

A subcharacter factorization of Korean character embedding tables leads to a 99% reduction in embedding parameters, no loss of quality, and no sequence length increase!

Problem	Korean Orthography	One-Hot Modeling
<p>Korean is a highly agglutinative language, making character-level modeling an attractive choice. A major problem is that Korean’s base vocabulary contains 11,172 unique <i>syllables</i>, so, unlike character-level models for other languages, many parameters need to be spent on the syllable embedding table.</p>	<p><i>Syllables (characters)</i> are made up of exactly three <i>jamo (subcharacters)</i>:</p> <ul style="list-style-type: none"><li>Initial Consonant (19)</li><li>Vowel (21)</li><li>(Optional) Final Consonant (28)</li><li><math>19 \times 21 \times 28 = 11172</math> unique syllables</li></ul> <div><div><div><math>\mathcal{V}_i</math></div><div>ㄱ ㅋ ㆁ ㄷ ㅌ ㄹ ㅂ ㅍ ㅅ ㅈ ㅊ ㅍ ㅎ</div><div>ㅃ ㅇ ㅆ ㅉ ㅊ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅟ ㅠ</div></div><div><div><math>\mathcal{V}_v</math></div><div>ㅣ ㅌ ㅎ ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅟ ㅠ</div><div>ㅊ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅟ ㅠ</div></div><div><div><math>\mathcal{V}_f</math></div><div>ㄱ ㅋ ㆁ ㄷ ㅌ ㄹ ㅂ ㅍ ㅅ ㅈ ㅊ ㅍ ㅎ</div><div>ㅃ ㅇ ㅆ ㅉ ㅊ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅟ ㅠ</div><div>ㅅ ㅈ ㅊ ㅍ ㅎ</div></div></div> <div><div>한 = (ㅇ, ㅣ, ㄴ)</div><div>무 = (ㅁ, ㅜ, ㅇ)</div><div>각 = (ㄱ, ㅏ, ㄱ)</div><div>펼 = (ㅍ, ㅔ, ㅇ)</div></div>	<p>Modeling syllables or jamo can be done with regular one-hot encoding, but both have downsides:</p> <ul style="list-style-type: none"><li>Naive Syllable<ul style="list-style-type: none"><li>Large vocabulary (11,172)</li><li>Only about 2.5k syllables are commonly used</li></ul></li><li>Jamo<ul style="list-style-type: none"><li>Only requires 68 embedding vectors</li><li>3x longer context length</li></ul></li></ul>
<p>We decompose syllables into <i>jamo</i> to model at the syllable level, but using jamo embeddings.</p>		

Three-Hot Modeling	Modeling Schemes Comparison	Three-Hot Embedding Layer
<p>Three-Hot models represent full syllables as triplets of jamo. This has many benefits:</p> <ul style="list-style-type: none"> <li>• Small vocabulary/embedding table           <ul style="list-style-type: none"> <li>– One embedding vector per jamo (same as jamo modeling)</li> </ul> </li> <li>• Same sequence length as syllable           <ul style="list-style-type: none"> <li>– Jamo-level models triple the sequence length</li> </ul> </li> </ul>	<p>Syllable: <math>\text{한,국} \rightarrow \text{어}</math>            Jamo: <math>\text{ㅎ, ㅏ, ㄴ, ㄱ, ㅊ, ㄱ} \rightarrow \text{ㅇ} \rightarrow \text{ㅎ}</math>            Three-Hot: <math>(\text{ㅎ}, \text{ㅏ}, \text{ㄴ}), (\text{ㄱ}, \text{ㅊ}, \text{ㄱ}) \rightarrow (\text{ㅇ}, \text{ㅎ}, \text{ㅇ})</math></p> <p>Jamo has 3x as many tokens to attend to, which scales poorly with attention.</p>	<p>Three-Hot models form syllable embeddings from jamo embeddings:</p> $emb_s = emb_i + emb_v + emb_f$ <p>Every syllable embedding can be formed from just the component jamo embeddings.</p>

