Received: 28 August 2018 | Revised: 19 March 2019 | Accepted: 20 May 2019

DOI: 10.4218/etrij.2018-0456

ORIGINAL ARTICLE

ETRI Journal WILEY

Linear-Time Korean Morphological Analysis Using an Action-based Local Monotonic Attention Mechanism

Hyunsun Hwang | **Changki Lee**

Department of Computer Science, Kangwon National University, Chuncheon, Rep. of Korea

Correspondence
Changki Lee, Department of Computer Science, Kangwon National University, Chuncheon, Rep. of Korea.
Email: leeck@kangwon.ac.kr

Funding information
This work was supported by the Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government (MSIT) (2018-0-00605, Artificial Intelligence Contact Center Solution) and Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government (MSIT) (2013-0-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services).

For Korean language processing, morphological analysis is a critical component that requires extensive work. This morphological analysis can be conducted in an end-to-end manner without requiring a complicated feature design using a sequence-to-sequence model. However, the sequence-to-sequence model has a time complexity of $O(n^2)$ for an input length n when using the attention mechanism technique for high performance. In this study, we propose a linear-time Korean morphological analysis model using a local monotonic attention mechanism relying on monotonic alignment, which is a characteristic of Korean morphological analysis. The proposed model indicates an extreme improvement in a single threaded environment and a high morphometric F1-measure even for a hard attention model with the elimination of the attention mechanism formula.

KEYWORDS
deep learning, korean morphological analysis, local attention mechanism, natural language processing, sequence-to-sequence learning

1 | INTRODUCTION

Morphological analysis is an important task for the natural language processing of Korean. Korean morphological analysis is difficult not only in part of speech (POS) tagging but also in morphological separation and restoration [1]. Recently, an end-to-end Korean morphological analysis method using a sequence-to-sequence model [2,3] was proposed [4]. The sequence-to-sequence model encodes an input sequence using a recurrent neural network (RNN) that processes a sequence and subsequently decodes the encoded information using another RNN to generate an output sequence. Using this model, morphological analysis models without complicated feature designs can be created using morphological analysis training data. In addition, to ensure the high performance of the sequence-to-sequence model, an attention mechanism [5] has been proposed to calculate the information required each time when decoding the input sequence. In addition, a study of copying mechanisms [6] to copy input words to the output sequence has suggested a method to improve the performance of Korean morphological analysis based on its characteristics, which mostly appears in the output sequence of input words [4]. However, the addition of these techniques introduces a time complexity of $O(n^2)$ into the sequence-to-sequence model, where n is the length of the input sequence. In Korean morphological analysis, the alignment of the input sequence to the output sequence is observed to be monotonic and most operations are unnecessary. Therefore, in this study, we propose a Korean morphological analysis method using an attention

mechanism, a copying mechanism, and a local attention mechanism to reduce unnecessary operations. We propose an action-based local monotonic attention mechanism that predicts where the position information should be concentrated by considering the monotonic relationship between the input and output sequence. Additionally, we use a hard monotonic attention mechanism that removes the attention mechanism formula, thus improving the morphological analysis speed through the sequence-to-sequence model.

2 | RELATED WORK

Traditional Korean morphological analysis was performed using three methods: morpheme separation, POS tagging, and morphological restoration. In morpheme separation, a dictionary-based method and a pre-designed linguistic rule-based method are used to divide the phrases separated by Korean word spacing into morphemes, which are the minimal semantic units. In POS tagging, various machine learning methods are used to approach the problem of sequence labeling to attach the parts of speech of separate morphemes. Morpheme restoration has mainly used a dictionary-based method to restore a modified morpheme to its original form according to the rules of Korean agglutination [7,8].

