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

A Supplemental Material

A.1 Implementation Details for SISG Model

We trained our models over 5 epochs with the following parameters: 300 word vector dimensions, 20m bucket size to cover all n-grams including Hanja, 0.0001 sampling threshold, 0.05 learning rate, 5 negative sampling size, 3-5 jamo n-gram sizes, and 5 context window size.

A.2 Ablation Results

In this supplementary material, we include results for different experimental settings such as the type of training corpus and the character n-gram sizes(Table 1).

A.3 Full Results for Word Analogy

Here we present the full results for the word analogy test, including all sections.

References

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.

Method	Dataset	Char. N-Grams	Analogy			Similarity	
			Sem.	Syn.	All	Pearson	Spearman
SISG(cj)†	Original	1-6	*0.478*	*0.385*	*0.432*	-	*0.677*
SG	Improved	-	0.423	0.495	0.459	0.608	0.627
SISG(c)	Improved	1-6	0.450	0.591	0.520	0.620	0.612
SISG(cj)	Improved	1-6	0.398	0.484	0.441	**0.665**	0.671
SISG(cj)	Improved	1-4	0.400	0.485	0.442	0.640	0.637
SISG(cj)	Original	1-6	0.399	0.488	0.444	0.659	0.661
SISG(cj)	Original	1-4	0.398	0.483	0.441	0.652	0.653
SISG(cj)‡	Original	1-6	0.414	0.487	0.451	0.654	**0.674**
SISG(cjh3)	Improved	1-6	0.340	**0.450**	**0.395**	0.634	0.633
SISG(cjh3)	Improved	1-4	**0.339**	0.453	0.396	0.612	0.605
SISG(cjh4)	Improved	1-6	0.349	0.456	0.402	0.624	0.617
SISG(cjh4)	Improved	1-4	0.349	0.456	0.402	0.640	0.638
SISG(cjhr)	Improved	1-6	0.355	0.462	0.409	0.650	0.647

† 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: Full results of our ablation studies. By conducting experiments on varying character n-gram lengths, we determine that character n-grams ranging from 1-6 yield better results for our model. The dataset columns shows two different types of datasets: `original` and `improved`. The `original` dataset is the corpus originally provided by the authors of (Park et al., 2018), and the `improved` dataset is the one that has been further data-cleansed from the original corpus. The results show that word vectors trained on the improved corpus achieve marginal but still meaningful improvement in quality.

Method	Semantic					Syntactic					All
	City	Sex	Name	Lang	Misc	Case	Tense	Voice	Form	Honor	
SISG(cj)†	*0.425*	*0.498*	*0.561*	*0.354*	*0.554*	*0.210*	*0.414*	*0.426*	*0.507*	*0.367*	*0.432*
SG	0.471	0.478	0.413	0.338	0.419	0.540	0.482	0.517	0.486	0.449	0.459
SISG(c)	0.492	0.512	0.436	0.401	0.408	0.645	0.573	0.597	0.554	0.584	0.520
SISG(cj)	0.430	0.466	0.384	0.331	0.377	0.591	0.473	0.485	0.489	0.384	0.441
SISG(cj)‡	0.449	0.468	0.400	0.341	0.412	0.576	0.479	0.485	0.484	0.413	0.451
SISG(cjh3)	**0.363**	**0.424**	**0.326**	**0.258**	**0.328**	**0.558**	**0.439**	**0.461**	**0.444**	**0.348**	**0.395**
SISG(cjh4)	0.377	0.423	0.333	0.270	0.340	0.563	0.448	0.463	0.457	0.351	0.402
SISG(cjhr)	0.389	0.432	0.343	0.274	0.338	0.569	0.449	0.468	0.466	0.355	0.409

† reported by Park et al. (2018)

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

Table 2: Full results on the word analogy test. Note that the evaluation results reported by the original authors (†) is largely different from our results. This might be due to differences in implementation details, hence we report and compare only with the results of the authors' embeddings run on our test script (‡).