고유명사 기호화를 통한 신경망기반 한영 번역
(Kor-Eng NMT using Symbolization of Proper Nouns)

김 명 진 † 남 준 영 † 정 희 석 † 최 희 열 §
(Myungjin Kim) (Junyeong Nam) (Heeseok Jung) (Heeyoul Choi)

요 약 신경 기계 번역 분야는 딥러닝의 발전과 함께 성능이 발전하고 있지만, 이름, 신조어, 특정 그룹 내에서만 통용되는 단어 등과 같이 고유명사들이 들어간 문장의 번역이 정확하지 않은 경우들이 있다. 본 논문은 고유명사가 들어간 문장의 번역 성능 개선을 위해 최근 제안된 번역 모델인 Transformer Model에 추가적으로 한영 고유명사 사전과 고유명사 기호화 방식을 사용한다. 제안된 방식은 학습에 사용되는 문장의 단어들 중 일부를 고유명사 사전을 이용하여 기호화하고, 기호화된 단어들을 포함한 문장들로 번역 모델을 학습시킨다. 새로운 문장 번역시에도 고유명사 사전을 이용하여 기호화하고 번역후 복호화 하는 방식으로 번역을 완성한다. 제안된 방식의 성능을 검증하기위해 고유명사 기호화를 사용하지 않은 모델과 함께 비교 실험하였고, BLEU 점수를 통해 수치적으로 개선되는 것을 확인했으며, 몇가지 번역 사례들도 상용서비스 결과들과 함께 제시했다.

키워드 : 신경 기계 번역, 고유명사 번역, 기호화, 고유명사 사전

Abstract In the field of neural machine translation, performance is advancing, but there are cases where the translation of sentences containing proper nouns, such as names, new words, and words that are used only within a specific group, is not accurate. To handle such cases, this paper uses the Korean-English proper noun dictionary and the symbolization method in addition to the recently proposed translation model, Transformer Model. In the proposed method, some of the words in the sentences used for learning are symbolized using a proper noun dictionary, and the translation model is trained with sentences including the symbolized words. When translating a new sentence, the translation is completed by symbolizing, translation, and desymbolizing. The proposed method was compared with a model without symbolization, and improvement was quantitatively confirmed based on the BLEU score. In addition, several examples of translation are also presented along with commercial service results.

Keywords : Neural Machine Translation, Proper Noun Translation, Symbolization, Proper Noun Dictionary

† 학생회원 : 한동대학교 전산전자공학부

§ 종신회원 : 한동대학교 전산전자공학부 교수
heeyoul@gmail.com (Corresponding author)

논문접수 : 2020년 08월 일
심사완료 : 년 월 일

1. 서론

인공지능기반 자연어 처리 성능이 급속히 발전하고 있는데, 특히 학습 가치가 높은 데이터의 사용과 딥러닝의 발전이 신경 기계 번역의 수준을 향상시키는데 크게 기여하였다[1][2][3][4]. 국내에서 많이 사용되고 있는 번역기인 네이버 파파고(papago.naver.com)와 구글 번역기(translate.google.com) 또한 수준 높은 데이터와 신경 기계 학습 기반의 번역 모델을 사용하여 뛰어난 번역 성능을 보여주고 있다. 하지만, 이 번역기들은 고유명사가 들어간 문장을 번역할 때, 정확하지 않게 번역하는 문제를 가지고 있다. 아래 예시는 2020년 6월 16일 기준 결과이다.

입력 문장: “루이싱 커피의 가치가 떨어졌다.”
파파고: “The value of Luixing coffee has fallen.”
구글: “The value of louise coffee fell.”

“루이싱 커피”의 영어표현은 “Luckin Coffee” 인데, 위 번역 결과에서 파파고 번역기는 “Luixing coffee”, 구글 번역기는 “louise coffee”로 잘못 번역하는 문제를 보인다. 두 번째 예시는 아래와 같다.

입력 문장: “신명기는 성경의 한 부분이다.”
파파고: “The Divine Spirit is a part of the Bible.”
구글: “Deuteronomy is part of the Bible”

위 번역 결과에서는 구글의 번역은 정확한 반면, 파파고 번역기가 “Deuteronomy” 를 “The Divine Spirit”으로 잘못 번역하는 문제를 보인다.

