Journal of The Korea Society of Computer and Information
Vol. 22 No. 4, pp. 41-49, April 2017

www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.04.041

A Study on Word Sense Disambiguation Using Bidirectional Recurrent Neural Network for Korean Language

Jihong Min*, Joon-Woo Jeon **, Kwang-Ho Song***, Yoo-Sung Kim****

Abstract

Word sense disambiguation(WSD) that determines the exact meaning of homonym which can be used in different meanings even in one form is very important to understand the semantical meaning of text document. Many recent researches on WSD have widely used NNLM(Neural Network Language Model) in which neural network is used to represent a document into vectors and to analyze its semantics. Among the previous WSD researches using NNLM, RNN(Recurrent Neural Network) model has better performance than other models because RNN model can reflect the occurrence order of words in addition to the word appearance information in a document. However, since RNN model uses only the forward order of word occurrences in a document, it is not able to reflect natural language's characteristics that later words can affect the meanings of the preceding words. In this paper, we propose a WSD scheme using Bidirectional RNN that can reflect not only the forward order but also the backward order of word occurrences in a document. From the experiments, the accuracy of the proposed model is higher than that of previous method using RNN. Hence, it is confirmed that bidirectional order information of word occurrences is useful for WSD in Korean language.

▸ Keyword : Word Sense Disambiguation, Neural Network Language Model, Context Vector, Bidirectional Recurrent Neural Network

I. Introduction

최근 동형이의어 처리를 위해 인공 신경망으로 단어 임베딩 벡터(Word Embedding Vector)나 맥락 벡터(Context Vector)를 계산하는 신경망 언어 모델(Neural Network Language Model)[1]을 동형이의어 중의성 해소에 사용하는 연구[2][3][4][5]들이 이루어지고 있다. 한국어 동형이의어 중의성 해소를 위해 신경망 언어 모델을 사용한 연구들은 맥락 벡터를 계산하는 방법에 따라 CBOW(Continuous Bag of Words)[2][3][4]를 이용한 동형이의어 중의성 해소 모델과 순환 신경망(Recurrent Neural Network: RNN)[4][5]을 이용한 동형이의어 중의성 해소 모델로 나눌 수 있다. 순환 신경망을 이용한 동형이의어 중의성 해소 모델은 단어 출현 정보뿐만 아니라 단어 출현 순서 정보도 동형이의어 중의성 해소 과정에 반영한다. 따라서 순환 신경망을 이용한 동형이의어 중의성 해소 모델은 단어 출현 정보만 동형이의어 중의성 해소에 반영하는 CBOW를 이용한 동형이의어 중의성 해소 모델보다 더 정확한 동형이의어 중의성 해소 결과를 얻을 수 있다[5]. 하지만,

• First Author: Jihong Min, Corresponding Author: Yoo-Sung Kim
*Jihong Min (newindia89@gmail.com), Dept. of Information and Communication Engineering, Inha University
**Joon-Woo Jeon (junu0723@gmail.com), Dept. of Information and Communication Engineering, Inha University
***Kwang-Ho Song (crossofjc@gmail.com), Dept. of Information and Communication Engineering, Inha University
****Yoo-Sung Kim (yskim@inha.ac.kr), Dept. of Information and Communication Engineering, Inha University
• Received: 2017. 01. 24, Revised: 2017. 03. 02, Accepted: 2017. 03. 20.
• This research project was supported by Ministry of Culture, Sports and Tourism (MCST) and from Korea Copyright Commission in 2016.

기존 순환 신경망 모델은 단어 출현 순서의 순방향 정보만 반영한다. 이 문제로 인해 기존 순환 신경망을 이용한 동형이의어 중의성 해소 모델들은 뒤의 단어가 앞의 단어의 의미에 영향을 미치는 자연어의 특성을 제대로 반영하지 못한다는 문제가 있다[6].

