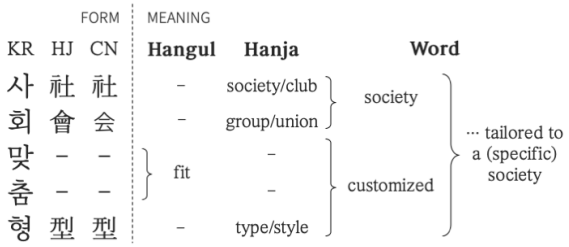

Figure 1: An example of a Korean word showing its form and multi-level meanings. The Sino-Korean word consists of Hangul phonograms (KR) and Hanja logograms (HJ). Although annotation of Hanja is optional, it offers deeper insight into the word meaning due to its association with the Chinese characters (CN).

To sum up, our contributions are threefold:

- We introduce a novel way of training Korean word embeddings with an expanded Korean alphabet vocabulary using Hanja in addition to the existing Korean characters, Hangul.
- We also explore the possibility of character-

*Equal contribution.

[1]Hanja and traditional Chinese characters are very similar but not completely identical. Some differences in strokes and language-exclusive characters exist.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3528–3533,
Hong Kong, China, November 3–7, 2019.

level knowledge transfer between two languages by initializing Hanja embeddings with Chinese character embeddings before training skip-gram.

- We prove the effectiveness of our method in two ways, one of which is intrinsic evaluation (the word analogy and similarity tests), and the other being two downstream tasks including a newly proposed Korean news headline generation task.

2 Model Description

2.1 Skip-Gram (SG)

Given a corpus as a sequence of words $(w_1, w_2, \ldots, w_T)$, the goal of a skip-gram model (Mikolov et al., 2013) is to maximize the log-probabilities of context words given a target word w_i:

$$\sum_{i \in \{1,\ldots,T\}} \sum_{c \in \mathcal{C}(i)} \log p\left(w_c | w_i\right), \tag{1}$$

where $\mathcal{C}$ returns a set of context word indices given a word index. The usual choice for training the parameterized probability function $p\left(w_c|w\right)$ is to reframe the problem as a set of binary classification tasks of the target context word c and other negative samples chosen from the entire dictionary (negative sampling). The objective is thus

$$L\left(s\left(w_i, w_c\right)\right) + \sum_{j \in \mathcal{N}(i,c)} L\left(-s\left(w_i, w_j\right)\right), \tag{2}$$

where $L\left(x\right) = \log\left(1 + e^{-x}\right)$, the binary logistic loss. $\mathcal{N}$ returns a set of negative sample indices given target word and context indices. In the vanilla SG model, the scoring function s is the dot product between the two word vectors: $\mathbf{v}_w$ and $\mathbf{v}_c$.

2.2 Subword Information SG (SISG) and Jamo-level SISG

In Bojanowski et al. (2017), the scoring function incorporates subword information. Specifically, given all possible n-gram character sets $\mathbf{G} = \{g_1, \ldots, g_G\}$ in the corpus and that the n-gram set of a particular word w is $\mathcal{G}_w \subset \mathbf{G}$, the scoring function is thus the sum of dot products between all n-gram vectors $\mathbf{z}$ and the context vector $\mathbf{v}_c$:

$$s\left(w, c\right) = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g \mathbf{v}_c. \tag{3}$$

However, due to the deconstructible structure of Korean characters and the agglutinative nature of the language, the model proposed by Bojanowski et al. (2017) comes short in capturing sub-character and inter-character information specific to Korean understanding. As a remedy, the model proposed by Park et al. (2018) introduces jamo-level n-grams $g^{(j)}$, whose vectors can be associated with the target context as well. Given that a Korean word w can be split into a set of jamo n-grams $\mathcal{G}_w^{(j)} \subset \mathbf{G}^{(j)} = \left\{g_1^{(j)}, \ldots, g_J^{(j)}\right\}$, where $\mathbf{G}^{(j)}$ is the set of all possible jamo n-grams in the corpus, the updated scoring function for jamo-level skip-gram $s^{(j)}$ is thus

$$s^{(j)}\left(w, c\right) = s\left(w, c\right) + \sum_{j \in \mathcal{G}_w^{(j)}} \mathbf{z}_j \mathbf{v}_c. \tag{4}$$