본 논문은 이와 같이 한영번역에서 고유명사를 오역하는 문제를 해결하기 위해 고유명사 사전과 기호화(Symbolization) [5] 방식을 사용한다. 고유명사 사전에 고유명사로서 가치가 높은 단어들이 많을 수록 번역 성능이 높아진다. 또한 본 논문은 웹 시스템에서 사용자로부터 고유명사를 입력받음으로써 지속적으로 고유명사 데이터를 추가하는 방식을 제공한다.

본 논문에서 사용한 번역 모델은 Self-Attention Mechanism을 사용한 Transformer모델 [4] 이고 학습에 사용된 데이터는 AI 통합 플랫폼 사이트 AI Hub(www.aihub.co.kr) 에서 제공하는 한국어-영어 번역(병렬) 말뭉치 AI 데이터를 사용한다. BLEU(Bilingual Evaluation Understudy) Score를 통해 고유명사 기호화를 사용하지 않은 모델과 고유명사 기호화를 사용한 모델의 성능을 비교하여 성능 향상을 증명하고자 한다. 또한 몇가지 번역 사례들을 파파고와 구글의 결과와 비교하여 제시한다.

본 논문은 배경 지식으로 시작하여 제안 모델, 실험 결과 그리고 결말을 통해 서론에서 언급한 문제에 대한 구체적인 해결 방안을 서술한다. 배경 지식에서는 연구에 사용된 다양한 기술에 대한 내용을 서술하고 제안 모델에서는 데이터 관련 내용과 본 논문의 핵심 기술인 고유명사 사전과 기호화에 대해 서술한다. 실험결과에서는 고유명사 기호화를 사용한 번역 모델의 실험 결과와 성능을 보여주고 결론부에서는 논문에 대한 전반적인 요약이 제시된다.

2. 배경 지식

신경망 기반 언어모델 (language model) 이나 기계번역 (neural machine translation) 과 같은 자연어 처리 관련하여 딥러닝 초기에는 주로 Recurrent Neural Networks (RNNs) 모델을 사용했다 [6][7][8]. 이후 Convolution Neural Networks (CNNs) 를 사용하는 모델이 제시되었고 [9], 최근에는 주의집중기법 (Attention Mechanism) 을 사용하는 Transformer 모델이 신경망 기반 번역 모델의 주류를 이루고 있다 [4].

RNN 모델은 문장이 길어질수록 성능이 떨어지는 문제를 해결하기 위해 Long Short-Term Memory(LSTM) 모델을 사용했다. 하지만 LSTM도 RNN 의 일종으로 문장내 단어 간의 관계를 모델링하기에는 한계가 있고 [3], 문장이 길어질 경우 학습 시간이 늘어나는 문제점도 가지고 있다. Transformer 모델은 기존의 Encoder와 Decoder 구조를 유지하지만 오로지 Attention 만을 사용하여 RNN 기반의 모델이 가진 단점을 개선하면서 학습시간이 더 빠르고 향상된 성능을 보여준다. 본 논문에서도 Transformer 모델을 기본 모델로 사용한다.

2.1. Transformer

Transformer 모델은 RNN 모델처럼 Encoder-Decoder 구조를 유지하고 있지만, RNN 모델에 비해 다음과 같은 장점들을 가지고 있다. 첫째, RNN 모델과 다르게 Encoder와 Decoder 각각 N 개의 계층을 가지고 있어서 입력과 출력 사이의 더 복잡한 관계를 모델링 할 수 있다. 둘째, Transformer 모델은 문장의 각 단어들이 순서대로 입력되는 것이 아니라 문장 전체가 한 번에 입력되기 때문에 병렬계산을 통해 연산 속도가 RNN 보다 빠르다는 장점이 있다. 마지막으로, Encoder와 Decoder에서 Recurrent connection 이 아니라 Attention을 통해 문장내 단어 간의 관계를 모델링하는데, 이는 멀리 떨어진 단어 간의 관계를 훨씬 더 효과적으로 모델링 할 수 있다. 최근 자연어 처리에 있어서 선행

학습된 모델인 BERT와 GPT 등에서도 Transformer 가 기본적으로 사용되고 있다 [10][11].