Recently, end-to-end processing techniques have been studied by converting morphological analysis into the sequence generation problem by examining the sequence-to-sequence model [4]. Table 1 displays an example of end-to-end morphological analysis using a sequence-to-sequence model. A sentence divided into syllable units is entered into the model, and a sentence is generated with morpheme separation, POS tagging, and completed morphological restoration in syllable units. In this case, "<sp>" represents the spacing information, and the output of the model is combined to yield the final result of the morphological analysis. The sequence-to-sequence model encodes the input sequence using different RNNs and subsequently decodes the encoded information to generate a new output sequence. However, a drawback of this model is that fixed encoding information is used for decoding. An attention mechanism [5] solves this problem by providing another neural network structure to obtain the score alignment of the input sequence during every decoding. However, it computes copious lengths of the input sequence during each decoding and can occasionally include unnecessary operations. In Ref. [9], a local attention mechanism was proposed to reduce unnecessary operations, and it was applied to machine translation work, where attention alignment is concentrated on a specific area. However, this method is unable to predict the position on which to concentrate. In Ref. [10], a hard attention mechanism was applied using an action to shift the focus to the morphological inflection generation task with monotonic alignment.

The main contributions of our work are as follows:

- We propose a Korean morphological analysis model with linear-time complexity by applying an action-based local monotonic attention mechanism.
- The local monotonic attention mechanism is applied to the copying mechanism as well as the attention mechanism of the sequence-to-sequence model to ensure high performance.
- We propose a local soft monotonic attention model for Korean morphological analysis and compare it with the hard monotonic attention model from [10].

TABLE 1 Example of Korean morphological analysis using the sequence-to-sequence model

Input sentence	오늘 경기에서도 정말 잘**했**다. (O-neul gyeong-gi-e-seo-do jeong-mal jal-**haet**-da.) "You played really well in the game today."
Model input	오 늘 <sp> 경 기 에 서 도 <sp> 정 말 <sp> 잘 **했** 다 . (O neul <sp> gyeong gi e seo do <sp> jeong mal <sp> jal **haet** da .)
Model output	오 늘 <NNG> <sp> 경 기 <NNG> 에 서 <JKB> 도 <JX> <sp> 정 말 <MAG> <sp> 잘 <MAG> **하** **<XSV>** **았** **<EP>** 다 <EF> . <SF> (O neul <NNG> <sp> gyeong gi <NNG> e seo <JKB> do <JX> <sp> jeong mal <MAG> <sp> jal <MAG> **ha** **<XSV>** **at** **<EP>** da <EF> . <SF>)
Morphological analysis result	오늘/NNG 경기/NNG 에서/JKB 도/JX 정말/MAG 잘/MAG **하/XSV** **았/EP** 다/EF ./SF (O-neul/NNG gyeong-gi/NNG e-seo/JKB do/JX jeong-mal/MAG jal/MAG **ha/XSV** **at/EP** da/EF ./SF)

3 | LINEAR-TIME END-TO-END KOREAN MORPHOLOGICAL ANALYSIS

3.1 | Action-based local monotonic attention mechanism

In this section, we introduce a local attention mechanism using the action modified from the model proposed in Ref. [10]. First, "<step>" is added to the output word dictionary of the sequence-to-sequence model for morphological analysis. This "<step>" is an action to change the position on which to concentrate. It is assumed that the task is a monotonic alignment relation. When one decoded output word is "<step>," the previously focused position is moved to the next word. The output of the model ($y'_{1:k}$ in Figure 1) is transformed into the final output ($y_{1:k}$ in Figure 1) without "<step>."

