



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공학석사 학위논문

종속절 및 형태소 분리를 이용한  
한국어 개체명 인식

Korean Named Entity Recognition  
Using Subordinate Clause and Morpheme Segmentation

지도교수 김재훈

2020 년 2 월

한국해양대학교 대학원

컴퓨터공학과  
윤 호

본 논문을 윤 호의 공학석사 학위논문으로 인준함

위원장 박 휴 찬 (인)

위 원 류 길 수 (인)

위 원 김 재 훈 (인)

2019년 12월 26일

한국해양대학교 대학원

# 목 차

<b>List of Tables</b> .....	<b>iv</b>
<b>List of Figures</b> .....	<b>v</b>
<b>Abstract</b> .....	<b>vi</b>
<b>초록</b> .....	<b>viii</b>
<b>제 1 장 서론</b> .....	<b>1</b>
<b>제 2 장 관련 연구</b> .....	<b>4</b>
2.1 개체명 인식 .....	4
2.2 형태소 분리와 부분단어 .....	7
2.3 종속절 분리 .....	8
<b>제 3 장 종속절 및 형태소 분리를 이용한 한국어 개체명 인식</b> .....	<b>10</b>
3.1 종속절 분리 .....	11
3.1.1 종속절 분리 학습말뭉치 제작 .....	13
3.1.2 종속절 분리 모델 .....	15
3.2 형태소 분리 .....	16
3.2.1 부분단어 사전 구축 .....	17
3.2.2 형태소 분리 모델 .....	18
3.3 개체명 인식 .....	22
<b>제 4 장 실험 및 평가</b> .....	<b>24</b>
4.1 종속절 분리 .....	24
4.1.1 실험 환경 및 실험 척도 .....	24
4.1.2 실험 결과 및 분석 .....	26

4.2 형태소 분리 .....	27
4.2.1 실험 환경 및 실험 척도 .....	27
4.2.2 실험 결과 및 분석 .....	28
4.3 개체명 인식 .....	31
4.3.1 실험 환경 및 실험 척도 .....	31
4.3.2 실험 결과 및 분석 .....	30
<b>제 5 장 결론 및 향후 연구 .....</b>	<b>34</b>
<b>참고문헌 .....</b>	<b>36</b>
<b>감사의 글 .....</b>	<b>42</b>



## List of Tables

<b>Table 3.1</b>	Kinds of subordinate connective endings .....	12
<b>Table 3.2</b>	The criterion for segmenting subordinate clauses .....	14
<b>Table 4.1</b>	Statistics of training corpus for segmenting subordinate clauses ...	24
<b>Table 4.2</b>	Parameters of Bi-LSTM/CRF for segmenting subordinate clauses	25
<b>Table 4.3</b>	2×2 contingency table for comparing the performance .....	25
<b>Table 4.4</b>	Performance of Bi-LSTM/CRF in subordinate clauses .....	27
<b>Table 4.5</b>	Statistics of the training corpus for morpheme segmentation .....	27
<b>Table 4.6</b>	Parameters of the transformer model .....	28
<b>Table 4.7</b>	Performance of the transformer model for segmenting morphemes	29
<b>Table 4.8</b>	Distribution of sentence lengths in test corpus for morpheme segmentation .....	29
<b>Table 4.9</b>	Types of errors of morpheme segmentation .....	30
<b>Table 4.10</b>	Training corpus for Korean named entity recognition .....	32
<b>Table 4.11</b>	Parameters of Bi-LSTM/CRF for Korean named entity recognition .....	32
<b>Table 4.12</b>	Performance of Korean named entity recognition .....	33

## List of Figures

<b>Figure 3.1</b> The structure of the proposed system for Korean named entity recognition using subordinate clause and morpheme segmentation .....	10
<b>Figure 3.2</b> An example sentence with subordinate clauses .....	11
<b>Figure 3.3</b> The process of building a training corpus for segmenting subordinate clauses .....	13
<b>Figure 3.4</b> The algorithm for checking subordinate clauses .....	14
<b>Figure 3.5</b> The structure of Bi-LSTM/CRF for segmenting subordinate clauses .....	16
<b>Figure 3.6</b> The process of building a subword dictionary .....	17
<b>Figure 3.7</b> The preprocessor of morpheme segmentation .....	18
<b>Figure 3.8</b> The structure of the transformer for morpheme segmentation .....	20
<b>Figure 3.9</b> The structure of Bi-LSTM/CRF for Korean NER based on morphemes as input .....	22
<b>Figure 3.10</b> The structure of Bi-LSTM/CRF for Korean NRE based on subwords as input .....	23

# **Korean Named Entity Recognition**

## **Using Subordinate Clause and Morpheme Segmentation**

Yoon, Ho

Department of Computer Engineering  
Graduate School of Korea Maritime and Ocean University

### **Abstract**

Named entity recognition (NER) is a subtask that seeks to locate and classify named entities in a given document into pre-defined categories such as person names, organizations, locations, and so on. NER can be applied to many applications related to natural language processing such as document summarization, question answering, machine translation, and chatbot etc. There is a notorious problem in NER called out-of-vocabulary (OOV). Many previous works have tackled the problem through extension of training corpus and various word representation in deep learning. In addition, most Korean NER systems have used morphological analysis as preprocessors, but Korean morphological analysis has the same problem of OOV of which errors are propagated to the NER system and cause the performance to deteriorate further.

In order to alleviate the problem, we propose a novel method for Korean NER using subordinate clause and subword segmentation. The proposal



method consists of three steps. The first step is to segment subordinate clauses from a given sentence using a recurrent neural network (RNN), especially Bi-LSTM/CRF. The second step is to segment morphemes from the segmented clauses using the Transformer model developed by Google. The model takes subwords as input in order to mitigate the OOV problem. The third step is to assign the most proper BIO tag to each morpheme using Bi-LSTM/CRF of RNNs. Through experiments, the proposed steps of subordinate clause and morpheme segmentation have been evaluated, achieving F1-scores of about 95% and 98%, respectively. For the proposed NER, experimental results show that our word outperforms the other Korean NER models, carrying out F1-score of about 72%.

In the future, we will do research on more accurate morpheme segmentation using the Transformer model with copy mechanism and also on subordinate clause segmentation or subsentence segmentation in linguistics.

KEY WORDS: Subordinate clauses segmentation, Morpheme segmentation, Named entity recognition

# 종속절 및 형태소 분리를 이용한 한국어 개체명 인식

윤 호

한국해양대학교 대학원  
컴퓨터공학과

초록

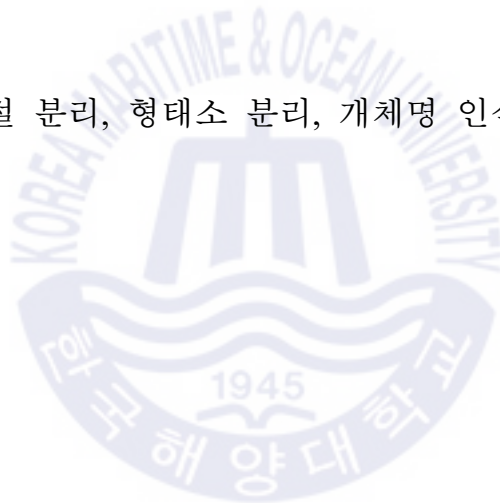
개체명 인식이란 주어진 문서에서 개체명의 범위를 찾고 개체명의 범주를 결정하는 것이다. 개체명 인식은 문서 요약, 질의응답, 기계번역, 잡담 처리와 같은 자연언어처리 전반에 사용된다. 개체명 인식의 고질적인 문제점 중 하나는 미등록어(out-of-vocabulary) 문제이다. 기존의 한국어 개체명 인식의 입력은 주로 형태소 분석 결과이다. 이 경우에는 미등록어로 발생된 오류가 개체명 인식에 그대로 전파되므로 여전히 미등록어로 인해 발생하는 문제가 완전히 해소되지 않는다.

이와 같은 문제를 다소 완화하기 위해 본 논문에서는 종속절 및 형태소 분리를 이용한 한국어 개체명 인식 방법을 제안한다. 제안된 방법은 세 단계로 구성된다. 첫 번째 단계는 종속절 분리 단계이며 순환신경망을 이

용하여 입력된 문장을 종속절 단위로 분리한다. 두 번째 단계는 형태소 분리 단계이며, Transformer 모델을 이용하여 각 종속절의 형태소를 분리한다. 이때 미등록어 문제를 완화하려고 형태소 분리 모델(Transformer 모델)의 입력으로 부분단어 정보를 이용한다. 세 번째 단계는 개체명 인식 단계이며, 순환신경망을 이용해서 분리된 형태소에 개체명 표지를 부착한다.

제안된 방법을 통하여 문장 분리에서는 95%의 문장 분리 정확률을 나타냈으며, 형태소 분리에서는 90%의 F1-점수를 나타내었으나 한글맞춤법을 고려할 경우 98.3%의 정확률을 보였다. 개체명 인식의 경우 72%의 F1-점수를 나타내었다. 위의 결과를 통해 제안하는 방법이 기존의 방법보다 성능이 우수함을 알 수 있다.

KEY WORDS: 종속절 분리, 형태소 분리, 개체명 인식



## 제 1 장 서 론

개체명이란 문서에서 나타나는 인명, 지명, 조직명, 시간, 날짜, 화폐 등 고유한 의미를 가지는 단어를 말한다. 개체명 인식이란 문서에 나타나는 개체명을 찾고 그에 해당하는 표지를 부착하는 작업을 말한다. 개체명 인식은 질의응답(구교정 외, 2018), 기계번역(이원기 외, 2017), 잡담처리(이창수 & 고영중, 2014)와 같은 자연언어처리에 두루 응용된다. 개체명 인식은 단어로 구성된 개체명이 문맥에 따라 다른 개체명으로 해석될 수 있는 중의성 문제와 시간의 흐름에 따라 새롭게 생성되는 미등록어(out-of-vocabulary) 문제 등을 가지고 있다(김재훈 외, 2010). 중의성 문제는 문맥을 반영한 자질을 추가하므로 다소 완화되었으며(이창기 외, 2006), 미등록어 문제는 말뭉치를 확장하여 일부 개선되었다(오교중 외, 2017). 하지만 미등록어는 시간에 따라 지속적으로 생성되므로 말뭉치도 또한 지속적으로 확장되어야 한다.

한국어 개체명 인식에서 미등록어 문제를 완화하기 위해서 입력 단위를 음절 단위로 분리해서 음절 단위에 개체명을 부여하는 방법이 연구되었다(천민아 외, 2018). 음절 단위의 개체명 인식 방법은 미등록어 문제에 유연하지만 입력 문장의 문맥을 반영하기 어렵다는 문제가 발생된다. 입력 문장의 문맥을 반영하기 위해서 한국어 개체명 인식에서는 입력 단위를 어절<sup>1)</sup> 또는 형태소로 지정하여 개체명을 인식하였다(정래정 & 김준태, 1996). 어절 단위 개체명 인식 방법은 어절 자체가 개체명이 될 수 없고 어절 안에서 다시 개체명의 범위를 찾아야 한다. 형태소 단위 개체명 인식 방법은 입력 문장에 대해 형태소를 분석한 후, 분석된 형태소열에 대

---

1) <https://github.com/naver/nlp-challenge>

해서 개체명을 인식한다. 형태소 단위 개체명 인식 방법은 입력 문장을 형태소로 분리하여 문맥을 어느 정도 반영할 수 있고, 어절 내에서 개체명의 범위를 별도로 찾을 필요가 없다. 하지만 형태소 단위 개체명 인식 방법의 문제점은 형태소 분석에서 미등록어로 발생하는 오류가 개체명 인식으로 그대로 전달된다는 것이다.