### Independent Three-Hot (Song et al., 18)

A prior approach to Three-Hot modeling approximates syllables as three independent probability distributions:

$$\mathcal{P}(s \mid h) \approx \mathcal{P}(i \mid h) \times \mathcal{P}(v \mid h) \times \mathcal{P}(f \mid h)$$

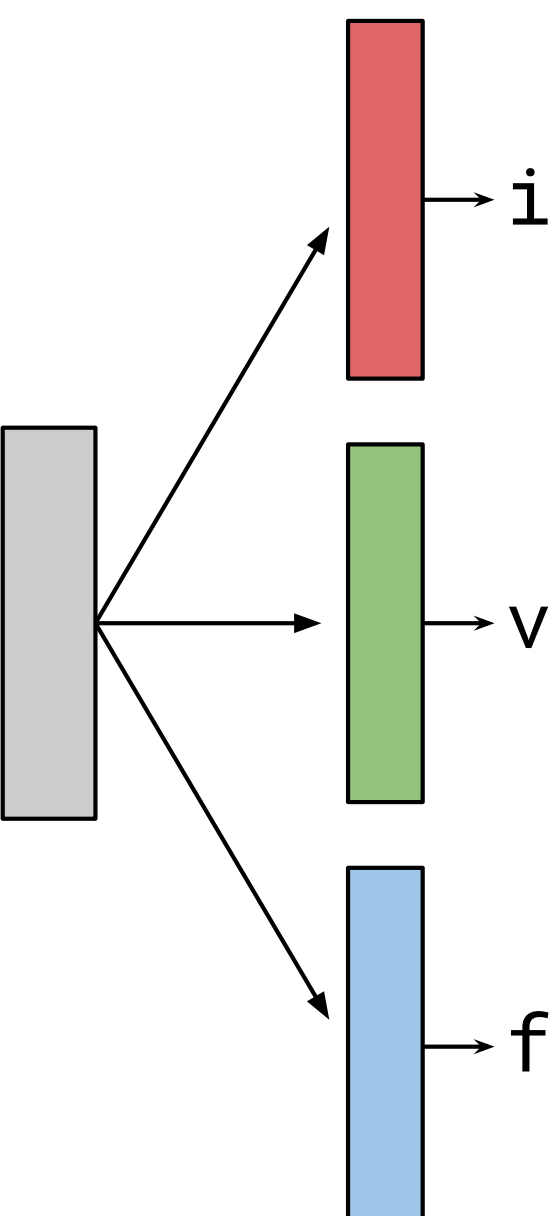
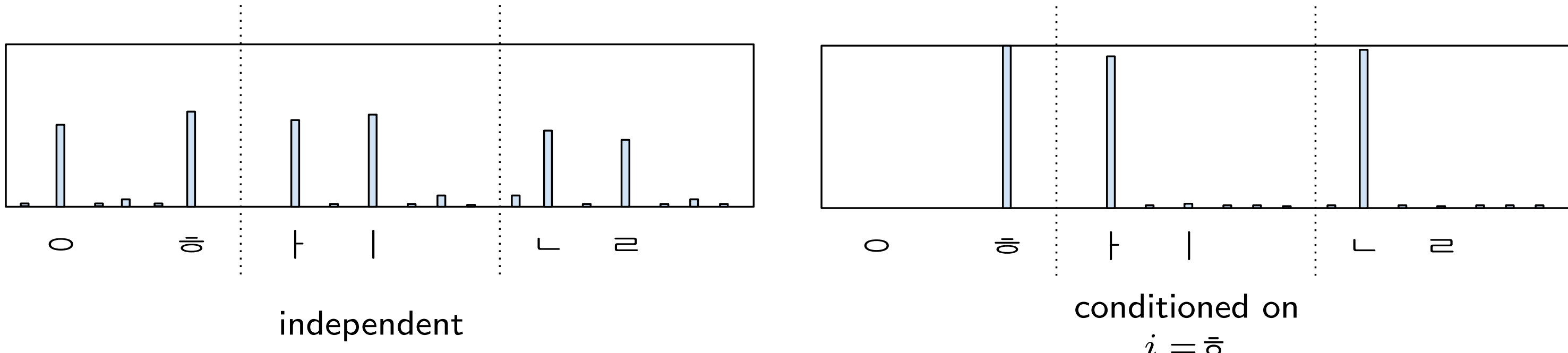
The Independent Three-Hot model can get confused when there are multiple high-probability syllable outputs. This makes beam search very difficult.

