



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](#)

홍 기 형 교수 지도

석사학위 청구논문

다중 의미를 갖는 보완대체의사소통
상징의 구문분석을 통한 딥러닝
기반 문장의 상징 시퀀스 변환

2023

성신여자대학교 대학원

미래융합기술공학과

유 세 희

다중 의미를 갖는 보완대체의사소통
상징의 구문분석을 통한 딥러닝
기반 문장의 상징 시퀀스 변환

홍기형 교수 지도

이 논문을 석사학위논문으로 제출함

2022년 11월

성신여자대학교 대학원

미래융합기술공학과


유 세 희


인준서

유세희의 석사학위 논문으로 인준함

2022년 11월

심사위원장 오 장민 

심사위원 윤 상혁 

심사위원 홍 기행 

성신여자대학교 일반대학원

논문개요

구어로 의사소통이 어려운 언어 장애인들이 사용하는 의사소통 도구에는 보완대체의사소통(Augmentative and Alternative Communication, AAC)이 있다. AAC 서비스는 모바일 서비스, 웹 서비스 등 다양한 형태로 개발되고 있으나, 현존하는 AAC 서비스는 주로 대면 의사소통에 활발히 사용되며, 전자우편, 메신저 등의 비대면 의사소통 상황에서는 활용이 어렵다. 인터넷의 보급 및 사용이 증가함에 따라 비대면 의사소통도 증가했으며, 비장애인과 장애인 사이의 의사소통 중재가 필요해졌다. 비대면 의사소통 상황에서는 주로 텍스트 대화가 오가는데, AAC 상징을 주로 사용하는 언어 장애인들은 글을 이해하는 것에 어려움이 있다.

본 논문에서는 상징이 가지는 다중 의미를 고려한 대화 문장의 AAC 상징 변환을 통해 비장애인이 입력한 텍스트를 언어 장애인에게 익숙한 AAC 상징으로 변환하는 모델을 제안한다. 대화 문장을 AAC 상징으로 변환하기 위해 한국어 대화 문장과 한국형 보완대체의사소통 체계집을 사용하였다. 한국어 대화 문장은 AI Hub의 공개 데이터인 한국어 대화 데이터에 유아용 동화책 문장을 추가로 수집하여 구성하였다. 문장은 상징 시퀀스로 대응시켜 나타낼 수 있으므로 AAC 상징 이미지의 특징을 고려하고, 문맥에 맞는 상징을 사용하여 상징 시퀀스 데이터를 구축하였다. 또, 상징이 가지는 다중 의미를 파악하여 문장에 대응되는 상징 시퀀스를 1개 이상 추가하는 작업을 수행하였다. 이에 따라, 다중 의미를 갖는 상징을 포함하는 문장은 최소 2개 이상의 상징 시퀀스를 갖는다. 문장을 AAC 상징으로 변환하는 모델은 기존 연구의 Sequence to Sequence

모델과 어텐션(Attention) 매커니즘 기반 Sequence to Sequence 모델, 그리고 케라스 임베딩과 Word2Vec 임베딩에 GloVe와 FastText 임베딩을 사용하는 모델을 추가하였다. GloVe 임베딩은 일정 범위 내의 단어만 고려하여 예측하는 Word2Vec의 단점을 보완하여 전체 말뭉치에서 단어가 등장하는 횟수를 반영한다. FastText 임베딩은 단어를 더 작은 단어로 나누어 학습을 수행하기 때문에 학습하지 않은 어휘에 대해서도 예측할 가능성이 높다. 두 임베딩을 사용한 모델을 추가 실험하여 문장의 상징 시퀀스 예측 변환 성능을 비교하였다. 성능은 BLEU 점수를 사용하여 측정하였고, 사전 훈련된 FastText 임베딩을 활용한 어텐션 매커니즘 기반 Sequence to Sequence 모델이 가장 좋은 성능을 보였다.

목 차

논문개요

| | |
|--|----|
| I. 서 론 | 1 |
| II. 관련 연구 | 4 |
| 1. 보완대체의사소통과 상징 | 4 |
| 2. 딥러닝 기반 대화 문장의 보완대체의사소통 상징 시퀀스 변환 | 8 |
| 3. 단어 임베딩(Word Embedding) | 10 |
| 1) Word2Vec(Word-to-Vector) | 12 |
| 2) GloVe(Global Vectors for Word Representation) | 13 |
| 3) FastText | 14 |
| III. 상징의 다중 의미를 반영한 상징 시퀀스 변환 | 15 |
| IV. 딥러닝 기반 문장의 상징 시퀀스 변환 설계 | 17 |
| 1. 학습 데이터 구축 | 19 |
| 2. 데이터 전처리 | 24 |
| 3. 다중 의미를 반영한 딥러닝 기반 문장의 AAC 상징 시퀀스 변환 모델 | 26 |

| | |
|--|----|
| V. 모델 실험 및 평가 | 30 |
| 1. 실험 설계 | 30 |
| 2. 실험 결과 | 31 |
| 3. 실험 평가 | 36 |
| 1) BLEU(Bilingual Evaluation Understudy) | 36 |
| 2) 평가 결과 | 37 |
| VI. 결론 및 향후 연구 | 40 |

참고문헌

ABSTRACT

표 목차

| | |
|---|----|
| [표 1] FastText 메커니즘의 $n=3$ 일 때의 n -gram | 14 |
| [표 2] 문장의 다중 상징 시퀀스 태깅 | 23 |
| [표 3] 최종 학습 데이터 수 (단위: 개) | 24 |
| [표 4] 어미와 조사 목록 일부 | 25 |
| [표 5] 실험 환경 | 30 |
| [표 6] 다중 의미를 반영한 딥러닝 기반 대화 문장의 한국형 AAC 상징 시퀀스 변환 모델의 학습 결과 | 32 |
| [표 7] “그건 얼마예요” 문장의 변환 결과 | 33 |
| [표 8] 단어 임베딩의 종류에 따른 문장의 상징 시퀀스 변환 결과 · | 35 |
| [표 9] BLEU 평가 결과 | 37 |

그림 목차

| | |
|--|----|
| [그림 1] AAC 상징 | 5 |
| [그림 2] 한국형 AAC 상징의 특징 1 | 6 |
| [그림 3] 한국형 AAC 상징의 특징 2 | 7 |
| [그림 4] “좋은 생각입니다” 문장의 상징 시퀀스로의 변환 결과 | 9 |
| [그림 5] 유사한 단어의 2 차원 공간에서의 벡터 표현 | 11 |
| [그림 6] Word2Vec 모델 구조 | 12 |
| [그림 7] 한국형 AAC 상징의 다중 의미 | 15 |
| [그림 8] 딥러닝 기반 대화 문장의 AAC 상징 시퀀스 변환 모델 설계도 | 17 |
| [그림 9] 학습 데이터 구축 과정 | 19 |
| [그림 10] “좋은 생각입니다.” 문장의 상징 시퀀스로의 변환 | 20 |
| [그림 11] 문맥에 따라 다르게 사용되는 상징 이미지 | 21 |
| [그림 12] “친구 어디에 있어요?” 다중 상징 시퀀스 | 22 |
| [그림 13] Sequence to Sequence 모델 구조 | 26 |
| [그림 14] 어텐션 매커니즘 기반 Sequence to Sequence 모델 구조 | 28 |
| [그림 15] 변환 모델 학습 그래프 | 31 |

I. 서 론

구어로 의사소통이 어려운 언어 장애인들이 사용하는 의사소통 도구에는 보완대체의사소통(Augmentative and Alternative Communication, AAC)[1]이 있다. AAC 상징은 사용자가 표현하고자 하는 어휘를 잘 나타내는 그림 또는 사진이다. 특정 장소와 상황에서 의사소통을 위하여 AAC 상징들을 묶어 AAC 보드를 구성할 수 있다. 사용자에게 맞게 구성된 보드와 상징은 전자적 또는 비전자적 도구를 통해 의사소통에 활용된다. 비전자적 도구는 종이, 책 등을 의미하며, 전자적 도구는 스마트폰, 태블릿 등의 도구를 의미한다[2].

전자적 도구의 보급 및 사용이 증가함에 따라 전자적 도구를 활용한 AAC 서비스의 연구가 활발히 이루어지고 있다. AAC 서비스는 모바일 서비스, 웹 서비스 등 다양한 형태로 개발되고 있다[2]. 현존하는 AAC 서비스는 사용자가 사용하고자 하는 AAC 상징의 표현을 음성으로 출력해주는 것이 주를 이루고 있다. 이는 대면 의사소통 상황에는 도움이 되지만 비대면 의사소통 상황에서는 사용하는 것에 한계가 있다.

인터넷의 보급과 스마트 기기의 사용이 증가하면서 비대면 의사소통이 가능해졌다. 최근에는 메신저, 소셜 네트워크 서비스(SNS), 전자우편 등에서부터 재택근무, 온라인 회의에 이르기까지 비대면 활동이 증가하면서 비대면 의사소통의 중요성이 커졌다. 비장애인뿐만 아니라 언어 장애인 또한 비대면 활동을 해야만 하는 상황이 있으며, 특히 글을 읽고 쓰는 것이 어려워 AAC 상징을 사용하는 언어 장애인은 비장애인과 비대면 의사소통에 어려움을 겪는다. 비장애인이 입력한 텍스트를 AAC 상징으

로 변환하거나, AAC 상징으로부터 텍스트를 생성하는 것은 AAC 사용자의 비대면 의사소통을 보조할 수 있다.

AAC 상징의 이미지는 그 자체로 여러 의미를 표현할 수 있다. 예를 들어, “과일” 상징은 “과일”, “과일 주세요.”, “과일 먹고 싶어요.” 등 다양한 의미를 내포한다. 본 논문은 언어 장애인이 비대면 의사소통 상황에서도 충분히 의사소통할 수 있도록 텍스트를 AAC 상징 시퀀스로 변환하는 모델을 제안한다. 변환 모델을 구현하기 위해 한국어 대화 문장과 한국형 보완대체의사소통 체계집[3]의 상징을 사용하였고, AAC 상징 이미지의 특징을 고려하여 학습 데이터를 구축하였다. 텍스트 문장의 상징 시퀀스로의 변환을 위해 Sequence to Sequence(Seq2Seq)[4]와 어텐션(Attention) 매커니즘[5]을 활용한 딥러닝 기반 변환 모델을 이용하였다. 케라스 임베딩[6]과 사전 학습 훈련된 Word2Vec[7-8], GloVe[9], FastText[10] 임베딩을 적용하여 총 8가지의 변환 모델을 구현하였다. GloVe 임베딩은 전체 말뭉치에서 단어의 등장 횟수를 고려하지 않는 Word2Vec 임베딩의 단점을 보완한 것이며, FastText 임베딩은 단어를 더 작은 서브 단어로 분리하여 모델 학습을 수행하므로 학습하지 않은 단어를 예측할 가능성이 높을 것이다. 이 두 임베딩을 사용한 모델을 추가 실험하여 문장의 상징 시퀀스 예측 성능을 높이고자 한다.

