Journal of the Korea Convergence Society
Vol. 12. No. 3, pp. 1-7, 2021

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2021.12.3.001

한국어 인공신경망 기계번역의 서브 워드 분절 연구 및 음절 기반 종성 분리 토큰화 제안

어수경[1], 박찬준[1], 문현석[1], 임희석[2*]
[1]고려대학교 컴퓨터학과 석·박사통합과정, [2]고려대학교 컴퓨터학과 교수

Research on Subword Tokenization of Korean Neural Machine Translation and Proposal for Tokenization Method to Separate Jongsung from Syllables

Sugyeong Eo[1], Chanjun Park[1], Hyeonseok Moon[1], Heuiseok Lim[2*]
[1]Master & Ph.D Combined Student, Department of Computer Science and Engineering, Korea University
[2]Professor, Department of Computer Science and Engineering, Korea University

요 약 인공신경망 기계번역(Neural Machine Translation, NMT)은 한정된 개수의 단어만을 번역에 이용하기 때문에 사전에 등록되지 않은 단어들이 입력으로 들어올 가능성이 있다. 이러한 Out of Vocabulary(OOV) 문제를 완화하고자 고안된 방법이 서브 워드 분절(Subword Tokenization)이며, 이는 문장을 단어보다 더 작은 서브 워드 단위로 분할하여 단어를 구성하는 방법론이다. 본 논문에서는 일반적인 서브 워드 분절 알고리즘들을 다루며, 나아가 한국어의 무한한 용언 활용을 잘 다룰 수 있는 사전을 만들기 위해 한국어의 음절 중 종성을 분리하여 서브 워드 분절을 학습하는 새로운 방법론을 제안한다. 실험결과 본 논문에서 제안하는 방법론이 기존의 서브 워드 분리 방법론보다 높은 성능을 거두었다.

주제어 : 기계번역, 전처리, 서브 워드 분절, 서브 워드, 딥러닝, 융합

Abstract Since Neural Machine Translation (NMT) uses only a limited number of words, there is a possibility that words that are not registered in the dictionary will be entered as input. The proposed method to alleviate this Out of Vocabulary (OOV) problem is Subword Tokenization, which is a methodology for constructing words by dividing sentences into subword units smaller than words. In this paper, we deal with general subword tokenization algorithms. Furthermore, in order to create a vocabulary that can handle the infinite conjugation of Korean adjectives and verbs, we propose a new methodology for subword tokenization training by separating the Jongsung(coda) from Korean syllables (consisting of Chosung-onset, Jungsung-neucleus and Jongsung-coda). As a result of the experiment, the methodology proposed in this paper outperforms the existing subword tokenization methodology.

Key Words : Machine Translation, Preprocessing, Subword Tokenization, Subword, Deep Learning, Convergence

*This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and this work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques).
*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)
Received January 13, 2021 Revised February 23, 2021
Accepted March 20, 2021 Published March 28, 2021

1. 서론

인공신경망 기계번역(Neural Machine Translation, NMT)은 소스 문장을 입력으로 넣어주고, 타겟 문장을 출력할 수 있도록 신경망 모델의 가중치들을 학습하게 하는 종단간 학습(End-to-End) 방법론이며 기존 규칙 및 통계 기반 기계번역의 성능을 뛰어넘었다. NMT 모델을 제작할 때 새로운 모델을 개발하거나 효율적인 알고리즘을 설계하는 것도 중요하나 고품질의 데이터를 이용하는 것, 데이터를 잘 정제하는 것, 컴퓨터가 데이터를 잘 이해할 수 있는 지식 표현 체계인 고차원의 벡터 형태로 잘 변환하는 것 등의 전처리 과정 역시 매우 중요하다[1].