이러한 문제를 다소 완화하기 위해 본 논문에서는 문맥을 최대한 반영하면서 미등록어 문제를 완화하는 개체명 인식 방법을 제안한다. 제안된 방법은 먼저 Transformer 모델(Vaswani *et al.*, 2017)을 이용하여 형태소를 분리한다. Transformer 모델은 인코더-디코더 모델(Sequence-to-Sequence) (Sutskever *et al.*, 2014)의 일종이며, 이 모델의 문제점은 인코더는 입력열의 모든 정보를 하나의 벡터로 압축하므로 모든 입력열의 정보를 충분히 표현할 수 없다는 문제가 발생된다는 것이다. 이와 같은 문제로 입력 문장의 길이가 길어지면 디코더에서 문장의 일부가 소실되거나 똑같은 단어를 반복하는 문제가 생긴다(Cho *et al.*, 2014).

이와 같은 정보 손실 문제를 완화하기 위해 본 논문에서는 주어진 입력 문장을 먼저 종속절로 분리하여 형태소를 분리하고 그 결과로부터 개체명을 인식하는 모델을 제안한다. 제안된 모델은 크게 세 단계로 구성된다. 첫 번째 단계는 종속절 분리 단계이며 주어진 입력 문장을 종속절 단위로 분리한다. 일반적으로 개체명은 종속적 연결어미를 포함하지 않으므로 종속절로 분리하여 개체명을 인식해도 문제가 없다. 두 번째 단계는 형태소 분리 단계이며 분리된 종속절을 입력으로 받아서 형태소를 분리한다. 이 단계에서 미등록어 문제를 완화하기 위해 부분단어(subword) 정보를 활용한다. 마지막 단계는 개체명 인식 단계이며 분리된 형태소를 입력으로 받아서 형태소에 개체명 표지를 부착한다. 모든 단계에서 심층학습 모델을 사용한다. 종속절 분리와 개체명 인식에서는 순환신경망 모델을 사용하고 형태소 분리에서는 Transformer 모델을 사용한다. 개체명 인식 말뭉치를 대상으로 실험하여 72%의 F1-점수를 보였으며 다른 모델들에 비해 높은 성능을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 개체명 인식 연구의 전반적인 내용과 형태소 분리와 부분단어, 종속절 분리에 관련된 기존 연구들을 설명한다. 3장에서는 본 논문에서 제안하는 종속절 및 형태소 분리를 이용한 한국어 개체명 인식 방법을 설명한다. 4장에서 실험을 통해서 제안된 방법과 기존의 방법의 성능을 비교하고 분석한다. 마지막으로 5장에서 결론을 맺고 향후 연구 방향을 제시한다.



## 제 2 장 관련 연구

이 장에서는 본 논문에서 제안하는 종속절 및 형태소 분리를 이용한 개체명 인식과 관련하여 개체명 인식, 형태소 분리 및 부분단어, 종속절에 대한 기존 연구들을 알아본다. 2.1절에서는 개체명 인식과 기존 연구에 대해 알아본다. 2.2절에서는 형태소 분리와 부분단어에 대한 연구들을 살펴보고, 2.3절에서는 종속절에 대해 설명한다.

### 2.1 개체명 인식

우리는 오늘날 빅데이터(big data)의 시대에 살고 있다. 넘쳐나는 정보 속에서 원하는 정보를 찾는다는 것은 쉬운 일이 아니다. 문서로부터 원하는 정보를 찾기 위해서 정보 추출 시스템을 이용한다. 정보 추출 시스템이란 신문 기사, 웹 문서, 전자우편 등과 같이 정형화되지 않은 문서를 입력으로 받아서 미리 정해놓은(찾기를 원하는) 정보를 찾아내는 시스템이다(김재훈, 2004). 많은 경우, 정보 추출의 대상은 개체명이다. 개체명(named entity)은 문서에서 등장하는 인명, 기관명, 지명 등을 뜻한다(Grishman & Sundheim, 1996). 주어진 문서에서 이와 같은 개체명을 찾아내는 과정을 개체명 인식이라고 한다. 개체명 인식은 크게 규칙 기반 방법(Humphreys et al., 1998), 통계 기반 방법(Bikel et al., 1999), 심층학습 기반 방법(Collobert et al., 2011)으로 분류된다.

#### (1) 규칙 기반 개체명 인식

규칙 기반 개체명 인식은 규칙과 사전을 이용하여 개체명 인식을 수행한다. 개체명 인식에서 사용되는 규칙은 사람이 개체명에서 자주 나타나



는 패턴을 보고 손으로 작성된다. 이런 규칙에는 개체명의 후보가 되는 단어 자체의 정보를 이용하는 단어 구성 규칙과 문장에서 개체명의 후보가 되는 단어의 주변 정보를 이용하는 문맥 규칙 등이 있다(Mikheev *et al.*, 1998). 개체명 인식에서 사용되는 사전은 개체명에 대한 직접적인 정보를 가지고 있는 개체명 사전과 개체명과 같이 붙어서 나타나는 단어들을 포함하는 결합 단어 사전 등이다(이경희 외, 2000). 한국어 개체명 인식에서는 어절 내의 단어 정보, 제한된 주변 문맥 정보, 용언의 하위범주화 정보와 개체명과의 관계, 개체명 간의 관계 정보를 고려한 네 단계로 이루어진 규칙 기반 개체명 인식 방법이 제안되었다(이경희 외, 2000).

## (2) 통계 기반 개체명 인식

통계 기반 개체명 인식 모델은 HMM(Hidden Markov Model)(Eddy, 1996), MEM(Maxium Entropy Model)(Kapur, 1989), SVM(Support Vector Machine)(Hearst *et al.*, 1998), CRF(Conditional Random Field)(Lafferty *et al.*, 2001) 등이 있다.

HMM은 WFSA(Weighted Finite-State Automata)의 일종으로 각 상태와 각 전이에 확률이 추가되어 있다(Rabiner, 1989). HMM 모델을 이용한 개체명 인식 방법으로는 단어, 단어 내에 있는 특수한 기호, 지명사전, 문맥 등 여러 가지 자질을 사용한 개체명 인식 방법이 제안되었고(Zhou & Su, 2002), 음절과 음절 n-gram을 자질로 이용한 개체명 인식 방법도 연구되었다(Klein *et al.*, 2003).

MEM 모델을 이용한 개체명 인식은 사람의 지식을 이용한 자질을 추가하여 개체명을 인식하는 방법이 연구되었고(Borthwick, 1999), 전체 말뭉치에서 추출한 unigram, bigram, 자주 등장하는 단어와 기능어와 같은 자질을 사용한 개체명 인식 연구도 진행되었다(Chieu & Ng, 2003).

SVM을 이용한 개체명 인식은 기존의 SVM을 개체명 인식에 맞게 효율적으로 개선한 모델을 제안하였다(Isozaki & Kazawa, 2002). CRF 모델을 이용한 연구로는 의미적 지식을 포함한 사전 자질을 사용한 개체명 인식



도 연구되었으며(Settles, 2004) 데이터 전처리를 통해서 성능을 향상한 모델도 제안되었다(Gliozzo *et al.*, 2005).

한국어 개체명 인식에서는 주로 CRF 모델과 SVM를 이용해서 연구가 이루어졌다. CRF를 이용한 연구로는 개체명 경계 인식에 CRF를 사용하고 경계 인식된 개체명의 클래스 분류에는 MEM을 사용한 연구도 진행되었고(이창기 외, 2006), 특히 개체명을 자동으로 인식하기 위해 CRF 기법을 사용한 연구도 진행되었다(이태석 외, 2016). SVM을 이용한 연구로는 입력을 형태소열로 형태소열과 품사에 대한 자질을 추출하여 SSVM(Structural SVM)와 수정된 Pegasos 알고리즘을 이용하여 한국어 개체명 인식을 연구하였다(이창기 & 장명길, 2010).

### (3) 심층학습 기반 개체명 인식

심층학습 기반 개체명 인식은 크게 합성곱신경망(CNN) 방법과 순환신경망(RNN) 방법을 이용하여 개체명 인식을 수행한다. 합성곱신경망을 이용한 개체명 인식 연구로는 입력이 단어인 문장에 대해 각각에 단어에 대해서 자질을 추출하고, 합성곱신경망을 이용하여 개체명을 인식하는 모델이 연구되었다(Collobert *et al.*, 2011). 한국어 개체명 인식에서는 입력 단위에 따라 형태소 기반 합성곱신경망을 이용해 개체명을 인식하는 방법이 제안되었고(유연수 & 박혁로, 2019), 음절 기반 합성곱신경망을 이용한 개체명 인식 모델이 제안되었다(박혜웅 & 송영숙, 2017).

순환신경망은 심층학습방법 중에서 순차적인 데이터를 학습하여 분류해 내는 작업에 특화되어 있는 심층학습방법이다(Schmidhuber, 1993). 순환신경망은 식 (2.1)과 같이 정의된다.

$$h_t = \sigma_h ( W_h x_t + U_h h_{t-1} + b_h )$$

$$y_t = \sigma_y (W_y h_t + b_y) \quad (2.1)$$

여기서,  $x_t$ 와  $h_t$ 는 각각 입력층과 은닉층을 나타내고,  $y_t$ 는 출력층을 나타낸다.  $W_h, W_y, U_h$ 은 가중치 행렬이며,  $b_h$ 는 편향(bias)이다.  $\sigma_h$ 와  $\sigma_y$ 는

각각 입력층과 은닉층의 활성화 함수(activation function)이다. 순환신경망을 이용한 개체명 인식은 입력 표상에 따라 다양한 연구가 제안되었다. 입력 표상이 단어인 경우는 주로 Bi-LSTM/CRF(Bidirectional LSTM/CRF) 모델을 사용한다(Huang *et al.*, 2015). 입력 표상이 음절일 경우는 LSTM을 이용하여 단어 표상을 만들어 Bi-LSTM/CRF에 적용하였다(Lample *et al.*, 2016). 입력 표상이 음절일 경우 또 다른 방법으로는 합성곱신경망과 기존의 단어 표상과 합쳐서 입력 표상을 만든 뒤, 순환신경망을 사용한 방법도 제안되었다(Ma & Hovy, 2016).

한국어 개체명 인식에서도 입력 표상을 만드는 방법에 따라 다양한 연구가 제안되었다. 순환신경망의 입력 표상을 형태소와 품사 그리고 음절에 대한 LSTM으로 확장하여 순환신경망에 적용시킨 방법이 제안되었다(유홍연 & 고영중, 2016). 형태소, 자음/모음 특징, 품사, 기구축 사전의 네 가지 입력을 하나의 입력 표상으로 만들어서 순환신경망의 입력으로 넣는 방법도 연구되었다(신유현 & 이상구, 2016). 단어가 가지는 개체명 비율을 결합하여 입력 표상을 확장한 Bi-LSTM/CRF 모델도 제안되었다(박동주 & 안창욱, 2019).

## 2.2 형태소 분리와 부분단어

한국어 개체명 인식기의 대부분은 형태소 분석의 결과를 입력으로 사용한다. 형태소 분석은 주어진 문장을 형태소로 분리하여 각 형태소에 품사를 부착하는 과정이다. 즉 형태소 분석은 형태소 분리와 품사 부착의 두 가지 단계로 이루어진다. 여기서 형태소 분리는 한 어절 내에 포함된 가능한 모든 형태소를 분리하는 것이며, 품사 부착은 분리된 형태소에 알맞은 품사를 부착하는 것을 말한다. 한국어 개체명 인식에는 반드시 형태소의 품사 정보가 필요하지 않으므로 효율적인 처리를 위해서 본 논문에서는 형태소 분리의 결과를 개체명 인식의 입력으로 사용하는 방법을 제안한다.