The sequence-to-sequence model for Korean morphological analysis uses the same model as [4] with the following formula:

$$\{\overleftrightarrow{h}_1, \ldots, \overleftrightarrow{h}_T\} = \text{biGRU}\left(x_1, \ldots, x_T\right). \tag{1}$$

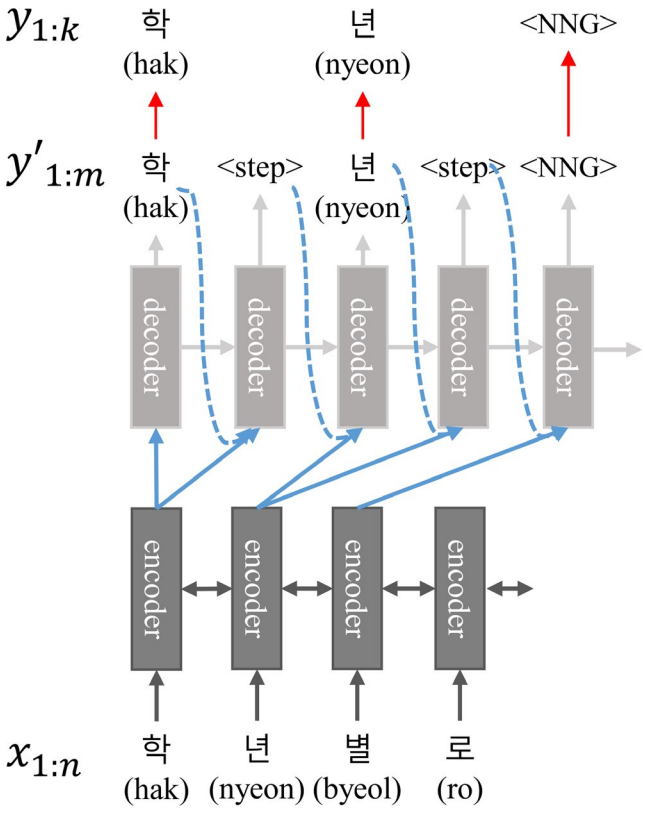

FIGURE 1 Example of action-based Korean morphological analysis

The encoder uses a bidirectional gated recurrent unit (biGRU) and encodes the input sequence $X = \{x_1, \ldots, x_T\}$ to create a hidden state sequence $\{\overleftrightarrow{h}_1, \ldots, \overleftrightarrow{h}_T\}$. The decoder subsequently uses the attention mechanism and copying mechanism applied by [4] to generate the output sequence. At this point, the position information p_t for the local attention mechanism is given by

$$p_t = \begin{cases} 0, & \text{if } t = 0, \\ p_{t-1} + 1, & \text{if } y_{t-1} = '<\text{step}>', \\ p_{t-1}, & \text{otherwise.} \end{cases} \tag{2}$$

In (2), t is the decoding time, and p_t is updated at each decoding time. The decoder model uses the attention mechanism and copying mechanism by applying the following equation and using the position p_t to be concentrated on.

$$\left(p_t - d\right) \le i \le \left(p_t + d\right), \tag{3}$$

$$e_i^t = \text{score}(y_{t-1}, h1_t, h2_t, \overleftrightarrow{h}_i), \tag{4}$$

$$a_i^t = \frac{\exp\left(e_i^t\right)}{\sum_{p_t-d}^{p_t+d} \exp\left(e_j^t\right)}, \tag{5}$$

$$c^t = \begin{cases} \sum_i a_i^t \overleftrightarrow{h}_i & \text{soft}, \\ \overleftrightarrow{h}_{p_t} & \text{hard}. \end{cases} \tag{6}$$

The attention mechanism computes the context vector c^t at time t by computing only the range of (3) in the input sequence. The score function is the score on which the attention mechanism should focus, $\overleftrightarrow{h}_i$ is the hidden state of the encoder, and $h1_t$ and $h2_t$ are the hidden states of the decoder ((7) and (8)). In this case, d is the window size for calculating the attention mechanism from the position to concentrate on, and the hard attention mechanism of (6) is $d = 0$.