본 논문 2장에서는 텍스트 문장을 상징 시퀀스로 변환하는 모델의 구성 개념과 관련 연구를 소개한다. 3장에서는 상징이 가지는 다중 의미에 대해 알아보며, 4장에서는 데이터 수집과 전처리, 실험을 위하여 구성된 변환 모델에 대해 설명한다. 한국어 대화 문장을 수집하여 각 문장에 대응하는 상징 시퀀스를 태깅한 후, 상징이 가지는 다중 의미를 반영하여

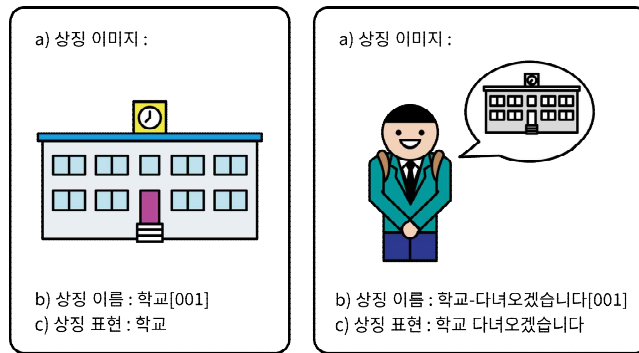
추가로 상징 시퀀스를 태깅한다. 이후, 전처리를 통해 최종 모델 학습 데이터를 구축하고 8가지 변환 모델을 사용해 문장의 상징 시퀀스 변환 예측을 수행한다. 5장에서는 제안 모델의 실험 결과를 BLEU 평가 지표를 통해 평가하고, 각 모델의 평가 결과를 비교 분석한다. 6장에서는 결론과 제안하는 모델의 활용을 위해 수행되어야 할 향후 연구에 대해 기술하였다.

II. 관련 연구

1. 보완대체의사소통과 상징

보완대체의사소통(AAC)[1]은 발화가 어렵거나 의사소통하기에 충분한 구어를 학습하지 못한 언어 장애인의 의사소통을 도와준다. AAC는 도구를 사용하는 AAC와 도구를 사용하지 않는 AAC로 분류된다. 도구적 AAC 시스템에는 문자, 상징, 이미지, 그림책 등의 자료가 있다. 최근에는 기술의 발전으로 전자기기에서 상징 표현을 음성으로 출력하는 기능이 널리 사용되고 있다. 비 도구적 AAC 시스템은 제스처나 수신호를 의미한다.

도구적 AAC 시스템 사용자는 상징을 이용해 자신의 의도를 표현함으로써 의사소통한다. AAC 상징은 [그림 1]과 같이 a) 상징 이미지, b) 상징 이름, c) 상징 표현으로 구성된다[2]. 상징의 이름은 상징을 관리하기 위한 것으로, 동일한 상징 표현이 여러 개 존재할 경우 상징 이름으로 구분한다. 상징 표현은 음성 합성기를 이용한 음성 출력을 위한 것으로, 특수문자 등 발화할 수 없는 것은 포함하지 않는다. 상징은 [그림 1]의 “학교”와 같이 하나의 단어를 표현하거나, [그림 1]의 “학교 다녀오겠습니다”와 같이 하나의 문장을 표현한다.

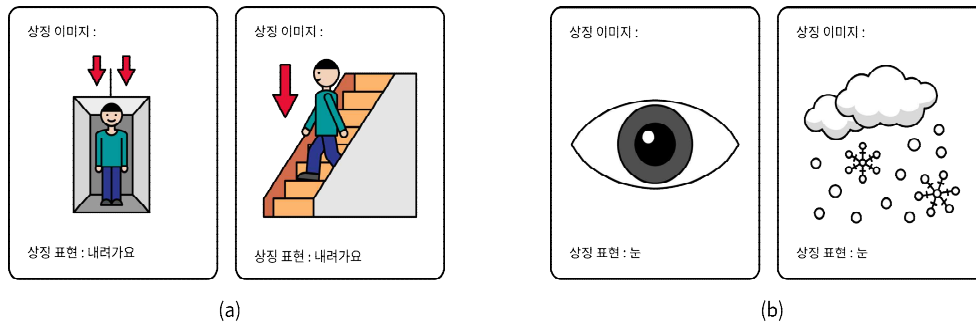


[그림 1] AAC 상징

AAC 도구는 간단한 그림카드부터 전자기기에서의 음성 출력까지 다양하다[11]. AAC 사용자는 직접 상징을 선택해 상대방에게 제시함으로써 자신의 의사를 표현하거나, 전자기기에 탑재된 AAC 도구를 활용하여 상징 표현을 음성으로 출력해 제시하는 방식으로 의사소통한다. 기술이 발전함에 따라 국내외에서 전자식 AAC 도구에 관한 연구가 활발히 이루어지고 있다. 국내의 전자식 AAC 서비스에는 WebAAC[2], 나의 AAC[12], 마이토키[13], GeoAAC[14] 등이 있다. 사용자는 이러한 AAC 서비스를 통해 대면 의사소통에 도움을 받을 수 있다.

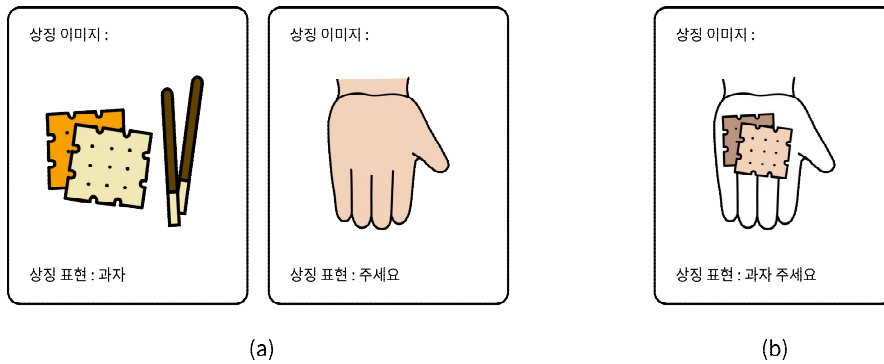
국내외 AAC 상징 중 아라비아 숫자나 인사 상징 이미지 등은 대부분의 국가에서 공통으로 사용할 수 있지만, 한국만의 문화나 전통에 관련된 상징은 존재하지 않아 표현에 제한이 있을 수 있다. 따라서, AAC 상징은 AAC 사용자가 자신의 의도를 명확하게 전달할 수 있도록 사용자가 속한 사회 문화를 반영해야 하며, 어휘와 이미지 또한 해당 사회 문화를 나타내기에 적합한 것으로 구성되어야 한다[3]. 국내에서 개발한 한국의 사회 문화를 반영한 AAC 상징에는 한국형 보완대체의사소통 상징 체계집[3]이 있다. 한국형 AAC 상징 체계집은 한국에서의 의사소통에 사용될

수 있는 “한복”, “삼일절” 등 한국 고유의 문화에 관련된 상징들을 포함한 일상생활에서 자주 사용되는 표현을 가지는 상징으로 구성된다. 한국형 AAC 상징 체계집에 속하는 상징들의 특징은 다음과 같다.



[그림 2] 한국형 AAC 상징의 특징 1

- 하나의 상징 표현이 서로 다른 상징 이미지를 가진다. [그림 2]의 (a)는 같은 상징 표현이지만 왼쪽 상징은 엘리베이터를 타고 내려가는 이미지를, 오른쪽 상징은 계단으로 내려가는 이미지로 구성된다.
- 상징 표현이 같지만 서로 다른 의미를 나타내는 상징이 존재한다. [그림 2]의 (b)는 상징 표현이 “눈”으로 같지만, 왼쪽 상징은 신체 부위인 ‘눈’을 의미하고, 오른쪽 상징은 날씨 중 하나인 ‘눈’을 의미한다.



[그림 3] 한국형 AAC 상징의 특징 2

- 상징 표현은 단어 또는 문장으로 구성된다. 사용자는 자신의 언어 능력에 따라 하나의 상징을 사용하거나 여러 개의 상징을 결합한 상징 시퀀스를 사용하여 의사소통할 수 있다. 예를 들어, [그림 3]의 “과자 주세요” 문장은 (a) “과자” 상징과 “주세요” 상징의 결합으로 표현하거나, (b) “과자 주세요”와 같이 하나의 상징으로 표현할 수 있다.

그 외에 한국형 AAC 상징 체계집에 존재하지 않는 어휘는 비슷한 의미를 가진 상징으로 대체하여 사용해야 하는 등의 특징이 있다. 예를 들어, ‘신상품’ 상징은 한국형 AAC 상징 체계집에 존재하지 않으므로, ‘새로운’과 ‘제품’ 상징의 결합으로 표현해야 한다.

AAC 사용자는 명확한 의사소통을 위해 한국형 AAC 상징 체계집 상징의 특성을 고려해야 한다. 맥락에 맞는 상징을 선택하여 사용해야 하며, 한국형 AAC 상징 체계집이 모든 한국어 어휘 상징을 포함하지 않는다는 것 또한 고려해야 한다. 이 경우 사용자는 자신만의 이미지나 사진으로 상징을 직접 만들어 사용하거나, 비슷한 의미를 가지는 상징의 조합

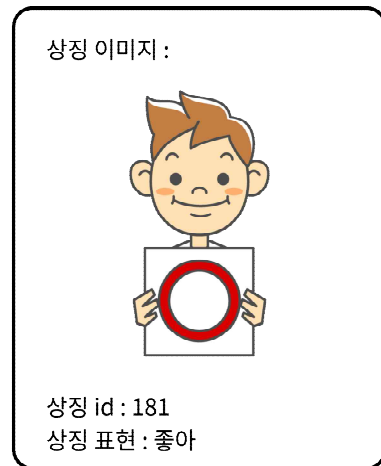
으로 대체하여 사용해야 한다.

구어를 통한 의사소통이 어렵고 글을 읽고 쓰는 것에 어려움을 느끼는 AAC 사용자들은 전화는 물론 문자나 소셜 네트워크 서비스(Social Network Service, SNS) 등을 사용하는 것에 제한을 받는다. 개인이 사용하는 전자기기의 보급과 사용이 증가함에 따라 AAC 사용자들의 AAC 서비스에 대한 수요도 함께 증가하고 있지만, 현존하는 AAC 서비스로는 대면 의사소통 외에 사용이 어렵다. AAC 어플리케이션과 일반 메신저 어플리케이션을 연동하여 AAC 사용자가 쉽게 메신저를 이용할 수 있도록 돕는 서비스의 구조를 제안하는 연구[15]는 있지만, 실질적인 개발까지 이어지지 않는 않았다. 더 나아가 인공지능을 AAC 서비스에 적용하는 연구도 이루어지고 있다. 여러 관련 연구 중 하나인 입력된 영어 문장을 교정하여 픽토그램을 생성하는 서비스[16]는 영어를 기반으로 제안되었으며, 한국어를 적용하기에 적합하지 않다.

2. 딥러닝 기반 대화 문장의 보완대체의사소통 상징 시퀀스 변환

한국어 텍스트를 AAC 상징 시퀀스로 변환하는 기존 연구로는 ‘딥러닝 기반 대화 문장의 보완대체의사소통 상징 시퀀스 변환’[17]이 있다. [17]에서는 문장을 한국형 AAC 상징 시퀀스로 변환하기 위한 방법으로 형태소 분석한 문장을 한국형 AAC 상징에 대응시켜 상징 시퀀스를 생성하는 것과 딥러닝을 적용하여 상징 시퀀스를 생성하는 것을 제안한다. 형태소 분석에는 코모란(Komorán)[18] 모델이 사용되었다.

“좋은 생각입니다” →



[그림 4] “좋은 생각입니다” 문장의 상징 시퀀스로의 변환 결과

형태소 분석기만을 사용해 문장을 상징 시퀀스로 변환하는 모델은 문장을 형태소 분석한 결과 형태소들과 한국형 AAC 상징이 일치하는 경우에 한해서 변환이 이루어진다. 서로 일치하는 형태소가 없는 경우 해당 상징을 제외한 상징 시퀀스가 생성되고, 이는 원래 문장의 의도를 전달하기에 부족하다. 예를 들어, [그림 4]는 “좋은 생각입니다” 문장을 상징 시퀀스로 변환한 결과이다. 한국형 AAC 상징 체계집에 존재하며 형태소 분석 결과가 일치하는 문장의 “좋은”이 “좋아” 상징으로 변환되었다. 그러나 한국형 AAC 상징 체계집에는 “생각” 상징이 없으므로 문장의 “생각”은 상징으로 변환되지 못하였다. 이는 숫자의 경우에도 마찬가지이다. “이만 칠천 원이에요” 문장은 “원”이라는 상징 시퀀스로 변환된다. 숫자 “이”에 해당하는 상징으로 “2” 상징이 존재하지만, 숫자 형식의 상징 표현이므로 일치하지 않는다고 판단하기 때문이다.

