# 한-영 관용구 기계번역을 위한 NMT 학습 방법

최민주○, 이창기

강원대학교 컴퓨터과학과
mjchoi0831@gmail.com, leeck@kangwon.ac.kr

# NMT Training Method for Korean-English Idiom Machine Translation

Min-Joo Choi○, Chang-Ki Lee
Dept. of Computer Science, Kangwon National University

**요 약**

관용구는 둘 이상의 단어가 결합하여 특정한 뜻을 생성한 어구로 기계번역 시 종종 오역이 발생한다. 이는 관용구가 지닌 함축적인 의미를 정확하게 번역할 수 없는 기계번역의 한계를 드러낸다. 따라서 신경망 기계 번역(Neural Machine Translation)에서 관용구를 효과적으로 학습하려면 관용구에 특화된 번역 쌍 데이터셋과 학습 방법이 필요하다. 본 논문에서는 한-영 관용구 기계번역에 특화된 데이터셋을 이용하여 신경망 기계번역 모델에 관용구를 효과적으로 학습시키기 위해 특정 토큰을 삽입하여 문장에 포함된 관용구의 위치를 나타내는 방법을 제안한다. 실험 결과, 제안한 방법을 이용하여 학습하였을 때 대부분의 신경망 기계 번역 모델에서 관용구 번역 품질의 향상이 있음을 보였다.

주제어: 기계번역, NMT, OpenNMT, 관용구, 관용구 번역

## 1. 서론

관용구는 두 개 이상의 단어로 이루어져 있으면서 그 단어들의 의미만으로는 전체의 의미를 알 수 없는, 특수한 의미를 나타내는 어구로, 단어 각각의 뜻을 직역하여 조합한 것과는 다른 의미를 가진다. 따라서 관용구를 번역할 때 관용구를 구성하는 단어 그대로 번역할 경우 그림 1의 예와 같은 오역이 발생한다.


**그림 1 관용구 포함 문장 기계번역 결과**

이러한 한계를 극복하기 위해서는 기계번역 모델이 관용구를 효과적으로 학습할 수 있도록 관용구 번역에 특화된 학습 방법이 필요하다.

본 논문에서는 기존에 구축한 관용구 기계번역을 위한 한-영 번역 쌍 데이터셋 'KISS[1]'를 활용하고, 특정 토큰을 이용해 문장에 포함된 관용구의 위치를 표기하여 기계번역 모델이 효과적으로 관용구를 학습시킬 수 있는 방법을 제안한다. 실험 결과, 제안한 방법을 이용하여 학습하였을 때 대부분의 신경망 기계 번역 모델에서 번역 품질의 향상이 있음을 보였다.

## 2. 관련 연구

신경망 기계번역 (Neural Machine Translation, 이하 NMT) 모델이 개발되기 이전에는 관용구가 포함된 문장을 번역할 때 문장 내에 관용구 사전에 존재하는 관용구가 포함되어 있으면 관용구로 번역하는 방법을 이용하였다[1]. 그러나 관용구 사전을 이용하는 방법은 구축하는 데에 많은 시간과 노력이 들고, End-to-end 방식인 NMT 모델에는 적합하지 않다.

따라서 NMT 모델이 관용구를 학습하기 위해 관용구 번역 쌍 데이터셋을 구축하기 위한 연구와[2], NMT 모델을 이용해 관용구를 번역한 결과를 효과적으로 평가하기 위해 블랙리스트를 구축하는 연구가 이루어졌다[3].

또한 이러한 선행연구를 참고하여 NMT 모델을 이용하여 한국어 관용구를 영어로 번역하기 위해 한국어 관용구의 영어 번역에 특화된 데이터셋을 구축하고, 데이터셋에 포함된 오역 쌍을 제거하고 번역 결과를 평가하는 방법에 대한 연구가 이루어졌다[4]. 이는 공개된 한-영 번역 쌍 말뭉치로부터 관용구가 포함된 번역 쌍을 추출하여 관용구 학습 데이터셋 'KISS'를 구축하는 방법과, 블랙리스트를 적용하여 오역 결과를 제거하고 번역 품질을 평가하는 방법을 포함한다.