NMT를 학습하기 전에는 미리 단어사전을 구축한다. 단어사전은 대부분 3만에서 20만 정도의 한정된 크기로 구축되는데, 이 크기로는 세상에 존재하는 모든 단어들을 수용할 수 없다. 더불어 NMT 모델은 대부분 열린 어휘(Open Vocabulary) 문제를 다루기 때문에 사전에 등록되지 않은 단어들이 입력으로 들어올 수 있다. 이렇게 단어 사전에 존재하지 않는 단어가 입력으로 들어오는 경우, 그 단어는 'unknown(UNK)' 토큰으로 처리되며 이를 Out of Vocabulary(OOV) 문제라고 한다. 위 문제를 다루고자 제안된 것이 바로 서브 워드 분절(Subword Tokenization)이다. 즉, 서브 워드 분절은 궁극적으로 NMT에서의 고정된 단어사전 개수로 인한 UNK 문제를 완화하기 위해 제안되었다. 서브 워드 분절은 "단어는 단어보다 작은 단위인 서브 워드들의 조합으로 구성된다." 라는 가정을 기반으로 하며, 문장의 각 토큰들을 단어보다 더 작은 서브 워드 단위로 분절한다. 서브 워드 단위로 분리를 진행하면 신조어, 오타 등으로 인한 UNK 문제의 해결에도 효과적이다. 이는 자연스레 NMT 모델의 성능 향상으로 이어지며, 해당 과정은 NMT 모델 제작 시 필수 과정으로 자리 잡고 있다.

이러한 필요성을 바탕으로 본 논문은 서브 워드 분절의 종류와 각각의 특징들을 설명한다. 더 나아가 한국어의 무한한 용언 활용을 더 잘 설명할 수 있는 사전을 만들기 위해 새로운 토큰화 방법을 제안하고 실험과 분석을 통해 그 효과를 증명한다.

2. 서브 워드 분절 알고리즘 연구

2.1 Byte Pair Encoding

Byte Pair Encoding(BPE)은 2016년에 발표된 대표적인 서브 워드 분절 알고리즘이다[2]. BPE는 본래 1994년에 제안된 기술인데, 빈도를 기준으로 가장 많이 등장하는 바이트(byte) 쌍들을 순차적으로 하나의 바이트로 통합해감으로써 데이터를 압축한다. 이에 착안하여 서브 워드 분절에서는 주어진 단어사전 학습 코퍼스에서 가장 많이 등장하는 문자 쌍들을 병합해나감으로써 단어사전을 구축해나간다.

BPE의 전반적인 절차는 다음과 같다. 우선 문장들은 어절, 또는 단어 단위로 미리 분절화(pre-tokenization)하거나 또는 Moses와 같은 규칙 기반 토크나이저를 이용하여 분절한다. 이후 단어 복원을 위해 단어 마지막에 띄어쓰기임을 표시하는 지표 '·'를 추가한다. 미리 분절한 어절 또는 단어들은 문자(character) 단위로 다시 분리되고, 구축하고자 하는 사전은 위 단어들을 구성하는 문자들로 초기화한다. 이후 가장 많이 등장하는 문자 쌍을 찾아 하나로 합치고, 병합한 쌍은 앞서 만든 사전에 새로 추가한다. 이 과정은 미리 지정한 사전의 크기에 다다를 때까지 진행하는데, 이때 단어사전의 크기는 매개변수(hyper-parameter)로 설정 가능하다. 사전의 크기는 처음 초기화한 단어사전의 크기에서 쌍을 병합하는 횟수를 더한 값과 같다.