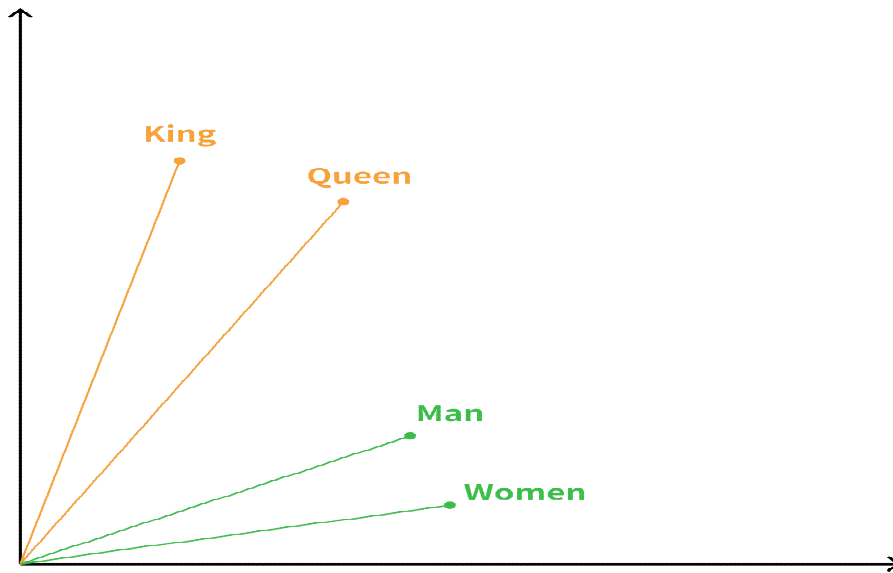
형태소 분석기만을 이용한 상징 시퀀스 생성의 한계점을 해결하기 위해 딥러닝을 적용한 모델을 구현하였다. 1차로 형태소 분석기만을 활용한

모델을 사용해 문장을 상징 시퀀스로 변환하였다. 결과 상징 시퀀스 중 잘 변환되지 않은 것은 적절한 상징을 찾아 대응시키는 작업을 진행했다. “좋은 생각이예요” 문장은 형태소 분석기를 통한 상징 시퀀스 생성 결과 “좋아”로 변환되는데, 추가 작업을 통해 “생각”과 비슷한 의미를 나타내는 어휘인 “아이디어” 상징을 추가해 원문과 비슷한 의미를 가지도록 하였다. “이만 칠천원이에요” 상징은 [“2”, “만원”, “7”, “천원”] 상징 시퀀스로 수정하였다.

[17]은 케라스 임베딩을 사용한 Seq2Seq 모델, 케라스 임베딩을 사용한 어텐션 기반 Seq2Seq 모델, 사전 훈련 임베딩을 사용한 Seq2Seq 모델, 사전 훈련 임베딩을 사용한 어텐션 기반 Seq2Seq 모델을 사용해 문장을 한국형 AAC 상징 시퀀스로 예측 변환하고 그 결과를 비교하였다. 실험 결과 평가에는 0 이상 1 이하의 값을 가지며, 1에 가까울수록 후보 문장과 참조 문장이 유사함을 의미하는 BLEU[19] 점수를 사용했다. 사용한 네 가지의 모델 중 사전 훈련 임베딩을 사용한 어텐션 기반 Seq2Seq 모델의 BLEU 점수가 0.8로 가장 높게 나타났다.

3. 단어 임베딩(Word Embedding)

단어 임베딩(Word Embedding)이란 단어를 고정된 길이의 벡터로 표현하는 것이다[20]. 단어나 문서를 벡터로 표현하면 직관적인 해석과 거리 측정 같은 계산이 가능해지며, 다양한 기계 학습 알고리즘에 사용할 수 있다. [그림 5]는 유사한 단어인 King-Queen과 Man-Woman을 2차원 공간에서 단어 벡터로 나타낸 것이다.



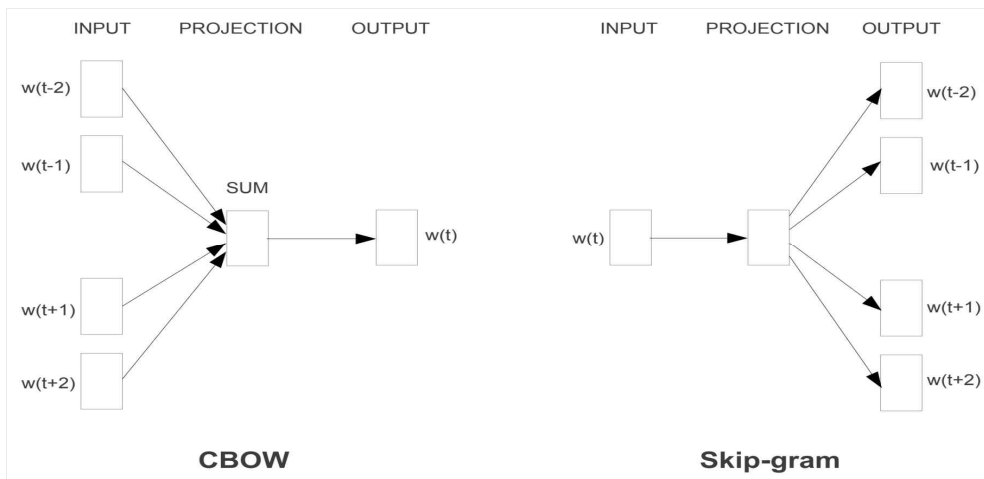
[그림 5] 유사한 단어의 2차원 공간에서의 벡터 표현
출처: <https://www.mygreatlearning.com/blog/word-embedding/>

최근에는 임베딩을 위해 신경망이나 임베딩 층(Embedding layer)에 의존하는 것이 아닌 단어의 벡터 표현을 위해 word-context matrix를 사용하기도 한다. 단어 임베딩은 임베딩을 유도하는 전략에 따라 prediction-based 모델과 count-based 모델 두 종류로 분류된다.

- Prediction-based model : 이 모델은 neural language model의 발전과 깊은 연관성을 가진다. 단어 임베딩은 임베딩 층인 첫 번째 층에 단어를 투영시키는 것이다. Prediction-based 모델에는 Word2Vec[7-8]과 skip-gram에서 더 발전된 방식인 페이스북[21]에서 제안하는 FastText[10] 등이 있다.
- Count-based model : 이 모델은 주어진 문맥에서 다음 단어를 예측하는 알고리즘을 학습하는 것으로 단어 임베딩을 생성하는 것이 아

닌, 말뭉치 내의 단어 동시 등장 횟수와 같은 정보를 사용하여 word-context matrix를 생성한다. Count-based 모델에는 GloVe[9] 단어 임베딩 방법이 있다.

1) Word2Vec(Word-to-Vector)



[그림 6] Word2Vec 모델 구조

출처: <https://mgubaidullin.github.io/deeplearning4j-docs/kr/word2vec>

Word2Vec의 학습 방식에는 CBOW(Continuous Bag-of-Words) 모델과 Skip-gram 모델이 있다[7]. 두 모델을 결합한 Word2Vec 모델의 구조는 [그림 6]와 같다.

CBOW는 주변의 단어들을 기반으로 중심 단어를 예측하는 방식이다. [그림 6]의 왼쪽에서 확인할 수 있듯이 CBOW는 은닉층이 여러 개인 딥러닝(Deep Learning)과 달리, 은닉층이 1개인 얇은 신경망(Shallow Neural Network)을 가진다. Skip-gram은 CBOW 모델과는 반대로, 중심

단어를 입력으로 받아 주변의 단어를 예측하는 방식이다[7].

예를 들어, “나는 오늘 간식으로 배를 먹었다.”와 같은 예문에서 “간식”을 제외한 나머지 단어로부터 “간식”을 예측하는 것이 CBOW의 역할이며, “간식”으로부터 나머지 단어를 예측하는 것이 Skip-gram의 역할이다.

2) GloVe(Global Vectors for Word Representation)

GloVe 모델은 prediction-base와 count-base 모델을 모두 사용하는 방법으로, 2014년에 개발되었다[9]. Count-based 모델인 LSA(Latent Semantic Analysis)와 Prediction-based 모델인 Word2Vec을 보완하려는 목적을 가지며, 뛰어난 성능을 보인다. GloVe 모델은 임베딩된 두 단어 벡터 간 유사도 측정을 효율적으로 수행하며 말뭉치 전체의 통계 정보를 효과적으로 반영한다. GloVe의 목적 함수(J)인 식 (1)을 최소화함으로써 임베딩된 “중심 단어 벡터와 주변 단어의 벡터 내적($w_i^T \tilde{w}_j$)”이 전체 말뭉치에서의 “동시 등장 확률($\log X_{ij}$)”이 되도록 할 수 있다[9]. 식 (1)에서 X는 동시 등장 행렬을 나타내며, X_{ij} 는 중심 단어 i가 나타났을 때 일정 범위 내에 주변 단어 j가 등장하는 횟수이다. w_i 는 중심 단어 i의 임베딩 벡터를, \tilde{w}_j 는 주변 단어 j의 임베딩 벡터를 의미한다. 상수항 b_i 와 \tilde{b}_j 는 각각 w_i 와 \tilde{w}_j 의 편향이다.

$$J = \sum_{i,j=1}^V f(X_{ij}) (w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \quad (1)$$

3) FastText

FastText 모델은 Word2Vec을 확장한 것이다. Word2Vec은 단어를 더 이상 나눌 수 없는 단위로 취급한다면, FastText는 하나의 단어를 n-gram의 구성으로 간주하여 서브 단어(Subword)를 고려한 학습을 수행한다[10]. 하나의 단어를 n개로 분리하여 구성된 벡터와 원래의 단어에 <과 >를 추가하여 시작과 끝을 나타낸다. FastText의 학습 후에는 모든 단어의 n-gram에 대해서 단어 임베딩이 이루어지며, 그로 인해 학습하지 않은 모르는 단어(Out of Vocabulary, OOV)에 대해서도 유사도를 계산할 수 있게 된다[10]. 한국어에 적용할 수 있는 FastText 메커니즘에는 음절 단위 임베딩과 자모 단위 임베딩의 두 가지 방식이 있다. 자모 단위 임베딩은 우선 단어를 자모 단위로 분리하고, 중성이 존재하지 않는 음절은 분리한 후 ‘_’ 토큰을 추가한다.

| 단어 - “자연어처리” | |
|--------------|-------------------------|
| 음절 단위 임베딩 | <자연, 자연어, 언어처, 어처리, 처리> |
| 자모 단위 임베딩 | <자, 자_, 자_ㅇ, ...> |

[표 1] FastText 메커니즘의 n=3일 때의 n-gram

[표 1]은 “자연어처리” 단어에 FastText 메커니즘을 적용한 예시이다. 음절 단위 임베딩을 적용한 FastText 모델이 “자연어처리”라는 단어를 학습하지 못하였고, 다른 문맥에서 “자연어”와 “처리”를 학습하였다면 FastText는 “자연어처리”에 대한 단어 벡터를 얻어낼 수 있다. 자모 단위 임베딩을 적용한 FastText 모델에서는 오타 또는 노이즈를 잘 반영한 임베딩을 기대해볼 수 있다.