2.3 Hanja-level SISG[2]

Another semantic level in Korean lies in Hanja logograms. Hanja characters can be mapped to Hangul phonograms, hence Hanja annotations are sufficient but not necessary in conveying meanings (Fig. 1). As each Hanja character is associated with semantics, their presence could provide meaningful aid in capturing word-level semantics in the vector space.

We propose incorporating Hanja n-grams in the learning of word representations by allowing the scoring function to associate each Hanja n-gram with the target context word. Concretely, given that a Korean word w contains a set of Hanja sequences $\mathcal{H}_w = (h_{w,1}, \ldots, h_{w,H_w})$ (e.g. for the example in Fig. 1 $h_{w,1}$ = "社會" and h_{w_2} = "型") and that $\mathcal{I}_h$ is the set of Hanja n-grams for a Hanja sequence h (e.g. $\mathcal{I}_{h_{w,2}}$ = { "<boh>型", "型", "型<eoh>"}), *all* Hanja n-grams $\mathcal{G}_w^{(h)}$ for word w are the union of Hanja n-grams of Hanja sequences present in w: $\bigcup_{h \in \mathcal{H}_w} \mathcal{I}_h$. The length of Hanja n-grams in $\mathcal{I}_h$ is a hyperparameter. Then the new score function for Hanja-level SISG is

$$s^{(h)}\left(w, c\right) = s^{(j)}\left(w, c\right) + \sum_{h \in \mathcal{G}_w^{(h)}} \mathbf{z}_h \mathbf{v}_c \tag{5}$$

where $\mathbf{z}_h$ is the n-gram vector for a Hanja sequence h.

[2]The code and models are publicly available online at https://github.com/kaniblu/hanja-sisg

2.4 Character-level Knowledge Transfer

Since Hanja characters essentially share deep roots with Chinese characters, many of them can be mapped by one-to-one correspondence. By converting Korean characters into simplified Chinese characters or vice versa, it is possible to utilize pre-trained Chinese embeddings in the process of training Hanja-level SISG. In our case, we propose leveraging advances in Chinese word embeddings from the Chinese community by initializing Hanja n-grams vectors $\mathbf{z}_h$ with the state-of-the-art Chinese embeddings (Li et al., 2018) and training with the score function in Equation 5.

3 Experiments

3.1 Word Representation Learning

3.1.1 Korean Corpus

For learning word representations, we utilize the corpus prepared by Park et al. (2018) with small modifications. We perform additional data cleansing (e.g. removing non-Korean sentences in the corpus and unifying number tags) to obtain a cleaner version of the corpus. All of our comparative studies are based on this corpus. For training Hanja-level SISG, we use Hanjaro[3] tagger to automatically annotate all of our datasets with Hanja. It is the state-of-the-art Hanja tagger available to the public to the best of our knowledge.

3.1.2 Models

We compare our approach with three other baselines. Skip-Gram (`SG`) model (Mikolov et al., 2013) does not incorporate any n-gram information in the training loss. Subword Information Skip-Gram (`SISG(c)`) (Bojanowski et al., 2017) uses character-level n-gram information to enrich word representations. For the Korean language, character-level n-grams correspond to syllable n-grams. Jamo-level SISG (`SISG(cj)`) (Park et al., 2018) uses jamo-level n-grams in addition to character-level n-grams. Our ablation studies confirm that the hyperparameter settings (character n-grams ranging from 1 to 6 and jamo n-grams ranging from 3 to 5) proposed by the authors are indeed optimal, hence our subsequent studies all employ the same settings. Our approach is denoted by `SISG(cjh)`). As for the specific hyperparameter settings of our approach, `SISG(cjh3)` denotes Hanja n-grams ranging from 1 to 3 while `SISG(cjh4)` denotes Hanja n-grams ranging from 1 to 4. `SISG(cjhr)` is a special version of our approach, where Hanja n-grams are randomly initialized as opposed to being pre-trained (Section 2.4). The details of our ablation studies and the model implementation are included in the supplementary material.

