|  |
| --- |
|  |
| **문제정의서(연구계획서)** |
|  |

|  |  |
| --- | --- |
| **과제명** | **신경망 기계번역에서 전문 용어 번역 정확도 향상에 대한 연구** |

|  |  |
| --- | --- |
| 조 | Lafuta조 |
| 지도교수 | 정상근 교수님 (서명) |
| 조원 | 201502039 김종운  201702073 정지수 |

**1. 연구의 필요성**

기계번역의 놀라운 성장에도 불구하고, 기계 번역과 인간 번역은 여전히 격차가 난다. 서보현, 김순영의 연구 (2018)에 따르면 기계번역은 원문, 번역문의 구문을 분석하여 만든 문법적 알고리즘을 활용하는 규칙 기반의 기계 번역(Rule-Based Machine Translation:RBMT)에서 출발한다. 이는 유지보수의 측면에서 적지 않은 시간이 든다는 한계가 있었고, 이를 극복하기 위해 통계기반 기계번역(Statiscal Machine Translation: SMT)가 개발되었다. 이는 원문과 번역문의 조합으로 이루어진 말뭉치를 이용하여 두 텍스트의 관계를 통계적으로 분석 후 결과를 적용하는 방식이다. 빅 데이터의 운용이 쉬워져 SMT는 빠른 속도로 발전한다. 그러나 텍스트의 맥락을 고려하지 못해 출발어와 도착어, 어순이나 문법 구조가 다른 경우에 품질이 저하된다는 단점이 있다. 요즘에는 인간의 신경망과 비슷한 구조의 네트워크를 만들어 활용하는 신경망 기계번역(Neural Machine Translation: NMT)를 개발하고 있다. 이는 눈에 띄는 성장세를 보이고 있으나 여전히 만족할 만한 수준은 아니라고 한다.

이러한 기계번역의 오류는 정확성(Accuracy), 가독성(Fluency), 통사구조(Syntax), 오탈자(Typo)의 네 개의 큰 틀로 분류할 수 있다. 이러한 것은 번역학 분야에서의 대표적인 기준인 충실성, 가독성과 함께 문법과 오탈자라는 기계 번역에서 지속적으로 발생하는 문제를 포함한 것이다. 이 네 개의 오류 유형 중 특히 부정확한 의미 오류라는 정확성의 문제가 다른 오류보다 많이 나온다. 이는 원문에서 주어진 어휘에 대응되는 올바른 번역이 이루어지지 않아 생기는 오류로, 다섯 가지의 하위 유형으로 분류한다. 문장의 의미를 해치는 경우, 기본 의미는 통하나 어휘 형태가 적절치 못한 경우, 다른 어휘가 추가된 경우, 번역문의 어휘 사용 방식이 문장 형식에 어긋나거나 가독성을 해치는 경우, 관용어를 인식하지 못하는 경우이다. 이는 언어 체계가 다르고 사용하는 철자도 다른 언어를 번역할 경우 자주 생기는 문제이다.

이러한 부정확한 의미 오류는 고유명사나 전문용어, 다의어 등의 등장으로 인해 번역된 단어에 대해 오류가 발생한다. 당연하게도 통상적으로 쓰이는 단어를 중점적으로 번역하다 보니, 이로 인해 특정 주제의 문맥에 맞게 단어가 번역되지 않는 문제가 있다. 이러한 문제는 다양한 분야에서의 글에 대해 제기되고 있다. 김장호, 최석범의 연구 (2018)에서는 무역 서신에서의 통상적인 표현들의 유형에서 전문용어의 번역이 적절치 못한 것에 대한 문제를, 전혜진의 연구(2019)는 문학, 특히 소설에서의 문맥에 맞지 않는 숫자, 날짜, 인명, 고유명사 등에 대한 해석 문제를 제기하고 있다.