---

[1] KISS : Korean-english Idioms in Sentences dataset
https://github.com/Judy-Choi/KISS

본 논문은 선행연구[4]의 연장선으로, 구축한 관용구 번역 쌍 데이터셋을 이용하여 NMT 모델이 관용구를 더욱 정확하게 번역할 수 있도록 효과적으로 학습시키는 방법에 대해 서술한다. 기존 논문[2, 5]에서 제안한 대로 문장에 포함된 관용구의 위치를 표기하는 다양한 방법을 한국어-영어 관용구 번역에 적용하고 다양한 NMT 모델간 성능 비교를 통해 기존 논문에서 제안한 방법의 효과를 검증한다.

## 3. 데이터셋 구성

### 3.1 데이터셋 구축

NMT 모델은 다량의 문장 번역 쌍을 이용하여 학습하므로 NMT 모델이 관용구를 학습하기 위해서는 관용구가 포함된 다량의 문장 번역 쌍이 필요하다. [4]에서는 관용구가 포함된 한-영 번역 쌍 7,500개로 이루어진 데이터셋 'KISS' 를 구축하였으며, 블랙리스트를 이용하여 오역을 제거하여 최종적으로 3,461개의 번역 쌍으로 이루어진 데이터셋을 생성하였다. 본 논문에서는 해당 데이터셋을 수작업으로 한번 더 검증하여 최종적으로 3,377개의 번역 쌍으로 구성된 관용구 번역 쌍 데이터셋을 구축하여 실험에 이용한다.

NMT 모델의 학습을 위해 AI Hub 에서 제공하는 한국어-영어 번역(병렬) 말뭉치 AI 데이터를 사용하였으며, 해당 말뭉치는 문어체 한영 번역 110만 문장(뉴스 80만, 정부 웹사이트 컨텐츠 10만, 조례 10만, 한국문화 10만)과 구어체 한영 번역 50만 문장(구어체 40만, 대화체 10만)으로 이루어져 있다(총 160만개의 한영 번역 문장). KISS 또한 AI Hub 데이터로부터 추출된 것이므로, 본 실험에서는 KISS로부터 오역을 제거한 총 3,377개의 관용구 포함 문장 쌍 데이터와, KISS를 제외하고 중복을 제거한 나머지 AI Hub 데이터를 함께 이용하였다.

### 3.2 데이터셋 분할

**표 1 데이터셋 문장 개수**

| | 관용구 포함 문장 쌍 | 관용구 미포함 문장 쌍 |
|---|---|---|
| 개발셋 | 500 | 5,000 |
| 학습셋 | 1,877 | 1,577,426 |
| 평가셋 | 1,000 | 10,000 |
| 합계 | 3,377 | 1,592,426 |

표1과 같이 관용구가 포함된 문장, 관용구가 포함되지 않은 일반 문장으로 각각 나누어 학습 및 평가를 위한 데이터셋으로 분할하였다. 관용구가 포함되지 않은 AI Hub 데이터셋은 학습셋 1,577,426 문장 쌍, 개발셋 5,000 문장 쌍, 평가셋 10,000 문장 쌍으로 분할하여 사용하였다. 3.1에서 추출한 관용구 포함 문장 쌍은 학습셋 1,877 문장 쌍, 개발셋 500 문장 쌍, 평가셋 1,000 문장 쌍으로 분할하여 사용하였다.