III. 상징의 다중 의미를 반영한 상징 시퀀스 변환

AAC 상징은 이미지와 그 이미지를 설명하는 단어 또는 문장으로 구성된다. AAC 상징의 사용자는 상징의 이미지와 표현을 보고 하나 이상의 상징을 조합하여 상대방에게 자신의 의도를 전달한다. “콜라 주세요.” 문장을 상징 시퀀스로 제시하기 위해, 사용자는 “콜라” 상징과 “주세요” 상징을 사용한다.



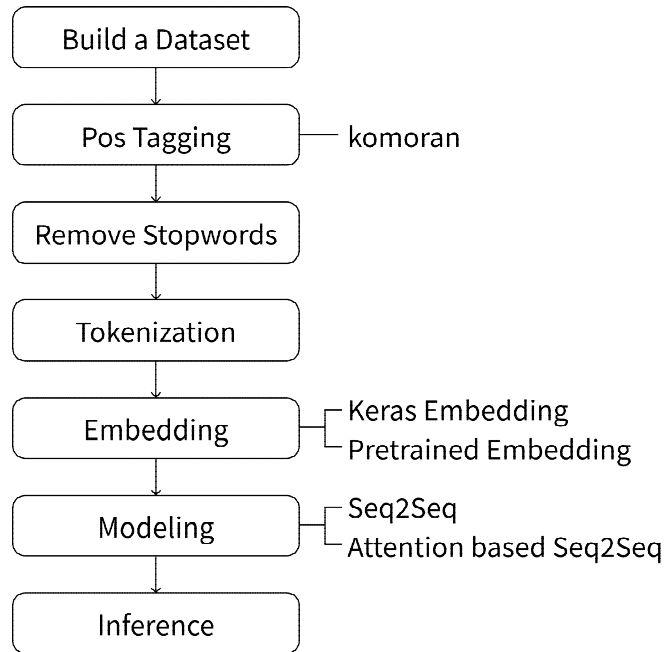
[그림 7] 한국형 AAC 상징의 다중 의미

대화와 같이 맥락이 이어지는 상황에서 상징은 다양한 의미로 사용될 수 있다. 예를 들어, 상대방이 질문을 하면 AAC 사용자가 그에 대한 답을 하는 상황이다. 상대방이 “어떤 음료 좋아하세요?”라는 질문을 한 경우 AAC 사용자는 그에 대한 답변으로 어떠한 음료를 좋아하는지에 대해 말해야 한다. 이때, [그림 7]의 (a)와 같이 “아메리카노” 상징과 “좋아해요” 상징을 조합한 상징 시퀀스를 제시할 수도 있지만, [그림 7]의 (b)와 같이 “아메리카노” 상징 하나만 사용해서 해당 음료를 좋아한다는 의도를 충분히 전달할 수 있다. AAC는 의사소통을 보조하는 시스템으로, AAC 상징은 일반적으로 의사소통 상황에서 사용된다. 의사소통 상황에

서는 맥락에 따라 생략된 말을 유추하여 해석하고 받아들일 수 있다. 따라서, AAC 사용자가 전달하려는 의도가 충분히 반영된다면 상징 시퀀스 구성 시 일부 상징을 생략하는 것이 가능하다.

AAC 상징의 다중 의미를 반영한 문장을 한국형 AAC 상징으로 변환하는 서비스의 개발은 AAC 사용자들이 대면은 물론 문자, SNS 등 비대면으로 의사소통하는 것에 느끼는 불편함을 감소시킬 것으로 기대한다. 문장을 구성하는 상징의 개수를 최소화한다면 장애 정도가 심하거나 AAC 상징이 익숙하지 않은 사용자는 쉽게 AAC 상징 시퀀스를 구성할 수 있을 것이다. 본 논문은 상징의 다중 의미를 적용하여 문장을 상징 시퀀스로 구성하였다. 구축된 데이터로 딥러닝 모델을 학습시키고, 문장을 상징 시퀀스로 예측 변환한 결과를 분석하고 평가하고자 한다.

IV. 딥러닝 기반 문장의 상징 시퀀스 변환 설계



[그림 8] 딥러닝 기반 대화 문장의 AAC 상징 시퀀스 변환 모델 설계도

본 논문은 기존 연구인 [17]의 데이터 셋과 모델을 구성하는 임베딩을 보완하여 개선하고자 한다. 기존 연구에서는 상징의 다중 의미를 고려하지 않은 데이터 셋을 구축하였다. 예측 변환 결과 모든 문장 요소가 하나의 상징으로 대응되어 상징 시퀀스를 구성하였다. 상징의 다중 의미를 반영하기 위해 데이터를 추가 수집하고, 모든 데이터에 대해 추가 상징 시퀀스 태깅 작업을 진행하여 데이터 셋을 보완하였다. 또, 기존 연구에서는 임베딩에 케라스 임베딩과 Word2Vec 임베딩 두 가지만을 사용하였는데, 여기에 GloVe 임베딩과 FastText 임베딩을 추가로 적용하여 모델을 구현하고 실험하였다.

본 논문은 네 가지 임베딩을 각각 사용한 Seq2Seq 모델과 어텐션 매커니즘을 적용한 Seq2Seq 모델을 설계하고 구현하였다. 모델의 설계 과정은 [그림 8]과 같다.

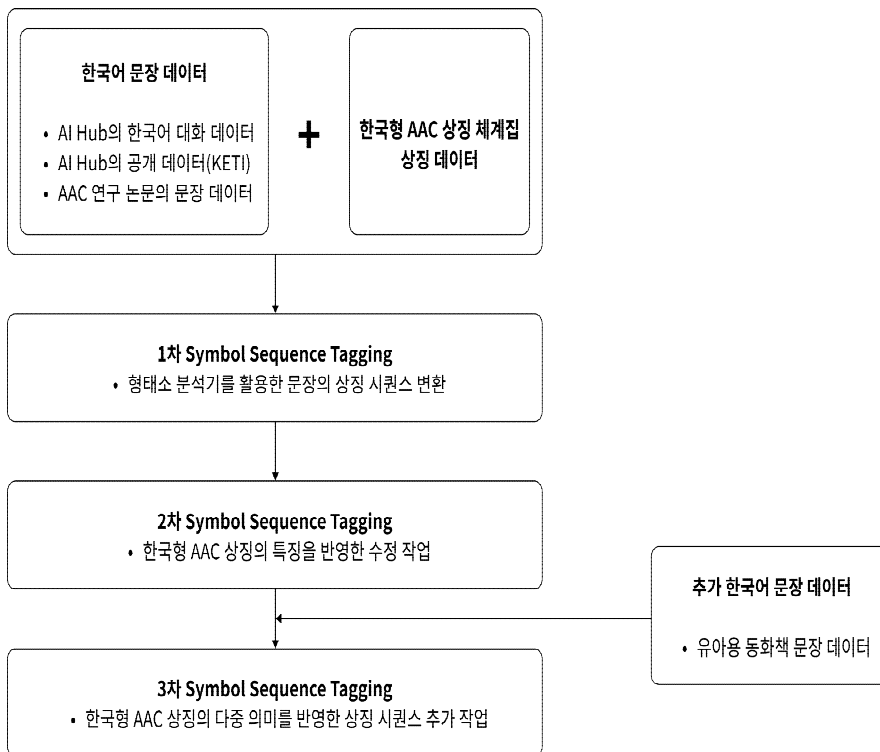
- Build a Dataset : 모델의 학습을 위한 학습 데이터를 구축한다. 학습 데이터는 한국형 AAC 상징 체계집의 상징과 한국어 대화 문장으로 구성된다[17].
- Pos Tagging : 구축한 데이터를 코모란 형태소 분석기를 사용하여 상징 표현과 한국어 대화 문장으로부터 형태소를 추출한다.
- Remove Stopwords : 형태소 추출한 데이터의 불용어를 제거한다.
- Tokenization : 형태소 추출, 불용어 처리를 포함한 전처리 과정을 거친 상징 표현 및 문장 데이터를 토큰화한다[17].
- Embedding : 토큰화된 데이터의 단어 임베딩을 진행한다. 이 과정에서는 케라스 제공 임베딩(Keras Embedding), 사전 훈련 임베딩(Pretrained Embedding)인 Word2Vec, GloVe와 FastText를 사용하여 딥러닝 모델을 구축하였다.

본 논문에서는 Seq2Seq에 단어 임베딩을 각각 적용한 네 가지 모델과 어텐션 매커니즘을 사용한 Seq2Seq에 단어 임베딩 각각을 적용한 네 가지 모델을 구현하였고, 총 8개 모델의 추론(Inference)을 통한 한국어 문

장의 상징 시퀀스 변환 결과를 비교하고 분석하였다.

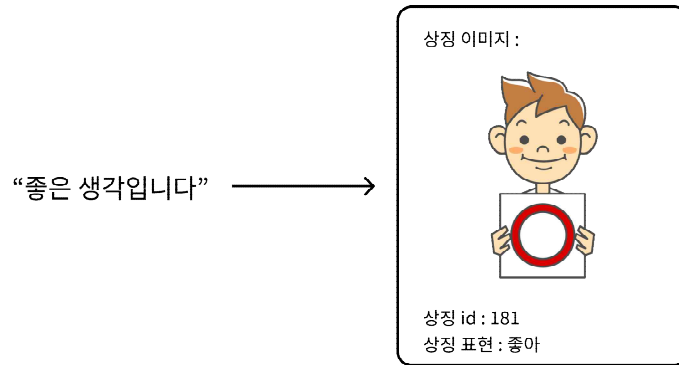
1. 학습 데이터 구축

본 논문에서는 AI Hub에서 제공하는 한국어 대화 데이터[22-23]와 AAC 중재를 위한 어휘 연구 논문[24-27]으로부터 발췌한 대화 문장, 추가 수집한 동화책 문장 데이터를 사용하였다. 학습을 위해 총 15,856개의 한국어 문장 데이터를 사용했다. 문장을 AAC 상징 시퀀스로 나타내기 위해 총 9,442개의 한국형 AAC 상징 체계집의 상징을 사용했다.



[그림 9] 학습 데이터 구축 과정

[그림 9]는 학습 데이터의 구축 과정이다. 한국어 문장 데이터를 3번에 걸친 수정 및 추가 작업을 통해 최종 학습 데이터로 만들었다.



[그림 10] “좋은 생각입니다.” 문장의 상징 시퀀스로의 변환

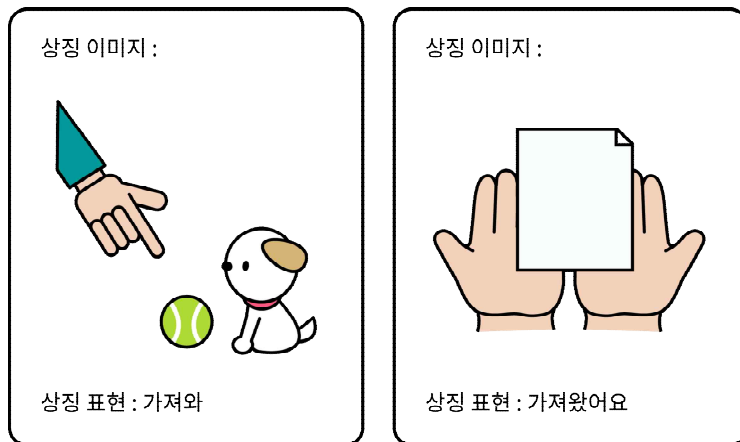
[17]에서는 1차로 형태소 분석기를 활용해 대화 문장의 어휘가 한국형 AAC 상징의 표현과 일치하는 경우 상징 시퀀스로 변환하였다. “다리 아파요.” 문장의 형태소 분석 결과인 “다리”, “아프”와 일치하거나 해당하는 상징이 있을 경우 상징 시퀀스로 변환한다. 형태소 분석기만을 활용한 상징 시퀀스 변환에는 한계점이 존재한다. “좋은 생각입니다.” 문장의 경우 [그림 10]과 같이 “좋아” 상징 시퀀스로 변환된다. “생각”에 해당하는 상징이 한국형 AAC 상징 체계집에 존재하지 않아 변환되지 않는다. 이처럼, 잘못 변환된 문장 데이터의 상징 시퀀스를 대상으로 2차 수정 작업을 진행하였다.