부분단어는 자연언어처리 분야 중 기계번역에서 먼저 연구되었다. 최근 기계번역의 성능이 크게 향상되었지만 저빈도어(rare word)에 대해서는 성능이 그다지 좋지 않다. 기계번역에서 저빈도어 문제를 다루기 위해 원문에 있는 단어를 번역 문장에 그대로 사용하는 모델이 제안되었고(Luong *et al.*, 2015), 주의 집중 모델을 기반으로 단어를 그대로 번역 문장에서 사용하는 방법도 제안되었다(Jean *et al.*, 2015). 위의 연구들은 기계번역에서 번역되는 단어와 번역할 대상이 되는 단어가 1:1 대응이 된다는 가정 하에는 번역의 품질이 좋게 동작하지만 실제로 문장을 번역할 때는 1:1 대응이 되지 않는 경우가 많다. 이런 단점들을 해결하기 위해 기계번역 문제에서 출력의 단위 또는 입력의 단위를 변경한 연구들이 진행되었다. 출력의 단위를 변경한 연구로는 출력 단위를 문자로 하는 방법이 제안되었고(Chung *et al.*, 2016), 출력 단위를 단어로 하되 미등록어에 대해서 문자로 출력하는 연구도 진행되었다(Luong *et al.*, 2016).

입력의 단위를 변경한 연구로는 말뭉치 내에서 등장하는 부분단어를 이용하여 문장을 분리하는 방법이 제안되었다(Sennrich *et al.*, 2016). 이 방법은 BPE(Byte Pair Encoding)(Gage, 1994)를 이용하여 학습말뭉치를 부분단어로 분절하여 사용하여 기계번역의 성능을 개선하였다. 또한 BPE 대신 WPM(Word Piece Model)을 이용한 방법도 제안되었다(Yonghui *et al.*, 2016). WPM은 BPE와 유사하지만 최적화 알고리즘을 이용해서 학습 데이터의 우도(likelihood)가 증가되는 부분단어를 선택한다. 또 다른 방법으로는 기존의 BPE와 달리 언어 모델(Bengio *et al.*, 2003)을 이용하는 ULM(Unigram Language Model)도 제안되었다(Kudo, 2018; Kudo & Richardson, 2018). WPM은 학습 데이터의 우도를 증가하는 부분단어를 선택하지만 ULM은 전체 데이터의 혼잡도(perplexity)가 감소되는 부분단어를 선택한다.

## 2.3 종속절 분리

종속절이란 한 문장에서 종속의 성분을 이루는 절이며 두 개의 절이 하나의 문장을 이룰 때 주절을 한정하는 절이다.

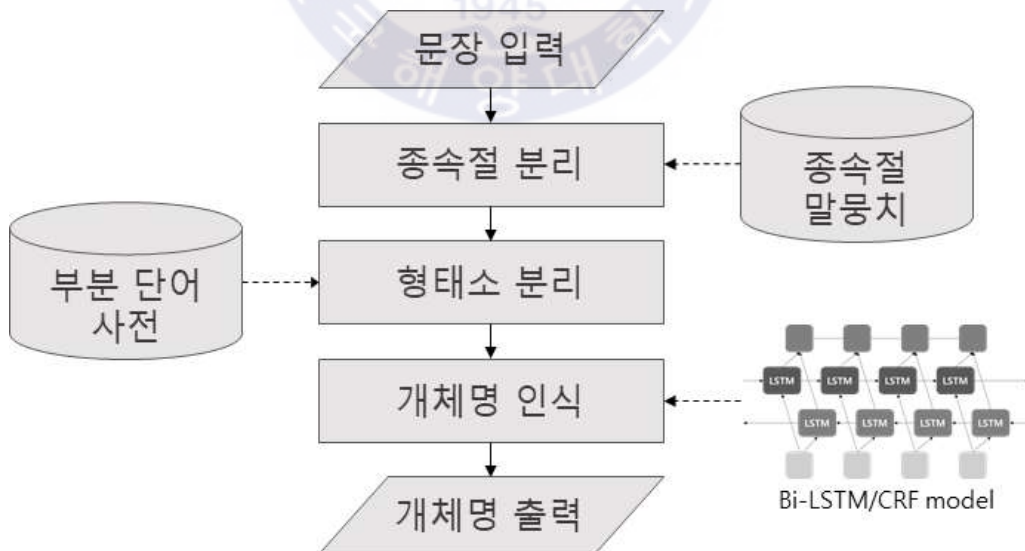
한국어 문장은 크게 홑문장과 겹문장으로 나뉜다. 홑문장은 문장 하나에 주어와 서술어가 각각 한 번만 나타나는 문장이다. 겹문장은 주어와 서술어가 두 번 이상 나타나는 문장을 말한다. 이때 하나의 주어와 서술어가 포함된 단위를 절이라고 한다. 겹문장은 문장에서 중심 역할을 하는 주절과 그 외의 절들로 이루어진다. 문장과 문장은 연결어미에 의해서 이어지며, 연결어미에는 대등적 연결어미, 종속적 연결어미 등이 있다. 대등적 연결어미로는 ‘고, 며, 자 거나’ 등이 있으며, 종속적 연결어미로는 ‘면, 니, 지만’ 등이 있다. 이때 대등적 연결어미로 이어진 문장을 대등절이라고 하며, 종속적 연결어미로 이어진 문장을 종속절이라 한다.

종속절 분리와 관련된 연구는 찾아볼 수 없으나 본 논문에서는 종속절 및 형태소 분리를 이용한 한국어 개체명 인식을 수행한다.



### 제 3 장 종속절 및 형태소 분리를 이용한 한국어 개체명 인식

본 논문은 종속절 및 형태소 분리를 이용한 한국어 개체명 인식 방법을 제안한다. 제안된 시스템의 전체 구조는 Figure 3.1과 같으며 종속절 분리, 형태소 분리, 개체명 인식의 세 단계로 진행된다. 종속절 분리는 주어진 문장을 종속절로 분리한다. 종속절을 분리하면 일부의 문맥정보가 손실될 수 있으나 순환신경망에서 문장이 길어질 때 발생하는 문제를 최소화하기 위해서 종속절로 분리한다. 형태소 분리는 분리된 문장을 형태소로 분리한다. 이때 미등록어 문제를 완화하기 위해 부분단어 정보를 이용한다. 개체명 인식에서는 생성된 형태소열에 대해서 한국어 개체명 인식을 수행한다.



**Figure 3.1** The structure of the proposed system for Korean named entity recognition using subordinate clause and morpheme segmentation.



이하의 절에서는 종속절 분리 학습말뭉치 제작 방법과 종속절 분리 모델을 기술하고 부분단어 사전 구축방법과 형태소 분리 모델을 설명한다. 그리고 나서 개체명 인식에 대해서 자세히 기술한다.

### 3.1 종속절 분리

종속절 분리는 주어진 문장을 종속절로 분리하는 것이다. 종속절은 주절을 한정하는 절이다. 예를 들면 Figure 3.2는 하나의 문장에 두 개의 절이 포함된 복문으로 앞의 절은 종속절이고, 뒤의 절은 주절이다.

철수는 서울로 갔으니,

철호는 부산으로 가라.

종속절(이유)

주절

**Figure 3.2** An example sentence with subordinate clauses.

종속절을 분리하는 이유는 3.2절에서 기술될 형태소 분리 단계에서 쓰이는 신경망 모델인 Transformer 모델에서 기인한다. Transformer 모델은 인코더-디코더 모델의 일종으로 입력 문장이 길어질 때, 특정 단어만을 반복해서 출력하거나 입력 성분 중 일부가 사라진다는 문제가 있다(Cho *et al.*, 2014). 이러한 문제를 완화하기 위해 입력 문장의 길이를 짧게 할 필요가 있다. 따라서 문장의 길이를 짧게 하기 위하여 본 논문에서는 종속절 분리를 제안한다. 종속절이 분리되더라도 주절과 종속절은 하나의 주어와 서술어를 가지는 독립적인 문장을 이루므로 본래의 의미가 크게 소실되지 않는다. 그러므로 형태소 분리 과정에서 문장의 의미를 고려해서 형태소를 분리할 수 있다.

종속절을 분리하려면 이어진 문장의 이해가 필요하다. 이어진 문장이란 둘 이상의 홑문장이 연결어미에 의해 결합된 문장이며, 연결어미의 종류에 따라 대등하게 이어진 문장과 종속적으로 이어진 문장으로 나눌 수 있다(윤여탁 외, 2014). 대등하게 이어진 문장은 대등적 연결어미에 의하여 대등한 관계로 결합된 문장이며, 종속적으로 이어진 문장은 종속적 연결

어미를 붙여 주절에 종속적인 문장이다. 특별히 종속적으로 이어진 문장을 종속절이라고 하며, 종속절의 말미에는 종속적 연결어미가 쓰인다. 종속적 연결어미란 종속절이 주절에 대해 동시, 이유, 양보, 조건, 의도, 배경, 미침, 필연, 비유, 더욱, 결과, 전환 등의 의미 관계를 갖도록 만들어 주는 어미이며, 종속적 연결어미의 종류는 Table 3.1과 같다.

**Table 3.1** Kinds of subordinate connective endings.

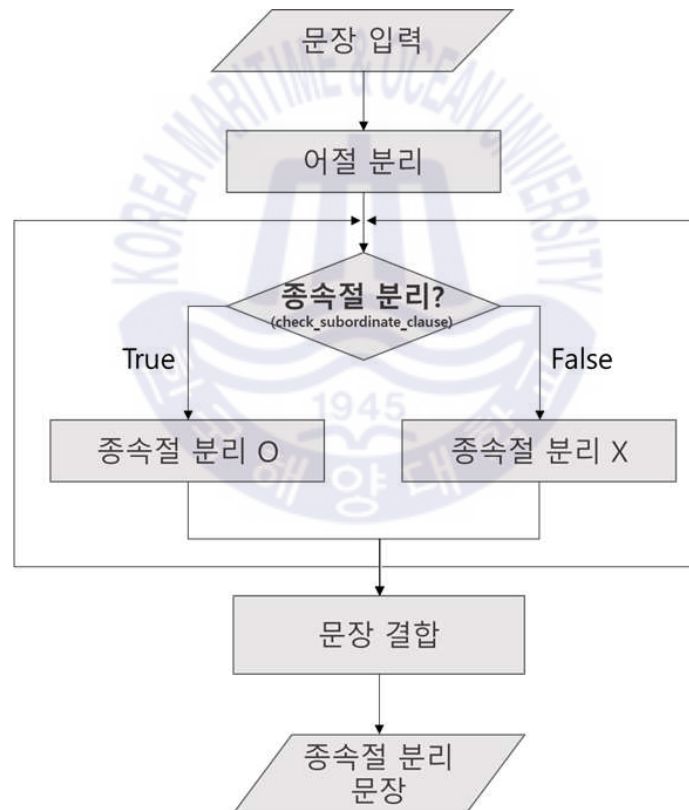
의미 관계	종속적 연결어미
동시	-자, -마자
이유(원인)	-아서, -어서, -으니까, -니까, -으므로, -므로, -느라고, -니
양보	-을망정, -르망정, -을지언정, -르지언정, -더라도, -은들, -니들, -아도, -어도
조건(가정)	-거든, -으면, -면, -라면
의도(목적)	-으러, -러, -고자, -으려고, -려고
배경	-는데, -은데, -니 데
미침	-게, -도록
필연(당위)	-아야, -어야, -야
전환	-다가, -다
비유	-듯, -듯이
더욱	-을수록, -르수록

본 논문에서는 종속절 분리 문제를 순차적 표지 부착(sequence labeling) 문제로 접근하여 심층학습 모델 중 하나인 Bi-LSTM/CRF 모델(Huang *et al.*, 2015)을 사용한다.

이하의 3.1.1절에서는 종속절 분리 모델을 학습시키기 위한 말뭉치를 제작하는 과정을 자세히 설명하고, 3.1.2절에서는 종속절 분리 모델의 구조와 입력 표상에 대해 기술할 것이다.

