Grapheme-level Awareness in Word Embeddings for Morphologically Rich Languages

Suzi Park, Hyopil Shin
Department of Linguistics, Seoul National University
Gwanakro 1, Gwanak-gu, Seoul, South Korea

{mam3b,hpshin}@snu.ac.kr

Abstract

Learning word vectors from character level is an effective method to improve word embeddings for morphologically rich languages. However, most of these techniques have been applied to languages that are inflectional and written in Roman alphabets. In this paper, we investigate languages that are agglutinative and represented by non-alphabetic scripts, choosing Korean as a case study. We present a grapheme-level coding procedure for neural word embedding that utilizes word-internal features that are composed of syllable characters (Character CNN). Observing that our grapheme-level model is more capable of representing functional and semantic similarities, grouping allomorphs, and disambiguating homographs than syllable-level and word-level models, we recognize the importance of knowledge on the morphological typology and diversity of writing systems.

Keywords: Character-level CNN, Word Embedding, Grapheme-level embedding

1. Introduction

Semantic word representations are essential to natural language processing. Most embedding models have learned the meaning of words from their contexts. These techniques work for English, in which word forms rarely change according to their contexts. Though, in morphologically rich languages, inflectional forms of the same base are considered and learned as distinct words. Even if the base is frequent in a large corpus, each word forms rarely occurs, which leads to data sparsity. This weakness has been overcome to a certain extent by composing representation of subword units such as morphemes or characters. Character-based models can be generalized in a sense that they do not require a pre-trained morphological analyzer, and they enable to calculate vector representations even for out-of-vocabulary words.

However, vector representations of characters present the same problem as those of words, due to a trade-off between vocabulary size and token frequency. Alphabet systems, like Roman alphabets, contain a comparatively small number of symbols that occur frequently enough in a corpus. On the contrary, syllabaries have a more extensive inventory of syllable characters each of whose frequency is lower, making an analogy with morphologically rich languages.

Our objective in this paper is to address the issue of data sparsity, which we argue is caused by not only the morphological system of a language and but also its writing system. We begin by remarking on the morphological typology and diversity of writing systems. We then report two phenomena of languages that are agglutinative, represented by non-alphabetic scripts, or both, Korean as a case study. We find that they cannot be captured by the default method to code Korean text on the syllable-character level. From this, we propose a character-internal or grapheme-level coding procedure for neural word embedding that utilizes word-internal features that are composed of characters. We end by observing that our grapheme-level model is capable of representing functional and semantic similarities, grouping allomorphs, and disambiguating homographs. We conclude that it can improve existing character-level approaches to Korean NLP tasks.

2. Challenges in character-based approaches to word embeddings for morphologically rich languages

2.1. Scalability and multidimensionality of richness

2.1.1. Richness in morphology

With the distribution hypothesis(Harris, 1954; Sahlgren, 2008) as their basis, most approaches to word vector representation use co-occurrence statistics. In order for frequencies of word co-occurrence to be counted efficiently and accurately, each word should have a form that is invariant with respect to its location in a sentence. This is not likely to be the case in morphologically rich languages, however. The resulting data sparsity of morphologically rich languages has been long discussed and examined in relation to a variety of NLP tasks. As far as is known, languages that have been categorized as morphologically rich have been non-English European languages such as German and Czech, with languages such as Turkish and Arabic occasionally being included as well. Character-based approaches have been applied to improve the performance of word embeddings(Ling et al., 2015; Kim et al., 2016), and are acknowledged as non-language-specific and not requiring linguistic knowledge, however these approaches were used comparatively less frequently with other morphologically rich languages, such as Japanese and Korean. In order to shed some light on the matter, recall that it is widely held among linguists that languages are classified into more than two morphological types. Comrie (1989), for example, distinguishes between fusional and agglutinative languages:

- *Isolating or analytic languages* have no or little inflectional morphology. Since a lemma does not have various word forms and word order plays a crucial role to determine grammatical features, N-gram models work well.