[3] http://hanjaro.juntong.or.kr/

Method	Analogy			Similarity	
	Sem.	Syn.	All	Pr.	Sp.
`SISG(cj)`†	*0.47*	*0.38*	*0.43*	-	*0.67*
`SG`	0.42	0.49	0.45	0.60	0.62
`SISG(c)`	0.45	0.59	0.52	0.62	0.61
`SISG(cj)`	0.39	0.48	0.44	**0.66**	0.67
`SISG(cj)`‡	0.41	0.48	0.45	0.65	**0.67**
`SISG(cjh3)`	**0.34**	**0.45**	**0.39**	0.63	0.63
`SISG(cjh4)`	0.34	0.45	0.40	0.62	0.61
`SISG(cjhr)`	0.35	0.46	0.40	0.65	0.64

† reported by Park et al. (2018)
‡ pre-trained embeddings provided by the authors of Park et al. (2018) run with our evaluation script

Table 1: Word analogy and similarity evaluation results. For word analogy test, the table shows cosine distances averaged over semantic and syntactic word pairs (lower is better). For word similarity test, we report on Pearson and Spearman correlations between predictions and ground truth scores. We observe that our approach (SISG(cjh)) shows significant improvements in the analogy test but some deterioration in the similarity test. Note that the evaluation results reported by the original authors (†) are largely different from our results. This might be due to differences in implementation details, hence we report and compare with the results of the authors' embeddings run on our test script (‡).

3.2 Intrinsic Evaluation

3.2.1 Word Analogy Test

Following the experimental protocol of Park et al. (2018), given an analogy pair $a : b \leftrightarrow c : d$, we compute the cosine distance between the word vectors $\mathbf{v}_a + \mathbf{v}_b - \mathbf{v}_c$ and $\mathbf{v}_d$ using the following equation: $1 - \cos(x, y)$, where cos is the cosine similarity. The Korean word analogy dataset (Park et al., 2018) consists of 10,000 quadruples, designed with a set of semantic and syntactic features specific to the Korean. We report our findings in Table 1. We observe that our approach improves significantly compared to the previous state-of-the-art models in both semantic and syntactic word analogy detection.

3.2.2 Word Similarity Test

The word similarity test aims to evaluate the correlation between word vector distances and

human-annotated scores. We use the Korean WS353 dataset provided by Park et al. (2018). Our results are shown in Table 1. Although our approaches do not outperform the jamo-level baseline (SISG(cj)), we observe that Hanja-level information provides some advantage compared to just using pure character-level n-grams (SISG(c)). Also note that WS353 is a relatively small dataset, which could be prone to statistical bias.

3.2.3 Analysis

How much of the improvement can be explained by transfer learning from Chinese? Word analogy test results show that word representations trained with pre-trained Chinese n-grams perform better than those trained without (SISG(cjhr)), supporting our claim that our approach is able to transfer relevant knowledge from the Chinese language for detecting analogical relationships. However, for the word similarity test, word vectors trained without Chinese embeddings perform better, suggesting that there are some trade-offs.

3.3 Downstream Tasks

In this section, we try to demonstrate the effectiveness of our approach to downstream tasks with two specific cases.