### 3.1.1 종속절 분리 학습말뭉치 제작

종속절 분리 모델을 학습시키기 위해서는 문장의 어느 지점에서 종속절을 분리할 수 있는지에 대한 정보가 포함된 학습말뭉치가 필요하다. 불행히도 이와 같은 말뭉치는 공개되어 있지 않아서 본 논문에서 세종 형태소 부착 말뭉치(김홍규 외, 2007)로부터 학습말뭉치를 구축한다. Figure 3.3은 그 학습말뭉치의 구축 과정을 보이고 있다.



**Figure 3.3** The process of building a training corpus for segmenting subordinate clauses.

Figure 3.3에서 보는 바와 같이 문장이 입력되면 먼저 문장을 어절로 분리하고, 각 분리된 어절에 대해서 종속절 여부를 판단하고 결과에 따라 문



장결합을 하여 학습말뭉치를 생성한다. 종속절 여부를 판단하는 알고리즘 (check\_subordinate\_clause)은 Figure 3.4와 같다. 이 알고리즘의 입력은 현재 어절(curr\_eojeol)과 다음 어절(next\_eojeol)이고, 출력은 주어진 어절이 종속 절로 분리되어야 하면 참(True)을 반환하고, 그렇지 않으면 거짓(False)을 반환한다.

```

def check_subordinate_clause(curr_eojeol, next_eojeol):
    if 'EC' not in curr_eojeol.pos: return False // 기준 (1)
    if ',' in curr_eojeol.morph: return True // 기준 (2)
    i = get_index(curr_eojeol.pos, 'EC')
    if curr_eojeol.morph[i] in 종속적_연결어미:
        if 'VX' in next_eojeol.pos: return False // 기준 (3)
        return True // 기준 (4)
    return False

```

**Figure 3.4** The algorithm for checking subordinate clauses.

Figure 3.4의 알고리즘은 최대한 Python 프로그래밍 언어의 문법을 따르고 있으며, 각 어절(curr\_eojeol, next\_eojeol)은 품사열(pos)과 형태소열(morph)로 구성되어 있다. 알고리즘에서 사용된 품사 표지 'EC'와 'VX'는 세종형태품사표지(김홍규 외, 2007)를 따른다. 또한 함수 get\_index(pos, 'EC')는 품사열(pos)에서 품사('EC')가 있는 색인을 반환한다. 이 알고리즘을 요약하면, Table 3.2와 같은 기능을 수행한다.

**Table 3.2** The criterion for segmenting subordinate clauses.

순번	분리 기준	분리 여부
(1)	연결어미를 포함하지 않음	False
(2)	연결어미 + 쉼표(,)	True
(3)	연결어미 + 보조용언	False
(4)	종속적 연결어미	True

Table 3.2에서 첫 번째 기준(1)은 종속적 연결어미를 포함하지 않는 어절은 분리할 필요가 없다. 두 번째 기준(2)은 종속적 연결어미 뒤에 쉼표(,)가 있으면 종속절로 간주한다. 예를 들어 ‘서울로 갔으니,’라는 문장에서 연결어미 ‘으니’ 다음에 ‘,’가 있다면 종속절로 분리하게 된다. 세 번째 기준(3)은 연결어미 다음에 보조용언이 오는 경우에서의 연결어미는 보조적 연결어미이므로 분리하지 않는다. 예를 들어, 문장 ‘보게 한다’에서 ‘-게’와 같은 경우 뒤 어절인 ‘한다’의 첫 형태소가 ‘하다’이고 품사는 보조용언이므로 ‘-게’는 보조적 연결어미이다. 이때는 종속절 분리가 수행되지 않는다. 네 번째 기준(4)은 그 외의 종속적 연결어미는 모두 분리한다. 세종형태품사표지 'EC'는 대등적 연결어미를 포함하고 있어서 Table 3.1과 같은 종속적 연결어미를 사전(종속적\_연결어미)으로 구축하여 사용한다. 이와 같은 기준으로 종속절을 분리하는 말뭉치를 구축하였다.

### 3.1.2 종속절 분리 모델

종속절 분리는 순차적 표지 부착(sequence labeling) 문제로 모델링할 수 있다. 즉 주어진 문장(어절열)에 대해서 종속절 표지를 부착하는 문제로 모델링하고 그 결과는 Figure 3.5와 같은 Bi-LSTM/CRF 모델을 사용한다. 이 모델은 입력자질은 어절자질, 음절자질, 자소자질로 구성되며, 각 자질을 합쳐서(concatenate) 모델의 입력자질로 사용한다. 어절 자질은 학습된 단어 표상(word embedding)이며, 세종말뭉치에서 GloVe(Pennington *et al.*, 2014)를 통해서 학습한다. 하지만 어절의 수가 매우 많으므로 모든 어절을 학습할 수 없다. 결과적으로는 저빈도 어절이 학습되지 않는다. 이와 같은 문제를 보완하려고 음절 자질과 자소 자질을 추가한다.

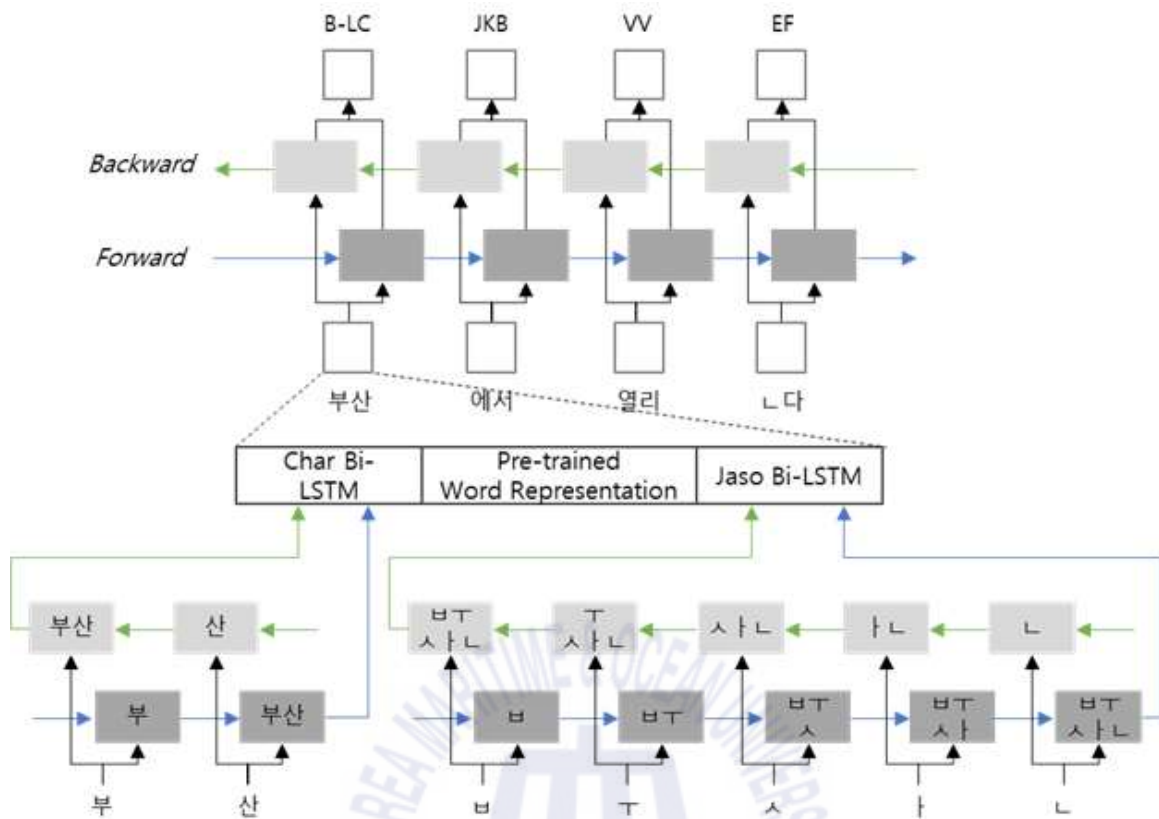


Figure 3.5 The structure of Bi-LSTM/CRF for segmenting subordinate clauses.

### 3.2 형태소 분리

일반적으로 개체명 인식 방법은 형태소를 분석한 후, 개체명을 인식한다(이창기 & 장명길, 2010; 신유현 & 이상구, 2016). 즉 형태소 분석이 개체명 인식의 전처리로 수행된다. 이 경우, 입력 문장에 미등록어가 포함되어 있다면 그 미등록어로 인하여 형태소 분석 단계에서 오류가 발생되며 그 오류가 개체명 인식으로 그대로 전달된다. 이와 같은 문제를 완화하기 위해 본 논문에서는 부분단어 정보를 이용하고 부분적인 형태소 분석을 수행한다. 즉 형태소 분리만 수행한다. 형태소 분리는 입력이 문장이고 출력은 주어진 문장에 가장 적합한 형태소열이다. 본 논문에서는 문장을 형태소열로 번역하는 기계번역 문제로 접근하여 Transformer 모델(Vaswani *et al.*, 2017)을 사용하며, 미등록어 문제를 완화하기 위해 이 모델의 입력

은 부분단어열을 사용한다.

이하의 절에서는 부분단어 사전을 구축하는 방법과 형태소 분리 모델을 구체적으로 기술한다.

### 3.2.1 부분단어 사전 구축

미등록어 문제를 완화하기 위해 형태소 분리 모델의 입력으로 부분단어를 사용하며 부분단어 사전을 구축하는 방법은 Figure 3.6과 같으며 기본적인 알고리즘은 BPE(Gage, 1994)를 따른다.

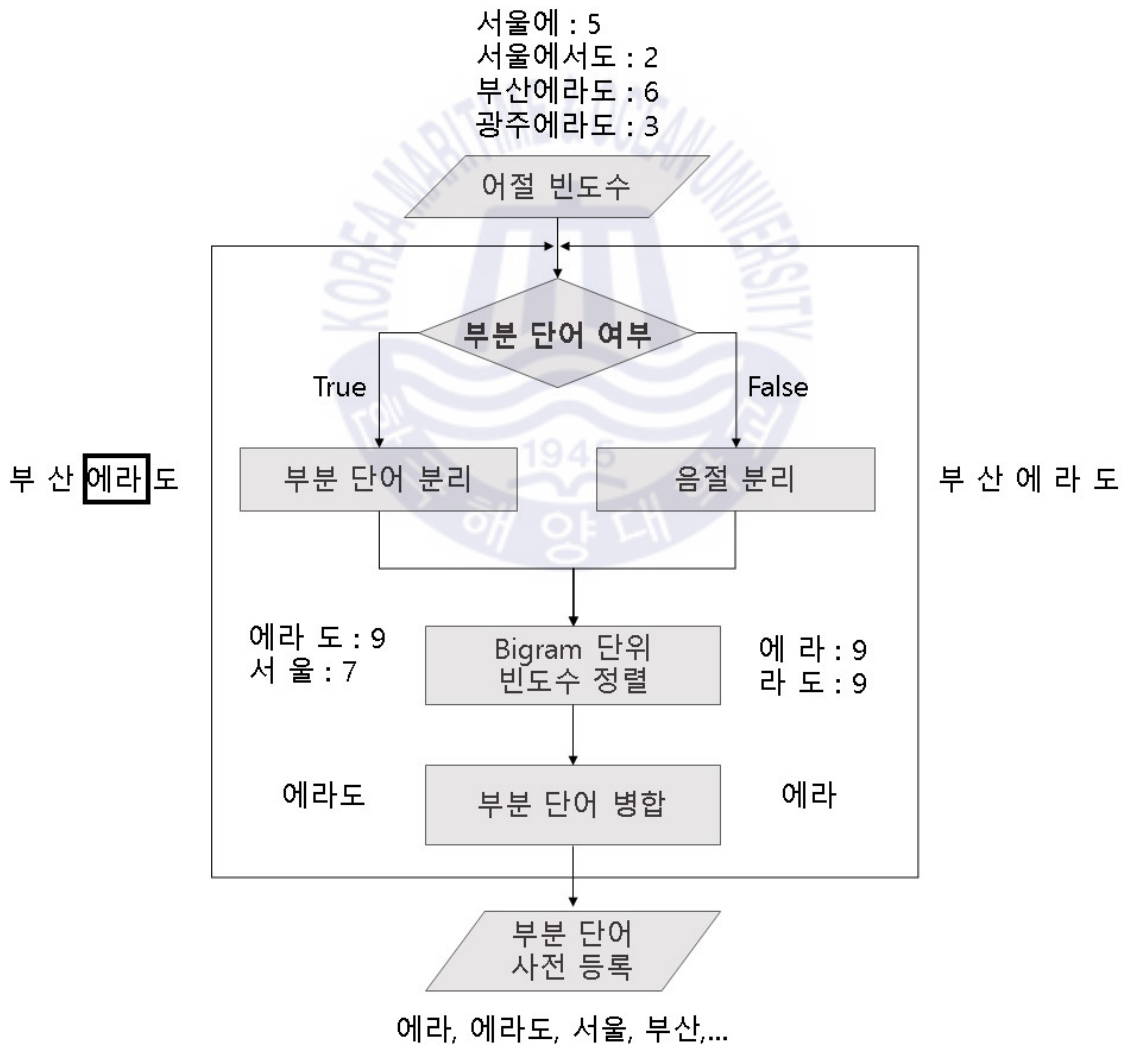
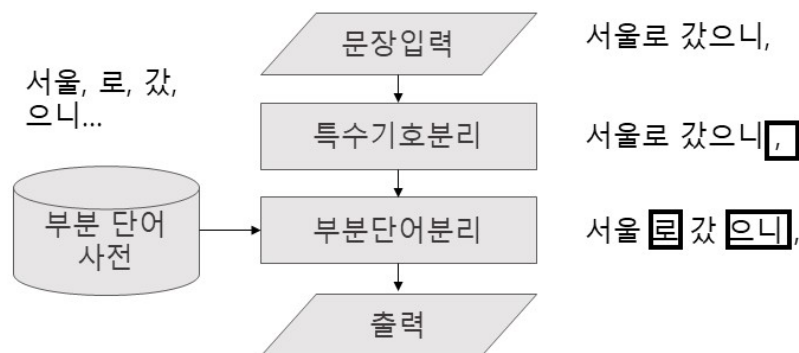