$$h1_t = \text{GRU}_{\text{decoder}}\left(E_{tgt}\left(y_{t-1}\right), c^t, h1_{t-1}, h2_{t-1}\right), \tag{7}$$

$$h2_t = \text{ReLU}\left(W_f h1_t + b_f\right), \tag{8}$$

$$s_t = \text{softmax}\left(W_{y,h1} h1_t + W_{y,h2} h2_t + W_{y,y} E_{tgt}\left(y_{t-1}\right) + W_{y,c} c^t + b_y\right), \tag{9}$$

$$e_{\text{copy}_i}^t = \tanh\left(W_{\text{copy}} \overleftrightarrow{h}_i\right) \times h1_t, \tag{10}$$

$$p\left(y_t | X\right) = \begin{cases} \frac{1}{z}\left(\exp\left(s_t\right) + \sum_{j:x_j = y_t} \exp\left(e_{\text{copy}_j}^t\right)\right), & y_t \in \{x_{p_t-d}, \ldots, x_{p_t+d}\}, \\ \frac{1}{z}\exp\left(s_t\right), & \text{otherwise.} \end{cases} \tag{11}$$

The copying mechanism calculates the scored copies of the input words at decoding time t, as in (4) of the attention mechanism. In (9), (10), and (11), s_t is the output score before the copying mechanism is applied.

3.2 | Data conversion for the training model

To apply the action-based local attention mechanism proposed herein to morphological analysis, the "<step>" action is required to be in the correct output sequence of training data. However, adding the "<step>" action by predicting the attention alignment that unsupervised learning would yield is difficult. Therefore, we used a global attention model [4]. First, the attention weights are obtained by inputting training sentences into the global attention model. At this time, the input of the decoder is not the output word of the previous instance but the correct word. Next, we assume that the position of the highest weight among the attention weights obtained is to be concentrated in the input sentence. This position is different for each decoding time instance. The position of the current decoding time is compared with the position of the previous decoding time to obtain an increment. Next, we create new training data by adding a "<step>" action between the previous output word and the current output word using an increment. At this instance, it is assumed that the position does not decrease because the actions only serve to increase the position.

In the newly converted training data, there is a "<step>" action in the output sentence. Our proposed model learns to generate an output sentence containing the "<step>" action, and the local attention mechanism of the model learns by changing the position based on the correct "<step>" action.

4 | EXPERIMENTS

We used Sejong Korean morphological analysis data to evaluate our proposed model. The total data included 97,410 words, and the learning, development, and evaluation datasets were 88,225, 1,000, and 8,185 sentences, respectively. The sequence-to-sequence model for morphological analysis was designed as done in [4], source embedding was 200 dimensions, and all hidden layers had 1,000 dimensions. The window size d of the local attention mechanism was specified as {1, 2, 3}. The beam size of the beam search for high performance of the sequence-to-sequence model was fixed at 10. The training data of the proposed local soft attention model and the hard monotonic model were converted using the trained global attention model.

TABLE 2 Comparison of morphological F1-measures in terms of the models

	Dev	Test
Global	97.50	96.80
Global + copying	97.64	96.97
Local soft ($d = 1$) + copying	95.77	95.49
Local soft ($d = 2$) + copying	96.40	96.13
Local soft ($d = 3$) + copying	95.52	95.20
Hard monotonic + copying	95.79	95.34

Table 2 displays the morphological analysis performance of the proposed model. The global attention mechanism model (Global + copying in Table 2) performed best, with 96.97% of the morphometric F1-measure for the test set, and the local attention mechanism model (Table 2, local soft + copying) showed a performance decrease of 0.84%–1.77%. However, in the case of the hard monotonic + copying model ($d = 0$) that removes the attention mechanism formula from the local soft + copying model and uses only one input word, this method seems to use only relevant information in Korean morphological analysis.

To confirm the improvement of the speed performance of the proposed local attention mechanism, a single-threaded CPU (Intel (R) Core (TM) i7-6700 CPU @ 3.40 GHz) that cannot use the GPU environment (NVIDIA GTX 1070) was used to measure speed.