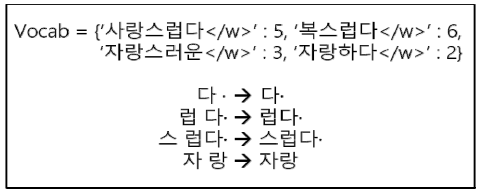

Fig. 1. Merging process of the BPE algorithm

Fig 1은 '사랑스럽다, 복스럽다, 자랑하고, 자랑하다'가 사전을 구축하기 위한 단어들이고, '자랑스럽다'가 새로운 입력으로 들어온 경우를 설명한 그림이다. 이 경우 Fig 1과 같이 '다'와 '·'이 '다·'로, '럽'과 '다·'가 '럽다·'로, '스'와 '럽다·'가 '스럽다·'로, '자'와 '랑'이 '자랑'으로 병합됨으로써 결국 '자랑스럽다'는 '자랑'과 '스럽다·'의 두 서브 워드들로 분절된다. 위 알고리즘을 이용하면 open vocabulary problem에서 사전에 등장하지 않은 단어들이 입력으로 들어왔다 하더라도 여러 서브 워드들의 조합을 통해 '사전에 없는 단어'가 아닌 '사전으로 설명할 수 있는 단어'로 만들 수 있다는 큰 장점이 있다. 즉 BPE 알고리즘은 등장 빈도가 가장 높은 쌍을 묶어나가는 방식으로 서브 워드 분절을 진행하게 되며 OOV 문제를 완

화하였다. 하지만 BPE 알고리즘을 이용하였을 때의 단점 역시 존재하는데, 등장 빈도가 높은 쌍들은 빠르게 병합되지만 드물게 등장하는 쌍에 대해서는 병합이 되지 않는 경우가 있다. 또한 다양한 확률에 의해 분절되지 않고 가장 많이 등장하는 쌍에 대해서만 병합하므로 잘못된 서브 워드 분절 결과가 나올 수 있다는 한계점이 존재한다. 즉, 본래 하나의 단어도 서브 워드에 따라 다양한 방법으로 분절될 수 있는데 BPE는 그리디(greedy) 방식으로 분절을 진행하므로 하나의 분절 결과밖에 얻을 수 없다. 뿐만 아니라 한국어의 경우에는 Fig 1의 예시에서 '-스럽다'와 '-스러운'처럼 기본형이 동일한 의미를 지니더라도 표현에 따라 형태가 다르다면 다른 단어로 간주된다는 문제가 발생한다. 한국어에서 나타나는 활용형은 숫자로 지정할 수 없을 만큼 무한하기 때문에 BPE의 방식을 이용한다면 제한된 양의 사전을 효율적으로 활용하기가 어렵다.

2.2 Unigram Language Model Tokenizer

Subword regularization 방법인 unigram language model tokenizer는 2018년에 제안된 서브 워드 분절 알고리즘이며[3] 기존 BPE 알고리즘의 단점을 지적하고 이에 대한 완화책을 제안하였다. Unigram language model tokenizer는 빈도가 아닌 확률을 기반으로 쌍을 병합하기 때문에 다양한 방법의 분절(multiple segmentation)결과를 얻을 수 있는 특징이 있다.

Unigram language model tokenizer는 "각 서브 워드들은 서로 독립이며 서브 워드의 나열로 구성된 문장(subword sequence)의 확률은 각각의 서브 워드가 나타날 확률들을 곱한 것과 같다."라는 가정을 기반으로 한다. 최적화된 사전을 구축하기 위해 학습 데이터로부터 적절한 크기의 근본 단어사전(seed vocabulary)을 생성한다. 이후 단어 사전의 개수가 지정해놓은 크기가 될 때까지 다음의 세 가지 과정을 반복한다. 첫째, 단어 집합을 고정하고 EM 알고리즘(Expectation Maximization Algorithm)을 이용해 서브 워드의 등장 확률을 최적화한다. 서브 워드의 등장 확률은 직접 알아내기 어렵기 때문에 EM 알고리즘을 이용하여 서브 워드가 등장할 기댓값을 최대화한다. 둘째, 각각의 서브 워드들에 대해 교차 엔트로피 손실 값(Cross-entropy loss)을 측정한다. loss 값은 특정 서브 워드가 현재 시점의 단어사전에서 제거되었을 때의 손실 값이다. 셋째, loss 값을 크기순으로 나열하고 상위 n%만을 남겨둔다. 사전에 등장하지 않은 단어들이 입력으로 들어올 수 있으므로 문자(single character)들은 남겨두어야 한다. 이 세 과정을 반복하여 미리 정한 사전의 크기에 도달하면 최적화된 사전 준비가 완료된다.