Figure 3.6 The process of building a subword dictionary.

부분단어 사전 구축을 위한 입력은 어절과 빈도수의 쌍이 들어온다. 처음에는 부분단어 사전에 아무것도 존재하지 않으므로 입력으로 들어온 어절과 빈도수에 대해서 어절을 각각의 음절로 분리한다. 예를 들어 ‘부산에라도’라는 1개의 어절이 들어왔을 때 ‘부 산 에 라 도’의 5개의 음절로 분리된다. 분리된 음절에 대해서 bigram 단위 빈도수 정렬을 하게 된다. 정렬을 했을 때 ‘에’와 ‘라’의 bigram인 ‘에라’가 전체 bigram 빈도수에서 9번 등장했으므로 가장 많은 빈도수를 가지게 된다. 가장 많은 빈도수를 가진 bigram ‘에라’를 부분단어 사전에 등록한다. 그리고 나서 ‘에라’를 하나의 음절로 간주하여 다시 bigram 빈도수를 구한다. Figure 3.6의 예에서는 다시 ‘에라도’가 가장 많은 빈도수를 가지므로 부분단어 사전에 등록된다. 이와 같은 방법을 반복하면 Figure 3.6의 예에서는 ‘에라, 에라도, 서울, 서울에, 부산, 광주’와 같은 순으로 부분단어가 생성되어 이들을 차례로 부분단어 사전에 등록한다. 이와 같은 방법으로 구축된 부분단어 사전을 이용해서 어절 ‘서울에라도’가 형태소 분리에 입력되면 부분단어 ‘서울’과 ‘에라도’로 분리할 수 있다. 부분단어 분리는 형태소 분리 모델의 전처리 과정으로 쓰인다.

### 3.2.2 형태소 분리 모델

형태소 분리 모델의 전처리는 Figure 3.7과 같이 두 단계로 구성된다.



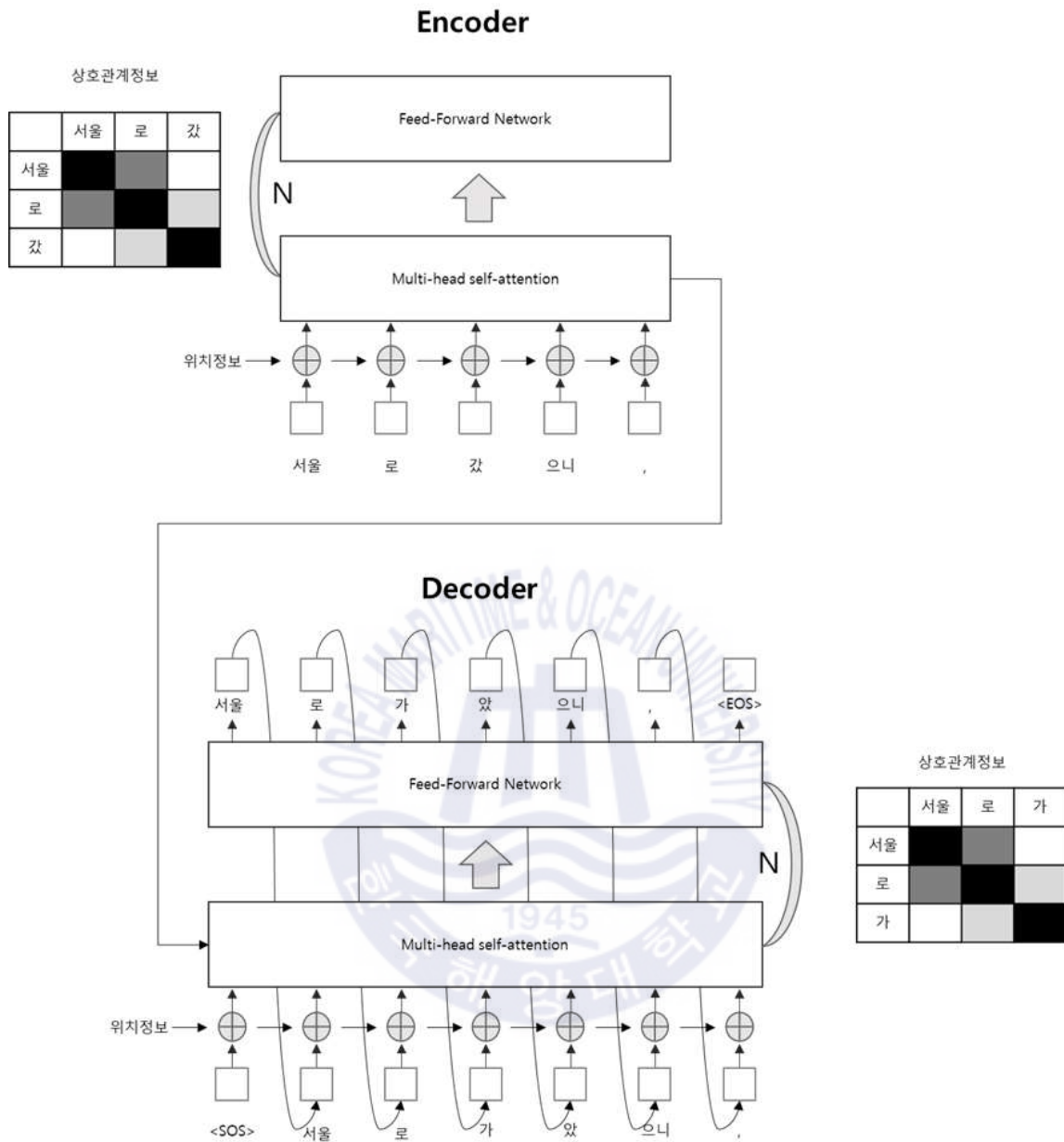
**Figure 3.7** The preprocessor of morpheme segmentation.

Figure 3.7에서 첫 번째 단계는 특수기호를 분리하는 단계이다. 예를 들면

문장 ‘서울로 갔으니,’에서 특수기호 ‘,’를 분리한다. 왜냐하면 특수기호 그 자체를 하나의 번역 단위로 간주하기 때문이다. 두 번째 단계는 부분 단어로 분리하는 단계이다. 이 단계에서는 3.2.1절에서 구축된 부분단어 사전을 이용하여 최장일치원리에 따라 분리한다. 예를 들면, 특수기호가 분리된 문장 ‘서울로 갔으니 ,’를 3.2.1절에서 구축된 사전을 이용하면 ‘서울 로 갔 으니 ,’와 같은 부분단어열을 얻을 수 있다.

Figure 3.8은 형태소 분리를 위한 Transformer 모델이다.



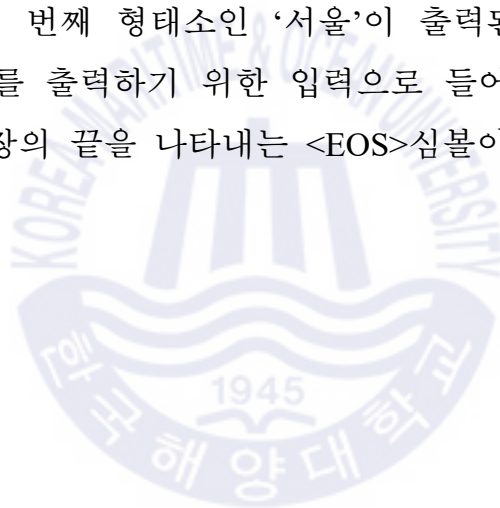


**Figure 3.8** The structure of the transformer for morpheme segmentation.

이 모델은 크게 인코더(encoder)와 디코더(decoder)로 나누어진다. 인코더는 부분단어의 표상과 그 단어의 위치정보가 입력된다. 인코더에서는 입력에 대한 상호관계정보(attention)를 Multi-head self-attention을 통해서 얻게 된다. 상호관계정보는 입력열에서 특정 입력과 나머지 입력에 대한 확률정보이다. 이 상호관계정보를 여러 겹으로 쌓아서 만든 것이 Multi-head



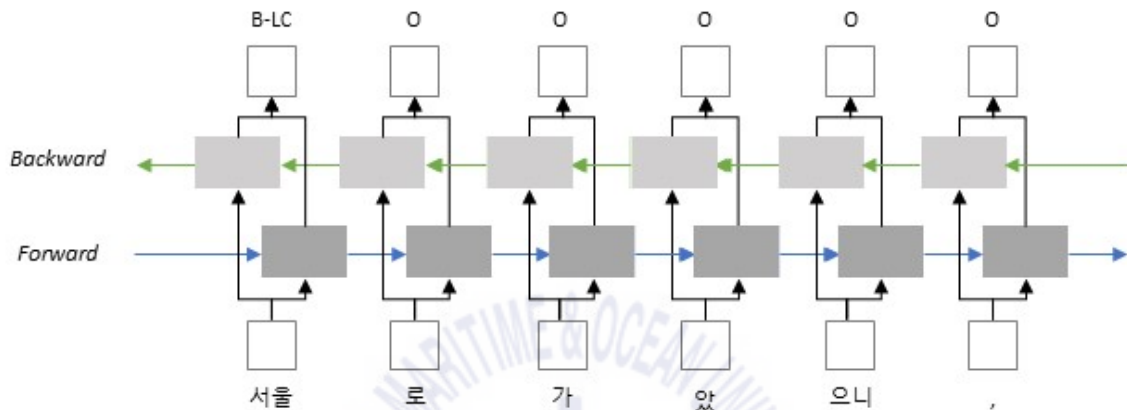
self-attention이다. 즉 Multi-head self-attention을 통해서 하나의 입력에 대해서 주변 문맥을 반영할 수 있다. Multi-head self-attention 위에 Feed-forward network를 구성하여 깊은 상호관계정보를 투영할 수 있게 된다. 그리고 인코더의 층을 N개만큼 쌓아서 더 깊은 의미를 투영할 수 있게 한다. 입력 문장이 인코더를 거치면 문장이 뜻하는 깊은 의미를 가지게 되고 이를 디코더로 보낸다. 디코더에서는 입력을 하나씩 받아서 하나씩 출력하는 구조를 가지고 있다. 그래서 문장의 제일 처음이 출력될 때에 디코더의 입력으로는 문장의 시작을 뜻하는 <SOS>심볼이 입력된다. 입력에 대한 포상과 단어의 위치정보가 입력으로 넣어준다. 입력은 인코더에서 온 정보와 Multi-head self-attention을 통해서 의미를 나타내게 되고 Feed-forward network를 통해서 첫 번째 형태소인 ‘서울’이 출력된다. 그리고 ‘서울’은 그대로 다음 형태소를 출력하기 위한 입력으로 들어가게 된다. 형태소가 하나씩 출력되고 문장의 끝을 나타내는 <EOS>심볼이 출력되면 형태소 분리를 종료하게 된다.