- *Fusional or inflected languages* have grammatical affixes, each of which encodes multiple grammatical features simultaneously. Typically, a noun declension includes gender, number, and case, and a verb conjugation includes tense, aspect, and voice. A Lemma has multiple word forms, but their number is restricted by the number of declension/inflection morpheme types. In Latin, a verb can have a maximum of 144 inflected forms(Schinke et al., 1996). This number does not seem so small, but would be reduced in its descendants, the Romance languages including French, Spanish, Portuguese, and Italian.

- *Agglutinating languages* also have a variety of grammatical morphemes, but each of them has no more than one feature, and more than two morphemes can be concatenated to a stem.[1] Therefore, the set of possible word forms of a stem is larger than in inflected languages. For instance, Karlsson (1986) reported that a Finnish noun can be inflected into 2,200 forms, and Argüelles and Kim (2004) introduced 623 (non-exhausted) forms of each Korean verb conjugation.

Though both inflectional and agglutinating languages are morphologically richer than an isolating language, the complexity of the latter should not be overlooked.

The examples in Table 1 show phrases consisting of more than four words in English that are translated into just one word in Korean, which consists of five morphemes: a noun stem 고양이 (*ko.yang.i*)[2] 'cat' has a masculinity prefix 수 (*swu*), a plural suffix 들 (*tul*) '-s', a dative marker 에게 (*ey.key*) 'to', and finally an emphasis marker 조차 (*co.cha*) 'even'. Also note that agglutination can occur beyond the noun declension or verb conjugation that can be found in fusional languages such as Portuguese. In English, the case and gender of the noun *cat* is realized lexically(*to* and *male*), and its number, morphologically(*-s*). Even if these three features can be encoded in a pure inflectional language such as Latin, the adverb *even*, which is not a part of declension, cannot be morphologically attached to the noun head. However, agglutination can encode features that would not have been realized morphologically but have become lexical items like adverbs, conjunctions, and light verbs.

2.1.2. Richness in writing system

By comparing an agglutinating language with a fusional one, we saw that the morphological richness of a language is scalable rather than binary. Now we turn our attention to another kind of richness that can be observed in natural language processing.

Analytic	EN	5	[Even] [to] [the] [male] [cats]
Fusional	PT	3	[até] [aos] [gatos]
agglutinating	KO	1	[수코양이들에게조차] (*swu.kho.yang.i.tul.ey.key.co.cha*)
Analytic	EN	4	[if] ... [had] [passed] [away]
Fusional	PT	3	[se] [tivesse] [falecido]
agglutinating	KO	1	[돌아가셨었다면] (*tol.a.ka.syess.ess.ta.myen*)

** EN: English, PT: Portuguese, KO: Korean

Table 1: Realization of grammatical features in three morphological types of languages

Since we are dealing with writing systems that are in the form of text, it is reasonable to take account of writing systems. As a language can be morphologically rich, a writing system, or script, can be graphonomically rich. These two properties do not depend on one another. The Roman alphabet, for example, is used in a variety of languages that span across a wide range of morphological richness: isolated languages, such as English and Vietnamese, inflectional languages, such as French and Polish, and agglutinative lanuguages, such as Basque and Hungarian.

The amount of information that a character can encode varies according to the writing system to which it belongs, as shown in Table 2. For example, a non-alphabetic character may not be the smallest unit of a writing system. From this, we can dempose a character into a greater number of sub-components just as words are processed as sequences of characters in various applications of NLP.

Writing system	What is assigned to each symbol
Logographic script	Morpheme
Logosyllabic script	Syllabic & semantic values
Syllabary	Syllable
Abjad	Consonant
Alphabet	Consonant or vowel

Table 2: Categorization of scripts(Daniels 1992)

2.2. Case study on a language both morphologically and graphonomically rich

We explore the possibility that the richness of languages used in language agnostic character-based approaches is restricted. Moving beyond the more commonly examined languages, Korean is a good case study as it is both a morphologically rich agglutinative language and graphonomically rich. The Korean script, Hangul, is alphabetic as each grapheme represents a phoneme. However, a grapheme occurs only as a component of a syllable. For example, the character 몸 (*mom*) is a syllable whose structure consists of an initial consonant ㅁ (*m*), a medial vowel ㅗ (*o*), and a final consonant ㅁ (*m*), each of which cannot be written separately in text.

