Learning Sub-Character level representation for Korean Named Entity Recognition

Yejin Kim* and **Yekyung Kim***
LG Electronics, AI Lab.
Yangjae R&D, 38, Baumoe-ro, Seocho-gu, Seoul, Korea 06763

Abstract

Most of the previous studies on the Korean Named Entity Recognition (NER) topic focused on utilizing morphological-level information because the language is rich in character diversity. This paper illustrates an improved unigram-level Korean NER model with sub-character level representation, *jamo*, which can represent a unique linguistic structure of Korean and its syntactic properties and morphological variations. The experimental result shows that exploiting sub-character gives us a boost of + (avg) 2 F1, also, our proposed C-GRAM model outperformed about 3 F1 comparing with the baseline.

Introduction

Most research efforts have focused on classifying entities in English, while other complex character languages (*e.g.*, Arabic, Chinese, Korean, etc.) have been widely used in real-world. As to Named entity recognition (NER), the lack of word boundary and the character richness is more complicated properties in complex character languages (Matteson et al. 2018). It leads to difficulties and limitations indirectly applying techniques from English research to the languages which have the different internal structure from English (Oh et al. 2018).

Previous studies of Korean NER systems focused on the morphological-level text that is made using morphological analysis and handcrafted features (Na et al. 2019; Yu and Ko 2017; Lee et al. 2016) while there are various studies about character-level tagger (Kuru, Can, and Yuret 2016) or using sub-character information in Chinese (Dong et al. 2016). Although morphological-level NER has advantages in utilizing richer linguistic information such as part-of-speech (POS) tags and word boundaries, it suffers from the potential issue of error propagation that error of morphological analysis results commonly lead to NER errors. (Chung et al. 2004; Kwon, Ko, and Seo 2019). Recent Korean NER work(Kwon, Ko, and Seo 2019) has looked at suggesting the character-level BLSTM-CRF model with bigram vector representation and Eojeol's positional prefix information. While bigram embedding shows an outperformed performance, it has a considerable number of vocabulary that could be lead to increase sparsity.

Also, unique linguistic structures have been neglected in existing modern Korean NER. Korean character is composed of a small and minuscule set of sub-characters, *jamo*, unlike character commonly been regarded as the minimal unit in Natural language processing (NLP). For example, Korean have 11,172 characters[1] which are composed of three sub-characters. In recent work, integrating Korean linguistic structure at the level of *jamo* (consonants, vowels) is shown to be useful for sentence parsing (Stratos 2017), POS tagging (Matteson et al. 2018) and word representation (Oh et al. 2018). (Stratos 2017) also shows the possibility of such data sparsity could be alleviated by *jamo*.

This work is motivated by the desire to build a compact character-level NER architecture utilizing the unique structure of sub-character and a wide range of context from character sequence. We present neural architectures for NER that use no externals resources or features beyond a small amount of supervised training data. Our models are designed with two ideas. First one is to try exploiting semantic information of sub-character in NER, which is difficult to access with the character-level units. To investigate the effectiveness of *jamo* representation, we compare the use of Bidirectional Long Short-Term Memory (BLSTM) (Chung et al. 2014), Bidirectional Gated Recurrent Unit (BGRU)(Gers, Schmidhuber, and Cummins 2000) and Convolutional Neural Network (CNN) based *jamo*-level representation in BLSTM-CRF NER model.

Second, since character-level input has a longer time sequence compared to the word-level, it is essential to capture the information of local context for improvement. To this end, We proposed a new neural network architecture in NER with two concepts. (i) a stacked BLSTM(Graves, Mohamed, and Hinton 2013) with a subsequent Conditional Random Field (CRF) layer above it (stack-BLSTM-CRF) and (ii) capturing the local information around each character using convolution with different two methods (GRAM-CNN (Zhu et al. 2018), C-GRAM), then using them as inputs BLSTM-CRF.

The results show that models with CNN-based *jamo*-level

* These two authors contribute equally to the work.

[1] This information is from the National Institute of the Korean Language.

character embeddings (CNN-*jamo*) achieve the highest F1 score of 78.35 among neural networks. Also, it shows that incorporating a *jamo*-based component in the NER model provides substantial performance improvements on baseline economically. At the NER method, our proposed methods C-GRAM-BLSTM-CRF shows the significantly outperformed on the character-level NER with 80.54 F1 scores. Also, it shows improvement over bigram based model and achieves competitive performance with the morphological-level model in the same dataset.