2.2. 기호화

자연어 처리에서 모든 단어를 포함하기 위해 단어의 리스트가 너무 길어지는 경우가 있는데, 그렇게 되면 모델의 크기가 커지는 문제로 이어진다. 또한 흔하지 않은 단어들의 경우 학습이 잘 되지 않는 문제도 발생된다.

또한 단어 리스트를 아무리 키우더라도 학습 코퍼스에 없었던 단어의 경우는 학습의 기회가 전혀 없어서 번역이 불가능하다. 이를 해결하기위해 몇가지 방법들이 있는데 출력문장의 단어를 출력할 때 입력 문장의 특정 단어를 가리키게 하거나 단어단위를 음절단위로 내려서 번역 모델을 학습하는 등의 방법들이 있다 [12] [13].

그런 방법들 중에 하나가 기호화 방식이다. 이는 여러 단어를 하나의 기호로 변환함으로써 모든 단어를 다룰 수 있도록 하는데, 예를 들어 사람의 이름을 '__NAME' 라고 변환함으로써 '__NAME' 이라는 하나의 기호가 모든 이름을 대신하게 한다. 고유명사나 숫자 등을 기호화 함으로써 단어 리스트의 길이를 대폭 축소 할 수 있어서 학습을 용이하게 하며, 새로운 단어에 대한 번역도 가능하게 한다. 자세한 내용은 [5]에서 확인할 수 있다.

기존의 인공 신경망을 활용한 번역기들은 숫자 번역에 있어서도 어려움을 겪게 되는데, 구글번역에서도 다음과 같은 사례에서 오역을 발견할 수 있었다 (2018년 6월1일 기준) [14].

> 입력 문장: "이 제품에는 약 3억 2백만 개의 다양한 미생물들이 있습니다."
> 구글 번역: "There are about 32 million different microorganisms in this product."

이와 같은 문제를 해결하기 위해 숫자를 기호화하는 방법을 사용한 연구가 있었고, 이를 통해 기호화하는 방법이 번역 성능을 개선하는데 효과가 있다는 것을 확인할 수 있었다 [14]. 하지만, 지금까지 인물의 이름이나 지명과 같은 고유명사의 한영사전을 이용한 기호화의 연구 사례는 없었다.

3. 제안모델

3.1. 데이터 설정

웹에서 크롤링을 통해 얻은 데이터는 유효하지 않은 데이터들이 많아 데이터 클렌징 작업이 필요하며, 클렌징작업에 성능이 많이 좌우된다. 본 논문에서는 '한국정보화진흥원 AI Hub' 에서 제공하는 한국어-영어 말뭉치 데이터를 사용하였는데, 이는 문어체, 구어체 등 160만 데이터를 사용함으로써 일반적인 번역 작업에 필요한 학습 데이터를 포함하고 있다.

학습 성능을 높이기 위해 데이터를 용도에 맞게 전처리 작업을 수행한다. 문자열을 여러 개의 조각인 토큰 단위로 쪼개기 위해 주어진 코퍼스에서 토큰단위로 나누는 작업인 Tokenize를 진행한다. 토큰에서 온점, 콤마, 물음표, 따옴표와 같은 문장 부호를 기준으로 문장을 띄워 쓰게 해준다. 본 논문에서는 한국어도 영어와 같은 방식으로 나눠주었다.

많은 데이터를 사용하여 학습할 때, 세상의 모든 단어를 컴퓨터에 입력할 수 없기에 OOV(Out of Vocabulary), 즉 UNK(Unknown Word) 문제가 발생한다. 이를 방지하기 위해 하나의 단어를 단어보다 작은 단위의 의미 있는 여러 Subword로 분리한다. 이때 사용한 알고리즘은 BPE(Byte Pair Encoding)로 가장 빈도수가 높은 유니그램의 쌍을 하나의 유니그램으로 통합하는 방식을 반복한다 [15]. 이를 통해 얻은 Subword들을 모아 단어장을 생성하였다. 본 논문에서는 10,000 개의 Subword들을 단어장으로 사용하였다.