2.2.1. Mismatch between morpheme boundary and character boundary

Korean poses additional challenges for word embeddings. First, some morpheme boundaries do not agree with character boundaries. One class of grammatical morphemes of high frequency includes morphemes such as ㅆ (*ss*) '-ed', ㅁ (*m*) '-ing', ㄴ (*n*) '(that) has/have ...ed', and ㄹ (*l*) '(that) will ...' that are all final consonants that are realized with

[1] Non-agglutinating languages also have multiple affixation (e.g. *un-accept-abil-ity* and *de-caffein-ate*), but it is restricted in derivation rather than inflection, so they are less productive than agglutinating languages.

[2] We use the Yale Romanization for Korean. It is a transliteration rather than a phonemic or phonological transcription, and is better suited to representing forms written in the Korean script.

different stems consisting of different syllabic characters. All words in Table 3 end with a gerund marker ㅁ (*m*), but they have no other characters in common.

잠	(_ca**m**_)	'sleeping'	세움	(_sey.wu**m**_)	'stopping'
뜀	(_ttwi**m**_)	'jumping'	부름	(_pwu.lu**m**_)	'calling'
달림	(_tal.li**m**_)	'running'	마침	(_ma.chi**m**_)	'finishing'

Table 3: Gerund formation by ㅁ (*m*) '-ing' in Korean

This phenomenon can be found also in Japanese. Table 4 shows that there are no characters in common between よむ (*yo.mu*) 'read' and いく (*i.ku*) 'go', even though they belong to the same grammatical category.

Pres. ind.		Volitional		Imperative		
いく	(_i.k**u**_)	いこう	(_i.k**o.u**_)	いけ	(_i.k**e**_)	'go'
とぶ	(_to.b**u**_)	とぼう	(_to.b**o.u**_)	とべ	(_to.b**e**_)	'fly'
よむ	(_yo.m**u**_)	よもう	(_yo.m**o.u**_)	よめ	(_yo.m**e**_)	'read'

Table 4: Examples of Japanese verb conjugations

2.2.2. Variety of allomorphs

The second challenge posed by the Korean is a result of its large variety of allomorphs. At the character level, two allomorphs of the same morpheme do not overlap each other. For words of Chinese origin [3] many characters have multiple readings, which are differentiated only by an initial, as exemplified in Table 5.

率	(_**l**ywul_)	확률 確率	(_hwak.**lywul**_)	'probability'
	(_ywul_)	비율 比率	(_pi.**ywul**_)	'ratio'
樂	(_**l**ak_)	쾌락 快樂	(_khoey.**lak**_)	'pleasure'
	(_nak_)	낙관 樂觀	(_**nak**.kwan_)	'optimism'

Table 5: Multiple readings of Chinese characters in Korean

Most frequent noun case markers and verb tense markers are also allophonic. For instance, two accusative case markers 을 (*ul*) and 를 (*lul*) are distincted by an initial ㅇ (∅) or ㄹ (*l*), and past tense markers 았 (*ass*) and 었 (*ess*), by a medial ㅏ (*a*) or ㅓ (*e*).

3. Word-internal and Character-internal Models

Building upon the case study of the previous section, we propose a character-internal coding procedure for word embeddings that utilizes word-internal features that are composed of characters. We use the character-based CNN-LSTM language model proposed by Kim et al. (2016)[4], and apply several methods to decompose characters into more finely grained units. In Korean text, the most superficial characters represent syllables; they consist of graphemes. We have observed that the contrasting unit is not a grapheme alone, but a grapheme along with a fixed position. For example, verbs of the same inflectional type share a final consonant, and allomorphs sharing a Chinese origin are distinguished by their initital consonants. This leaves us with two options when considering how to approach the decomposition of syllables.

Graphemes with Initial-Medial-Final distinction The syllable 몸 (*mom*) has three grapheme types ᄆ (Unicode: U+1106)(*m*), ᅩ (U+1169)(*o*), and ᆷ (U+11B7)(*m*).

Graphemes with Consonant-Vowel distinction 몸 (*mom*) consists of two ㅁ (U+3141)(*m*)'s and one ㅗ (U+3157)(*o*).