따라서 본 논문에서는 기존 순환 신경망을 이용한 동형이의어 중의성 해소 연구의 단점을 개선하기 위해 단어 출현 순서의 순방향 정보 뿐 아니라 역방향 정보도 반영하는 신경망 모델인 양방향 순환 신경망(Bidirectional Recurrent Neural Network: BRNN)[7]을 이용한 동형이의어 중의성 해소 모델을 제안한다. 본 논문에서 제안하는 모델은 단어 출현 순서의 순방향 정보 뿐 아니라 역방향 정보를 동형이의어 중의성 해소 과정에 같이 반영하여 앞의 단어의 뜻을 파악하는데 뒤의 단어가 영향을 줄 수 있는 자연어의 특성을 기존 모델보다 잘 반영할 수 있는 특징이 있다. 또한 본 논문에서 제안하는 동형이의어 중의성 해소 모델은 기존 순환 신경망을 이용한 동형이의어 중의성 해소 모델[4][5]과 달리 LSTM(Long Short Term Memory)[8]과 유사하지만 구조가 간단한 GRU(Gated Recurrent Unit)[9]를 적용해 LSTM을 사용한 기존 순환 신경망 모델보다 적은 파라메터로 더 정확한 중의성 해소가 가능하도록 하였다. 제안하는 모델의 성능을 검증하기 위한 동형이의어 중의성 해소 실험 결과, 본 연구에서 제안하는 모델이 기존 순환 신경망 모델보다 더 정확한 결과를 얻음으로써 동형이의어 중의성 해소에 단어 출현 순서의 양방향 정보를 반영하는 것이 효과적임을 확인하였다.

본 논문의 구성은 다음과 같다. 2절에서는 본 연구와 연관이 있는 관련 연구들에 대해 소개한다. 3절에서는 본 연구에서 제안하는 동형이의어 중의성 해소 모델에 대해 설명한다. 4절에서는 제안하는 동형이의어 중의성 해소 모델에 대한 성능을 검증한다. 5절에서는 본 연구의 결론을 맺는다.

II. Related Works

1. Word Sense Diambiguation Using Neural Network Language Model

신경망 언어 모델(Neural Network Language Model)[1]을 사용한 동형이의어 중의성 해소 연구는 인공 신경망을 이용해 단어를 표현하는 단어 임베딩 벡터나 문장을 대표하는 맥락 벡터(Context Vector)를 계산하는 신경망 언어 모델을 동형이의어 중의성 해소에 사용한 연구들이다. 신경망 언어 모델을 이용한 동형이의어 중의성 해소 연구들은 신경망 모델을 사용해 문장들의 집합인 말뭉치에 포함된 단어들에 대한 단어 임베딩 벡터를 계산하는 과정, 문장에 포함된 단어들의 단어 임베딩 벡터들로부터 문장을 표현하는 벡터인 맥락 벡터를 계산하는 과정, 그리고 맥락 벡터들 간 코사인 유사도 비교를 이용해 동형이의어의 중의성을 해소하는 의미 결정 과정까지 총 3단계의 과정을 통해 동형이의어의 의미를 결정한다.

신경망 언어 모델 기반 동형이의어 중의성 해소 연구의 3단계중 2번째 단계인 맥락 벡터를 계산하는 단계에서 이용하는 맥락 벡터를 계산 방법에 따라 CBOW(Continuous Bag-of-Words)를 이용한 동형이의어 중의성 해소 모델(이하 CBOW 모델)[2][3][4]과 단어의 출현 정보와 단어 출현 순서 정보를 계산 과정에 반영하는 신경망인 순환 신경망(Recurrent Neural Network)을 이용한 동형이의어 중의성 해소 모델(이하 순환 신경망 모델)로[4][5] 나눌 수 있다. 이들 중 단어의 출현 정보만 반영하는 CBOW 모델보다 단어의 출현 정보 뿐 아니라 단어 출현 순서 정보도 반영한 순환 신경망 모델이 더 정확하게 동형이의어의 중의성을 해소하는 것으로 알려져 있다[4][5].

2. Word Sense Disambiguation Using Recurrent Neural Network

순환 신경망을 사용한 동형이의어 중의성 해소 모델은 입력 값 뿐 아니라 입력 값들의 출현 순서 정보를 반영해 결과를 계산하는 신경망 모델인 순환 신경망[10]으로 맥락 벡터를 계산하는 모델이다. 순환 신경망 모델의 구조는 Fig. 1과 같다.