Table 3 shows the execution time results for each model. The execution time of each model was measured by randomly extracting 10 sentences from the test set. The local soft + copying model demonstrated a 6.79× improvement in speed in the single-threaded CPU environment over the global + copying model, although it was rather slow in the GPU environment due to the output sequence length being increased by 63% through the addition of the "<step>" action to the output word dictionary. In addition, at the pure decoder

TABLE 3 Average data length and execution time in CPU, GPU environments

	Output length (average syllable)	CPU (s/sent)	GPU (s/sent)
Input length (average syllable): 73.5			
Global + copying	115.6	234.27	2.65
Local soft ($d = 1$) + copying	187.8	34.45	2.95
Local soft ($d = 2$) + copying	188.9	34.48	2.88
Local soft ($d = 3$) + copying	187.4	52.95	3.10
Hard monotonic + copying	187.6	19.72	1.93

speed, there was a speed improvement of approximately 50% on the GPU. In the GPU environment, the difference in implementation of the attention mechanism seems to be due to the processing of the input sequence score alignment calculated at each decoding time instance in the form of a matrix operation. However, in an environment where the GPU cannot be used, it can be expected to increase the effective speed.

Figure 2 shows an example of Korean morphological analysis using a sequence-to-sequence model. Figure 2A shows the alignment of the global model, indicating that some input contextual context information is used in morphological analysis. Figure 2B and 2C show the soft attention ($d = 2$) and alignment of the hard attention model of the local model. The "<step>" action, which moves the focus position