Syllables as characters 몸 (*mom*), whose unicode designation is U+BAB8, is regarded as a single character with no internal structure.

4. Language Modelling

4.1. Experiments

4.1.1. Datasets

To learn our language model, we utilize the Korean Sejong Corpus(The National Institute of the Korean Language, 2010). Though the corpus contains morphologically and syntactically parsed texts, but we use just raw texts to make our models work without morphological knowledge. During preprocessing, we categorize words with a frequency of less than 5 instance as `<unk>`, and replace characters from the Chinese script and Roman alphabet with H and R, respectively. Digits are replaced with N. Our training set consists of 1,562,497 word tokens, 102,532 word types, 1,766 syllable types, and 164 grapheme types (with IMF distinction).

4.1.2. Setup

In order to determine a baseline, we compare three character-level CNN models (GraphIMF, GraphCV, and Syl) and two word-level Word2Vec models (Word-CBOW and Word-Skip) as baseline. For CNN models we borrow the Char-CNN algorithm(Kim et al., 2016).

We use a default parameter setting of each architecture except for Syl-CNN. Kim et al. (2016) set the character embedding dimension as $d = 15$ and the filter numbers as $h = 1, 2, \ldots, 7$, but we modify them into $d = 150$ and $h = 1, 2, \ldots, 4$ for our Korean syllable vocabulary of size 1,766.

4.2. Inspection on Learned Vector Spaces

4.2.1. Visualizing Vector Spaces

To survey our results in Figure 1, we visualize vectors representing the most frequent 1,000 words for each model using t-Distributed Stochastic Neighbor Embedding(Van Der Maaten and Hinton, 2008). If each vector space reflects similarity between words properly, similar word vectors will form a cluster and different word vectors will be distant from each other.

In the first two word-level models, most vectors are scattered. We see small clusters on the periphery, but they are restricted to inflections of verb stems such as 있 (*iss*) 'exist'

[3] In spite of etymology, these words are almost always written in the Korean script, not Chinese characters.

[4] `https://github.com/yoonkim/lstm-char-cnn`

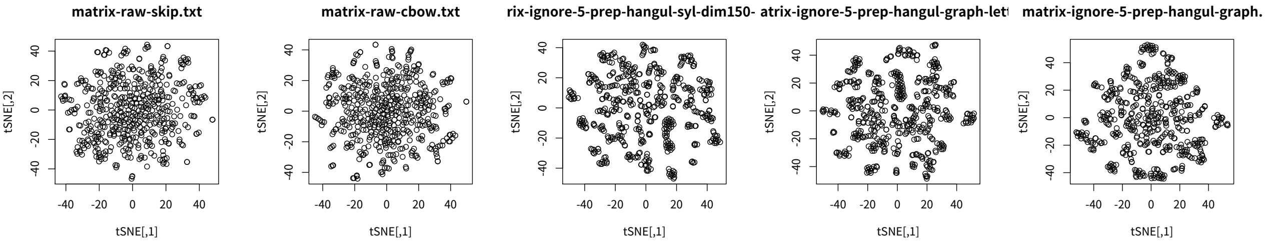

Figure 1: t-SNE visualization (Word-Skip, Word-CBOW, Syl-CNN, GraphCV-CNN, and GraphIMF-CNN)

and 않(*anh*) 'not', whose frequency is very high and whose meaning is functional rather than lexical.

The other three character-level models reveal clusters more clearly. Some of the clusters represent noun and pronoun cases, which each share a case marker: nominative 이(*i*) or 가(*ka*), accusative 을(*ul*) or 를(*lul*), locative 에(*ey*) and so on. Verbs are also clustered with respect to markers like a sentence final 다(*ta*) and relative clause 는(*nun*). We also note that our GraphIMF-CNN model gathers some negative polarity items including 아무(*a.mwu*) 'any', 결코(*kyel.kho*) 'ever', and 전혀(*cen.hye*) 'at all'.

4.2.2. Clustering Vectors

By visualizing vector spaces, we may be able to find salient groupings of words. To expand our observation to the rest, we identify what words each cluster consists of and how homogeneous it is using the K-means clustering algorithm(Hartigan and Wong, 1979).