### 3.3 실험 조건별 학습 데이터셋 구축

3.2에서 분할한 데이터셋에 실험 조건별로 다르게 전처리를 적용하여 실험을 수행하였다. [2]에서 문장 내의 관용구 위치를 나타내기 위해 특수한 토큰 '<idm>' 을 관용구 앞에 부착하는 방법을 제안하였으며, [5]에서는 <idm> 대신 <fig> 토큰을 관용구 앞에 부착하는 방법과 함께 관용구의 시작과 끝을 나타내는 <start_fig>, <end_fig> 토큰을 관용구 앞뒤에 각각 부착하는 방법을 제안하였다. 이러한 선행 연구를 참고하여 실험 조건을 다음 4가지로 설정하여 조건에 맞는 학습셋을 구성하였다.

1. 관용구가 포함되지 않은 문장으로만 이루어진 학습셋: 1,577,426 문장 쌍
2. 관용구를 포함하는 문장과 관용구가 포함되지 않은 문장으로 이루어진 학습셋: 1,577,426 + 1,877 문장 쌍
3. 2에서 문장 내 관용구의 시작 위치를 <idm> 으로 표기한 학습셋
4. 2에서 문장 내 관용구의 시작과 끝 위치를 각각 <idm>, </idm> 으로 표기한 학습셋

**표 2 관용구 학습 데이터셋 예시**

| | |
|---|---|
| 1 | 관용구 미포함 |
| 2 | **발을 동동 구르는** 상황이다. |
| 3 | **<idm>발을 동동 구르는** 상황이다. |
| 4 | **<idm>발을 동동 구르는</idm>** 상황이다. |

## 4. 실험 및 평가

### 4.1 실험

3.3 에서 구축한 데이터셋에 형태소 분석을 적용하고 Byte Pair Encoding[2][6]을 적용하였다(32,000회 적용). 관용구를 학습하지 않는 실험의 경우 문장 내에 관용구가 포함되지 않은 총 1,577,426개의 문장 쌍을, 관용구를 학습하는 경우에는 1,577,426개의 문장 쌍에 관용구가 포함된 문장 쌍 1,877개를 추가로 더하여 학습에 이용하였다.

다양한 NMT 모델로 실험하기 위해 OpenNMT[7]를 이용하였으며, NMT 모델은 총 4가지로 OpenNMT 의 기본 모델인 2-Layer LSTM 모델, [5]에서 이용한 4-Layer LSTM 모델, 그리고 4-Layer Transformer 모델과 6-Layer Transformer 모델이다. 이 중 6-Layer Transformer 모델은 학습 시 5개의 Random Seed를 주어 학습시킨 후, 각각 다른 5개의 seed값으로 학습된 모델의 BLEU 점수의 평균값을 구하여 다른 모델의 BLEU 점수와 비교하였다.

[2] https://google.github.io/seq2seq/nmt/

**표 3 각 번역 모델의 Hyperparameter**

| | 2 Layer LSTM (OpenNMT Default) | 4 Layer LSTM | Transformer |
|---|---|---|---|
| Encoder hidden state size | 500 | 1,000 | 512 |
| Batch size | 64 | 100 | 2,048 |
| Dropout | 0.3 | 0.1 | 0.1 |
| Train steps | 100,000 | 100,000 | 100,000 |

**표 4 실험 결과 BLEU 점수 [8] 비교**

| NMT 모델 | | 2 Layer LSTM (OpenNMT Default) | | 4 Layer LSTM | | 4 Layer Transformer | | 6 Layer Transformer | |
|---|---|---|---|---|---|---|---|---|---|
| 데이터셋 | | 관용구 포함 평가셋 | 관용구 미포함 평가셋 | 관용구 포함 평가셋 | 관용구 미포함 평가셋 | 관용구 포함 평가셋 | 관용구 미포함 평가셋 | 관용구 포함 평가셋 | 관용구 미포함 평가셋 |
| 1 | 관용구 미포함 학습 (Baseline) | 24.45 | **30.32** | 26.05 | **32.76** | 28.28 | 33.58 | 28.37 | **34.45** |
| 2 | 관용구 포함 학습 | 24.37 (-0.08) | 30.23 (-0.09) | 26.45 (+0.40) | 32.63 (-0.13) | 29.09 (-0.81) | 34.09 (+0.51) | **28.73 (+0.36)** | 34.22 (-0.23) |
| 3 | 관용구 포함 학습 + <idm> 태그 | 24.32 (-0.13) | 30.16 (-0.16) | **26.76 (+0.71)** | 32.62 (-0.14) | **29.46 (+1.18)** | **34.71 (+1.13)** | 28.42 (+0.05) | 34.02 (-0.43) |
| 4 | 관용구 포함 학습 + <idm> </idm> 태그 | **24.51 (+0.06)** | 30.00 (-0.32) | 25.89 (-0.16) | 32.68 (-0.08) | 27.63 (-0.65) | 33.88 (+0.30) | 27.16 (-1.21) | 33.88 (-0.57) |