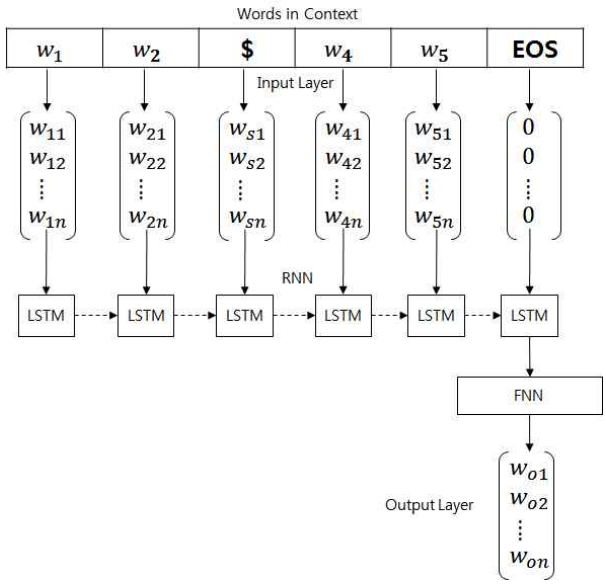


Fig. 1. Structure of Recurrent Neural Network[5]

Fig. 1과 같이 단어 w_1, w_2, w_3, w_4, w_5로 이루어진 문장이 있고, 이 중 동형이의어는 w_3라고 가정할 때, 순환 신경망 모델이 문장의 맥락 벡터를 계산하는 과정은 다음과 같다. Fig. 1의 순환 신경망 모델은 입력 계층에서 문장의 단어들을 단어 임베딩 벡터로 매핑하는 과정을 진행한다. 이 과정에서 중의성 해소 대상 동형이의어 w_3는 특정한 심볼 $를 가리키는 단어 임베딩 벡터로 매핑하고, 문장의 끝을 나타내는 심볼인 EOS(End of

Sentence)들은 0벡터로, 나머지 단어들은 각 단어에 해당하는 단어 임베딩 벡터로 매핑한다.

Fig. 1의 순환 신경망은 입력값의 출현 순서에 따라 상태 정보를 계산, 저장, 전달을 가능하게 하는 recurrent unit[9]의 일종이며, 상태정보를 계산하는 게이트와 상태정보를 저장하는 메모리 셀로 이루어진 LSTM(Long Short Term Memory)[8]을 사용해 단어 임베딩 벡터와 단어 출현 순서의 순방향 정보를 맥락 벡터 계산 과정에 반영한다. Fig. 1의 LSTM은 먼저 첫 단어인 w_1에 대한 단어 임베딩 벡터를 입력으로 받아 초기 상태 정보를 계산한다. 이후 다음 단어인 w_2에 대한 단어 임베딩 벡터와 초기 상태 정보로부터 w_2까지의 상태 정보를 계산하며 마지막 단어의 단어 임베딩 벡터에 의한 상태 정보를 반영할 때 까지 반복한다. 순환 신경망에 의해 계산된 상태 정보는 이후 완전 연결 신경망(Fully Connected Neural Network: FNN)에서 추가 연산을 진행하고 동형이의어 중의성 해소에 사용하는 맥락 벡터를 반환하게 된다.

순환 신경망 모델은 단어의 출현 정보 뿐 아니라 단어 출현 순서 정보도 맥락 벡터 계산 과정에 반영하기 때문에 CBOW 모델보다 동형이의어 중의성 해소 정확도가 더 높은 것으로 확인되었다. 하지만, 순환 신경망 모델은 순방향의 단어 출현 순서 정보만 반영하기 때문에, 뒤의 단어가 앞의 단어의 뜻을 확인하는데 영향을 줄 수 있는 자연어의 특성을 온전하게 반영하지 못한다는 문제가 있다[6].

3. Bidirectional Recurrent Neural Network

앞서 언급한 바와 같이 순환 신경망 모델은 입력의 순서 중 순방향의 순서만을 고려하며, 자연어처럼 뒤의 단어가 앞의 단어의 뜻이나 역할에 영향을 미치는 경우가 존재하는 경우에는 [6] 계산 과정에서 정보의 손실이 있을 수 있다. 이러한 순환 신경망의 결점을 해결하기 위해 입력 순서의 순방향과 역방향을 동시에 고려하는 양방향 순환 신경망[7]이 최근 자연어 처리 연구들에 활발하게 적용되고 있다.