2차 Symbol Sequence Tagging 작업은 잘못 변환된 상징 시퀀스 수정과 동시에 변환된 상징 시퀀스가 한국형 AAC 상징의 특징을 반영하도록 하였다. 한국형 AAC 상징은 하나의 상징 표현이 여러 상징 이미지를

갖거나, 동음이의어 상징이 존재한다. [17]에서는 크게 다섯 가지의 특징을 고려하여 수정 작업을 진행하였다.

첫째, 행위 또는 상태를 나타내는 표현은 원형(기본형) 대신 활용형을 사용하였다. 예를 들어, 한국형 AAC 상징 체계집에 없는 상징인 “가다” 대신 활용형 표현인 “가요” 상징을 사용한다. 높임말 표현도 마찬가지로 활용형을 사용하였다. “드려요” 대신 “주세요” 상징을 사용하여 변환하였다.

둘째, 표현할 수 없는 상징은 비슷한 상징 시퀀스로 대체하였다. “초등학생”은 상징으로 존재하지 않으므로 “초등학교”와 “재학생” 상징을 결합하여 “초등학생”이라는 의미를 나타내는 시퀀스를 구성하였다.



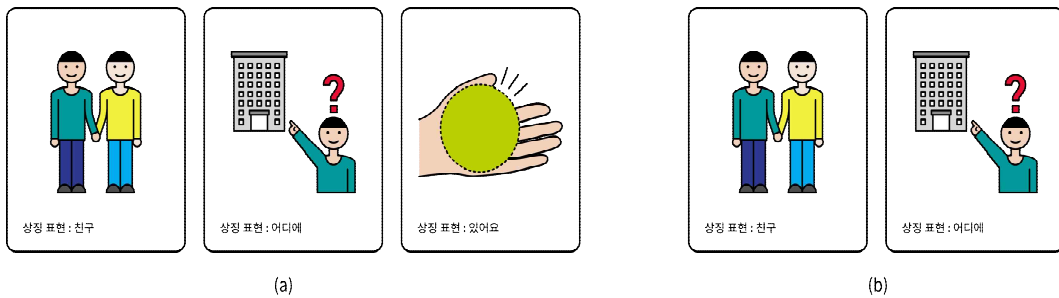
[그림 11] 문맥에 따라 다르게 사용되는 상징 이미지

셋째, 문맥에 맞는 상징 이미지를 사용하도록 하였다. [그림 11]의 “가져와” 상징은 동물에게 지시하는 상황에서 사용될 수 있는 상징이고, “가져왔어요” 상징은 사람과 대화하는 상황에서 사용될 수 있는 상징이다. 이처럼, 상황과 문맥에 따라 적합한 상징 이미지를 가지는 상징 시퀀스로 수정하였다. 네 번째로, 한국형 AAC 상징 체계집의 상징으로는 0부터

100까지 표현할 수 있는데, 그 이상의 숫자는 상징으로 존재하지 않는다. 101 이상의 숫자는 0에서 9까지의 숫자의 조합으로 표현하였다. 예를 들어, 숫자 123은 상징 “1”, 상징 “2”, 상징 “3”을 결합한 상징 시퀀스로 표현한다.

마지막으로, 돈을 나타내는 상징 시퀀스는 화폐 단위 상징을 사용하였다. 예를 들어, “만 칠천 원” 문장은 상징 “만원”, 상징 “7”, 상징 “천원”의 조합으로 표현하였다.

[17]에서 수집한 데이터 외에 추가로 데이터를 수집하여 모델 학습에 활용하였다. 추가 데이터의 수집 방식은 크게 두 가지로 나뉜다. 첫 번째로, 유아용 동화책 문장을 수집하였다. 두 번째로, 기존 문장 데이터와 추가 수집한 문장 데이터에 다중 의미 상징 시퀀스를 추가하는 작업을 수행하였다. 기존의 상징 시퀀스 외에 다른 상징 시퀀스를 가질 수 있는 문장인 경우 해당 문장에 새로운 상징 시퀀스를 추가한다.



[그림 12] “친구 어디에 있어요?” 다중 상징 시퀀스

예를 들어, “친구 어디에 있어요?” 문장은 [그림 12]의 (a)와 같이 “친구” 상징, “어디에” 상징과 “있어요” 상징을 조합한 상징 시퀀스로 표현할 수 있다. 이 중 “어디에” 상징만으로도 “어디에 있어요” 의미를 나타

낼 수 있으므로, “친구 어디에 있어요?” 문장은 [그림 12]의 (b)와 같이 “있어요” 상징을 제외한 상징 시퀀스로도 표현할 수 있다.

| 문장 | 상징 시퀀스 1 | 상징 시퀀스 2 | 상징 시퀀스 3 |
|----------------------|--|------------------------------|------------------------------|
| 어디에 있나요 | [[어디에], {있어요}] | [[어디에]] | |
| 폼 얼마나 오래 걸려요? | [[과마], {시간}, {얼마나오래걸려요}] | [[과마], {얼마나오래걸려요}] | |
| 여기는 몇 시에서 몇 시까지 인가요? | [[여기], {몇시},{시작}, {몇시}, {끝}, {몇시에끝나요}] | [[여기], {몇시},{시작}, {'몇시},{끝}] | [[여기], {몇시}, {시작}, {몇시에끝나요}] |

[표 2] 문장의 다중 상징 시퀀스 태깅

3차 Symbol Sequence Tagging 작업에서는 [표 2]와 같이, 하나의 상징이 가지는 다중 의미를 추가하는 작업을 수행하여 데이터를 구축하였다. 상징 시퀀스 1은 형태소 분석기를 활용한 변환을 수행한 후 한국형 AAC 상징의 특징을 반영하도록 수정한 데이터이다. 상징 시퀀스 2와 3은 상징이 가지는 다중 의미를 반영하여 상징 시퀀스 1을 수정한 것이다. “여기는 몇 시에서 몇 시까지 인가요?” 문장의 경우, 형태소 분석기를 활용한 변환 결과 [[여기], {몇시}, {시작}, {몇시}, {끝}, {몇시에끝나요}]로 변환된다. 원본 상징 시퀀스에서 ‘몇시에끝나요’ 상징을 제거하여도 문장의 의미를 충분히 표현할 수 있으므로 상징 시퀀스 2는 [[여기], {몇시}, {시작}, {몇시}, {끝}]으로 구성한다. 또한, 원본 상징 시퀀스에서 ‘몇시’와 ‘끝’ 상징을 제거하고 ‘몇시에끝나요’ 상징만 포함하여도 문장의 의미를 나타낼 수 있다. 상징 시퀀스 3은 [[여기], {몇시}, {시작}, {몇시에끝나요}]로 구성한다. 모든 문장이 다수의 상징 시퀀스를 가지지는 않는다. 학

습에 사용할 최종 데이터의 수는 [표 3]과 같다. 중복되는 상징 표현을 구분하기 위해 상징 시퀀스는 모두 상징 id로 표현하였다.

| | AI Hub | KETI | 논문 | 동화책 | 합계 |
|----------|--------|-------|-------|-------|--------|
| 문장 | 7,638 | 2,600 | 1,665 | 3,953 | 15,856 |
| 상징 시퀀스 1 | 6,951 | 2,500 | 1,615 | 3,953 | 15,019 |
| 상징 시퀀스 2 | 3,117 | 710 | 311 | 15 | 4,153 |
| 상징 시퀀스 3 | 186 | 26 | 14 | 0 | 226 |

[표 3] 최종 학습 데이터 수 (단위: 개)

2. 데이터 전처리

앞서 구축한 모든 학습 데이터는 전처리를 통해 모델 학습에 적합한 형태로 만들었다. 상징 시퀀스 1이 존재하지 않는 문장 데이터는 제거하였고, 상징 시퀀스 2와 3이 있는 경우에는 각각을 새로운 문장-상징 시퀀스 쌍으로 데이터에 추가하였다. 중복된 문장-상징 시퀀스 쌍을 가지는 데이터는 제거하였다. 상징 시퀀스에는 어미, 조사 등의 형태소들이 사용되지 않는다. 따라서, 문장에서 조사, 어미 등의 형태소를 제거하여 AAC 상징 시퀀스로 변환하였다. 코모란 형태소 분석기를 사용하여 어미 742개, 조사 427개를 판별하고 삭제하여 불용어 처리하였다. [표 4]는 불용어 처리 대상인 어미와 조사의 일부이다.

| 어미 | 조사 |
|----------|----------|
| 거나 | 가 |
| 거늘 | 같이 |
| 거니 | 같이나 |
| 거니와 | 같이는 |
| 거드면 | 같이는야 |
| 거드면은 | 같이는커녕 |
| 거든 | 같이도 |
| 거들랑 | 같이만 |
| 거들랑은 | 같이 |
| ... (중략) | ... (중략) |
| 지마는 | 하고에게 |
| 지만 | 하고의 |
| 지요 | 하고조차 |
| 진 | 하고조차도 |
| 질 | 하곤 |

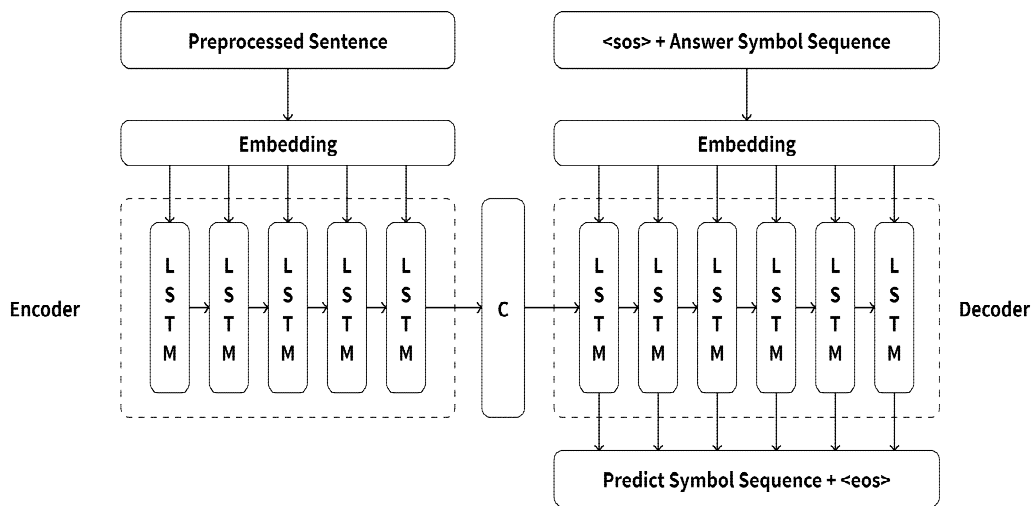
[표 4] 어미와 조사 목록 일부

한국형 AAC 상징 체계집의 상징은 하나의 표현에 대해 상징 이미지가 같거나 다른 여러 상징이 존재한다. 예를 들어, “좋아”를 나타내는 상징에는 “좋아(상징 id 181)”, “좋아요(상징 id 182)”, “좋아요(상징 id 183)”가 있으므로 “좋은 생각이야”는 서로 다른 상징으로 구성된 상징 시퀀스 3개를 가질 수 있다. 이처럼 여러 상징이 존재하는 상징 시퀀스를 개별 데이터 행으로 분리하여 추가하고, 여러 개의 상징 시퀀스를 갖는 문장도 각각의 상징 시퀀스를 별도의 행으로 추가하였다. 중복 문장-상징 시퀀스 쌍을 제거한 후의 최종 학습 데이터는 총 158,539개이다.