3.3.1 Korean News Headline Generation

To show that our approach helps supervised models in gaining deeper understanding of Korean texts, we devise a new task using news articles, which requires an understanding of the semantics of Korean texts. We collected Korean news articles published from 2017 January to February. The dataset covers balanced categories (e.g. politics, sports, world, etc.) and contains the headline title as the annotation, akin to the CNN/Daily Mail Dataset (Hermann et al., 2015). The dataset is relatively large - containing 840,205 news article and title pairs

We use the encoder-decoder architecture supported by OpenNMT (Klein et al., 2017) framework for the task, where the encoder is a bidirectional LSTM cell and the decoder is an LSTM cell (the hidden size is 512 for both) with a bridging feed-forward layer between the two RNNs. We employ soft attention (Bahdanau et al., 2014) to generate news headlines given the first three sentences of an article body. We tokenize each headline using a Korean morpheme tokenizer[4] such that decoder tokens are relatively dense. The copy mechanism (Gu et al., 2016), a popular technique of sequence-to-sequence models for translation and summarization, cannot be applied directly to our model, as the encoder tokens (space tokenized) and the decoder tokens (morpheme tokenized) do not share the same token space. Word vectors obtained from both baselines and our approach are used to initialize the encoder embeddings before the training begins. The dataset is split into training, validation, and test sets (8:1:1). We report performances on the test set after validating our training models on the validation set. We evaluate the titles generated by our models with the BLEU (Papineni et al., 2002)[5]. We also report per-word perplexity obtained from the models.

Embeddings	BLEU				PPL
	1	2	3	4	
None	26.02	7.76	3.08	1.38	5.335
SG	30.33	10.20	4.29	1.98	4.122
SISG(c)	31.34	10.96	4.69	2.19	3.942
SISG(cj)	31.78	11.17	4.80	2.25	3.938
SISG(cj)†	31.77	11.16	4.81	2.27	3.940
SISG(cjh3)	**32.03**	11.25	4.83	2.27	3.941
SISG(cjh4)	32.02	**11.34**	**4.92**	**2.30**	**3.909**

† pre-trained embeddings by Park et al. (2018)

Table 2: Korean News Headline Generation results.

As shown in Table 2, models trained with our embeddings perform better and have better language modeling capability than those trained with the previous state-of-the-art embeddings. This improvement can be partially explained by the fact that formal Korean texts, such as news articles, are more likely to contain Sino-Korean words, allowing models with awareness of Chinese to gain an edge in Korean understanding.

3.3.2 Sentiment Analysis

To estimate the performance of our approach on a different task then the previous experiment, we conduct an experiment on Naver Sentiment Movie Corpus (NSMC)[6]. This dataset consists of 200K movie reviews, each of which is labeled with its sentiment, i.e. positive or negative. We split the corpus into training (100K), validation (50K), and test sets (50K). It is worth noting that this

[4] https://github.com/shin285/KOMORAN

[5] Although ROUGE is a more widely adopted measure in English literature, no equivalent exists for Korean.

[6] https://github.com/e9t/nsmc

Embeddings	Model: LSTM			
	Acc.	P	R	F1
SISG(c)	77.43	75.89	80.41	78.08
SISG(cj)	**83.16**	82.36	**84.66**	**83.50**
SISG(cjh3)	81.61	81.23	82.28	81.75
SISG(cjh4)	82.25	**82.57**	81.77	82.17

Table 3: Naver Sentiment Movie Corpus results. We report accuracy (Acc.), precision (P), recall (R), and F1-score (F1) following Park et al. (2018). Each reported result is the average of several runs initialized with different random seeds.

case study is designed to figure out to what extent our embeddings are generally applicable for even cases, where most of the sentences in the dataset consist of spoken words rather than written words and often do not contain Sino-Korean words.

We utilize a basic LSTM (Hochreiter and Schmidhuber, 1997) module as a sentence encoder to exclude possible exceptional gains from sophisticated encoders, concentrating on the usefulness of input word embeddings. Likewise the previous experiment, word vectors obtained from both baselines and our approach are used as input for the encoder. We regard the last hidden state of the LSTM as the sentence representation, which is consumed by a feed-forward network followed by a softmax classifier. The dimension of the LSTM cells is fixed as 300. Each result from the baselines and our models is the average of 3 independent runs initialized with different random seeds, and the outcome of each run is chosen by the performance (F1-score) on the validation set.