본 연구진은 이런 다양한 전문용어 번역에 대한 정확도를 높이는 목적으로, 전문용어에 대한 정확한 번역이 필요한 논문이라는 글의 갈래에 대해 번역의 정확도를 높일 수 있도록 중점적으로 연구해보고자 한다. 현재는 이러한 기계번역 결과물을 수정, 개선하여 활용도를 높이는 방안으로 포스트에디팅(post-editing) 작업을 하고 있다. 포스트에디팅이란, 기계번역을 이용한 선번역 작업을 뜻하는 프리에디팅(pre-editing)의 반의어로, 작업기계번역 결과물의 오류를 수정하여 더욱 개선된 최종 결과물을 만들어내는 작업이다. 그러나, 빈번하게 쓰이는 고유명사와 전문용어를 모두 다 고치며 문맥에 맞게 수정하는 일은 번거로운 일이다. 본 연구는 이런 불편함을 해소하기 위해 전문 용어 번역의 정확도를 높이는 방법을 고안하여 번역의 질을 높이고, 기계번역을 통해 시간과 비용을 줄이면서 질 높은 영어 산문을 집필하는 데에 기여할 것으로 예상된다. 특히 논문이라는 학문적 갈래의 글에 대해 영어로 번역하는 것이 쉬워진다면 광역적인 학문적 소통이 더욱 편리해질 수 있다.

**2. 연구의 목표 및 내용**

본 연구는 논문 번역이라는 특징적 과제에 대하여 정확한 번역을 제공하는 것을 목표로 한다. 특히 신경망 기계번역에 대한 고유 명사, 전문 용어에 대한 정확도 향상을 중점적으로 연구한다. 신경망 기계번역에는 다양한 모델이 연구되고 있다. 대부분은 메시지를 기호화하는 인코더-디코더 모델을 기반으로 한다. 이런 모델은 단어 예측을 위해서는 방대한 데이터를 이용한 학습이 필요한데, 이러한 메모리나 시간 경량화를 위해 어텐션 벡터(attention vector)를 적용한다. (송연석, 2018)

이후, 입력 시퀀스를 하나의 벡터 표현으로 압축하는 인코더와 이를 통해 출력 시퀀스를 만들어내는 디코더로 구성되어 문맥을 고려한 번역이 가능한 sequence to sequence 모델이 있다. 이러한 sequence to sequence모델은 입력 시퀀스가 하나의 벡터로 압축되는 과정에서 정보 손실이 발생하기 때문에 이를 보정하기 위해 attention을 사용한다.

이러한 모델에서 한 단계 더 나아가, Transformer 모델은 sequence to sequnce의 구조를 따르면서도, 어텐션만을 구현하여 학습속도를 빠르게 만든다. RNN을 사용하지 않고, 인코더-디코더가 여러 개 존재할 수 있다는 특징이 있다. (Vaswani, 2017)

이후 등장한 사전학습 모델은 이러한 모델의 훈련을 용이하게 한다. ELMo (Peters et al., 2018), GPT/GPT-2 (Radford et al., 2018; 2019), BERT (Devlin et al., 2019)와 같은 사전학습(pre-trained) 모델은 먼저 대량의 말뭉치를 이용하여 학습을 진행한 후, 용도에 맞게 조정(Fine-tuning)하여 전이학습을 통해 정해진 과제에 더욱 더 집중된 학습을 진행시킨다. 이는 사전학습부터 시작하는 것보다 훨씬 저렴하며, 정확한 결과물을 얻을 수 있도록 한다.

Transformer 모델을 사전학습으로서 더욱 발전시켜 양방향으로 학습하여 문맥의 유추를 더욱 용이하게 만든 BERT (Bidirectional Encoder Representations from Transformers)가 있다. (J Devlin, 2018) BERT는 사전학습과 전이학습으로 학습 과정이 나뉜다. 사전학습된 모델은 용도에 맞게 조정하여 전이학습을 진행하여 사용할 수 있다.

이러한 강력한 사전학습 양방향 모델인 BERT는 여러 자연어 처리 연구 분야에 파급력을 미치는데, 신경망 기계 번역 분야도 예외는 아니다. (Zhu, J, 2020) BERT와 인코더, 디코더를 이용하여 점진적으로 처리하는 BERT fused NMT 모델과 같은 방법론이 제시되기도 한다.