3. 다중 의미를 반영한 딥러닝 기반 문장의 AAC 상징 시퀀스 변환 모델

텍스트 대화 문장을 한국형 AAC 상징 시퀀스로 변환하기 위해 본 논문은 Seq2Seq와 어텐션 메커니즘, 네 가지의 단어 임베딩을 적용한 변환 모델을 사용하였다. 각 모델은 한국어 대화 문장을 입력받아 한국형 AAC 상징 시퀀스를 출력한다. 동일한 표현을 가진 상징을 구분하기 위해 한국형 AAC 상징 체계집의 상징 아이디로 상징 시퀀스를 나타낸다. 입력 데이터는 전처리하여 토큰화한다. 토큰화한 데이터는 수치형 텐서를 입력받는 딥러닝 모델을 위해 임베딩 과정을 진행한다.

본 논문은 다음과 같이 8가지의 모델을 사용하여 딥러닝 기반 한국어 대화 문장의 한국형 AAC 상징 시퀀스로의 변환을 학습하였다.

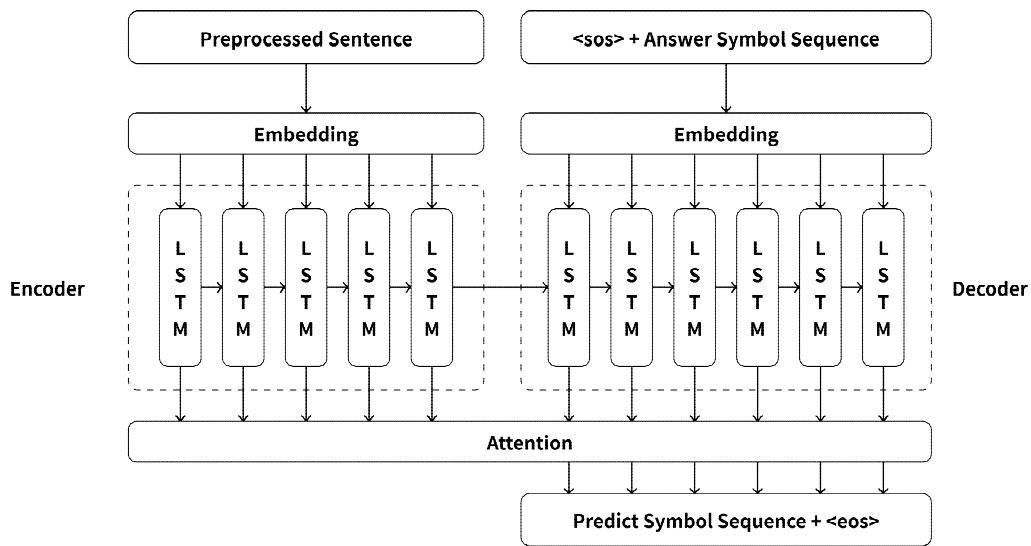


[그림 13] Sequence to Sequence 모델 구조

[그림 13]은 Seq2Seq 모델의 구조이다. Embedding에 케라스 임베딩, Word2Vec, GloVe, FastText를 사용하여 구현한 네 가지 모델은 다음과 같다.

- Seq2Seq : Sequence to Sequence를 사용한 모델로, 케라스 임베딩을 사용하였다.
- Seq2Seq using Word2Vec : Sequence to Sequence를 사용한 모델로, 사전 훈련 학습된 Word2Vec 임베딩을 사용하였다.
- Seq2Seq using GloVe : Sequence to Sequence를 사용한 모델로, 사전 훈련 학습된 GloVe 임베딩을 사용하였다.
- Seq2Seq using FastText : Sequence to Sequence를 사용한 모델로, 사전 훈련 학습된 FastText 임베딩을 사용하였다.

Seq2Seq 모델은 LSTM으로 구현된 인코더와 디코더의 연결 구조를 가진다. 수치화된 텐서는 인코더에 순차적으로 입력되며, 인코더는 입력된 데이터의 정보를 압축한 컨텍스트 벡터를 출력한다. 출력 벡터는 디코더에 전송되고, 디코더는 전달받은 벡터에 근거하여 예측된 한국형 AAC 상정을 출력한다.



[그림 14] 어텐션 매커니즘 기반 Sequence to Sequence 모델 구조

[그림 14]는 어텐션 매커니즘 기반 Seq2Seq 모델의 구조이다. Embedding에 케라스 임베딩, Word2Vec, GloVe, FastText를 사용하여 구현한 네 가지 모델은 다음과 같다.

- Attention based Seq2Seq : 어텐션 매커니즘 기반 Sequence to Sequence 모델로, 케라스 임베딩을 사용하였다.
- Attention based Seq2Seq using Word2Vec : 어텐션 매커니즘 기반 Sequence to Sequence 모델로, 사전 훈련 학습된 Word2Vec 임베딩을 사용하였다.
- Attention based Seq2Seq using GloVe : 어텐션 매커니즘 기반 Sequence to Sequence 모델로, 사전 훈련 학습된 GloVe 임베딩을 사용하였다.

- Attention based Seq2Seq using FastText : 어텐션 매커니즘 기반 Sequence to Sequence 모델로, 사전 훈련 학습된 FastText 임베딩을 사용하였다.

어텐션 매커니즘 기반 Seq2Seq 모델도 LSTM을 사용한 인코더와 디코더의 연결 구조를 가진다. 수치화된 텐서는 순차적으로 인코더에 입력된다. 디코더는 예측을 위해 인코더에 입력된 모든 문장을 참고하고, 예측 단어와 연관된 부분에 집중하며 최종 예측을 수행한다.

V. 모델 실험 및 평가

1. 실험 설계

Linux OS의 Jupyter 노트북 환경에서 개발하였다. 실험 환경은 아래 [표 5]과 같다.

| | 구분 | 사양 및 버전 |
|-----|------------|--|
| H/W | CPU | Intel(R) Xeon(R) Silver 4116 CPU @ 2.10 GHz |
| | RAM | 128.0GB |
| | GPU | Nvidia Titan Xp |
| S/W | OS | Ubuntu 20.04.5 LTS |
| | Python | 3.8.12 |
| | Tensorflow | 2.10.0 |
| | CUDA | 11.4 |

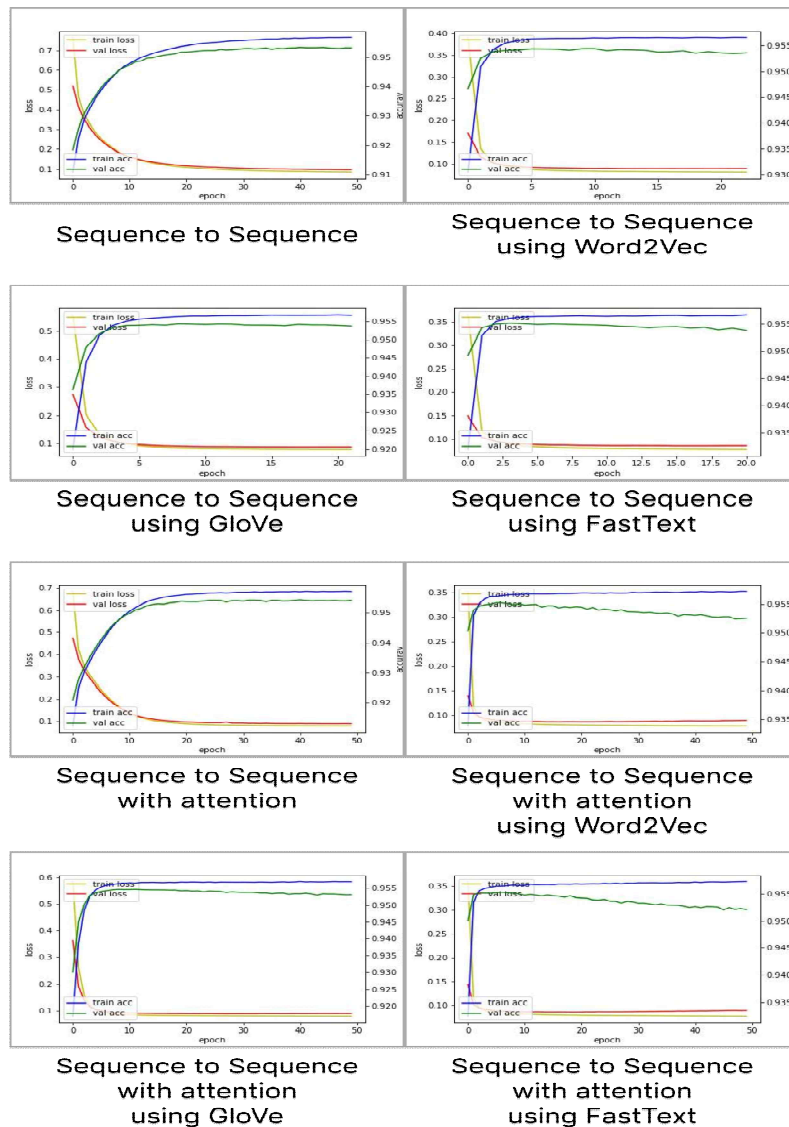
[표 5] 실험 환경

학습을 위해 AI Hub의 한국어 대화 데이터와 유아용 동화책 문장 데이터를 수집하였고, 상정의 다중 의미를 더하는 작업을 진행하여 최종 데이터를 구축하였다. 구축한 데이터 중 무작위의 80%는 학습에, 10%는 검증에, 나머지 10%는 테스트에 사용하였다.

본 논문에서는 4.3에서 설명한 총 여덟 개의 모델을 사용하여 결과를 도출하고 비교하였다. Adam[29]을 옵티마이저로 사용하였고, Sparse Categorical Cross Entropy를 손실 함수로 지정하여 사용했다. 에폭(Epoch)은 50으로 하고 과적합 방지를 위한 조기 종료를 사용했다. 여덟 개의 모델의 실험 결과를 비교 분석하여 다중 의미를 적용한 딥러닝 기

반 대화 문장의 한국형 AAC 상징 시퀀스 변환에 가장 적합한 모델을 확인하였다.