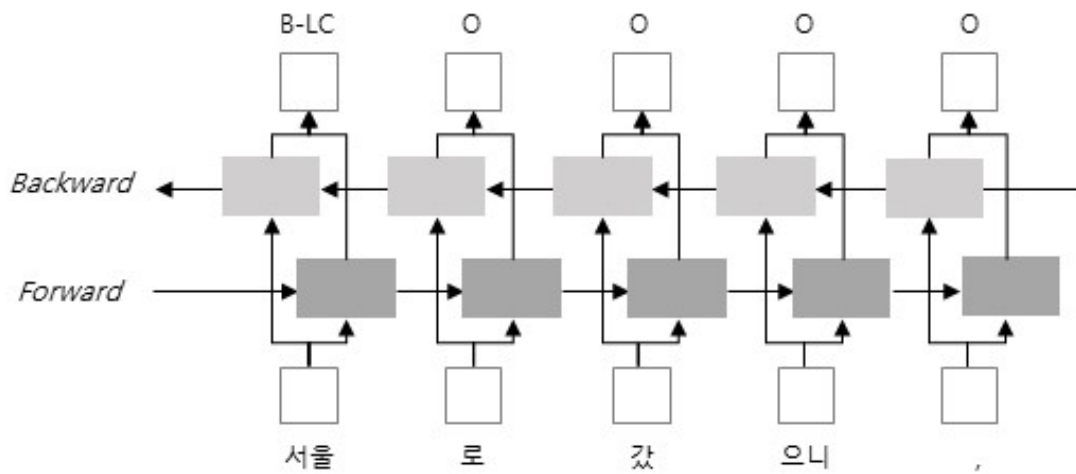
### 3.3 개체명 인식

개체명 인식도 종속절 분리와 같은 Bi-LSTM/CRF 모델을 사용한다. 성능 비교를 위해 본 논문에서는 형태소 기반 개체명 인식 모델(Figure 3.9)과 부분단어 기반 개체명 인식 모델(Figure 3.10)을 제안한다.



**Figure 3.9** The structure of Bi-LSTM/CRF for Korean NER based on morphemes as input.

예를 들어 입력이 ‘서울로 갔으니,’라서 형태소 기반 개체명 인식 모델은 그 입력의 형태소 분리 결과인 ‘서울 로 가 았 으니 ,’가 개체명 인식의 입력이 된다. 반면에 같은 입력에 대해서 부분단어 기반 개체명 인식 모델은 그 입력의 부분단어 분리 결과인 ‘서울 로 갔 으니 ,’가 개체명 인식의 입력이 된다. 주어진 예에서 두 모델 모두 개체명 ‘<서울: 지명>’을 찾을 수 있지만 Bi-LSTM의 문맥정보는 전자(형태소 기반 개체명 인식 모델)가 좀 더 정확하게 반영될 수 있다. 그리고 입력 표상의 초기 가중치를 Xavier Uniform Initializer(Glorot & Bengio, 2010)를 통해서 초기화하게 되고 개체명 말뭉치를 학습하면서 입력 표상도 학습하도록 한다.



**Figure 3.10** The structure of Bi-LSTM/CRF for Korean NRE based on subwords as input.



## 제 4 장 실험 및 평가

이 장에서는 종속절 및 형태소 분리 모델의 성능을 평가하고 각 모델의 문제점을 논의한다. 그리고 나서 한국어 개체명 인식을 평가하고 그 결과를 분석한다.

### 4.1 종속절 분리

#### 4.1.1 실험 환경 및 실험 척도

종속절 분리를 위한 말뭉치는 세종형태표지부착말뭉치(김홍규 외, 2007)를 사용하였다. 실험을 위해 세종형태표지부착말뭉치를 학습말뭉치와 개발말뭉치 그리고 평가말뭉치로 분리하였으며, 각각의 문장 수와 어절 수는 Table 4.1과 같다.

**Table 4.1** Statistics of training corpus for segmenting subordinate clauses.

말뭉치	문장 수	어절 수
학습말뭉치	524,004	5,427,324
개발말뭉치	74,810	777,173
평가말뭉치	74,833	773,207

종속절 분리 모델의 실험 환경은 아래 Table 4.2와 같다. 기본 모델인 Bi-LSTM/CRF<sup>2)</sup>는 최대한 기본 매개변수(default parameters)를 따랐다. char LSTM(음절 기반 LSTM)과 jaso LSTM(자소 기반 LSTM)의 LSTM 유닛의 크기와 출력 차원이 작다. 왜냐하면 음절이나 자소는 미등록어절이 발생

2) [https://github.com/guillaumegenthial/tf\\_ner](https://github.com/guillaumegenthial/tf_ner)

되었을 경우에 이를 보조하기 위한 수단으로 사용되기 때문이다. 따라서 전체 어절의 표상에서 어절 자체의 의미를 크게 반영하기 위해서이다. 따라서 어절의 입력 표상의 차원이 300, 음절과 자소의 출력 차원이 각각 100이다. 입력되는 차원이 500으로써 크기 때문에 LSTM 유닛을 500으로 크게 하여 문맥을 잘 반영할 수 있도록 하였다.

**Table 4.2** Parameters of Bi-LSTM/CRF for segmenting subordinate clauses.

모델	매개변수
Bi-LSTM/CRF	dropout = 0.5
	batch_size = 20
	LSTM_size = 500
	eojeol_input = 300
char LSTM, jaso LSTM	output_dim = 100
	LSTM_size = 25

종속절 분리 모델의 평가척도는 정밀도(precision)와 재현율(recall)과 F1-점수(F1-score)를 사용하며, 각각 식 (4.1)과 식 (4.2)와 식(4.3)과 같이 정의된다(Manning *et al.*, 2008). 각 식에서  $TP$ ,  $FN$ ,  $FP$ ,  $TN$ 은 Table 4.3과 같이 정의된다.  $TP$ 과  $TN$  은 시스템의 예측과 정답이 각각  $P$ 와  $N$ 로 일치하는 경우의 수이며,  $FN$ 과  $FP$ 는 시스템의 예측과 정답이 서로 다른 경우의 수이다.

**Table 4.3** 2×2 contingency table for comparing the performance.

		시스템 예측		합계
		$P$	$N$	
실제 정답	$P$	True Positive ( $TP$ )	False Negative ( $FN$ )	$TP+FN$
	$N$	False Positive ( $FP$ )	True Negative ( $TN$ )	$FP+TN$
합계		$TP+FP$	$FN+TN$	$Total$

정밀도는 시스템이 종속절을 예측한 개수( $TP+FP$ ) 중에서 실제 정답( $TP$ )과 같은 비율이며, 재현율은 실제 정답( $TP+FN$ ) 중에서 시스템이 예측한 종속절과 같은( $TP$ ) 비율이다. F1-점수는 정밀도와 재현율의 조화평균(harmonic average)이다.

$$Precision = \frac{TP}{TP+FP} \quad (4.1)$$

$$Recall = \frac{TP}{TP+FN} \quad (4.2)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4.3)$$

#### 4.1.2 실험 결과 및 분석

종속절 분리 실험 결과는 Table 4.4와 같다. 정밀도와 재현율은 각각 94.88%와 96.13%이며 따라서 F1-점수는 95.50%이다. 이 성능은 그다지 높지 않아서 오류를 분석해 보았다. 오류 중에서는 어절이 명사와 어미가 모호할 때 오류가 자주 나타났다. 예를 들어 “지휘 책임을 물어 경찰서장을 서면 경고했다고 24일 밝혔다.”라는 문장에서 ‘서면’은 “일정한 내용을 적은 문서”의 뜻을 가지는 명사인데 반해 실제로 종속절 분리 모델은 ‘서면’을 “사람이나 동물이 발을 땅에 대고 다리를 쭉 뻗으며 몸을 곧게 하다”라는 의미의 동사 ‘서다’로 인식하였다. 하지만 종속절 분리는 오류는 형태소 분리에 큰 영향을 주지 않을 것으로 판단되어 종속절 분리 모델의 성능을 개선하기 위한 노력을 기울이지 않았다. 이는 차후에 종속절 분리에 대한 더 깊은 연구가 필요할 것이다.

**Table 4.4** Performance of Bi-LSTM/CRF in subordinate clauses.

평가척도	결과
정밀도	94.88
재현율	96.13
F1-점수	95.50

## 4.2 형태소 분리

### 4.2.1 실험 환경 및 실험 척도

형태소 분리 실험의 학습말뭉치는 세종형태표지말뭉치를 이용하였고 평가말뭉치1은 “2016년 국어 정보 시스템 경진 대회”<sup>3)</sup>에서 배포한 개체명 인식 말뭉치에서 형태소 분석 및 정렬 오류를 수정하여 사용하였다. 평가말뭉치2는 KMOU 개체명 말뭉치<sup>4)</sup>를 사용하였다. 각 말뭉치에 대한 정보는 Table 4.5와 같다. 형태소 분리 모델은 Transformer 모델(Vaswani *et al.*, 2017)을 사용하였으며, 그 모델의 세부 매개변수는 Table 4.6과 같다. Transformer의 블록 개수는 6개이고 모델의 입력은 512차원이며, 마지막 feedforward 신경망의 출력은 2048차원이며 Multi-head self-attention에서 head는 16개, drop out은 0.3으로 설정하였다. 그리고 출력된 결과는 정답과 편집거리 알고리즘을 이용해서 일치 여부를 판단하였다.

**Table 4.5** Statistics of the training corpus for morpheme segmentation.

말뭉치	문장 수	어절 수	형태소 수
학습말뭉치	671,977	8,616,562	15,521,232
평가말뭉치1	4,092	72,583	158,777
평가말뭉치2	23,964	345,739	764,403

3) <https://ithub.korean.go.kr/user/contest/contestIntroView.do>

4) <https://github.com/kmounlp/NER>

**Table 4.6** Parameters of the transformer model.

시스템	매개변수
Transformer	num_hidden_layers = 6
	hidden_size = 512
	filter_size = 2048
	num_heads = 16
	dropout = 0.3

#### 4.2.2 실험 결과 및 분석

형태소 분리 실험 결과는 Table 4.7과 같다. 평가말뭉치1의 경우 형태소 분리만 시행했을 경우, 정밀도는 90.06%이고, 재현율은 88.45%이며, F1-점수는 89.25%이다. 또한 종속절 분리를 하고 형태소 분리를 시행했을 경우, 정밀도는 90.35%이고, 재현율은 90.05%이며, F1-점수는 90.20%이다. 그리고 평가말뭉치2의 경우 형태소 분리의 정밀도는 86.31%이고, 재현율은 83.38%이며, F1-점수는 84.82%이다. 종속절 분리와 형태소 분리를 시행했을 경우, 정밀도는 86.33%이고, 재현율은 85.27%이며, F1-점수는 85.79%이다. 실제로 종속절 분리를 시행했을 때, 성능이 크게 개선되지 않았다. 그 이유를 살펴보자. 형태소 분리 모델은 인코더-디코더 모델의 일종으로 문장의 길이가 50 이상부터 성능이 크게 하락하기 시작한다(Cho *et al.*, 2014). 문장의 길이가 50 이상일 때 종속절 분리를 통해서 형태소 분리 성능 개선을 유도할 수 있다. 하지만 Table 4.8과 같이 평가말뭉치1에서는 길이가 50 이상인 문장이 전체 말뭉치에서 0.35% 밖에 포함되지 않아서 성능이 크게 개선되지 않았다. 종속절 분리가 잘 적용되고 형태소 분리가 잘 되었을 경우, 정답과의 일치하는 비율인 재현율이 상승해야 하는데 정밀도의 상승률보다 재현율의 상승률이 높아서 종속절 분리의 영향이 형태소 분리에도 미쳤다고 판단할 수 있다.