이러한 다양한 모델을 조합하여 새로운 모델을 구상하고, 시도해보며 한국어 텍스트를 번역했을 때의 부정확한 의미 오류를 최소화할 수 있도록 한다. 이에 덧붙여 최소한의 텍스트 데이터를 효율적으로 사용하여 모델을 경량화할 방법을 고안해보고자 한다.

**3. 연구의 추진전략 및 방법**

NMT에는 다양한 모델이 존재한다. 이 중 특히 BERT에 주목하여, 모델의 전이 학습에서 신경망 기계번역의 프리에디팅(pre-editing)에 대해 고유명사나 전문용어에 대해 attention을 주어 정확도를 높일 수 있을 것이라는 가설을 두고 연구를 진행하고자 한다. 기본적으로 transformer 모델의 구조에 양방향성을 띄어 attention을 적용시 큰 효과를 얻을 것으로 추측된다. 데이터는 문장 단위의 태깅된 논문 데이터를 사용하고자 한다. 배포되고 있는 한국어 데이터 중 논문에 관한 데이터는 없어 직접 데이터를 수집할 가능성이 크다. 필요한 데이터는 논문에서 쓰이는 문장과 그에 대한 번역 시 정답 문장이다.

이 이후 수집한 데이터에 대해 Preprocessing을 진행한다. 번역 정답 문장과 고유명사와 전문용어에 대한 태그를 씌운다. 이렇게 태그된 데이터는 워드 임베딩을 통해 단어를 벡터로 표현한다. 이를 위해 먼저 하나의 Token으로 만드는데, 한국어의 경우 띄어쓰기를 기준으로 단순 split을 할 수 없다. 한국어는 조사, 어미 등을 붙여 쓰는 교착어로, 단순히 띄어쓰기 단위로 나눈다면 이는 단어 토큰화와 큰 차이가 생긴다. 그래서 형태소 단위 토큰화를 사용한다. KoNLPy라는 파이썬 패키지를 이용하여 형태소 단위 토큰화를 이용해야 한다.

이렇게 토큰화한 문장을 벡터로 표현하기 위한 방법에는 원-핫 인코딩과 워드 임베딩이 있다. 원-핫 인코딩은 단어의 인덱스값을 1로, 나머지는 0으로 주는 희소 벡터를 이용한 방법이다. 이러한 표현 방법은 각 단어간의 유사성을 표현할 수 없다는 단점이 있어, 단어의 의미를 다차원 공간에 벡터화하기 위한 방법을 필요로 한다. 이러한 방법을 고려하는 것이 워드 임베딩이다. 기본적으로 분산 표현이라는 방법으로, 비슷한 위치에 등장하는 단어는 비슷한 의미를 가진다는 가정을 한다. 다시 말해, 단어들 간의 의미들로 연산을 할 수 있다는 것이다. 단어 벡터가 단어 간 유사도를 반영할 수 있기 때문이다. 워드 임베딩은 다양한 방법이 존재하는데, LSA, Word2Vec, FastText, Glove 등의 방법론이 존재한다. 그 중 Word2Vec가 가장 대표적이다.

NMT에 대한 모델은 BERT를 주로 이용하나, 이를 비롯한 XLM, RoBERTa 등의 다양한 사전학습 모델과 seq2seq, transformer 등을 적용하여 사용해보고자 한다. 다양한 모델을 시도하여 정확도를 측정하는 실험을 통해 최적의 방법을 찾아보고자 한다. 앞에서 말했듯, 직접 데이터를 수집할 가능성이 크므로, 이는 모델의 정확성에 시간적, 비용적 한계로 작용할 것으로 예상된다. 이를 보완하기 위해 최대한 효율적인 모델을 설계하여야 한다.