이렇게 미리 분절한 단어 사전을 준비한 후 서브 워드 샘플링(subword sampling) 과정을 진행한다. 먼저 Forward-DP Backward-A 알고리즘을 이용하여 확률에 따라 가장 높은 L개의 (l-best) 분할 후보를 얻는다. 이후 정확한 샘플링을 위해서는 Forward-Filtering and Backward-sampling algorithm(FFBS)을 이용한다. FFBS에서 서브 워드 분절 후보들은 격자 구조로 나타나 있으며, 각 서브 워드들은 노드들로 표현되어 있다. 먼저 모든 서브 워드들의 확률을 계산한 이후 문장의 끝부터 처음까지 노드들을 순회하면서 확률에 따라 재귀적으로 샘플링을 진행한다.

해당 방법론은 서브 워드 분절 과정에서 발생할 수 있는 noise에 대해서도 견고하게 대처할 수 있는 특징이 있다. 그러나 각 서브 워드들의 확률을 예측하기 위한 unigram language model, 단어사전을 최적화하기 위한 EM 알고리즘, 서브 워드 분절 샘플을 만들기 위한 Viterbi 알고리즘을 따로 학습해야 하므로 복잡하다는 단점이 존재한다.

2.3 BPE-Dropout

BPE-Dropout은 BPE의 성능을 개선하고 unigram language model tokenizer의 단점을 보완하고자 제안한 방법이다[4]. BPE-Dropout은 기존의 BPE 알고리즘을 기반으로, BPE merge table 중 무작위로 정해진 비율만큼 dropout을 실행하는 새로운 서브 워드 정규화(subword regularization) 방법론이다. 해당 방법론은 동일한 단어에 대해서도 다양한 서브 워드 분절이 가능해진다는 특징이 있다. 기존 BPE에 dropout을 추가하는 간단한 방법이면서도 드물게 나타나는 단어들에 대해 처리를 잘하며, 오타에 강하다는 장점을 보였다.

BPE-Dropout의 알고리즘은 기존 BPE 알고리즘과 동일하며 p% 만큼 무작위로 dropout을 진행한다. 만약 dropout을 진행할 확률 p가 0이라면 기존 BPE의 알고리즘과 동일하고 p가 1이라면 모든 word들이 구별되는 문자들로 나뉜다. 대개는 병합이 가능한 쌍들의 10% 정도를 dropout한다.

2.4 Word Piece Model

WPM(WordPieceModel)은 2012년에 처음 제안된 알고리즘[5]이며, 2016년 기계번역 분야에 이용되었다

[6]. 단어 사전을 문자 단위로 나눈 후 쌍들을 병합하는 방식은 앞선 BPE 알고리즘과 동일하다. 한 가지 다른 점은 BPE 방식은 빈도를 기반으로 단어를 병합하지만, WPM은 우도(likelihood)를 높이는 방식으로 단어를 병합한다.

먼저 학습 코퍼스의 문장에 대해 임의로 띄어쓰기 또는 규칙 기반의 토크나이저를 이용하여 pre-tokenization 과정을 거친 후 분절한 어휘들을 문자 단위로 나눈다. 쌍을 병합하는 모델링 과정을 시작하기 전 본 문장의 띄어쓰기된 부분에는 특별한 토큰(special token) '__'을 추가하며 문장을 복원하는 과정에서 본래 문장의 띄어쓰기와 서브 워드 분절을 구분하기 위함이다. 이후 쌍을 병합하는 과정에서는 각각 '서브 워드 1'과 '서브 워드 2'의 등장 확률을 곱한 후 '서브 워드 1, 서브 워드 2 pair'의 등장 확률로 이를 나눈 값이 가장 큰 쌍을 병합한다. 즉 병합하였을 때 우도를 가장 높이는 쌍을 병합해나가는 방식이다. 병합하는 과정은 미리 지정한 단어사전의 크기에 다다를 때까지 계속해서 반복한다.