### Conditional Three-Hot (ours)

In reality, the jamo for a syllable are not independent. Instead, we should properly model the joint probability distribution  $\mathcal{P}(s \mid h) = \mathcal{P}(i, v, f \mid h)$ :

- Predict all three jamo together, but in a three-step process
- Each jamo is conditioned on previously generated jamo
- We decode it similarly to a 3-stage, unrolled RNN

$$\mathcal{P}(s \mid h) = \mathcal{P}(i \mid h) \times \mathcal{P}(v \mid i, h) \times \mathcal{P}(f \mid i, v, h)$$

### Conditional Decoding

The diagram illustrates a Conditional Decoding RNN. An encoder (gray box) processes input embeddings (red, green, blue) and a decoder (yellow box) generates output embeddings (red, green, blue) for labels  $i$ ,  $v$ , and  $f$ . The decoder is connected to an RNN block.

### Parameter Reduction

Total Parameters =  $\underbrace{d|\mathcal{V}_{i,v,f}|}_{\text{Encoder Embedding}} + \underbrace{d|\mathcal{V}_{i,v}|}_{\text{Embedding}} + \underbrace{d|\mathcal{V}_{i,v,f}|}_{\text{Output}} + \underbrace{2d^2}_{\text{RNN}}$

To further reduce parameters, we:

- Share embedding tables between embedding and decoding layers
  - Reduces embedding parameters from  $3d|\mathcal{V}_{i,v}| + 2d|\mathcal{V}_f|$  to  $d|\mathcal{V}_{i,v,f}|$
- Replace dense internal RNN matrices with diagonal matrices
  - Reduces RNN parameters from  $2d^2$  to  $2d$

Final Parameter Count =  $d|\mathcal{V}_{i,v,f}| + 2d$

Results				Summary		
	Embedding Params	Decoding Params	Total	BPJ	BLEU	chrF
Syllable (unshared)	5.7M	5.7M	11.4M	0.339	14.1	38.1
Syllable (shared)	5.7M	-	5.7M	0.342	14.0	38.1
Jamo (unshared)	35k	35k	70k	0.355	13.7	37.8
Jamo (shared)	35k	-	35k	0.356	14.1	38.0
Three-Hot (Ind., unshared)	35k	35k	70k	0.555	7.9	28.9
Three-Hot (Ind., shared)	35k	-	35k	0.556	8.3	29.4
Three-Hot (IVF, unshared)	35k	579k	614k	0.287	14.3	38.1
Three-Hot (IVF, shared)	35k	524k	559k	0.292	14.1	38.1
Three-Hot (IVF, diag., unshared)	35k	71k	106k	0.293	14.2	38.2
Three-Hot (IVF, diag., shared)	35k	1k	36k	0.306	14.0	38.0
Three-Hot (FIV, unshared)	35k	579k	614k	0.289	14.0	37.9
Three-Hot (FIV, shared)	35k	524k	559k	0.294	14.1	37.8
Three-Hot (FIV, diag., unshared)	35k	71k	106k	0.294	14.0	37.9
Three-Hot (FIV, diag., shared)	35k	1k	36k	0.304	13.8	37.8

These research results were obtained partially from the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), Japan.