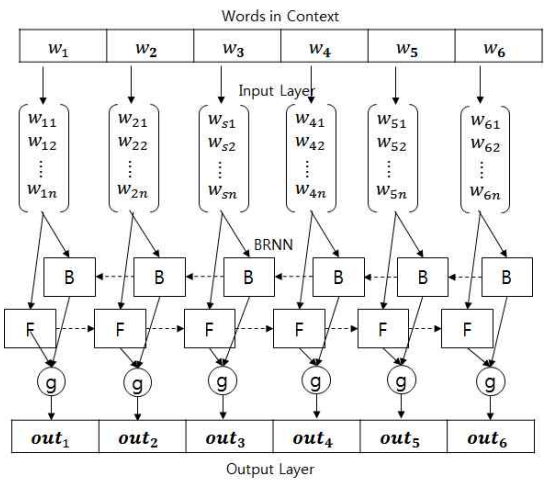


Fig. 2. BRNN for Korean Language Processing

한국어에 대해 형태소 분석이나 의미역 결정을 위해 사용하는 양방향 순환 신경망[6][11]들은 Fig. 2와 같이 모든 입력 상태마다 도출하는 출력값을 사용해 자연어 처리에 활용한다. Fig. 2와 같은 한국어 처리에 사용되는 양방향 순환 신경망 모델은 Fig. 1의 순환 신경망 모델과 달리 입력의 순방향(그림의 F)으로 상태 정보를 전달하는 recurrent unit과 입력의 역방향(그림의 B)으로 상태 정보를 전달하는 recurrent unit으로 이루어져있다. 순방향 recurrent unit들은 이전 단어 임베딩 벡터에 대한 상태 정보 계산 결과를 현재 단어의 상태정보 계산에 반영한다면, 역방향 recurrent unit들은 다음 단어 임베딩 벡터에 대한 계산 결과를 현재 단어의 상태 정보 계산에 반영한다. 출력 계층에서는 동일한 단어를 입력 값으로 받는 순방향 및 역방향 recurrent unit의 결과를 별도의 연산 g를 거쳐 결과를 반환한다.

4. Recurrent Unit

순환 신경망은 오류 전파를 위해 계산하는 편도함수 값이 0에 가까워져 모델 학습에 장애를 주는 그래디언트 소멸(vanishing gradient)아니면 편도함수 값이 매우 커지는 그래디언트 발산 (exploding gradient) 문제가 발생 할 수 있다[12]. 이와 같이 현재 입력으로부터 멀리 떨어진 입력의 영향이 현재 입력에 의한 상태를 계산할 때 거의 영향을 주지 못하는 현상을 장기 의존성(long term dependency)문제라고 한다[13]. 순환 신경망을 적용하는 연구들은 순환 신경망의 은닉 계층에 단순한 활성화 함수 대신 상태 정보와 입력 값의 입출력 반영 비율을 계산하는 여러개의 비선형 활성화 함수 집합인 게이트와 상태 정보를 저장하기 위한 메모리 셀로 이루어진 recurrent unit을 적용해 장기 의존성 문제를 해결하였다. Recurrent unit의 종류로는 LSTM(Long Short Term Memory)[8]과 GRU(Gated Recurrent Unit)[9]가 있으며, LSTM과 GRU의 개략적인 구조는 Fig. 3과 같다.

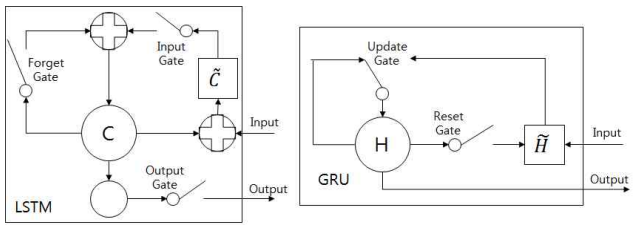


Fig. 3. Structure of LSTM and GRU[16]

LSTM은 입력 값의 가중치 합과 비선형 활성화 함수를 상태 정보로 기억하는 메모리 셀 C와 과거 상태의 정보를 현재 상태 계산을 위해 전달할 정보의 양을 결정하는 forget 게이트, 현재 상태를 계산하기 위해 현재의 입력과 이전 상태 C 사이의 연산 결과인 $\tilde{C}$ 를 현재 상태 계산에 반영할 양을 결정하는 input 게이트, 현재 상태 C를 출력 값으로 반환하기 위한 output 게이트로 구성된 recurrent unit이다. GRU는 과거의 상태와 현재의