**Table 4.7** Performance of the transformer model for segmenting morphemes.

시스템		정밀도	재현율	F1-점수
평가말뭉치1	형태소 분리	90.06	88.45	89.25
	종속절 분리 + 형태소 분리	90.35	90.05	90.20
평가말뭉치2	형태소 분리	86.31	83.38	84.82
	종속절 분리 + 형태소 분리	86.33	85.27	85.79

**Table 4.8** Distribution of sentence lengths in test corpus for morpheme segmentation.

길이	빈도수	전체 말뭉치에 대한 비율
0~19	2,810	68.67%
20~29	969	23.68%
30~39	256	6.25%
40~49	43	1.05%
50이상	14	0.35%
합계	4,092	100%

형태소 분리 실험 결과에서 실제로 미등록어에 잘 반응하는지를 살펴보았다. 평가말뭉치2에서 등장하는 ‘러블리즈’라는 아이돌 그룹이 나타날 때, 형태소 분리는 ‘러블리즈’를 그대로 형태소라고 분리하였으나, 형태소 분석기(신준철 & 옥철영, 2012)에서는 ‘러 블 리즈’로 분석을 하여 각각 고유명사로 이루는 형태로 출력이 나왔다. 또 다른 예로는 ‘서머너즈워’라는 게임이 존재하는데 형태소 분리에서는 ‘서머너즈워’ 그대로 게임 이름이 나온 반면에 형태소 분석기에서는 앞서서와 마찬가지로 ‘서머너즈워’ 이렇게 각각 고유명사라고 품사가 부착되었다. 두 경우 모두 실제로 형태소 분석기나 형태소 분리 안에 들어 있지 않은 미등록어이다. 이와 같은 예를 통해서 형태소 분석기가 미등록어에 얼마나 취약한지를 알 수 있었다. 한편 형태소 분리는 이와 같은 미등록어에 능동적으로 대



처함을 관찰할 수 있었다.

형태소 분리의 모든 결과를 수동으로 직접 확인하는 것은 시간과 노력이 많이 들어가므로 평가말뭉치1로부터 100문장을 임의로 추출하여 오류를 분석하였다. Table 4.9는 그 분석 결과를 보이고 있다. 오류의 유형은 크게 띄어쓰기 불일치 오류, 모음조화 미적용 오류, 복합용언 분리 불일치, 형태소 분리 오류로 나눈다.

**Table 4.9** Types of errors of morpheme segmentation.

형태소 개수	빈도수	전체 형태소에 대한 비율
정답	3,278	91.72%
띄어쓰기 불일치	153	95.97% (4.25%p)
모음조화 미적용	42	97.14% (1.17%p)
복합용언 분리 불일치	41	98.29% (1.14%p)
형태소 분리 오류	61	100% (1.71%p)
전체 형태소	3,575	100%

첫 번째 유형인 띄어쓰기 불일치 오류는 정답인 원문과 형태소 분리의 결과가 띄어쓰기로 발생한 오류이다. 예를 들어 정답 ‘리조트이용료’에 대해서 형태소 분리에서는 ‘리조트 이용료’로 인식하였다. 그러나 한글맞춤법통일안(한글학회, 1989)의 제49항 “성명 이외의 고유 명사는 단어별로 띄어 씀을 원칙으로 하되, 단위별로 띄어 쓸 수 있다.”에 따라 ‘리조트 이용료’도 잘못된 결과라고 볼 수 없다.

두 번째 유형인 모음조화 미적용 오류는 모음조화의 원리에 따라서 분석 결과가 불일치하는 오류이다. 예를 들면 ‘했다’는 모음조화 규칙에 따라서 ‘하 았 다’로 분리되어야 하지만 ‘하 었 다’로 분리되어도 잘못된 분리로 볼 수 없다. 왜냐하면 ‘았’과 ‘엿’은 모두 과거를 나타내는 선어말어미로 같은 의미를 나타내기 때문이다. 다만 발음 규칙에 해당하는 모음조

화 규칙에는 부적합하다.

세 번째 유형으로는 복합용언 분리 불일치는 복합용언의 과도하게 분석되어 분석과의 정답이 불일치하는 오류이다. 예를 들어 ‘괴로워하는 아들’이라는 구가 입력되었을 때, 형태소 분리는 ‘괴로워하 는 아들’으로 분석되었다. 그러나 정답에는 ‘괴롭 어 하 는’으로 되어 불일치 오류가 발생했다. 두 경우 어느 것이 잘못이라고 판단할 수 없다. 왜냐하면 ‘괴로워하’이라는 복합용언도 표준국어대사전에 등재되어 하나의 표제어로 인정되기 때문이다. 이처럼 복합용언도 하나의 사전 표제어로 인정되는 경우에는 그 분리 결과를 반드시 오류로 판단할 수 없다.

따라서 위의 세 가지 유형은 반드시 오류라고 판단할 수 없으므로 정답으로 인정되어야 한다. 이 경우를 반영하여 다시 평가하면 98.29%의 정확률을 보이며 실질적인 오류는 1.71%에 불과하다.

실질적인 오류에 해당하는 네 번째 유형의 예를 살펴보자. 입력 ‘낭떠러지 끝에라도’에 대해 형태소 분리를 한 결과는 ‘낭떠러지 끝 에 라도’였다. 그러나 정답은 ‘낭떠러지 끝 에 이 라도’이다. 즉 긍정지정사인 ‘이’를 추가로 생성하지 못한 경우이다. 이 경우는 명백한 오류이다.

## 4.3 개체명 인식

### 4.3.1 실험 환경 및 실험 척도

개체명 인식의 말뭉치는 “2016년 국어 정보 시스템 경진 대회”에서 배포한 개체명 인식 말뭉치<sup>5)</sup>에서 형태소 분석 및 정렬 오류를 수정하여 사용하였으며 이 말뭉치의 정보는 Table 4.10과 같다. 개체명 인식 모델의 매개변수는 Table 4.11과 같으며 음절과 자소 정보를 추가하지 않은 이유는 음절, 부분단어 모델과의 비교를 위해서 형태소 표상만을 입력으로 넣

5) <https://ithub.korean.go.kr/user/contest/contestIntroView.do>

은 모델을 구성하였다. 그리고 종속절 분리 모델의 입력 유닛 크기가 500 인데 비해 개체명 인식 모델의 입력은 음절과 자소 정보를 포함하지 않으므로 입력 유닛 크기를 축소하였다.

**Table 4.10** Training corpus for Korean named entity recognition.

말뭉치	문장 수	형태소 수	개체명 수
학습말뭉치	3,241	126,073	6,890
개발말뭉치	425	16,398	894
평가말뭉치	426	16,306	886

**Table 4.11** Parameters of Bi-LSTM/CRF for Korean named entity recognition.

시스템	매개변수
Bi-LSTM/CRF	dropout = 0.5
	batch_size = 20
	LSTM_size = 100
	input_dim = 300

#### 4.3.2 실험 결과 및 분석

개체명 인식 결과는 Table 4.12와 같다. 실제로 음절, 부분단어, 형태소 분리 순으로 개체명 인식률이 상승하였다. 음절과 부분단어보다는 형태소 분리가 월등히 상승하였으며, 이는 형태소를 분리하고 개체명을 인식하는 방법이 좋다고 판단할 수 있다. 실험에 대한 결과 분석에서는 ‘레비 스트로스’라는 인명이 존재한다고 할 때, 음절단위 개체명 인식기에서는 하나도 인식을 못하였고, 부분단어 개체명 인식기에서는 ‘레 비 스트로스’로 인식하여 ‘스트’와 ‘로스’에 대해서 개체명을 인식되었다. 그리고 형태소 분리에서는 ‘레비 스트로스’로 인식되어 ‘레비 스트로스’라는 인명을 모두

인식하였다. 또 다른 예로는 ‘지난해보다’라는 문장에서 ‘지난해’가 바로 낱짜를 나타내는 개체명인데 여기에서 음절 모델은 마찬가지로 인식하지 못하였고, 부분단어는 ‘지난해보다’를 하나의 단어로 인식하여서 개체명으로 인식되지 않았다. 하지만 형태소 분리에서는 ‘지난해 보다’라고 인식하여 ‘지난해’에 대해서 알맞은 개체명 표지를 부착하였다.

**Table 4.12** Performance of Korean named entity recognition.

평가척도	정밀도	재현율	F1-점수
음절	68.48	54.46	60.67
부분단어	69.33	63.31	66.18
형태소 분리	73.84	68.51	71.07
종속절 분리 + 형태소 분리	74.66	69.52	71.99



## 제 5 장 결론 및 향후 연구

개체명 인식이란 주어진 문서에서 개체명의 범위를 찾고 개체명의 범주를 결정하는 것이다. 개체명 인식에서 어려운 문제는 미등록어 문제와 중의성 문제가 존재한다. 두 가지 문제를 해결하기 위한 기존의 방법에서는 입력 단위를 바꾸어 가면서 중의성 문제와 미등록어 문제를 해결하였다. 하지만 각각의 장단점이 존재하였고 각각의 장점을 합쳐서 모델을 제안하였다.

본 논문에서는 입력에 대해 종속절 분리, 형태소 분리, 개체명 인식이라는 세 단계의 개체명 인식 방법을 제안했다. 종속절 분리는 순환신경망을 통해서 입력인 원문에 대해 종속절로 분리된 문장을 각각 만들게 되고, 형태소 분리에서는 Transformer 모델을 이용해서 기계번역 문제로 접근하여 수행하였다. 그리고 형태소 분리된 형태소열을 통해서 개체명 인식을 수행하였다. 제안하는 방법의 종속절 분리를 통해 형태소 분리 단계에서 발생하는 오류를 줄여 주었고 형태소 분리를 통해 중의성 문제를 Multi-head self-attention 기법으로 해결하였고, 미등록어 문제는 입력에 대해 부분단어 정보를 이용하여 분절하는 방법을 통해서 미등록어 문제를 해결하였다.

종속절 분리 모델에서는 종속절 분리에 대해 94.88%의 정밀도, 96.13%의 재현율, 95.50%의 F1-점수를 보였다. 종속절 분리에서는 명사와 어미의 모호성에서 생기는 오류들이 발생하는 문제가 존재하였지만 형태소 분리 단계에 영향을 크게 미치지 않았다. 형태소 분리 모델에서는 90.35%의 정밀도, 90.05%의 재현율, 90.20%의 F1-점수를 보였다. 형태소 분리에서 미등록어에 대한 분리에 대해서 다른 형태소 분석기들보다 유연하게 반응하

는 것을 확인하였다. 형태소 분리에서는 실제로 편집 거리를 이용하여 측정을 하였는데 실제 오류로 측정된 것들에 대해 분석을 진행한 결과 네 가지의 유형을 확인하였고 이 중 세 가지 유형인 띄어쓰기 불일치와 모음 조화 미적용, 복합용언 분리는 오류에 해당하지 않았다. 해당하지 않는 오류를 포함하지 않고 성능을 측정했을 시 98.29%에 해당하는 F1-점수를 보여주었다. 또한 개체명 인식에서의 성능은 74.66%의 정밀도와 69.52%의 재현율, 71.99%의 F1-점수를 보여줬으며 음절, 부분단어, 형태소 분리의 순서대로 성능이 향상됨을 보였다.