GraphIMF-CNN	GraphCV-CNN	Syl-CNN
1. N(cit.) 2. N(nom.) 3. N(acc.) 4. N(loc., top., ...) 5. N(top.), V(rel.) 6. A(att.), V(rel.) 7. Adverbs 8. Nominal Adverbials?? 9. V with connectives 10. Sentence finals	1. N(cit.) 2. N(nom.) 3. N(acc.) 4. N(gen.) 5. N(top.), V(rel.) 6. A(rel.), V(rel.) 7. Nominal Adverbials?? 8. Adverbs and V with connectives 9. V with connectives 10. Sentence finals	1. N(cit.) 2. N(cit.), Demonstratives, A and V with connectivs 3. N(nom.) 4. N(acc.) 5. N(gen.) V(rel.) A(att.) 6. N(top.) V(rel.) 7. Nominal and verbal adverbials?? 8. Adverbs? 9. V(rel.) V with connectives 10. Sentence finals

The symbol ?? means that a marked cluster is noisy and dubious.

Table 6: 10-means clustering over the most frequent 1,000 words

The results with $K = 10$ for three character-level models are shown in Table 6. Even though we started with texts without part-of-speech tagging, we can see that the clusters we have generated largely overlap with syntactic categories. Using word-level models, on the other hand, we cannot find a common property among words in the same cluster. The least heterogeneous cluster might be tagged as Medical because 62 out of 151 words, such as 근육(*kun.ywuk*) 'muscle(citation)', 장애가(*cang.ay.ka*) 'disability(nominative)', 질환을(*cil.hwan.ul*) 'disease(accusative)', are relative with the topic. The other 89 words include more generic terms and conjunctions like 변화(*pyen.hwa*) 'change(citation)' and 또는(*tto.nun*) 'or'. Setting $K = 30$, citation forms of nouns are subcategorized into six clusters in the GraphIMF-CNN model, three clusters in the GraphCV-CNN model, and five clusters in the Syl-CNN model. We can manually cluster in the GraphIMF-CNN model as in Table 7, which generates results that are less noisy than those generated by the other models.

Citation forms of nouns
Concrete and basic morning, water, tree, road, room, rice, book, fruit, ...
Pronouns and personal he, I, grandmother, mom, uncle, mother, dad, child, author, ...
Collective Korea, USA, Japan, China, North Korea, Our country, France, party, Western, humankind, ...
Abstract (1) kind, history, literature , culture, market, cinema, environment, development, ...
Abstract (2) public, movement, school, society, liberation, economy, ...
Temporal year, minute, afternoon, this year, next year, every year, one year, spring, ...

We translated all examples from Korean to English.

Table 7: Part of 30-means clustering from the GraphIMF-CNN model

4.3. Discussion

We examine the nearest neighbors of words using cosine similarity. We select words representing the characteristics of morphologically rich languages such as those we mentioned in Section 2.

4.3.1. The same morpheme in different syllables

As we saw in Table 3, a non-syllabic morpheme can appear in syllables of various forms. Two such examples are the final consonant -ㄴ (*-n*) '(that) -' and the medial vowel -ㅓ (*-e*) '-ing.' They combine with a stem 기쁘(*ki.ppu*) 'be glad' and make the words 기쁜(*ki.ppun*) '(that) is/are glad' and 기뻐(*ki.ppe*) 'being glad' respectively. Table 8 shows what each model predicts as the most similar words to these two targets.

For 기쁜(*ki.ppun*) '(that) is/are glad', at least 4 out of 5 words have the right morpheme ㄴ (*n*) in all of the models, on both the character- and word- level. One error in the Syl-CNN model is 홈런(*hom.ren*) 'home run', which is morphologically totally different from the target. Moreover, the sixth nearest neighbor is 타이슨(*tha.i.sun*) 'Tyson.' The syllable-level model's prediction wrongly includes loanwords that end with ㄴ (*n*).