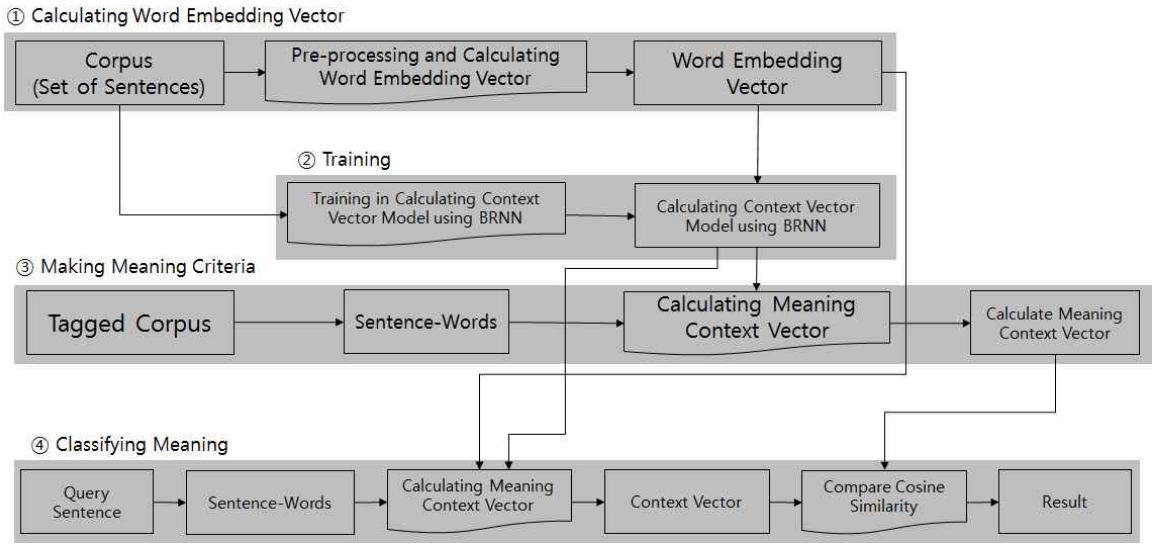


Fig. 4. Flowchart of Word Sense Disambiguation for Korean Using BRNN

입력의 상호 보간을 통해 메모리 셀에 저장할 현재 상태를 계산하는 recurrent unit이다. GRU의 상태 정보를 저장하는 메모리 셀 H와 이전 상태를 현재 상태 계산에 얼마나 반영할지 계산하는 reset 게이트, 현재 상태를 계산하기 위해 reset 게이트를 통과한 과거의 상태와 현재의 입력의 반영 비율을 결정하는 update 게이트로 이루어져있다.

LSTM과 GRU 모두 장기 의존성 문제를 해결할 수 있다는 공통점이 있다. 다만 LSTM은 output 게이트의 존재로 셀이 저장하는 상태와 외부에 출력되는 값이 다르지만, GRU는 셀이 저장하는 상태와 외부에 출력되는 값이 동일하다는 차이점이 있다. 또한 현재까지의 연구 결과로는 둘 중 어느 것이 우월한지 알려져 있지 않으며, 사용하는 데이터와 실험 환경에 따라 적합한 recurrent unit이 다른 것으로 알려져 있다[9].

본 논문에서는 기존 순환 신경망 모델이 단어 출현 순서의 순방향 정보만을 이용하기 때문에 발생하는 정보의 손실을 개선하기 위해 양방향 순환 신경망을 이용해 맥락 벡터를 계산하는 동형이의어 중의성 해소 모델을 제안한다. 또한 한국어에 대한 동형이의어 중의성 해소 정확도 실험 결과와 모델 학습에 필요한 시간을 고려해 제안하는 모델에 적합한 recurrent unit 선정한다.