본 연구에서는 종속절 분리 이후 형태소 분리, 개체명 인식의 세 가지 단계로 진행을 하였다. 하지만 형태소 분리에서 출력되는 형태소열은 많은 한국어 자연언어처리에 이용될 수 있다. 형태소 분리에서 출력되는 형태소열을 구문음과 같은 다른 영역에 적용할 예정이다.



## 참고문헌

- Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003). “A neural probabilistic language model”, *Journal of Machine Learning Research*, vol. 3, pp. 1137-1155.
- Bikel, D. M., Schwartz, R., and Weischedel, R. M. (1999). “An algorithm that learns what’s in a name. *Machine Learning*”, vol. 34, pp. 211–231.
- Borthwick, A. (1999). “A maximum entropy approach to named entity recognition”, Ph.D. Dissertation, New York University.
- Chieu, H., and Ng, H. (2003). “Named entity recognition with a maximum entropy approach.” *Proceedings of Conference on Computational Natural Language Learning*, pp. 160–163.
- Cho, K., Merrienboer, B., Bahdanau, D. and Bengio, Y. (2014). “On the properties of neural machine translation: Encoder–Decoder approaches.”, *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*.
- Chung, J., Cho, K. and Bengio, Y. (2016). “A character-level decoder without explicit segmentation for neural machine translation”, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 1693-1703.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. (2011). “Natural language processing (almost) from scratch”, *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537.
- Eddy, R. S. (1996). “Hidden markov models,” *Current Opinion in Structural*



- Biology*, vol. 6, no. 3, pp. 361-365.
- Gage, P. (1994). "A new algorithm for data compression", *Journal of the C Users*, vol. 12, no.2, pp. 23-38.
- Gliozzo, A., Giuliano, C. and Rinaldi, R. (2005). "Instance filtering for entity recognition", *ACM SIGKDD Explorations Newsletter*, vol. 7, pp. 11-18.
- Glorot, X. and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks", *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 249-256.
- Grishman, R. and Sundheim, B. (1996). "Message understanding conference-6: A brief history," *Proceedings of the 16th conference on Computational linguistics*, vol. 1, pp. 466-471.
- Hearst, A. M., Dumais, T. S., Osuna, E., Platt, J. and Scholkopf, B. (1998). "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18-28.
- Huang, Z., Xu, W. and Yu, K. (2015). "Bidirectional lstm-crf models for sequence tagging", arXiv:1508.01991.
- Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., and Wilks, Y. (1998). "Univ. of Sheffield: Description of the LaSIE-II system as used for MUC-7." *Proceedings of the 7th Message Understanding Conference*.
- Isozaki, H., and Kazawa, H. (2002). "Efficient support vector classifiers for named entity recognition." *Proceedings of the Conference on Computational Linguistics*. vol. 1, pp. 1-7.
- Jean, S., Cho, K., Memisevic, R. and Bengio, Y. (2015). "On using very large target vocabulary for neural machine translation", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol. 1, pp. 1-10.

- Kapur, N. J. (1989). *Maximum-entropy Models in Science and Engineering*. John Wiley & Sons.
- Klein, D., Smarr, J., Nguyen, H., and Manning, C. (2003). “Named entity recognition with character-level models.”, *Proceedings of 7th Conference of the North American Chapter of the Association for Computational Linguistics*, vol. 4, pp. 180-183.
- Kudo, T. (2018). “Subword regularization: Improving neural network translation models with multiple subword candidates”, *Proceedings of the Association for Computational Linguistics*, pp. 66-75.
- Kudo, T. and Richardson, J. (2018). “SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing”, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66-71.
- Lafferty J., McCallum A. and Pereira F. (2001). “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, *Proceedings of the 18th International Conference on Machine Learning*, pp. 282-289.
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016). “Neural architectures for named entity recognition”, *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 260-270.
- Luong, M., Sutskever, I., Le, Q., Vinyals, O. and Zaremba, W. (2015). “Addressing the Rare Word Problem in Neural Machine Translation”, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 11-19.
- Luong, M. and Manning, C. D. (2016). “Achieving open vocabulary neural machine translation with hybrid word-character models”, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 1,

- pp. 1054-1063.
- Ma, X. and Hovy, E. (2016). "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF", *Proceedings of the Association for Computational Linguistics*, pp. 1064-1074.
- Manning, C. D., Raghavan, P and Schütze, H. (2008). "Evaluation in information retrieval", *Introduction to Information Retrieval*, Cambridge University Press, pp. 154-158.
- Mikheev, A., Grover, C., and Moens, M. (1998). "Description of the LTG system used for MUC-7", *Proceedings of 7th Message Understanding Conference*.
- Pennington, J., Socher R. and Manning D. C. (2014). "GloVe: Global vectors for word representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532-1543.
- Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the Institute of Electrical and Electronics Engineers*, vol. 77, no. 2, pp. 257-286.
- Schmidhuber, J. (1993). Habilitation thesis: *System Modeling and Optimization*, The Technical University of Munich.
- Sennrich, R., Haddow, B. and Birch, A. (2016). "Neural machine translation of rare words with subword units", *Proceedings of the Association for Computational Linguistics*. pp. 1715-1725.
- Settles, B. (2004). "Biomedical named entity recognition using conditional random fields and novel feature sets", *Proceedings of Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 104-107.
- Sutskever, I., Vinyals, O. and Le, Q. (2014). "Sequence to sequence learning with neural networks", *Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol. 2, pp. 3104-3112.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. and Polosukhin, I. (2017). "Attention is all you need", *Advances in*

*Neural Information Processing Systems*, pp. 5998–6008.

Zhou, G. and Su, J. (2002). “Named entity recognition using an HMM-based chunk tagger”, *Proceedings of 40th Meeting of Association of Computational Linguistics*, pp. 473–480.

구교정, 홍정아, 서아정, 차지원, 여운영, 김종우 (2018). “개체명 인식을 통한 이커머스 상에서의 질의 특징 파악”, *한국지능정보시스템학회 춘계학술대회 논문집*, pp. 7-8.

김재훈 (2004). “정보추출의 기술 현황”, *정보과학회지*, 제22권, 제4호, pp. 35-46.

김재훈, 김형철, 최윤수 (2010). “기계학습 기반 개체명 인식을 위한 사전 질 생성”, *정보관리연구*, 제41권, 제2호, pp. 31-46.

김흥규, 강범모, 홍정하 (2007). “21세기 세종계획 현대국어 기초말뭉치: 성과와 전망”, *제19회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 311-316.

박동주, 안창욱 (2019). “개체명 비열 사전을 결합한 Bidirectional LSTM-CRF 기반 개체명 인식”, *한국컴퓨터종합학술대회 논문집*, pp. 721-723.

박혜웅, 송영숙 (2017). “음절 기반의 CNN을 이용한 개체명 인식”, *제29회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 330-332.

신유현, 이상구 (2016). “양방향 LSTM-RNNs-CRF를 이용한 한국어 개체명 인식”, *제28회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 340-341.

신준철, 옥철영 (2012). “기분석 부분 어절 사전을 활용한 한국어 형태소 분석기”, *정보과학회논문지 : 소프트웨어 및 응용*, 제39권, 제5호, pp. 415-424.

오교중, 이동진, 최호진, 권성태, 홍사욱 (2017). “대화 문장의 기계학습 기반 개체명 경계 인식 및 미등록어 개체명 사전 확장에 관한 연구”, *한국컴퓨터종합학술대회 논문집*, pp. 657-659.

유연수, 박혁로 (2019). “CNN-CRFs를 이용한 한국어 개체명 인식기”, *제31회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 77-79.

유홍연, 고영중 (2016). “품사 임베딩과 음절 단위 개체명 분포 기반의 Bidirectional LSTM CRFs를 이용한 개체명 인식”, *제28회 한글 및 한국어*

- 정보처리 학술대회 논문집, pp. 105-110.
- 윤여탁, 구분관, 정정순, 김기훈, 장성남, 소정섭, 신호은, 최혜민, 김정은, 송소라 (2014). *독서와 문법*, 서울 : 미래엔
- 이경희, 이주호, 최명석, 김길창 (2000). “한국어 문서에서 개체명 인식에 관한 연구”, *제12회 한글 및 한국어 정보처리 학술대회 발표논문집*, pp. 292-299
- 이원기, 김영길, 조승우, 권홍석, 이의현, 조형미, 이종혁 (2017). “개체명 인식과 단어 정렬을 이용한 통계적 기계번역의 성능 향상”, *한국컴퓨터종합학술대회 논문집*, pp. 615-617.
- 이창기, 황이규, 오효정, 임수중, 허정, 이충희, 김현진, 왕지현, 장명길 (2006). “Conditional random fields를 이용한 세부 분류 개체명 인식”, *제18회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 268-272.
- 이창기, 장명길 (2010). “Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식”, *인지과학*, 제21권, 제4호, pp. 655-667.
- 이창수, 고영중 (2014). “대화형 개인 비서 시스템의 언어 인식 모듈을 위한 개체명 및 문장목적 동시 인식 방법”, *정보과학회논문지 : 소프트웨어 및 응용*, 제41권, 제4호, pp. 295-301.
- 이태석, 신수미, 강승식 (2016). “조건부 랜덤 필드를 이용한 특허 문서의 개체명 인식”, *정보처리학회논문지*, 제5권, 제9호, pp. 419-424.
- 정래정, 김준태 (1996). “고유 명사 출현 패턴을 이용한 색인의 성능 향상에 관한 연구”, *한국정보과학회 언어공학연구회 학술발표 논문집*, pp. 68-72.
- 천민아, 김창현, 박호민, 김재훈 (2018). “Multi-head attention 방법을 적용한 문자 기반의 다국어 개체명 인식”, *제30회 한글 및 한국어 정보처리 학술대회 발표논문집*, pp. 167-170
- 한글학회 (1989). "한글 맞춤법 통일안(1933~1980, 처음판 및 고침판 모음)", 서울 : 한글학회



## 감사의 글

본 논문을 작성하기 위해 부족한 저에게 수많은 조언과 지도를 해주신 김재훈 지도교수님께 이 자리를 빌려 깊은 감사를 전합니다. 교수님의 지도편달 덕분에 저의 연구역량이 많이 향상됨을 느끼고 교수님은 늘 배울 점이 많은 분이라고 여깁니다. 논문의 심사를 맡아주신 박휴찬 교수님과 류길수 교수님 그리고 부족한 저에게 많은 배움을 주신 다른 교수님들께도 감사드립니다.

연구를 하면서 많은 시련과 고난을 겪는 저를 보고 참아주신 자연언어처리 실험실 여러분께도 깊은 감사를 표합니다. 동고동락하면서 정말 많은 일화들과 늘 주고받는 유머들이 석사 과정동안 많은 도움이 되었습니다. 그리고 같이 석사 과정동안 수업을 들었던 김정래 형님, 여동규, 이정욱에게도 감사드립니다.

가끔 실험실에서 벗어나서 쉴 때 많은 도움이 되었던 친구 최광욱과 바쁘게 사는 양원녕 그리고 항상 늘 티키타카하는 학사팸 친구들과 한 번씩 부산을 와서 재밌게 노는 강동훈과 중학교 때부터 같이 늑어가고 있는 고창우, 윤지영, 강길현과 학사 과정 때 같이 수업을 들었던 한국해양대 친구들에게도 감사합니다.

늘 말을 듣지 않아도 아낌없는 사랑으로 보듬어주시는 부모님과 어릴 적부터 저를 정말 오랫동안 키워주시고 무엇보다도 큰 사랑으로 저를 아껴주시는 저의 외할머니 한순죽 여사와 늘 매일 자고 일어나며 형의 짜증을 받아주는 동생 윤충현과 가족들에게 정말 정말 큰 감사를 전합니다.