2. 실험 결과



[그림 15] 변환 모델 학습 그래프

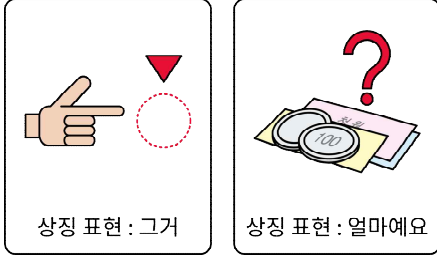

158,539개의 데이터로 학습한 후의 학습 그래프는 [그림 15]와 같다. 왼쪽 세로축은 손실 정도를, 오른쪽 세로 축은 학습 정확도 수치를, 가로축은 학습 에포크 수를 나타낸다. 8가지 모델 모두 0.95 이상의 학습 정확도를 보인 후 학습이 종료되었다. 본 논문에서 구현한 모델의 실험 결과인 위 그림을 아래 [표 6]와 같이 정리하였다.

| 모델 | train | | validation | |
|---------------------------------------|----------|--------|------------|--------|
| | accuracy | loss | accuracy | loss |
| Seq2Seq | 0.9565 | 0.0842 | 0.9532 | 0.0963 |
| Seq2Seq using Word2Vec | 0.9565 | 0.0803 | 0.9532 | 0.0963 |
| Seq2Seq using GloVe | 0.9568 | 0.0793 | 0.9537 | 0.0870 |
| Seq2Seq using FastText | 0.9567 | 0.0793 | 0.9538 | 0.0863 |
| Seq2Seq with attention | 0.9569 | 0.0796 | 0.9540 | 0.0900 |
| Seq2Seq with attention using Word2Vec | 0.9572 | 0.0777 | 0.9526 | 0.0894 |
| Seq2Seq with attention using GloVe | 0.9569 | 0.0784 | 0.9529 | 0.0887 |
| Seq2Seq with attention using FastText | 0.9574 | 0.0771 | 0.9522 | 0.0894 |

[표 6] 다중 의미를 반영한 딥러닝 기반 대화 문장의 한국형 AAC 상징 시퀀스 변환 모델의 학습 결과

[표 6]에서 확인할 수 있듯이 네 가지 모델 모두 약 0.95의 정확도를 나타내며 학습이 잘 이루어졌음을 알 수 있다. 어텐션 매커니즘을 사용한 모델이 어텐션 매커니즘을 사용하지 않은 모델보다 좋은 결과를 보였다. 또, 케라스 임베딩을 사용한 모델보다 사전 훈련 임베딩을 사용한 모델이 더 나은 결과를 보인다. GloVe 임베딩을 사용한 모델보다 Word2Vec, FastText를 사용한 모델이 더 나은 결과를 보인다. Word2Vec 임베딩을 사용한 모델 보다는 FastText 임베딩을 사용한 모델이 더 나은 결과를 보임을 알 수 있다.

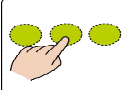


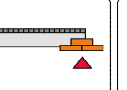

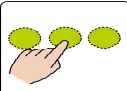
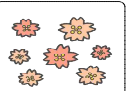

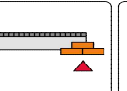

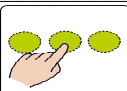
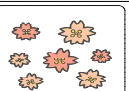

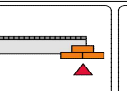

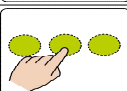


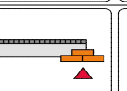

어텐션 매커니즘 기반 Seq2Seq 모델은 Seq2Seq 모델보다 높은 성능을 보인다. 사전 훈련된 단어 임베딩에 한국어 데이터를 사용하였고, 한국어 단어의 문맥적 쓰임을 포함한 수치형 벡터를 생성한다. 사전 훈련된 단어 임베딩을 적용한 모델은 중요한 부분에 가중치를 부여하여 순차적으로 예측을 수행하기 때문에 케라스 제공 임베딩을 적용한 모델보다 높은 성능을 보인다. 또, FastText 임베딩을 사용한 모델의 경우 훈련되지 않은 단어에 대해서도 임베딩 벡터 생성이 가능하므로 Word2Vec, GloVe 단어 임베딩보다 높은 성능을 보임을 알 수 있다.

| Sentence | | 그건 얼마예요 |
|----------|--|---|
| Model | Seq2Seq, Seq2Seq using GloVe, Seq2Seq with attention |  |
| | Seq2Seq using Word2Vec, Seq2Seq using FastText, Seq2Seq with attention using Word2Vec, Seq2Seq with attention using GloVe, Seq2Seq with attention using FastText |  |

[표 7] “그건 얼마예요” 문장의 변환 결과

8가지 모델 모두 높은 학습 정확도와 낮은 손실 정도를 보였지만, 그 중 Seq2Seq 모델은 상대적으로 낮은 성능을 보였다. [표 7]은 8가지 모델의 “그건 얼마예요” 문장에 대한 변환 예시이다. 본 논문은 상징의 다중 의미가 반영된 예측을 수행하는 것을 목적으로 한다. 다중 의미를 반영한 변환을 수행하였을 경우, “그건 얼마예요” 문장은 “얼마예요” 상징 하나만으로 구성된 상징 시퀀스로 변환될 수 있다. 8가지 모델 모두 문장을 적절한 상징 시퀀스로 변환하였지만, Word2Vec 임베딩을 사용한 Seq2Seq, FastText 임베딩을 사용한 Seq2Seq, Word2Vec 임베딩을 사용한 어텐션 매커니즘 기반 Seq2Seq, GloVe 임베딩을 사용한 어텐션 매커니즘 기반 Seq2Seq, FastText 임베딩을 사용한 어텐션 매커니즘 기반

Seq2Seq 모델이 본 논문의 목적에 가장 부합한 추론을 수행한 것을 확인할 수 있다.

| 문장 : “이거는 봄 신상품 아니에요” | | | | | |
|---------------------------------------|---|--|--|---|---|
| Seq2Seq with attention |  상징 표현: 이거 |  상징 표현: 봄 |  상징 표현: 새로운 |  상징 표현: 제품 |  상징 표현: 아니오 상징 id: 95 |
| Seq2Seq with attention using Word2Vec |  상징 표현: 이거 |  상징 표현: 봄 |  상징 표현: 새로운 |  상징 표현: 제품 |  상징 표현: 아니오 상징 id: 92 |
| Seq2Seq with attention using GloVe |  상징 표현: 이거 |  상징 표현: 봄 |  상징 표현: 새로운 |  상징 표현: 제품 |  상징 표현: 아니오 상징 id: 92 |
| Seq2Seq with attention using FastText |  상징 표현: 이거 |  상징 표현: 봄 |  상징 표현: 새로운 |  상징 표현: 제품 |  상징 표현: 아니오 상징 id: 92 |

[표 8] 단어 임베딩의 종류에 따른 문장의 상징 시퀀스 변환 결과

[표 8]은 Seq2Seq 모델 보다 높은 성능을 보인 어텐션 매커니즘 기반 Seq2Seq에 서로 다른 단어 임베딩을 적용한 모델의 문장을 상징 시퀀스로 변환한 결과를 비교한 것이다. 각 모델은 한국형 AAC 상징 체계집에 존재하지 않는 “신상품” 상징을 “새로운”과 “제품” 상징의 조합으로 표현하였다. “아니다”의 활용형인 “아니에요”는 “아니오(상징 id 92, 95)” 상징으로 적절히 변환되었다. 케라스 임베딩을 사용한 모델과 사전 훈련된 단어 임베딩을 사용한 모델은 “아니오” 상징을 서로 다른 이미지의 상징으로 변환하였다. 예측 상징 시퀀스를 구성하는 상징 이미지의 차이만 있을 뿐, 4가지 임베딩 모두 적절한 번역을 수행하는 것을 알 수 있다.

3. 실험 평가

1) BLEU(Bilingual Evaluation Understudy)

BLEU(Bilingual Evaluation Understudy)는 기계 번역과 사람이 번역한 결과의 유사도를 비교하여 번역 성능을 측정하는 방법이다[19]. 측정 기준은 n-gram에 기반한다. BLEU 측정에는 기계 번역 결과와 사람이 직접 번역한 결과가 사용된다. BLEU는 보정된 유니그램 정밀도(Modified Unigram Precision)를 사용한다. 기계 번역 결과를 기반으로 각 단어의 최대 출현 개수를 지정하고, 사람이 번역한 결과에 기계 번역 결과의 단어가 있는지 확인한 후 점수를 계산한다. 문장의 단어 순서를 고려하지 않는 유니그램에서 확장하여 n-gram을 사용해 보정 정밀도를 계산한다. 단어 순서를 고려하는 것에서 나아가 기계 번역 결과의 길이에 대한 것을 추가로 고려하도록 한다. 기계 번역 결과가 사람이 번역한 결과보다 짧을 경우 점수에 패널티를 부여하는데, 이를 브레버티 패널티(Brevity Penalty, BP)라고 한다. 기계 번역을 후보로, 사람이 직접 번역한 결과를 참조로 표현하여 나타낸 BLEU 계산식은 식 (2)와 같다. c 는 기계 번역 결과의 길이를, r 은 기계 번역 결과와 길이 차가 가장 작은 사람이 번역한 결과의 길이를 의미한다. P_n 은 보정된 정밀도를 나타내고, W_n 은 보정된 정밀도에 부여하는 가중치를 나타낸다.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \quad (2)$$

$$BLEU = BP \times \exp\left(\sum_{n=1}^N W_n \log P_n\right)$$

2) 평가 결과

본 논문에서는 모델의 성능 평가를 위해 BLEU를 사용하였다. 참조 문장은 예측 대상 문장에 태깅된 상징 시퀀스로 사용하였다. 본 논문에서 구현하고 실험한 모델의 BLEU 평가 결과는 [표 9]과 같다.

| Model | BLEU Score | BLEU Score without considering multiple sequences |
|---------------------------------------|------------|---|
| Seq2Seq | 0.6449 | 0.9060 |
| Seq2Seq using Word2Vec | 0.6627 | 0.9330 |
| Seq2Seq using GloVe | 0.6636 | 0.9345 |
| Seq2Seq using FastText | 0.6670 | 0.9392 |
| Seq2Seq with attention | 0.6631 | 0.9328 |
| Seq2Seq with attention using Word2Vec | 0.6671 | 0.9394 |
| Seq2Seq with attention using GloVe | 0.6686 | 0.9406 |
| Seq2Seq with attention using FastText | 0.6728 | 0.9472 |

[표 9] BLEU 평가 결과

하나의 문장이 두 개 이상의 상징 시퀀스를 가지더라도 각 상징 시퀀스는 궁극적으로 하나의 문장을 표현한다. 이 점을 고려하여 BLEU 점수를 두 가지 방법으로 산출하였다. BLEU Score는 문장과 상징 시퀀스 쌍의 개수를 통해, BLEU Score without considering multiple sequences는 중복되지 않는 문장 데이터 수를 통해 계산되었다.

Seq2Seq 모델의 BLEU 평가 점수는 나머지 7개의 모델보다 낮은 점수를 나타낸다. 어텐션 기반 Seq2Seq 모델의 점수가 Seq2Seq의 점수보다 높은 것으로 볼 때, 어텐션 매커니즘을 사용한 모델이 더 좋은 성능을 낼 수 있다. 사전 훈련된 임베딩을 사용한 어텐션 매커니즘 기반 Seq2Seq 모델의 점수는 비슷하게 나타났으나, Word2Vec 임베딩을 사용한 모델보다는 GloVe 임베딩을 사용한 모델이 더 높은 평가 점수를 보이며, FastText 임베딩을 사용한 모델의 BLEU 평가 점수가 가장 높게 나왔다. 결과적으로, 사전 훈련된 임베딩과 어텐션 매커니즘을 모두 사용한 모델인 FastText 임베딩을 사용한 어텐션 매커니즘 기반 Seq2Seq 모델의 BLEU 점수가 가장 높으며, 가장 높은 번역 성능을 나타냄을 알 수 있다.

BLEU 점수는 0과 1사이의 값을 가진다. 1에 가까울수록 후보 문장과 참조 문장이 유사함을 의미한다. [19]에서는 BLEU 점수가 0.4 이상일 경우 번역이 잘 된 것으로 해석한다. 본 논문의 네 가지 모델의 BLEU 점수는 모두 0.6 이상으로 대체로 높은 번역 품질을 보이는 것을 확인하였다.

[17]에서 제안하는 변환 모델의 가장 높은 BLEU 점수는 0.8로 본 논문

에서 제안하는 BLEU Score 보다 높다. BLEU 점수를 계산할 때 전체 평가 데이터에 대해 점수를 계산하여 합하고 이를 문장 수로 나누어 주었다. 본 논문에서는 상징의 다중 의미를 반영한 상징 시퀀스가 추가되었으므로, 나누어주는 수인 학습 데이터 수를 조정할 필요가 있다. 여러 개의 상징 시퀀스를 갖는 문장은 하나로 간주하여 계산한 BLEU Score without considering multiple sequences의 경우 0.9 이상의 매우 높은 번역 성능 점수를 보였다.