From Table 3, we confirm our approach is comparable to the best previous one and general enough to be employed in most of the downstream tasks, even though the performance of our approach is somewhat unsatisfactory. While not strictly proven, we conjecture one possible reason for the unsatisfying performance is the low reliability of the Hanja tagger we leveraged, which is not guaranteed to work well with spoken language. Specifically, we have manually inspected the tagging results and observed that there are some errors which can lead to performance degradation. This observation points out a limitation of our current approach, that is, the dependence on external taggers, encouraging us to develop an integrated approach, leaving as future work not resorting to the taggers.

4 Related Work

Although the word is usually regarded as the smallest and basic unit for most NLP pipelines, there is a recent trend of utilizing subword information for enriching word representations (Sennrich et al. 2016; Bojanowski et al. 2017, to name a few), or considering subwords themselves as input directly for NLP models (Zhang et al., 2015; Kim et al., 2016; Peters et al., 2018; Devlin et al., 2018).

As Korean is agglutinative (Song, 2006), the current literature in Korean word representations mainly focus on subword structures such as morphemes (Edmiston and Stratos, 2018), syllables (Choi et al., 2017) and Jamo (Choi et al., 2016; Stratos, 2017; Park et al., 2018). We here move forward one step further by incorporating Hanja information explicitly together with the aforementioned subword information.

On the other hand, cross-lingual representations is a trending topic in literature (Lample et al. 2018; Conneau et al. 2018; Lample and Conneau 2019, to name a few). Nevertheless, to the best of our knowledge, our work is the first to introduce character-level cross-lingual transfer learning based on etymological grounds. Furthermore, the novelty of our work lies of the fact that we use separated vocabulary instead of shared vocabulary such as Byte Pair Encoding (BPE) (Sennrich et al., 2016).

5 Conclusion

We have presented a method of training Korean word representations with Hanja. In our extensive experiments, we have demonstrated that our approach is effective in infusing more semantics into Korean word embeddings. One potential issue with our method, as already mentioned, is that it relies on external Hanja annotation, even though this can be mitigated to some extent by off-the-shelf taggers, as in this work. Thus, it would be an attractive direction to take as future work developing an end-to-end system.

Acknowledgments

We thank Daniel Edmiston and anonymous reviewers for their helpful feedback. This work was supported by BK21 Plus for Pioneers in Innovative Computing (Dept. of Computer Science and Engineering, SNU) funded by the National Research Foundation of Korea (NRF) (21A20151113068).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Jihun Choi, Jonghem Youn, and Sang-goo Lee. 2016. A grapheme-level approach for constructing a korean morphological analyzer without linguistic knowledge. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3872–3879. IEEE.

Sanghyuk Choi, Taeuk Kim, Jinseok Seol, and Sang-goo Lee. 2017. A syllable-based technique for word embeddings of korean words. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 36–40.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Daniel Edmiston and Karl Stratos. 2018. Compositional morpheme embeddings with affixes as functions and stems as arguments. In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 1–5.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *ICLR*.

Ki-Moon Lee and S Robert Ramsey. 2011. *A history of the Korean language*. Cambridge University Press.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the ACL)*, volume 2, pages 138–143.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on ACL*, pages 311–318. Association for Computational Linguistics.

Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho, and Alice Oh. 2018. Subword-level word vector representations for korean. In *Proceedings of the 56th Annual Meeting of the ACL*, volume 1, pages 2429–2438.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the NAACL: HLT, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of ACL*, volume 1, pages 1715–1725.

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

Jae Jung Song. 2006. *The Korean language: Structure, use and context*. Routledge.

Karl Stratos. 2017. A sub-character architecture for korean language processing. In *Proceedings of the 2017 Conference on EMNLP*, pages 721–726.

Insup Taylor. 1997. Psycholinguistic reasons for keeping chinese characters in korean and japanese. *Cognitive processing of Chinese and related Asian languages*, 319.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.