	Graphemes (IMF)		Graphemes (CV)		Syllable characters		Word2Vec (CBOW)		Word2Vec (Skip-gram)	
기**쁜**	슬**픈**	'sad'	슬**픈**	'sad'	어설**픈**	'poor'	하자**는**	'suggesting to do'	바**쁜**	'busy'
'(that) is/are glad'	즐거**운**	'pleasant'	가**쁜**	'breathless'	홈런	'home run'	억울**한**	'suffering'	반가**운**	'welcomed'
	무서**운**	'dreadful'	힘**든**	'hard'	슬**픈**	'sad'	힘**든**	'hard'	슬**픈**	'sad'
	가**쁜**	'breathless'	궂**은**	'sad'	즐거**운**	'pleasant'	정작	'but ... actually'	못**된**	'wicked'
	나**쁜**	'bad'	옳**은**	'right'	쉬**운**	'easy'	그**땐**	'then'	간다**는**	'insisting on going'
기**뻐**	**써**서	'writing'	추**워**서	'cold'	꽈르릉	(onomatopoeia)	보**아**서	'looking'	평상시와	'usual'
'being glad'	꺼**내**서	'taking out'	주**워**서	'picking up'	아슬아슬	(memetic word)	결단을	'decision(acc.)'	부러워하는	'envicious'
	갈**아**서	'changing'	꺼**내**서	'taking out'	쟁그랑	(onomatopoeia)	아내도	'also a wife'	잊어버리기	'to forget'
	떠서	'lifting'	아까**워**서	'sparing'	까악까악	(onomatopoeia)	당부를	'request(acc.)'	하라니까	'I told you, do it!'
	장가**가**서	'marrying'	누**워**서	'lying'	소리니	'Is it a sound?'	않았소	'did not(indicative)'	자기한테	'to oneself'

** Bold face denotes success in finding similarity, and an underline denotes a notable error.

Table 8: Nearest neighbors of the targets 기쁜(*ki.ppun*) '(that) is/are glad' and 기뻐(*ki.ppe*) 'being glad' respectively

The results of the latter 기뻐(*ki.ppe*) 'being glad' get worse in the Syl-CNN, Word-CBOW, and Word-Skip. Only GraphIMF-CNN and GraphCV-CNN find all true neighbors. We also note that the first four candidates from the Syl-CNN model are all onomatopoeic. This is unexpected as the syllables that make up 기뻐(*ki.ppe*) 'being glad' are rarely used in onomatopoeic words.

The results so far suggest that grapheme-level granularity captures morphological (and then functional) similarity more effectively than syllable-level and word-level granularity.

4.3.2. Sino-Korean polyphony and homophony

We also indicated that multiple Korean readings of Sino-Korean characters become a challenge for character-aware word embeddings. For example, in the structure of a word 확률(*hwak-lywul*) 'probability', 률(率) *lywul* is analyzed as a head but it can appear as 율(*ywul*) in other words. Table 9 shows that each model predicts as the nearest neighbors of the nominative and accusative forms of the stem.

In the GraphIMF-CNN model, the first four for nominative case and the first three for accusative case have the target syllable 률(*lywul*) or its allomorph 율(*ywul*). The GraphCV-CNN and Syl-CNN models find one or two 율(*ywul*) or 률(*lywul*) and their results are less satisfactory both in quality and quantity. They include "false friends" such as 규율이(*kywu.ywul.i*) 'discipline(nominative)' and 운율은(*wun.ywul.un*) 'rhyme(topic)', where 율(*ywul*) is not 率 but its homomorph 律. Moreover, 효율(*hyo.ywul*) 'efficiency' is the only 率-type word that the Syl-CNN model found, but its meaning is less similar with other words of form X率 'rate or probability of X.' This observation indicates that graphemes with IMF distinction capture semantic similarity represented by formal similarity better than other units and they also and they also clarify ambiguities caused by homonymy and polysemy.

5. Related Work

5.1. Character-level Approaches to Word Embeddings

Our work is based on studies of constructing vector representations of words through their constituent characters. Ling et al. (2015) first introduced this model by composing characters into representations of words using bidirectional long-short term memory networks, and then looked at language modeling and part-of-speech tagging performance on English, Portuguese, Catalan, German and Turkish. Kim et al. (2016) used convolutional neural networks and highway layers to capture morphological similarities and semantic similarities between words, and claimed that their language model is language non-specific. Their model outperformed word and morpheme-level baseline performance on Arabic, Czech, French, German, Spanish, and Russian, which are morphologically richer than English.