III. Word Sense Disambiguation Using Bidirectional Recurrent Neural Network

1. Introduction

본 절에서는 양방향 순환 신경망모델을 이용한 동형이의어의 중의성 해소 모델을 제안한다. 본 논문에서 제안하는 모델은 크게 학습에 사용할 말뭉치를 전처리하고 맥락 벡터를 계산하기 위해 사용할 단어별 단어 임베딩 벡터를 계산하는 단어 임베딩 벡터 계산, 맥락 벡터를 계산하기 위한 맥락 벡터 계산 모델을 만들고 말뭉치로부터 맥락 벡터 계산 모델을 학습하는 학습 과정, 의미 정보가 부착된 말뭉치로부터 동형이의어의 의미를 결정하는 기준으로 사용하기 위한 맥락 벡터들을 계산하는 의미 판정 기준 결정 과정, 의미 판정 기준들의 맥락 벡터들과 질의 문장의 맥락 벡터간 유사도 비교를 통해 동형이의어의 중의성을 해소하는 의미 결정 4단계로 구분할 수 있다. 이 논문에서 제안하는 모델의 흐름도는 Fig. 4와 같다.

2. Computing Word Embedding Vector

본 논문에서 제안하는 양방향 순환 신경망 모델의 단어 임베딩 벡터 계산 과정은 말뭉치 전처리 과정과 단어 임베딩 벡터 계산으로 나뉜다. 전처리 과정은 다시 형태소 분석과 저빈도어 치환으로 구분할 수 있다. 형태소 분석과정은 말뭉치의 문장들에 형태소 분석기를 적용해 내용어인 명사, 동사, 형용사, 부사들만 추출하는 과정이다. 형태소 분석 없이도 내용어[14]를 추출할 수 있는 굴절어인 영어와는 달리, 교착어인 한국어는 내용어가 조사나 접사와 결합해 어절을 구성한다. 따라서 불필요한 조사나 접사들을 제거하기 위해 별도의 형태소 분석 과정을 통해 말뭉치의 어절들이 내용어만 포함하도록 학습용 말뭉치를 가공한다. 저빈도어 치환은 학습 말뭉치에서 적게 출현하는 단어를 unknown 토큰으로 치환하고, 단어 임베딩 벡터의 표현력을 높이는 과정이다. 본 연구에서는 단어 임베딩 벡터 라이브러리[15]의 기본 값인 5회를 최저 출현 빈도수로 정하였고, 이 빈도수보다 낮은 단어들은 unknown으로 치환해 5회 이상 출현하는 단어와 unknown에 해당하는 단어 임베딩 벡터를 만들었다. Unknown에 해당하는 단어 임베딩 벡터는 맥락 벡터 계산 모델의 입력 계층에서 동형이의어 심볼 $을 위한 벡터로 사용한다.

단어 임베딩 벡터 계산 과정은 전처리가 된 말뭉치로부터 개

별 단어의 단어 임베딩 벡터를 계산하는 과정이다. 본 논문에서 제안하는 양방향 순환 신경망 모델은 단어 임베딩 벡터를 만들기 위해 사용하는 알고리즘으로 단어 임베딩 벡터를 계산하고자 하는 특정 단어의 인접단어들의 단어 임베딩 벡터들로부터 특정 단어의 단어 임베딩 벡터를 계산하는 CBOW[16]를 사용하였다.

3. Context Vector Computation Model Using Bidirectional Recurrent Neural Network

본 연구에서 맥락 벡터를 계산할 때 사용하는 양방향 순환 신경망을 사용한 맥락 벡터 계산 모델은 입력 계층, 양방향 순환 신경망, 완전 연결 신경망, 그리고 출력 계층으로 이루어져 있으며, 그 구조는 Fig. 5와 같다.

Fig. 5는 단어 w_1, w_2, w_3, w_4, w_5와 동형이의어 w_3로 이루어져 있는 문장의 맥락 벡터를 계산하는 양방향 순환 신경망을 사용한 맥락 벡터 계산 모델이다. 입력 계층은 중의성을 해소하고자 하는 동형이의어 w_3를 동형이의어를 나타내는 심볼인 $로 변환한다. 이후 이루어지는 매핑 과정에서는 단어 w_1, w_2, w_4, w_5는 각 단어에 해당하는 단어 임베딩 벡터로 매핑 진행한다. 반면, 동형이의어를 나타내는 심볼 $는 3.2절에서 계산한 unknown 토큰에 대한 벡터로 매핑하며, 문장의 끝을 나타내는 심볼인 EOS는 0벡터로 매핑 한다.