2.5 Sentence Piece

Sentence Piece는 2018년에 제안된 비지도 학습 기반 토크나이저(tokenizer) 및 디토크나이저(detokenizer)이다[7].

앞서 소개한 서브 워드 분리 방법들은 모두 문장을 단어 단위로 분리하는 미리 분절화(pre-tokenization) 과정을 거쳐야 했다. 그러나 띄어쓰기가 없는 중국어, 일본어 또는 교착어의 특성을 지닌 한국어의 경우는 단어로 미리 분절하는 것 자체가 까다롭다. Sentence Piece에서는 위 문제를 지적하고, BPE, Unigram language model 등을 언어에 무관하게 원 문장(raw text)을 바로 학습에 사용할 수 있도록 보완했다. 즉, 언어마다 특화된 전, 후처리 방식을 이용하지 않고 end-to-end 시스템을 이용함으로써 NMT 구조와도 쉽게 통합할 수 있도록 만들었다.

Sentence Piece는 Normalizer, Trainer, Encoder, Decoder의 네 가지로 구성된다. Normalizer는 의미적으로 동일한 문자들을 표준 형식으로 정규화한다. Trainer에서는 앞서 정규화를 완료한 말뭉치로부터 서브 워드 분절 모델을 학습한다. Encoder에서는 Trainer로 학습된 모델을 이용하여 Normalizer로부터 정규화된 말뭉치를 서브 워드의 나열로 분절하며 Decoder에서는 서브 워드의 나열을 문장 형태로 변환한다. Sentence Piece는 입력 문장에 대해 빠른 속도로 단어를 id로 변경하고, 속도 역시 1초에 5만 개의 문장들을 분절할 수 있다는 장점이 존재한다.

3. 한국어에 특화된 음절 기반 종성 토큰화 방법론

3.1 기존 연구

한국어의 서브 워드 분절과 관련한 연구들은 다양하게 이루어지고 있다. [8]에서는 과거의 의미를 더하는 'ㅆ'과 같이 자모음이 통사적 속성을 나타내는 경우를 고려하기 위해 자모 단위로의 분절을 진행함으로써 data sparsity 문제를 완화할 수 있었다. 이와 비슷한 전략으로 [9]에서도 한국어의 단어를 자모음으로 분리한 후 임베딩을 진행하는 방식인 서브 워드 단위 인코딩 기법을 이용하였다. [10]에서는 고전번역을 진행하면서 엔티티에 대한 정보를 보존하기 위해 개체 정보에 대해 서브 워드 분리 작업에서 Restrict를 진행하였으며 이를 기반으로 BPE를 적용하였다. [1]에서는 교착어의 특성을 고려하여 SentencePiece를 이용한 서브 워드 분절 시 MeCab-ko를 이용하여 형태소 분절을 함께 진행했다. 이를 확장하여 [11]에서는 자모음, 음절, 형태소, 서브 워드, 형태소 분석과 서브 워드, 단어 단위 등 다양한 토큰화 전략들을 적용하여 실험했고, 형태소를 고려한 서브 워드 분절이 가장 좋은 성능을 냈음을 보였다.