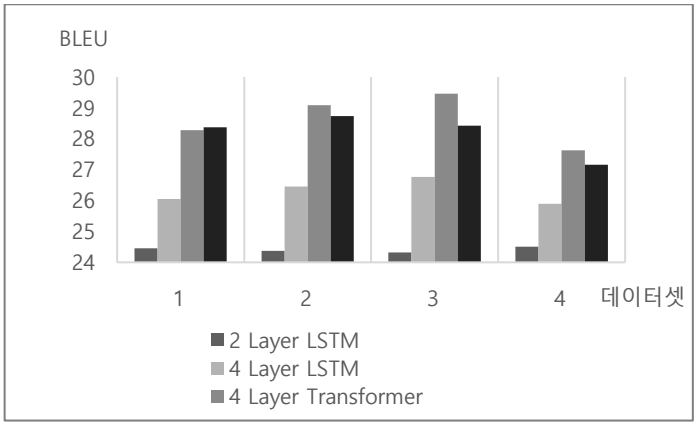


**그림 2 관용구가 포함된 문장의 BLEU 점수 비교**

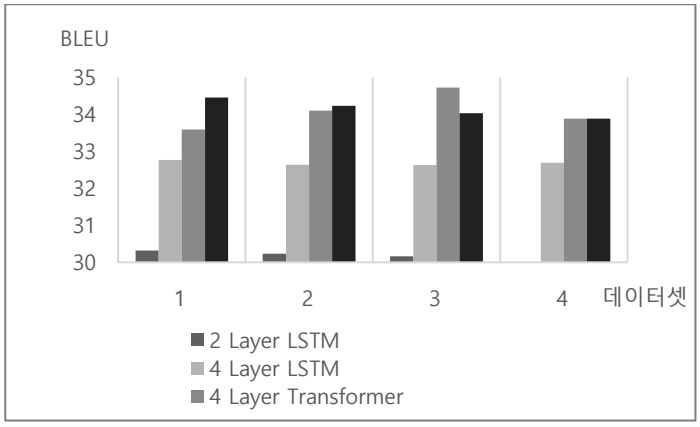


**그림 3 관용구가 미포함된 문장의 BLEU 점수 비교**

또한 기존의 선행연구에서는 관용구가 포함된 문장의 번역 성능만을 평가하였으나, 본 논문에서는 관용구가 포함되지 않은 일반 문장의 번역 성능까지 함께 평가하기 위해 관용구가 포함된 문장 쌍 1,000개와 관용구가 포함되지 않은 문장 쌍 10,000개를 각각 평가에 이용하였다.

### 4.2 평가

실험 결과 문장 내에 토큰을 삽입하여 관용구 위치를 표기하였을 때(3번 및 4번 조건) 관용구가 포함된 문장의 번역 성능이 대부분의 NMT 모델에서 향상되었다. 특히, 3번 조건(관용구 포함 학습 + <idm> 태그 사용)에 4 Layer Transformer 모델을 적용했을 경우에 모든 평가셋(관용구 포함 및 미포함 평가셋)에서 가장 좋은 성