To summarize, our contributions are as follow.

- To our best knowledge, it is the first attempt to apply sub-character ('*jamo*') representation in neural Korean NER model. We found that utilizing CNN based *jamo*-representation shows a clear performance boost. It allows us to train an informative character embedding with a very small vocabulary.
- We propose methods for efficiently leveraging the local contexts based on n-gram character and perform extensive experiments on NER to verify the utility of the approach. Then, we show that using C-GRAM methods for character-level NER outperforms from the baseline by a large margin.
- The models are much more compact than previous Korean NER model and do not rely on the pipeline framework as the morphological analysis.

Model

In this section, we describe the methods to build a character-level input representation with *jamo* and propose our new neural based NER models.

Decomposition of Korean Character

Korean character can be decomposed into two or three sequences of *jamo*. Whole *jamo* set is composed fixed 51-element of units. The *jamo* elements of a character have a structure as head consonants J_h, vowels J_v, tail consonants J_t. For example, "곰" (bear) ∈ C(character) is composed of ㄱ ∈ J_h, ㅗ ∈ J_v, and ㅁ ∈ J_t. The tail consonants can be omitted, but the others are necessary. In our work, we fill an empty symbol (ϕ) when a character lack tail consonant such that a character always has 3 *jamo* elements.

Input Character Embeddings

A BLSTM network or CNN for extracting character-level representation of words has been previously explored by (Lample et al. 2016; Ma and Hovy 2016). we applied three networks to a *jamo*-level representation of characters. Each network takes as input three sequences of *jamo* of a character and returns an embedding vector.

BLSTM and BGRU concatenate the last outputs of forward and backward networks. In the case of CNN, the *jamo*-level representation of characters is computed with convolution layer and max-pooling. The outputs of each three networks concatenate to an embedding vector from a character lookup-table to obtain representation for each character.The output of each embedding networks is an input vector of an

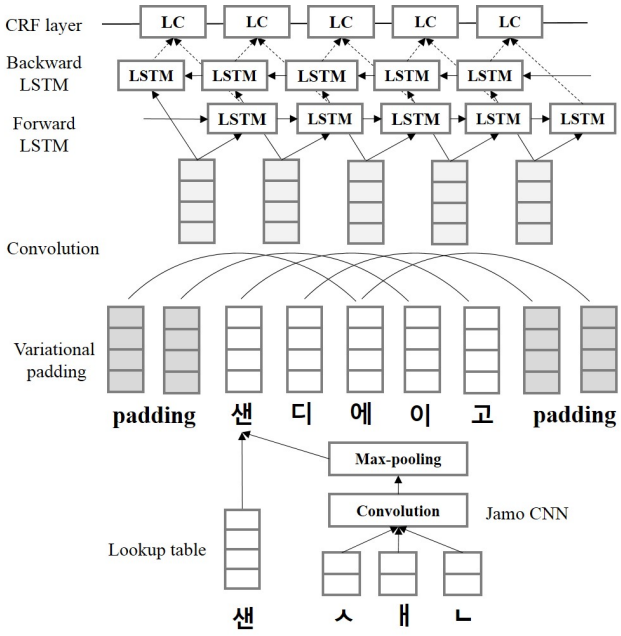

Figure 1: Architecure of C-GRAM-BLSTM-CRF models with *jamo*-level character representation. (means 'San Diego'in english)

NER network. To feed character-level inputs, we abandon the position indicator prefixes (*e.g.*, B-, I-) and use entity types (*e.g.*, PER, LOC, etc.) directly as labels.

NER Network Architectures

We adopt a BLSTM-CRF network as a baseline for character-level embeddings and a CRF layer for label decoding. The BLSTM-CRF network is commonly used in sequence labeling tasks (Huang, Xu, and Yu 2015; Lample et al. 2016; Ma and Hovy 2016). Recent Korean NER model (Kwon, Ko, and Seo 2019) applied it either. Based on the BLSTM-CRF network, we introduce three extended neural architectures.

The CharNER(Kuru, Can, and Yuret 2016) consists of stacked BLSTM which outputs of previous BLSTM layers are fed into the next layer, and the same process is carried on for the additional BLSTM networks and decode with a Viterbi. Bring this idea, we apply multiple layers of BLSTM, and the outputs are fed through CRF(stack-BLSTM-CRF) instead of Viterbi without using transition matrices which are manually-built only to allow tags consistent within a word.