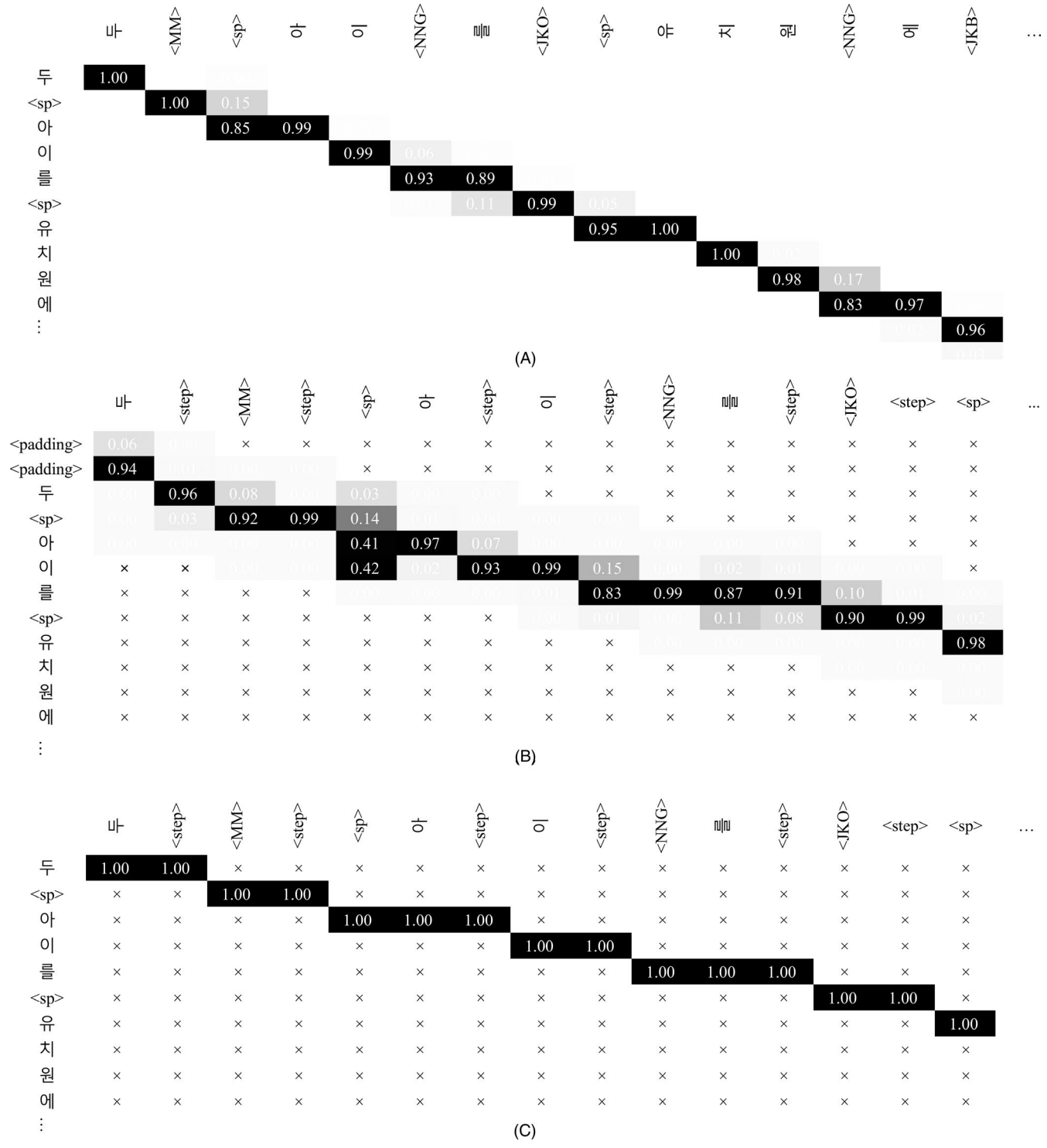

FIGURE 2 Example of Korean morphological analysis alignment: (A) Global model, (B) soft attention model, and (C) hard attention model

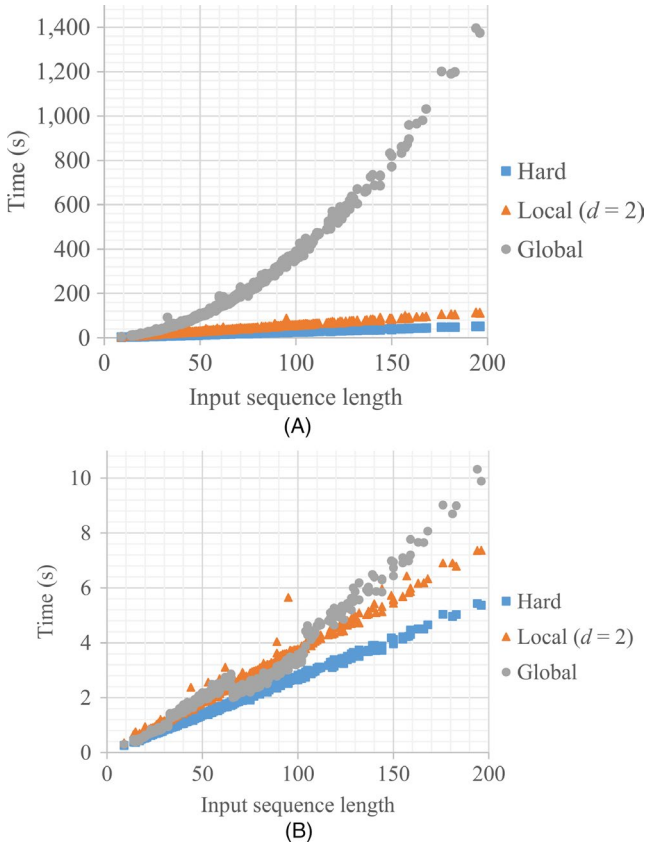

FIGURE 3 Execution time by sentence length: (A) CPU and (B) GPU environments

information, works excellently and focuses on the position similar to the global model. The hard attention model with the limit of the context window is also well focused, indicating that the model is efficient for morphological analysis.

Figure 3 shows the amount of time spent by the input sentence length when processing the development set data in the CPU (Figure 3A) and GPU (Figure 3B) environments. As displayed, the global model has a nonlinear time complexity with respect to the length of the input sentence, but it is found that the proposed model, local ($d = 2$), has linear time complexity. Compared to the GPU environment, the speed difference is less in the CPU environment. However, as the length of the input sentence increases, the global model still has nonlinear time complexity.