5.2. Subword information on Chinese

The importance of characters has been a focus for approaches dealing with Chinese text, a language in which characters are written without spaces (Sproat et al., 1996). Because each Chinese character has an internal structure consisting of radicals, units that are more fine-grained than characters have been exploited in research utilizing word embeddings(Yin et al., 2016; Xu et al., 2016) and character embeddings(Sun et al., 2014; Li et al., 2015; Peng and Cambria, 2017).

5.3. Korean Word Embeddings

To the best of our knowledge, in most studies on Korean word embeddings, characters were coded at the syllable level.Cinarel and Zhang (2016) incorporated a word with syllable N-grams using a subword model proposed by Bojanowski et al. (2017). Choi (2017) used convolutional neural networks to extract a word vector from syllable vectors and a skip-gram model to learn the distribution of each word. Choi evaluated the model by measuring training losses and cosine similarities on WS353(Finkelstein et al., 2001), and implementing out-of-vocabulary tests on a several nonce words and morphology tests on two types of noun forms.

Yu and Ko (2017) added a syllable embedding to a word representation as an input of bidirectional LSTM CRF models for Korean named entity recognition. As in our grapheme model the nearest neighbors of the word 마르크스 *ma.lu.khu.su* 'Marx' consist mostly of loanwords from German. We can therefore expect that a grapheme-level approach would capture a typical phonotactic pattern of proper names better than a syllable-level approach does in NER tasks.

5.4. Adoptation of graphemes for Korean

Choi et al. (2016) adopted a grapheme-level approach for constructing a Korean morphological analyzer. They did not specify from which unicode block their morphemes were encoded, but we deduce that they used CV distinction.

	Graphemes (IMF)		Graphemes (CV)		Syllable characters		Word2Vec (CBOW)		Word2Vec (Skip-gram)	
확**률**이	환**율**이	'exchange rate'	환**율**이	'exchange rate'	효**율**이	'efficiency'	평상시와	'usual'	저렴하게	'cheaply'
'probability'	투표**율**이	'voter turnout'	바벨탑이	'Tower of Babel'	규율이	'discipline'	부러워하는	'envious'	음미될	'to be appreciate'
(nominative)	성장**률**이	'growth rate'	용수량이	'water capacity'	확산이	'diffusion'	잊어버리기	'to forget'	말해볼	'to try telling'
	손해**율**이	'loss rate'	예측이	'expectation'	수법이	'method'	하라니까	'I told you, to it!'	정도라면	'if (it is) that much'
	진덕왕이	'Queen Jindeok'	성장**률**이	'growth rate'	변증법이	'dialectic'	자기한테	'to oneself'	높다고	'(say that it is) high'
확**률**을	수익**률**을	'profit rate'	환**율**을	'excange rate'	품격을	'dignity'	일치할	'to coincide'	참여는	'participation(topic)'
'probability'	합격**률**을	'acceptance rate'	운율을	'rhyme'	징역을	'imprisonment'	빠져들	'to sink into'	아래에서는	'under(topic)'
(accusative)	성장**률**을	'growth rate'	경영을	'management'	변별력을	'discrimination'	참조할	'to refer to'	가시적인	'visible'
	옘병을	'epidemic'	관우를	'Guan Yu'	효**율**을	'efficiency'	규정될	'to be defiend'	허용될	'to be allowed'
	원형을	'prototype'	집행을	'execution'	경계심을	'wariness'	제시할	'to suggest'	완벽에	'to perfection'

** Bold face denotes success in finding similarity, and an underline denotes a notable error.

Table 9: Nearest neighbors of the targets 확률이 (*hwak.lywul.i*) 'probability(nominative)' and 확률을 (*hwak.lywul.ul*) 'probability(accusative)' respectively

They reported that the non-CRF version of their model had misanalyzed 차우세스쿠 (*cha.wu.se.su.kwu*) 'Ceauşescu', 창 (*chang*) and ㅜ세스크 (*wu.se.su.ku*). From Table 10, we note that this error could have been avoided by IMF distinction, in which 창 (*chang*) cannot be analyzed as a substring of 차우세스쿠 (*cha.wu.se.su.kwu*), and ㅜ세스크 (*wu.se.su.ku*) is not a valid string.