The GRAM-CNN (Zhu et al. 2018) is a CNN model to extract local information between a target word and its n-gram features. The representation of an input word corresponds to each output of GRAM-CNN. A max-pooling is only applied over the correlated features of words, resulting in a vector of size one for each word we feed the outputs of GRAM-CNN into BLSTM network (GRAM-CNN-BLSTM).

The Convolutional-GRAM (C-GRAM) network is moti-

vated by GRAM-CNN. A traditional CNN model in a pooling layer has a possibility to lose useful features (Zhang et al. 2017). To overcome this, We apply various size of pads corresponds to output sequences length after passing through the convolutional layer. Each input sequences padded to front and rear of input sequences alternatively. The output of the convolutional layer is directly fed into BLSTM-CRF. The pooling layer is eliminated in C-GRAM-BLSTM-CRF to prevent from losing feature information. Figure 1 describes the overall architecture of our model.

Experiment

Dataset

The experiments evaluated the effectiveness of our methods on 2016 and 2017 Klpexpo NER task dataset[2] which both have five defined labels: Person, Location, Organization, Date, and Time. We used only 2016 dataset when we compared with previous work in Korean which tested on 2016 dataset only. Except for it, we used both years dataset to evaluate a general performance. Klpexpo datasets (2016, 2017) include 9301 sentences. The unique characters are 2166. We divided the datasets into a train set, development set, and test set to a ratio of 8:1:1. The training, development, and test data contains 22656, 2670, and 2125 NEs, respectively. Each category of NEs contains 7236 Person, 4394 Location, 5501 Organization, 4388 Date, and 771 Time.

Experiment Conditions

Our experiments are conducted in a two-step. The first step is exploring diverse embedding methods based on neural networks to maximize the effectiveness of *jamo*-level representations, and the other step is demonstrating the performance of our new NER models applying *jamo*-level representation of characters. The training epochs processed up to 50 and character vector of the lookup table as 50-dimensional vectors arbitrarily initialized. The dimension of the *jamo* vector was set up 25. The hidden unit size of BLSTM for *jamo* embedding was 50 each from forward and backward *jamo*-level representation, which resulted in 100 dimensions. The BGRU network had the same dimension with the BLSTM. We applied windows length 2 and 3 for CNN to compute the *jamo*-level representation of characters and each filter size was 50. To demonstrate the performance of our new NER models, we used Adam optimizer with a learning rate of 0.001, decay 0.95, clipping 5.0 and dropout rate 0.5 for BLSTM. The number of hidden unit of all BLSTM for NER step is 100. The CNN based NER networks (GRAM-CNN, C-GRAM) are applied a combination of window sizes 2 to 6. The total number of filters is 200, and the number of each filter is determined according to the number of window size combinations. All models are performed with a CRF layer on top.

Results and Discussion

Table 1 compares the performance of three different methods for *jamo* embeddings in baseline NER model. CNN based *jamo*-level representation obtains better performance of 78.35 F1 scores than BLSTM and BGRU networks. Table 2 shows the performance of our proposed NER models and baseline model in unigram. The best performance of GRAM-CNN was performed on [2,3,4,5] windows, and C-GRAM was performed on [3,4,5,6] windows. As the empirical results in Table 2 show that integrating character with *jamo*-level representation contribute to improving average 2.4 pt of F1 the over-all model. As we can see in the tables, adding depth with LSTM is beneficial for the performance compared to the single layer model(baseline, GRAM-CNN, C-GRAM). We are supposed that added layer learn a more informative high-level feature due to our feature is relatively low-level. Throughout the entire models, our C-GRAM-BLSTM with *jamo* outperformed with 80.54 of F1 score. These results demonstrate that C-GRAM method is a versatile approach to character-level NER model. Table 3 reveals the results of comparison with previous work. Our model outperforms all other character-level models, even achieve competitive performance with morphological-level models which are tagged by morphological analysis and human. From the point of view feature representation, our model has very small vocabulary size 1476, while the best of previous work have a 33339 vocabularies. Furthermore, if pretrained vectors were used, it would be significantly larger. Compared with the character model which have similar vocabulary size, our model shows the significantly improved performance nearly 5 pt of F1 score.

Embedding Layer	F1 (%)
BLSTM	77.86
BGRU	77.31
CNN	**78.35**

Table 1: Performance of each *jamo* embedding layer in baseline (BLSTM-CRF) NER model.