3.2 제안하는 방법론

기존 연구 중 자모 단위로의 분리를 진행하면 초성, 중성, 종성이 모두 분리되게 되는데 [9]의 결과에 따르면 자모 단위로의 분리가 최적의 성능을 내지는 못했다. 본 논문은 초성과 중성, 종성을 모두 분리하지 않고 종성만을 초성과 중성에서 분리하는 방법을 제안한다. 용언에 해당하는 수많은 단어들이 있지만 하나의 예시를 들면 형용사 '예쁘다'는 '예쁘다, 예쁜, 예쁠, 예쁨'의 형태로 변환될 수 있다. 이에 대해 우리가 제안하는 방법론으로 분리를 하면 '예쁘다, 예쁘ㄴ, 예쁘ㄹ, 예쁘ㅁ'으로 분리가 되고, 서브 워드 사전에서 '예쁘'라는 하나의 단어로 세 활용형을 동일한 의미로 간주할 수 있게 된다. 용언은 모음 활용도 더러 있지만 '가-, 간-, 갔-, 갈-'(go), '잘-, 잔-, 자-, 잠'(sleep), '빨갛, 빨간, 빨가'과 같이 자음 활용으로 형태가 변환되는 경우들이 많다. 따라서 초성과 중성을 병합하는 과정을 먼저 진행하지 않고도 서브 워드 분절

을 통해 데이터 희소성 문제를 해결할 수 있다.

4. 실험

4.1 데이터 및 모델

학습을 위한 데이터 셋으로 OpenSubtitles[12], AI HUB[1]에서 제공하는 한-영 기계번역 말뭉치를 활용하였고[13], 테스트 데이터 셋으로는 Iwslt-16[2]을 이용하였다. 모델은 Fariseq에서 제공하는 Transformer[14] 모델을 이용해 번역을 학습했고 BLEU[15]를 이용하여 번역에 대한 평가를 진행했다. 실험을 위한 GPU 환경으로는 RTX 8000 2장을 이용했다.

4.2 실험 결과

Table 1. Subword tokenization experiment result

Subword-Tokenization	BLEU
Subword-nmt[2]	15.71
SP-BPE[7]	16.03
SP-Unigram[7]	**16.93**

실험 결과 BPE 알고리즘을 이용하였을 때 가장 성능이 안 좋았고, SentencePiece(SP)에서 모델 유형을 unigram language model로 설정했을 때 가장 성능이 좋았다. 이는 앞서 언급한 바와 같이 한국어는 미리 분절화가 어렵기 때문에 Subword-nmt를 이용하였을 때는 단순 어절 단위로 진행되었지만 SentencePiece는 미리 분절화 없이 바로 학습을 했기 때문에 더 좋은 성능을 보였다고 해석할 수 있다. 따라서 본 논문은 종성 분리 데이터 셋에 대한 실험에서 SentencePiece의 unigram language model로 실험을 진행했다. 실험 결과는 Table 2와 같다.

Table 2. Experimental results of the proposed methodology

Subword-Tokenization	BLEU
SP-Unigram+Syllable	16.93
SP-Unigram + Jongseong tokenization	**17.15**

실험 결과 본 논문에서 제안한 종성만 분리한 방법론이 음절단위 방법론보다 0.22 BLEU만큼 더 높은 성능을 보였다. 이를 통해 음절에서 종성을 분리한 후 서브 워드 학습을 진행하면 더 좋은 성능을 낼 수 있음을 확인했다. 지나치게 초성과 중성, 종성을 모두 분절하지 않음으로써 서브 워드 학습의 본질을 흐리지 않았고 동시에 용언의 다양한 활용형들을 종성 분리로 인해 잘 커버했다.

더 나아가 더 높은 기계번역의 성능을 얻기 위하여 beam size를 각각 다르게 한 후 실험을 진행했고, 실험 결과는 Table 3과 같다.

Table 3. Experiment result according to beam size

Beam Size	JongSung	Syllable
1	16.30	15.88
2	16.60	16.59
3	17.07	16.85
4	17.32	16.89
5	17.15	16.93
6	17.24	**17.05**
7	17.33	16.94
8	**17.36**	16.94
9	17.35	17.00
10	17.30	16.96

실험 결과 Beam size를 기본 5로 설정했을 때보다 종성 분리 시 0.21 BLEU 만큼 성능이 향상되었고 음절에서도 0.12 BLEU 만큼 점수를 높일 수 있었다. 종성 분리 시에는 beam size를 8로 설정했을 때가 17.36으로 기존 서브 워드 분절 방법의 최대 성능보다 0.31 BLEU 만큼 향상되었다. 이처럼 같은 모델의 동일한 체크포인트를 이용했을 때 beam size를 조절하는 것만으로도 성능을 크게 향상시킬 수 있었다.