사용하는 데이터에는 뉴스, 구어체, 대화체 등이 순서대로 저장되어 있다. 학습 성능을 높이기 위해 서로 관련된 문장 세트를 순서대로 학습시키지 않기 위해 랜덤으로 섞어 주었다.

본 논문에서 사용한 문장 데이터는 학습 데이터 1,096,311, 검증 (Valid) 데이터 2,500, 테스트 데이터 2,500 개이다.

3.2. 고유명사 사전

고유명사 사전은 고유명사 기호화 과정에서 고유명사를 판별 하기 위한 기준이 된다. 그렇기에 학습에 앞서 고유명사 사전을 먼저 만들어야 하는데, 초기 고유명사 사전을 만드는 과정은 다음과 같다.

1) 한국어로 사람 이름과 회사 이름 및 신조어 등을 포함하는 리스트를 만든다 (인명사전 등 활용).
2) 리스트 중 번역데이타에서 자주 나오는 254개를 영어 이름과 함께 고유명사 사전에 추가한다.
3) 나라 이름 150개를 추가로 고유명사 사전에 등록하였다.
4) 추출한 고유명사들을 파일에 저장한다.

이렇게 만들어진 초기 고유명사 사전은 고유명사 기호화를 이용한 모델 학습에 사용되고 학습이 완료된 모델로 새로운 문장을 번역 할 때도 사용된다.

본 논문은 한번 구축한 고유명사 사전에 추가로 사전을 확장하기 위해 웹 인터페이스에서 사용자로부터 고유명사 한영 쌍을 입력 받는다. 하지만 사용자로부터 추가된 단어들이 유효한지에 대한 검증이 필요하므로 실제 고유명사들이 저장된 파일에는 저장되지 않고, 서버 내 사용자 별로 고유명사 사전이 관리되고, 실제 번역시에는 원본 사전과 함께 사용자 별 사전을 이용하여 기호화한다. 이는 번역서비스의 개인화라고 볼 수 있다. 추후 사용자별 사전에 저장된 단어 쌍들은 관리자의 검증을 거쳐 원본 파일에 추가될 수 있다.

3.3. 고유명사 기호화 및 학습

고유명사가 들어간 문장의 번역 성능을 높이기 위해 고유명사 기호화를 사용한다. 고유명사 기호화는 학습과 번역 과정에서 모두 사용되며 고유명사 사전을 이용하여 기호로 치환하는 방식을 사용한다.

고유명사 기호화를 사용하여 모델을 학습시키는 과정은 다음과 같다. 고유명사 사전에는 {'테슬라: Tesla', '일론 머스크: Elon Musk', '스페이스X: SpaceX'} 가 들어있다고 가정한다. 아래 예시는 한영 번역 모델을 학습시키는 과정으로 지도학습을 위해 한영 문장 쌍이 입력으로 들어간다.

1) Train, Valid, Test 데이터의 단어들 중 고유명사 사전의 단어와 매칭되는 것들을 문장내 고유명사의 순서에 따라 '__P0', '__P1', ..., '__PN' 으로 기호화한다. 그 예시는 다음과 같다.

> 한국어: '일론 머스크는 테슬라와 스페이스X를 창립했다.'
> 기호화: '__P0는 __P1와 __P2를 창립했다.'
> 영어: 'Elon Musk founded Tesla and SpaceX.'
> 기호화 '__P0 founded __P1 and __P2.'

2) 기호화 된 Train 데이터들을 Transformer 모델에 넣어 학습시킨다. 모델이 기호화 된 문장들을 학습할수록, 번역 시 모델의 출력 문장은 입력 문장의 기호들에 대응되는 기호들을 출력한다.
3) 학습중에 검증데이터를 번역하고 BLEU Score를 얻어 early stop 방법으로 학습을 종료한다.
5) 학습이 완료된 모델에 Test 데이터를 이용하여 번역하고 BLEU Score를 얻는다.