Model	F1 (%)	
	Char	Char +*jamo*
Baseline(Kwon, Ko, and Seo 2019)	75.77	78.35
stack-BLSTM	78.54	80.20
GRAM-CNN	73.79	76.65
GRAM-CNN-BLSTM	78.06	80.13
C-GRAM	74.20	78.28
C-GRAM-BLSTM	**78.77**	**80.54**

Table 2: Performance of our NER model and comparison models with unigram input. We applied CNN as a *jamo* embedding method and a CRF layer is added on top for joint decoding in all experiments.

Conclusion and Future Work

In this paper, We have shown that our neural network model which benefits from leveraging a CNN based *jamo*-level representation and capturing n-gram context through C-GRAM methods. Due to the small vocabulary size, our models are

[2]This is dataset was disseminated by the National Institute of the Korean Language for the Klpexpo contest.

Model	F1 (%)	Feature Representation
char	73.64	1425 × 200
bigram	74.12	31914 × 200
char+bigram	78.15	33339 x 200
char+bigram+eojeol	77.64	37359 x 200
morphological*	80.51	-
our best	**78.77**	**1476 x 200**

Table 3: F1 score of the previous work(Kwon, Ko, and Seo 2019) when they only use NER training data in Klpexpo 2016 dataset. our best model is C-GRAM-BLSTM-CRF. (char: character embedding vector. bigram: character-level representation of bigram. char+bigram: character-level representation of bigram + bigram embedding. char+bigram+eojeol: added Korean spacing unit prefix features to "char+bigram." morphological: embedding morphological analysis of input text performed by a human. the morphological-level NER is marked as *.)

very compact. However, it shows a similar or better performance than character-level NER when using only the provided training datasets.

There are several potential directions for future work. First, our model can be further improved by approaches to expand with more useful and correlated information. Another direction is to apply our model to noisy data such as social media. Since our model does not require any preprocessing, it might be effortless to apply it to these data.

References

Chung, E.; Lim, S.; Hwang, Y.; and Jang, M. 2004. Hybrid named entity recognition for question-answering system. In *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*.

Chung, J.; Gülçehre, Ç.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* abs/1412.3555.

Dong, C.; Zhang, J.; Zong, C.; Hattori, M.; and Di, H. 2016. Character-based LSTM-CRF with radical-level features for chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2-6, 2016, Proceedings*, 239–250.

Gers, F. A.; Schmidhuber, J.; and Cummins, F. A. 2000. Learning to forget: Continual prediction with LSTM. *Neural Computation* 12(10):2451–2471.

Graves, A.; Mohamed, A.; and Hinton, G. E. 2013. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 6645–6649.

Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR* abs/1508.01991.

Kuru, O.; Can, O. A.; and Yuret, D. 2016. Charner: Character-level named entity recognition. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, 911–921.

Kwon, S.; Ko, Y.; and Seo, J. 2019. Effective vector representation for the korean named-entity recognition. *Pattern Recognition Letters* 117:52 – 57.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural architectures for named entity recognition. *CoRR* abs/1603.01360.

Lee, S.; Song, Y.; Choi, M.; and Kim, H. 2016. Bagging-based active learning model for named entity recognition with distant supervision. In *2016 International Conference on Big Data and Smart Computing, BigComp 2016, Hong Kong, China, January 18-20, 2016*, 321–324.

Ma, X., and Hovy, E. H. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Matteson, A.; Lee, C.; Kim, Y.; and Lim, H. 2018. Rich character-level information for korean morphological analysis and part-of-speech tagging. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 2482–2492.

Na, S.-H.; Kim, H.; Min, J.; and Kim, K. 2019. Improving lstm crfs using character-based compositions for korean named entity recognition. *Computer Speech Language* 54:106 – 121.

Oh, A.; Park, S.; Byun, J.; Baek, S.; and Cho, Y. 2018. Subword-level word vector representations for korean. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 2429–2438.

Stratos, K. 2017. A sub-character architecture for korean language processing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 721–726.

Yu, H., and Ko, Y. 2017. Expansion of word representation for named entity recognition based on bidirectional lstm crfs. *Journal of KIISE* 44:306–313.

Zhang, T.; Li, C.; Cao, N.; Ma, R.; Zhang, S.; and Ma, N. 2017. Text feature extraction and classification based on convolutional neural network (CNN). In *Data Science - Third International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2017, Changsha, China, September 22-24, 2017, Proceedings, Part I*, 472–485.

Zhu, Q.; Li, X.; Conesa, A.; and Pereira, C. 2018. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics* 34(9):1547–1554.