본 논문은 추가적으로 질적 분석을 통해 과연 음절 단위로 서브 워드 학습을 진행했을 때보다 종성을 분리하여 학습했을 때 사전이 더 문장을 잘 설명할 수 있고 일반화 효과를 줄 수 있었는지 확인해 보았다.

Table 4. Qualitative analysis of vocabulary obtained by the proposed methodology

Jongsung Tokenization word dictionary
서ㄴ저ㅇ/겨ㅇ기/
ㅆ스ㅂ니다/ㅆ어요/ㅆ고/ㅆ어/ㅆ더ㄴ/ㅆ지마ㄴ/ㅂ시다/ㅂ니다/ㄹ러/ㄹ라/ㄹ까/ㅁ으ㄴ/ㅁ으로써/ㄴ드
/사거ㄴ이/도ㄹ아가/되기/내려가/느ㄹ어나/채ㅇ겨/드ㄹ려/마ㅌ기/노려ㄱ해/

1)http://www.aihub.or.kr/aidata/87

2)https://sites.google.com/site/iwsltevaluation2016/mt-track

사전을 확인한 결과 Table 4의 '서ㄴ저ㅇ(선정), 겨ㅇ기(경기)'처럼 명사들은 온전하게 본래의 의미를 복원한 것을 확인할 수 있다. 또한 종성 분리로 인해 'ㅆ습니다, ㅂ니다, ㄹ까, ㅁ으로써'와 같은 용언 활용들을 단어에서 분리해낼 수 있었고, '사거ㄴ이, 도ㄹ아가, 내려가'와 같은 단어 뒤에 활용형들을 덧붙일 수 있도록 함으로써 더 많은 용언 활용을 커버할 수 있었다. 데이터 셋에 대해서도 서브 워드 분절을 적용한 결과를 확인해보았고 '모/으/ㅂ니다, 추워/ㅆ으니까요, 부르ㄴ다/ㅂ니다, 구ㅇ그ㅁ하/시/ㄹ까봐, 새ㅇ가ㄱ하시/ㄹ'과 같이 나뉜 결과를 확인했다. '모읍니다'의 경우 음절 단위로의 서브 워드 분절을 진행했었더라면 '모으-'와 '모읍-'이 각각 다른 의미로 적용되었겠지만 종성을 분리한 결과 '모으-'의 하나의 단어로도 이들을 잘 설명할 수 있다. 결국 음절에서 종성을 분리해내어 단어사전을 구성함으로써 사전에 있는 하나의 단어로도 더욱 많은 활용형들에 이용될 수 있음을 확인했다.

5. 결론

하나의 단어는 의미 있는 여러 내부 단어들의 조합으로 구성된 경우가 많기 때문에, 서브 워드 분절(Subword Tokenization)은 하나의 단어를 서브 워드들로 분리해서 단어를 이해한다는 의도로 접근한 전처리 작업이다. 본 논문은 다양한 서브 워드 분절 알고리즘들에 대해 자세히 소개했다. 또한 한국어에서, 무한한 용언 활용형으로 인해 동일한 단어가 각각 다른 단어로 취급되는 경우를 최소화하기 위해 음절 중 종성을 분리하는 새로운 서브 워드 분절 방법을 제안했다. 정량적 분석을 통해 제안하는 방법론에 대한 효과를 증명했으며, 정성적 분석을 통해 실제 한국어의 용언 활용을 효율적으로 다룰 수 있도록 사전이 구축되었음을 확인했다.