능을 보였다(그림 2와 3 참고). 그림 2와 같이 관용구가 포함된 문장의 각 모델별 번역 성능 비교 결과 관용구를 함께 학습시키거나 <idm> 태그만 사용하였을 경우 성능이 가장 좋음을 알 수 있다. 다만 그림 3과 같이 관용구가 포함된 문장을 함께 학습시킬 경우 관용구가 포함되지 않은 일반적인 문장의 번역 성능이 4 Layer Transformer 모델 을 제외한 대부분의 NMT 모델에서 저하되었다. 또한 [5]에서는 영어-독일어 관용구의 시작과 끝 위치를 모두 표기하였을 때에 가장 높은 번역 성능을 보였으나 한-영 관용구 번역에서는 관용구의 시작 위치만을 표기하였을 때에 가장 높은 번역 성능을 보인다.

**표 5 관용구 학습 방법에 따른 번역 결과 비교**

| | 관용구 학습 방법 | 번역 결과 |
|---|---|---|
| 원문 | 학부모만 **발을 동동 구르는** 상황이다. | only the parents are **anxious**. |
| 1 | 관용구 미포함 학습 (Baseline) | only parents are **rolling their feet**. |
| 2 | 관용구 포함 학습 | only parents are **rolling their feet**. |
| 3 | 관용구 포함 학습 + <idm> 태그 | only parents are **struggling**. |
| 4 | 관용구 포함 학습 + <idm> </idm> 태그 | only parents are **jumping**. |

## 5. 결론

NMT 모델을 이용한 기계번역 시 관용구가 포함된 문장은 그렇지 않은 문장에 비해 번역의 정확도가 낮다. 이를 보완하기 위해 본 논문에서는 관용구 기계번역에 특화된 데이터셋을 이용하여 NMT 모델에 관용구를 효과적으로 학습시키기 위해 특수한 토큰을 삽입하여 문장 내 관용구의 위치를 표기하였고, 실험 결과, 대부분의 NMT 모델에서 번역 품질이 상승하였다. 또한 한국어의 경우 관용구의 시작과 끝 위치를 모두 표기할 경우보다 시작 위치만을 표기할 경우 번역 정확도가 가장 크게 증가하였다.

향후 연구로는 NMT 모델에 관용구뿐만이 아니라 숙어, 속담과 같이 함축적인 의미를 지닌 어휘를 효과적으로 학습시키는 연구를 진행할 예정이다. 또한 NMT 모델에 관용구 포함 문장 및 관용구의 위치를 나타내는 토큰을 삽입하여 학습시킬 경우 관용구가 포함되지 않은 문장의 번역 성능이 저하되므로 관용구가 포함되지 않은 일반적인 문장의 번역 품질을 유지하면서 관용구 번역 품질을 개선할 수 있는 다른 학습방법을 연구할 필요가 있다.

## 참고문헌


[1] 이호석, 김영택. 영어-한국어 기계번역을 위한 연어와 숙어 트랜스퍼 사전. (구) 정보과학회논문지. 20.7: 976-987. 1993.

[2] Fadaee Marzieh, et al. Examining the tip of the iceberg:A data set for idiom translation. arXiv preprint arXiv:1802.04681. 2018.

[3] Shao Yutong, et al. Evaluating machine translation performance on Chinese idioms with a blacklist method. arXiv preprint arXiv:1711.07646. 2017.

[4] 최민주. 기계번역을 위한 한-영 관용구 데이터셋 구축 및 평가 방법. 2020 한국컴퓨터종합학술대회 논문집. pp 380-382. 2020.

[5] Liu, Changsheng. Toward Robust and Efficient Interpretations of Idiomatic Expressions in Context. Diss. University of Pittsburgh. 2019.

[6] Sennrich Rico, et al. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909. 2015.

[7] Klein, Guillaume, et al. Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810. 2017

[8]Papineni Kishore, et al. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics. 2002.



감사의 글 : 이 논문은 현대 자동차 AIR Lab의 "신경망 기계번역 모델을 위한 지식 증류 기술 연구" 과제의 지원을 받아 연구되었음