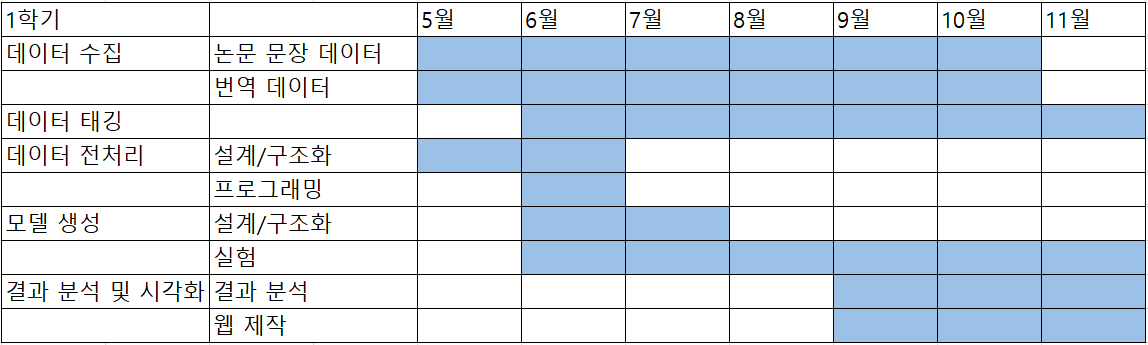
**4. 연구 팀의 구성 및 과제 추진 일정**

연구를 진행하는 전체적인 절차는 다음과 같을 예정이다. 논문의 원문과 번역 문장 데이터를 수집하고, 이 데이터에 대해 일정 단어에 적절한 태깅을 하여 전처리를 한 후, 임베딩을 포함한 기계번역 학습을 한다. 데이터를 이용한 학습 모델을 만들기 위해 방대한 데이터가 필요하여 데이터 수집은 상당히 많은 시간이 걸릴 것으로 예상된다.

1학기에는 데이터 수집 및 전처리에 더욱 중점을 두어 진행하며, 동시에 기계번역 학습 모델 생성을 위한 사전 학습 및 모델 구조 설계를 진행할 예정이다. 논문 문장 데이터를 수집하고, 이에 대해 번역된 정답 데이터를 만들거나 수집한다. 이후 어느정도 수집되면, 데이터에 대해 일정 단어에 적절한 태깅을 하여 기계번역 모델 훈련 시 정확도를 높일 수 있도록 정제한다. 데이터를 집중적으로 모아 훈련이 가능할 최소한의 데이터 이상을 모으려고 한다.

2학기에는 1학기에 전처리한 데이터를 이용하여 모델을 생성, 구조화하고 실험을 진행하며 정확도를 높이는 방안을 계속해서 시도해볼 예정이다. 데이터는 1학기와 같이 지속적으로 모으며, 최대한 많이 모아 훈련을 할 계획이다. 모델은 다양한 시도를 하며 최적화의 방법을 찾아내, 이를 시각화하고 웹 형태의 툴을 제작하여 상용화할 계획이다.

간트 차트로 나타내면 다음과 같다.



**- 참고문헌(Reference)**

서보현, 김순영. 20183기계번역 결과물의 오류유형 고찰

김장호, 최석범 (2018). 전자무역서신 작성을 위한 NMT기반 번역프로그램 연구. e-비즈니스연구, 19(5), 251265

전혜진 (2019). AI 시대, 문학번역에서 기계번역과 인간번역 비교분석 연구. 노어노문학, 31(1), 111-154

Vilar, D & Xu, J & D’Haro, L & Haro, D & Ney, H (2006) ‘Error analysis of statistical machine translation output’, Proceedings of the Fifth International Conference on Language Resources and Evaluation: 697-702.

송연석 (2018). 기계번역 담론에 대한 비판적 고찰. 번역학연구, 19(1), 119-145

Sutskever, Ilya, Oriol Vinyals and Quoc V. Le (2014) ‘Sequence to Sequence Learning with Neural Networks,’ in Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS), 3104-3112

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., ... & Liu, T. Y. (2020). Incorporating BERT into neural machine translation. *arXiv preprint arXiv:2002.06823*.

참고 자료

딥 러닝을 이용한 자연어 처리 입문/15장 트랜스포머/1)트랜스포머 <https://wikidocs.net/31379>