Consonant-Vowel distinction	
차우세스쿠 (*cha.wu.se.su.kwu*)	[ㅊ ㅏ ㅇ]ㅜ ㅅ ㅔ ㅅ ㅡ ㅋ ㅜ *U+314A 314F 3147* 315C 3145 3154 ...
창 (*chang*)	[ㅊ ㅏ ㅇ] *U+314A 314F 3147*
Initial-Medial-Final distinction	
차우세스쿠 (*cha.wu.se.su.kwu*)	[ᄎ ᅡ] ᄋ ᅮ ᄉ ᅦ ᄉ ᅳ ᄏ ᅮ *U+110E 1161* **110B** 116E 1109 1166 ...
창 (*chang*)	[ᄎ ᅡ] ᆼ *U+110E 1161* **11BC**

Table 10: Different substring relations in different grapheme decomposition methods

6. Conclusion

We have presented a grapheme-level approach to character-aware word embeddings for languages represented by syllabaries. We merely decompose syllable characters into graphemes in the framework of current neural word representations, without the cost of morphological analysis. Our grapheme-level model disambiguates homographs and finds both functional and semantic similarities among allomorphs of the same morpheme, which syllable-level and word-level models fail to capture. These results show that our model represents various word forms well even in the complex morphological system of agglutinative languages. We believe that our results apply to other languages with different morphological systems and writing systems.

Acknowledgement

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A5A2A01023336).

7. Bibliographical References

Argüelles, A. and Kim, J.-R. (2004). *A Handbook of Korean Verbal Conjugation*. Dunwoody Press.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Choi, J., Youn, J., and Lee, S.-g. (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, dec.

Choi, S.-H. (2017). The modeling and training methods for syllable-based korean word embeddings. Master's thesis, Seoul National University.

Cinarel, C. and Zhang, B.-T. (2016). Better Word Embeddings for Korean. In *Proceedings of the Korea Information Science Society Winter Conference*, pages 627–629.

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Karlsson, F. (1986). Frequency considerations in morphology. *STUF-Language Typology and Universals*, 39(1-4):19–28.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2016). Character-Aware Neural Language Models. *AAAI*, pages 2741–2749.

Li, Y., Li, W., Sun, F., and Li, S. (2015). Component-Enhanced Chinese Character Embeddings. *Emnlp*, (September):829–834.

Ling, W., Luís, T., Marujo, L., Astudillo, R. F., Amir, S., Dyer, C., Black, A. W., and Trancoso, I. (2015). Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. (September):1520–1530.

Peng, H. and Cambria, E. (2017). Radical-Based Hierarchical Embeddings for Chinese Sentiment Analysis at Sentence Level. In *The Thirtieth International Florida Artificial Intelligence Research Society conference*.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, 20:33–53.

Schinke, R., Greengrass, M., Robertson, A. M., and Willett, P. (1996). A Stemming Algorithm for Latin Text Databases. *Journal of Documentation*, 52(2):172–187.

Sproat, R., Gale, W., Shih, C., and Chang, N. (1996). A stochastic finite-state word-segmentation algorithm for chinese. *Computational linguistics*, 22(3):377–404.

Sun, Y., Lin, L., Tang, D., Yang, N., Ji, Z., and Wang, X. (2014). Radical-Enhanced Chinese Character Embedding. In *International Conference on Neural Information Processing*, pages 279–286.

The National Institute of the Korean Language. (2010). 21st Century Sejong Project.

Van Der Maaten, L. J. P. and Hinton, G. E. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Xu, J., Liu, J., Zhang, L., Li, Z., and Chen, H. (2016). Improve Chinese Word Embeddings by Exploiting Internal Structure. *Naacl*, pages 1041–1050.

Yin, R., Wang, Q., Li, R., Li, P., and Wang, B. (2016). Multi-Granularity Chinese Word Embedding. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, pages 981–986.

Yu, H. and Ko, Y. (2017). Expansion of Word Representation for Named Entity Recognition Based on Bidirectional LSTM CRFs. *Journal of Korean Institute of Information Sciencentist and Engineers*, 44:306–313.