3.4. 학습후 고유명사 번역

학습이 완료된 모델을 이용하여 번역하는 과정은 다음과 같다. 고유명사 사전에는 {'테슬라: Tesla', '일론 머스크: Elon Musk', '스페이스X: SpaceX', '페라리: Ferrari'} 가 들어있다고 가정한다.

1) 번역하고자 하는 문장을 입력한다.
2) 입력 문장의 단어들과 고유명사 사전을 비교하여 기호화한다.
3) 이 때, 기호화된 단어들을 고유명사 사전을 통해 해당 단어를 '번역 사전' 에 따로 저장한다.
4) 모델에 기호화된 입력 문장을 입력하고 출력 문장을 받는다.
5) 출력문장내의 기호들에 대해서는 번역 사전으로부터 각 기호들에 해당하는 출력 언어의 단어를 찾아내고 복호화 (Desymbolization) 하여, 최종 번역 결과를 얻는다.

아래는 번역 과정의 예를 보여준다.

> 입력 문장: '나는 페라리보다 테슬라를 선호한다.'
> 기호화된 문장: '나는 __P0보다 __P1를 선호한다.'
> 번역 사전: {'__P0: Ferrai', '__P1: Tesla'}
> 출력 문장: 'I prefer __P1 to __P0'
> 번역 결과: 'I prefer Tesla to Ferrari'

위 예에서 번역 사전은 각 기호별로 고유명사 사전을 활용하여, __P0 의 '페라리'에 해당하는 'Ferrai', __P1 의 '테슬라'에 해당하는 'Tesla' 를 등록한다.

4. 실험결과

4.1. 데이터

본 논문에서는 모델을 학습하기 전에 데이터를 고유명사 기호화하였다. 전체 1,096,331개의 학습 데이터에서 224,209개, 2,500개의 검증 데이터에서 490개, 2,500개의 테스트 데이터에서 526개의 고유명사를 기호화했다.

기호화 된 데이터들은 데이터 전처리 작업을 거쳤다. 데이터 전처리 순서는 Tokenize, Subword, Shuffle 순으로 진행한다. 다음은 고유명사 기호화와 데이터 전처리 작업을 거친 한영 문장 쌍 중 하나의 예를 보여준다.

- 기호화 전 영어 문장: There will be a chance to see the winning films of the 7th DMZ International Documentary Film Festival again.
- 기호화 후 영어 문장: There will be a chance to see the winning films of the 7th __P0 International Documentary Film Festival again.
- 전처리 후 영어 문장: There will be a chance to see the winning films of the 7th __P0 International Doc@@ um@@ entary Film Festival again .
- 기호화 전 한국어 문장: 제7회 DMZ국제다큐영화제 수상작을 다시 볼 수 있는 기회가 온다.

- 기호화 후 한국어 문장: 제7회 __P0 국제다큐영화제 수상작을 다시 볼 수 있는 기회가 온다.
- 전처리 후 한국어 문장: 제7@@ 회 __P0 국제@@ 다큐@@ 영화@@ 제 수상@@ 작@@ 을 다시 볼 수 있는 기회가 온다 .

위와 같이 데이터에 고유명사 기호화를 하여 모델이 고유명사를 받아들이고 출력할 수 있도록 학습시킬 수 있었다.

4.2. 번역 결과

본 논문에서 한영/영한 번역 모델의 성능 향상을 위해 사용한 방법은 최근 번역 모델인 Transformer 모델을 사용하고 고유명사를 기호화하는 것이다. Table 1은 RNN 모델에서 한영(Kr-En) 번역시 고유명사 기호화를 적용하기 전과 적용한 후의 BLEU Score 결과이다 [16]. 고유명사 기호화를 적용하였을 때 BLEU Score는 Valid에서 1.84점 , Test에서는 1.93점 상승하였다.

표 1. RNN 모델에서 고유명사 기호화를 통한 한영 번역의 성능 (BLEU score) 개선
Table 1. The performance (BLEU score) improvement of Korean-English translation based on symbolization of proper nouns in RNN model.