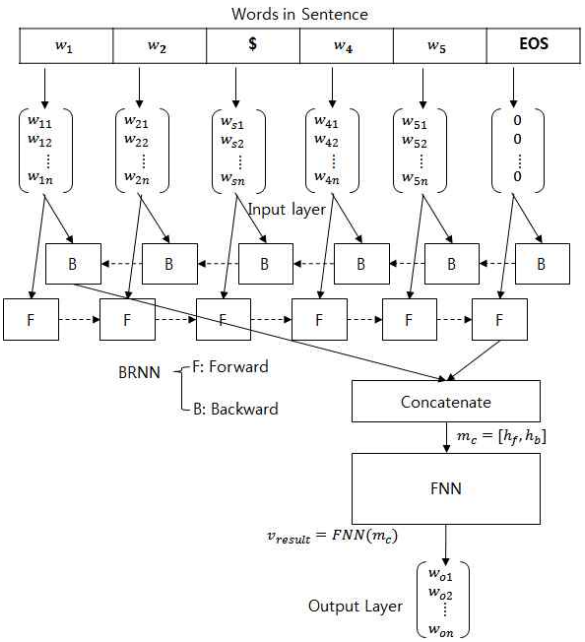


Fig. 5. Calculating Context Vector Using BRNN

단어의 출현 정보와 단어의 순서정보를 반영해 맥락 벡터를 계산하는 양방향 순환 신경망은 단어 임베딩 벡터에 따른 상태 정보 전달 방향이 다른 recurrent unit으로 이루어진 은닉 계층을 가지고 있다. 순방향 F의 recurrent unit은 메모리 셀에 저장되어 있는 첫 단어부터 전 단어까지의 상태정보와 현재 단어 임베딩 벡터를 사용해 메모리 셀에 저장되어 있는 상태 정보를 갱신하며, 계산된 현재 상태 h_{f_n}은 식 (1)과 같다. 반면 역방향 B의 recurrent unit은 메모리 셀에 저장되어 있는 마지막 단어부터 다음 단어까지의 상태 정보와 현재 단어의 단어 임베딩 벡터를 사용해 메모리 셀에 저장되어 있는 상태 정보를 갱신하며, 계산된 현재 상태 h_{b_n}은 식 (2)와 같다. 식 (1)과 (2)에서 z_f, z_b는 출력값에 영향을 주는 recurrent unit의 게이트를, x_n는 특정 위치 n에서의 입력 값을, h_{f_n}은 순방향 recurrent unit이 출력으로 내보내는 상태, h_{b_n}는 역방향 recurrent unit이 출력으로 내보내는 상태, f는 단어순서의 순방향, b는 단어순서의 역방향을 의미한다.

$$h_{f_n} = z_f(x_n, h_{f_{n-1}}) \quad (1)$$

$$h_{b_n} = z_b(x_n, h_{b_{n+1}}) \quad (2)$$

순방향 recurrent unit은 w_1부터 EOS까지 반영된 상태 정보 h_f를 반환하고, 역방향 recurrent unit은 EOS부터 w_1까지 반영된 상태정보 h_b를 반환한다. 본 논문에서 제안하는 양방향 순환 신경망 모델은 상태정보 h_f와 h_b로부터 하나의 맥락 벡터를 계산하기 위해 식 (3)과 같이 상태 정보들을 Concatenate해 하나의 행렬을 만들고 완전 연결 신경망의 입력으로 전달한다.

$$m_c = [h_f, h_b] \quad (3)$$

완전 연결 신경망은 2개의 은닉 계층으로 이루어져 있으며, m_c를 입력 값으로 사용해 동형이의어 w_3에 대한 맥락 벡터 v_{result}를 계산하고 그 결과를 반환한다. 완전 연결 신경망을 통과한 맥락 벡터 v_{result}는 식 (4)와 같다.