VI. 결론 및 향후 연구

본 논문에서는 언어 장애인들이 문장을 받아들이기 쉽도록 문장을 보완대체의사소통(AAC) 상징 시퀀스로 변환하는 딥러닝 기반 번역 모델을 제안한다. 언어 장애인들은 주로 AAC 상징을 대면 의사소통에 사용해왔다. 비대면 의사소통이 중요해짐에 따라 비장애인과 장애인의 비대면 의사소통 또한 증가하였다. 비대면 의사소통 상황에는 전자기기를 통한 글 또는 목소리를 통해 소통이 이루어진다. 언어 장애인들은 비대면 의사소통 상황에 필요한 글을 읽고 쓰는 능력이 낮기 때문에 원활한 의사소통이 어렵다. 현존하는 모바일 의사소통 보조 서비스는 같은 공간에 있는 상대에게만 적용 가능하다는 제한점이 있다. 대화 문장을 언어 장애인들에게 익숙한 AAC 상징 시퀀스로 변환함으로써 문자, 소셜 네트워크 서비스 등 비대면 의사소통에 도움을 줄 수 있을 것이다.

상징 이미지는 그 자체로 여러 의미를 가질 수 있다. 본 논문에서는 상징의 이러한 특징에 주목하여 기존 연구의 데이터에 상징이 표현할 수 있는 다중 의미를 고려하여 추가한 상징 시퀀스 데이터를 구축하였다. 또한, AI Hub에 공개된 데이터 외에 유아용 동화책의 문장 데이터를 추가하여 데이터의 다양성을 높였다.

본 논문에서는 딥러닝을 기반으로 하는 한국어 문장의 한국형 AAC 상징으로의 변환 모델을 이용해 실험하였다. 실험에는 한국어 대화 문장들과 한국형 AAC 상징 체계집의 상징들이 사용되었다. 코모란 형태소 분석기를 사용해 불용어 처리하고 전처리 하였다. 케라스 제공 임베딩, 사전 훈련된 Word2Vec 임베딩, GloVe 임베딩, FastText 임베딩과

Sequence to Sequence(Seq2Seq), 어텐션 매커니즘 기반 Seq2Seq를 조합하여 총 여덟 가지의 모델을 구현하였다. GloVe 임베딩은 전체 말뭉치에서 단어의 등장 횟수를 고려하며, FastText 임베딩은 단어를 작게 나누어 학습하므로 학습하지 않은 단어를 잘 예측할 것이므로, 두 가지 임베딩을 사용한 모델을 추가 실험하여 결과를 비교하고 평가하였다. 평가에는 BLEU 점수를 사용하였다. BLEU 점수 해석에 따르면 여덟 가지 모델 모두 좋은 번역 성능을 나타내는데, 그 중 FastText 임베딩을 활용한 어텐션 매커니즘 기반 Seq2Seq 모델이 BLEU 점수 0.67로 가장 높은 성능을 보였다.

다양한 문장을 상징으로 변환하기 위해서는 상징의 추가적인 개발이 필요하다. 한국형 AAC 상징 체계집의 상징은 약 만 개의 유아/초등 저학년 수준의 어휘로 구성된다. 일상생활의 단어를 모두 표현하기에는 그 수가 적다. 부족한 일상생활에서 자주 사용되는 단어뿐만 아니라 “메타버스”와 같이 기술이 발전함에 따라 새로 등장하는 단어를 상징으로 제작하여 AAC 사용자가 더 많은 단어를 번역할 수 있도록 해야 한다. 또한, 중/고등 수준의 어휘 상징을 개발하여 넓은 연령대의 AAC 사용자가 사용하는 문장을 AAC 상징 시퀀스로 번역할 수 있도록 지원해야 한다.

본 논문에 사용된 데이터는 공개 한국어 데이터 문장과 동화책 문장, AAC 중재를 위한 문장 연구 논문에서 발췌한 문장 데이터를 사용하였다. 논문에서 발췌한 문장 데이터 외에 공개 데이터와 동화책 문장 데이터는 타당도 검증이 이루어지지 않았다. 추후 연구로 검증되지 않은 데이터에 대한 검증이 필요하다. 또, 사용자의 선호 상징을 파악하는 연구가 필요하다. AAC 사용자가 자주 사용하는 상징으로 문장을 번역한다면, 사

용자가 쉽게 번역 결과를 받아들일 수 있을 것이다. 본 논문에서 제한하는 모델을 실생활에 적용하기 위해서는 우선 실제 언어 장애인들이 자주 사용하는 문장으로 해당 모델을 평가하고 보완해야 한다.

참 고 문 헌

- [1] D. R. Beukelman and P. Mirenda, “Augmentative and alternative communication: Supporting children and adults with complex communication needs,” 4th ed. Baltimore, MD: Brookes Publishing, 2020.
- [2] Y. Jang and K.-H. Hong, “An HTML5 based AAC board making system,” *KIISE Transactions on Computing Practices*, vol. 21, no. 5, pp. 365–372, 2015.
- [3] E. Park, Y. Kim, K.-H. Hong, S. Yeon, K. Kim and J. Lim, “Development of Korean Ewha-AAC symbols: validity of vocabulary and graphic symbols,” *AAC Research & Practice*, vol. 4, no. 2, p. 19–40, 2016.
- [4] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with Neural Networks,” *arXiv [cs.CL]*, 2014.
- [5] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv [cs.CL]*, 2014.
- [6] “Embedding Layers - Keras Documentation,” *Keras.io*. [Online]. Available: <https://keras.io/ko/layers/embeddings/>. [Accessed: 10-Nov-2022].
- [7] T. Mikolov, K. Chen, G. Corrado and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv [cs.CL]*, 2013.
- [8] K. Park, “wordvectors,” [Online]. Available: <https://github.com/Kyubyong/wordvectors>. [Accessed: 10-Nov-2022].
- [9] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors

- for Word Representation,” In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Proceedings(EMNLP), p. 1532-1543, 2014.
- [10] M. V. Mahoney, “Fast Text Compression with Neural Networks,” In proceedings of the 30th International Florida Artificial Intelligence Research Society Conference, AAAI press, p. 230-234, 2000.
- [11] E. Park, “Using AAC in Special Education,” AAC Research & Practice, vol. 2, no. 2, p. 131, 2014.
- [12] “나의 AAC 종합정보사이트,” myaac.co.kr. [Online]. Available: <https://www.myaac.co.kr/>. [Accessed: 01-Nov-2020].
- [13] “마이토키,” mytalkie.co.kr. [Online]. Available: <http://www.mytalkie.co.kr/>. [Accessed: 01-Nov-2020].
- [14] H. Cho and K.-H. Hong, “GeoAAC, A Location-based Augmentative and Alternative Communication Mobile Application,” AAC Research & Practice, vol. 8, no. 1, p. 87-117, 2020.
- [15] H.-J. Yun and H.-J. Park, “Use of Mobile Messengers through Connection of AAC with Kakao Talk,” AAC Research & Practice, vol. 3, no. 2, p. 167, 2015.
- [16] C.-J. Park, Y.-H. Kim, Y. Jang, U. G.R, and H.-S. Lim, “An AI service to support communication and language learning for people with developmental disability,” Journal of the Korean Convergence Societ, vol. 11, no. 6, p. 52-57, 2020.
- [17] 이주현, “딥러닝 기반 대화 문장의 보완대체의사소통 상징 시퀀스 변환,” 국내석사학위논문, 성신여자대학교, 2021. 서울특별시
- [18] “Komoran,” [Online]. Available: <https://docs.komoran.kr/>. [Accessed:

14-Nov-2022].

- [19] K. Papinei, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL*, 2021.
- [20] F. Almeida and G. Xexeo, “Word Embeddings: A Survey,” *arXiv [cs.CL]*, 2019.
- [21] “Meta(Facebook),” [Online]. Available: <https://about.meta.com/ko/>. [Accessed: 28-Nov-2022].
- [22] “AI Hub,” *Aihub.or.kr*. [Online] Available: <https://aihub.or.kr/>. [Accessed: 08-Nov-2022].
- [23] “인식기술-언어지능,” *Aihub.or.kr*. [Online] Available: https://aihub.or.kr/keti_data_board/language_intelligence. [Accessed: 08-Nov-2022].
- [24] 천춘경, “보완·대체 의사소통(AAC)체계 활용을 위한 지역 사회 중심의 기초 어휘 및 문장 조사,” 단국대학교, 서울, 2000.
- [25] 김수미, “AAC를 활용한 함께 책 읽기 중재가 복합의사소통장애 학생의 의미 관계 표현과 어휘다양도 변화에 미치는 효과,” 국내석사학위논문, 창원대학교, 2019. 경상남도
- [26] E.-H. Park, “Core vocabulary for nonverbal elementary school students with cerebral palsy,” *Korean Journal of Special Education*, vol. 13, no. 1, p. 91-115, 1996.
- [27] 이정은, 박은혜, “보완·대체의사소통체계 적용을 위한 상황 중심 핵심어 위 개발 연구,” *Journal of Rehabilitation Research*, vol. 4, no. 1, p. 96-122, 2000.

- [28] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv [cs.LG]*, 2014.

ABSTRACT

Transformation of Text Sentences to AAC Symbol Sequences Based on Deep Learning using Syntactic Information of AAC Symbols with Multiple Meaning

Ryu Se Hui
Department of
Future Convergence
Technology Engineering
Graduate School of
Sungshin University

People who have difficulty with communicating in spoken language use a communication tool, Augmentative and Alternative Communication(AAC). AAC services for those people have been developed in various forms such as mobile services and web services. Existing AAC services mainly used in face-to-face communication, but they are difficult to use in non-face-to-face communication like e-mail, messenger, or etc. As Internet usage increases significantly, the demand for non-face-to-face communication is also increasing significantly, and intervention for non-face-to-face communication between the non-disabled and the language impaired people has

become urgent. People generally send and receive text dialogue in non-face-to-face communication situation, but people with language disability have difficulty understanding texts.

In this paper, we proposed deep learning models to translate text sentences into AAC symbol sequences to help the people with language disability in non-face-to-face communication situation with AAC symbols which they are familiar with. In order to convert text sentences into AAC symbols, we use Korean dialogue sentences from AI Hub open data, additional text sentences extracted from Korean fairy tail books and Korean AAC symbols. The Korean text sentences are possible to be transformed to a Korean AAC symbol sequences, we built a data set by reflecting the characteristics of the Korean AAC symbol images and considering the communication context. Since there may be several AAC symbols corresponding to one text vocabulary and AAC symbols may have multiple meanings, we tagged one or more symbol sequences to each text sentences in the data set.

Using the data set, we constructed eight translation models by combining two deep learning models (sequence to sequence and attention based sequence to sequence models) with four word embedding techniques (Keras, Word2Vec, GloVe and FastText). Word2Vec and Keras embedding was already used in a previous word to translate text sentences into AAC symbol sequences.

We conducted additional experiments using two embeddings: GloVe and FastText. By reflecting the numbers of times words appear in the

whole data set, the GloVe embedding compensates for the shortcoming of Word2Vec, which predicts only words within a certain scope. The FastText embedding has higher possibility to predict untrained words, since it performs learning by dividing words into sub-words. We compared the performance of the previous work using Keras and Word2Vec embeddings with the models using GloVe and FastText embeddings. We used BLEU score to measure the model performance. Sequence to sequence with attention mechanism using pretrained FastText embedding showed the highest performance.