RNN (Kr-En)	Valid	Test
Conventional RNN base NMT	21.42	26.55
+ Symbolization of proper nouns	23.26	28.48

표 2. Transformer 모델에서 고유명사 기호화를 통한 한영/영한 번역의 성능.
Table 1. The performance of Korean and English translation through Transformer and proper noun symbolization

Transformer (Kr-En/En-Kr)	Kr-En		En-Kr	
	Valid	Test	Valid	Test
Transformer	30.75	30.81	13.78	14.13
+ Symbolization of proper nouns	31.16	30.36	14.3	14.16

Table 2는 Transformer 모델에서 한영, 영한 번역시 고유명사 기호화를 적용하기 전과 적용한 후의 BLEU Score 결과이다. 한영 번역시 고유명사 기호화를 적용하였을 때 BLEU Score는 Valid 에서 0.41, 영한 번역시 고유명사 기호화로 Valid 에서 0.52점 상승하였다.

Transformer 모델의 BLEU Score가 RNN 모델보다 고유명사 기호화 적용 시 점수 차가 적은 것을 확인할 수 있는데, 그 이유는 Transformer 모델이 RNN 모델보다 성능이 좋기 때문에 고유명사가 들어간 문장을 더 잘 번역하기 때문이다.

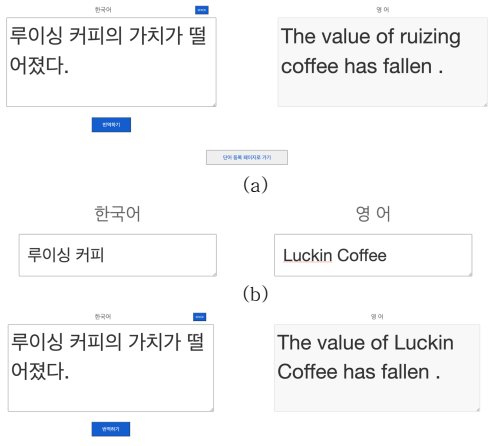


그림 1. 고유명사 사전에 등록하기 전후의 번역 비교. (a) '루이싱 커피' 를 고유명사로 등록 전 번역 결과 (아래 '단어 등록 페이지로 가기' 버튼으로 고유명사 사전 등록 페이지로 이동. (b) 단어 등록 화면에서 '루이싱 커피: Luckin Coffee' 를 고유명사 사전에 등록. (c) 단어 등록후 다시 번역한 결과.

Fig. 1. Translation before and after registering a proper noun pair into a dictionary. (a) translation before registration, (b) registration, and (c) translation after registration.

고유명사 기호화를 적용할 경우 BLEU Score만으로는 확인하기 힘든 부분을 실제 번역 예시를 통해 질적인 차이를 확인할 수 있다. 앞서 파파고나 구글 번역기와 같은 상용 번역기에서 보이는 고유명사 번역 문제점을 예시를 통해 지적했었는데, 제안하는 모델의 성능 개선을 동일한 예시의 번역을 통해 확인하고자 한다. 참고로, 네이버 파파고 번역기는 '루이싱 커피' 를 'Luixing coffee' 로 오역한다. 구글 번역기도 '루이싱 커피' 를 'louise coffee' 로 오역한다.

본 논문의 제안된 방법으로 Fig. 1. 과 같이 웹기반 번역 시스템을 구축했다. 번역기에서 '루이싱 커피'를 고유명사로 등록을 하기 전에는 Fig. 1(a) 에서 처럼 '루이싱 커피' 를 'ruizing coffee' 로 오역했지만, Fig. 1(b)와 같이 단어 등록 페이지에서 '루이싱 커피' 와 'Luckin Coffee' 를 고유명사 사전에 등록하면, 추가학습 없이 Fig. 1(c) 에서 보여진 것 처럼 '루이싱 커피' 를 'Luckin Coffee'로 올바르게 번역할 수 있다.

5. 결론

본 논문은 상용 번역기에서 고유명사가 들어간 문장의 오역 문제를 확인하고 이를 해결하기 위해 고유명사 사전을 구축하고, 고유명사를 기호화 하는 방법을 제안했다. Transformer 모델 기반의 실험에서 고유명사 기호화를 적용했을 때 BLEU Score 상승은 크지 않았지만 질적 비교를 통해 고유명사가 포함된 문장의 번역 오류가 해결된 것을 확인할 수 있었다.