5 | CONCLUSION

In this study, we propose a linear-time Korean morphological analysis model using an action-based local monotonic attention mechanism using the monotonic alignment characteristic of morphological analysis. The experimental results demonstrate that the performance of the local model drops by 0.84%–1.77% compared to the global model. The hard attention mechanism model with the attention mechanism formula removed shows a performance drop of 1.72%, minimizing and dramatically improving speed. However, all proposed models have the disadvantage that the output sequence length is increased by adding a "<step>" action to it. In [11], a monotonic attention model that separated the action selector was proposed. This model was able to train end-to-end without the help of the trained global attention model. In future research, we aim to study on reducing the length of the output sequence effectively by processing the position information to be concentrated on in parallel.

ACKNOWLEDGEMENTS

This study was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea Government (MSIT) (2018-0-00605, Artificial Intelligence Contact Center Solution) and Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea Government (MSIT) (2013-0-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services).

ORCID

Hyunsun Hwang https://orcid.org/0000-0001-8651-0244

REFERENCES

1. D.-G. Lee and H.-C. Rim, *Probabilistic models for Korean morphological analysis*, in Proc. Int. Joint Conf. Natural Language Process., Jeju Island, Rep. of Korea, Oct. 2005, pp. 197–202.
2. K. Cho et al., *Learning phrase representations using RNN encoder-decoder for statistical machine translation*, arXiv preprint, 2014, arXiv:1406.1078.
3. I. Sutskever, O. Vinyals, and Q.V. Le, *Sequence to sequence learning with neural networks*, Adv. Neural Inf. Process. Syst. **27** (2014), 3104–3112.
4. S. Jung, C. Lee, and H. Hwang, *End-to-end Korean part-of-speech tagging using copying mechanism*, ACM Trans. Asian Low-Resource Language Inf. Process. **17** (2018) no. 3, 19:1–28.
5. D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, arXiv preprint, 2014, arXiv:1409.0473.
6. G. Jiatao et al., *Incorporating copying mechanism in sequence-to-sequence learning*, arXiv preprint, 2016, arXiv:1603.06393.
7. S.-H. Na, *Conditional random fields for Korean morpheme segmentation and POS tagging*, ACM Trans. Asian Low-Resource Language Inf. Process. **14** (2015) no. 3, 10:1–10.
8. C. Lee, *Joint models for Korean word spacing and POS tagging using structural SVM*, J. Korean Inf. Sci. Soc.: Softw. Applicat. **40** (2013), no. 12, 826–832. (in Korean).
9. M.-T. Luong, H. Pham, and C.D. Manning, *Effective approaches to attention-based neural machine translation*, arXiv preprint, 2015, arXiv:1508.04025.

10. R. Aharoni and Y. Goldberg, *Morphological inflection generation with hard monotonic attention*, arXiv preprint, 2016, arXiv:1611.01487.
11. C.C. Chiu and C. Raffel, *Monotonic chunkwise attention*, in Proc. Int. Conf. Learning Representations, Vancouver, Canada, 2018, pp. 1–16.

AUTHOR BIOGRAPHIES

Hyunsun Hwang received his BS and MS degrees in computer science from Kangwon National University, Chuncheon, Rep. of Korea, from 2010 to 2017. He is now a PhD student at Kangwon National University. His research interests include natural language processing, deep learning, word embedding, information extraction, and dialogue systems.

Changki Lee received his BS degree in computer science from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 1999. He received his MS degree and PhD in computer engineering from POSTECH, Pohang, Rep. of Korea, in 2001 and 2004, respectively. From 2004 to 2012, he was a researcher with the Electronics and Technology Research Institute, Daejeon, Rep. of Korea. Since 2012, he has been a professor of computer science at Kangwon National University, Chuncheon, Rep. of Korea. His research interests include natural language processing, machine learning, and deep learning.