REFERENCES

[1] C. Park, Y. Yang, K. Park & H. Lim. (2020). Decoding strategies for improving low-resource machine translation. *Electronics, 9(10)*, 1562.
DOI : 10.3390/electronics9101562

[2] R. Sennrich, B. Haddow & A. Birch. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909.*
DOI : 10.18653/v1/P16-1162

[3] T. Kudo. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959.*
DOI : 10.18653/v1/P18-1007

[4] I. Provilkov, D. Emelianenko & E. Voita. (2019). Bpe-dropout: Simple and effective subword regularization. *arXiv preprint arXiv:1910.13267.*
DOI : 10.18653/v1/2020.acl-main.170

[5] M. Schuster & K. Nakajima. (2012, March). Japanese and korean voice search. *In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 5149-5152.
DOI : 10.1109/ICASSP.2012.6289079

[6] Y. Wu et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144.*

[7] T. Kudo & J. Richardson. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226.*
DOI : 10.18653/v1/D18-2012

[8] K. Stratos. (2017). A sub-character architecture for Korean language processing. *arXiv preprint arXiv:1707.06341.*
DOI : 10.18653/v1/D17-1075

[9] S. Moon & N. Okazaki. (2020, May). Jamo Pair Encoding: Subcharacter Representation-based Extreme Korean Vocabulary Compression for Efficient Subword Tokenization. *In Proceedings of The 12th Language Resources and Evaluation Conference,* 3490-3497.

[10] C. Park, C. Lee, Y. Yang & H. Lim. (2020). Ancient Korean Neural Machine Translation. *IEEE Access, 8,* 116617-116625.
DOI : 10.1109/ACCESS.2020.3004879

[11] K. Park, J. Lee, S. Jang & D. Jung. (2020). An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks. *arXiv preprint arXiv:2010.02534.*

[12] P. Lison & J. Tiedemann. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

[13] C. Park & H. Lim. (2020). A Study on the Performance Improvement of Machine Translation Using Public Korean-English Parallel Corpus. *Journal of Digital Convergence, 18(6),* 271-277.
DOI : 10.14400/JDC.2020.18.6.271

[14] A. Vaswani et al. (2017). Attention is all you need. *In Advances in neural information processing systems,* 5998-6008.

[15] K. Papineni, S. Roukos, T. Ward & W. J. Zhu. (2002, July). BLEU: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting of the Association for Computational Linguistics,* 311-318.
DOI : 10.3115/1073083.1073135

어 수 경(Sugyeong Eo) [학생회원]

· 2020년 8월 : 한국외국어대학교 언어인지과학과, 언어와공학전공 (문학사, 언어공학사)
· 2020년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정

· 관심분야 : Neural Machine Translation, Quality Estimation, Deep Learning
· E-Mail : djtnrud@korea.ac.kr

박 찬 준(Chanjun Park) [학생회원]

· 2019년 2월 : 부산외국어대학교 언어처리창의융합전공 (공학사)
· 2018년 6월 ~ 2019년 7월 : SYSTRAN Research Engineer
· 2019년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정

· 관심분야 : Machine Translation, Grammar Error Correction, Deep Learning
· E-Mail : bcj1210@naver.com

문 현 석(Hyeonseok Moon) [학생회원]

· 2021년 2월 : 고려대학교 수학과 (이학사)
· 2021년 3월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정

· 관심분야 : Neural Machine Translation
· E-Mail : glee889@korea.ac.kr

임 희 석(Heuiseok Lim) [종신회원]

· 1992년 : 고려대학교 컴퓨터학과(이학학사)
· 1994년 : 고려대학교 컴퓨터학과 (이학석사)
· 1997년 : 고려대학교 컴퓨터학과 (이학박사)
· 2008 ~ 현재 : 고려대학교 컴퓨터학과 교수

· 관심분야 : 자연어처리, 기계학습, 인공지능
· E-Mail : limhseok@korea.ac.kr