본 논문에서는 빈도수가 높은 고유명사를 고유명사 사전에 추출하였지만 다양한 방식의 고유명사 사전을 시도해볼 수 있다. 또한 외부 지식을 사용하는 다른 자연어처리 모델에도 본 논문에서 제시된 방법을 적용해볼 수 있다 [17].

사 사

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2018-0-00749,인공지능 기반 가상 네트워크 관리기술 개발)

참고문헌

[1] H. Choi, and Y. Min “Introduction to Deep Learning and Major Issues,” Korea Information Processing Society Review, Vol. 22, No. 1, p.7-21, 2015.
[2] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” ICLR 2015.
[3] H. Choi, “Understanding Neural Machine Translation,” Communications of the Korean Institute of Information Scientists and Engineers, Vol. 37, No. 2, p.16-24, 2019.
[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. “Attention Is All You Need,” arXiv 2017.
[5] H. Choi, K. Cho, and Y. Bengio, “Context-dependent word representation for neural machine translation,” Computer Speech & Language, Vol. 45, p. 149-160, 2017.
[6] T. Mikolov, M. Karaat, L. Burget, J. Cernocky, S. Khudanpur, “Recurrent Neural Network based Language Model,” INTERSPEECH, 2010.
[7] S. Hochreiter, J. Schmidhuber, “Long short-term memory,” Neural computation 9 (8), pp. 1735-1780, 1997.
[8] K. Cho, B. van Merrienboer, D. Bahdanau, Y. Bengio, “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches,” Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014.
[9] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin, “Convolutional Sequence to Sequence Learning,” arXiv Preprint, arXiv:1705.03122, 2017.
[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” NAACL, 2019.
[11] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving Language Understanding with Unsupervised Learning,” Technical report, OpenAI, 2018.
[12] C. Gulcehre, S. Ahn, R. Nallapati, B. Zhou, Y. Bengio, “Pointing the unknown words,” arXiv Preprint 1603.08148, 2016.
[13] W. Ling, I. Trancoso, C. Dyer, A. W. Black, “Character-based neural machine translation,” arXiv Preprint, arXiv:1511.04586, 2015.
[14] C. Kang, Y. Ro, J. Kim, H. Choi, “Symbolizing Numbers to Improve Neural Machine Translation," Journal of Digital Contents Society, Vol. 19, No. 6, p.1161-1167, 2018.
[15] R. Sennrich, B. Haddow, A. Birch, “Neural machine translation of rare words with subword units,” 54th Annual Meeting of the Association for Computational Linguistics, 2016.
[16] J. Nam, M. Kim, H. Jeong, H. Choi, “Kor-Eng Neural Machine Translation System Using Proper Noun Dictionary,” KCC, 2020.
[17] S. Ahn, H. Choi, T. Pärnamaa, Y. Bengio, “A Neural Knowledge Language Model,” arXiv:1608.00318, 2016.

김 명 진

2014년~현 재: 한동대학교 전산전자공학부 재학
관심분야: 머신러닝, 딥러닝, 인공지능

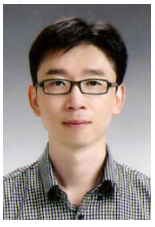
남 준 영

2014년~현 재: 한동대학교 전산전자공학부 재학
관심분야: 머신러닝, 딥러닝, 인공지능

정 희 석

2014년~현 재: 한동대학교 전산전자공학부 재학
관심분야: 머신러닝, 딥러닝, 인공지능

최 희 열



2010년 Texas A&M University, Computer Science and Engineering (PhD), 2010년~2011년 Indiana University, Cognitive Science Program (Post-Doc), 2011년~2016년 삼성전자 종합기술원 (전문연구원), 2015년~2016년 University of Montreal (Visiting Scholar), 2016년~현재 한동대학교 전산전자공학부 (조교수), 관심분야: Deep Learning, Cognitive Science