$$v_{result} = fnn(m_c) \quad (4)$$

4. Learning Context Vector Computation Model

동형이의어의 의미 정보가 부착된 말뭉치는 수동으로 의미 정보를 부착해야 해 제작에 많은 시간이 필요하며 그 양이 적을 수밖에 없다. 이러한 점 때문에 의미 정보가 부착된 말뭉치로부터 많은 내부 파라메터가 존재하는 양방향 순환 신경망 모델을 과적합 없이 학습하기에는 학습 데이터의 양이 부족하다는 문제점이 있다. 따라서 본 연구에서는 순환 신경망 모델을 사용한 [5]의 연구처럼 양방향 순환 신경망 모델을 학습하기 위해 단어 임베딩 벡터를 만들기 위한 학습 데이터들을 사용해 양방향 순환 신경망을 이용한 맥락 벡터 계산 모델을 학습한다. 양방향 순환 신경망을 이용한 맥락 벡터 계산 모델은 문장의 단어들 중 중의성 해소 대상 단어를 제외한 나머지 단어들의 단어 임베딩 벡터로 중의성 해소 대상 단어의 벡터를 계산한다. 이를 활용해 본 연구에서도 양방향 순환 신경망 기반 맥락 벡터 계산 모델을 학습 데이터에 있는 모든 문장들과 문장이 가지고 있는 모든 단어에 대해 Fig. 6과 같이 매핑을 진행한다.

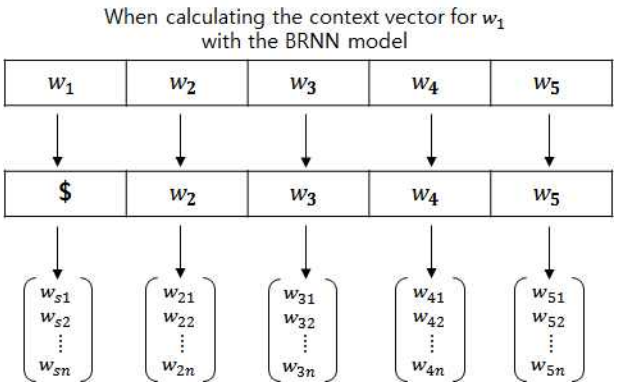


Fig. 6. Using BRNN to Calculate a Context Vector for w_1

Fig. 6은 w_1, w_2, w_3, w_4, w_5로 이루어진 문장을 양방향 순환 신경망을 이용한 맥락 벡터 계산 모델의 학습을 위해 변환하는 과정이다. Fig. 6과 같이 w_1을 중의성 해소 대상으로 하는 맥락 벡터를 계산할 경우, w_1은 동형이의어를 나타내는 심볼인 $로 치환해 unknown 토큰을 가리키는 벡터로 변환하고, 나머지 단어들은 각 단어에 해당하는 벡터로 변환한 후, 양방향 순환 신경망의 입력값으로 전달해 맥락 벡터를 계산해 맥락 벡터와 실제 w_1의 단어 임베딩 벡터와의 차이가 줄어드는 방향으로 학습을 진행한다. 이후 w_2를 대상으로 하는 맥락 벡터를 계산할 경우 w_2를 심볼 $로 치환해 unknown 토큰을 가리키는 벡터로 변환하고, 나머지 단어들은 각 단어에 해당하는 벡터로 변환한다. 예시로 설명한 과정들을 말뭉치의 모든 문장들에 대해 동일하게 적용해 맥락 벡터 계산 모델을 위한 학습데이터로 사용한다.

5. Word Sense Disambiguation

본 연구에서 의미 판정 기준으로 사용하는 동형이의어의 의미별 예문들과 그 맥락 벡터들은 Table 1과 같이 문장에 부착된 동형이의어의 단어와 의미번호에 따른 집합을 구성해 의미 판정 기준으로 사용한다.

본 연구에서는 맥락 벡터를 계산해 동형이의어의 의미를 판정하는 연구들과[2][3][5] 마찬가지로, 질의 문장의 맥락 벡터와 동형이의어의 의미별 맥락 벡터간 코사인 유사도[17]를 사용해 동형이의어의 뜻을 결정하였으며, 식 5와 같다.

$$i = \arg max \cos(v_{query}, v_{h_{ij}}) \quad (5)$$

식 5에서 v_{query}는 질의 문장의 맥락 벡터를, i는 동형이의어의 의미 번호를, $v_{h_{ij}}$는 특정 동형이의어 h의 i번째 의미의 j번째 예제 문장의 맥락 벡터를 뜻한다. 즉, 질의 문장의 맥락 벡터 v_{query}와 같은 동형이의어를 가지는 의미별 예시 문장의 맥락 벡터 $v_{h_{ij}}$간 코사인 유사도 비교를 진행하고, v_{query}와 코사인 유사도 결과가 가장 큰 맥락벡터가 표현한 동형이의어의 의미별 예시문장의 의미 i를 질의문장에 포함된 동형이의어의 의미로 결정한다.

Table 1의 예문들로 중의성 해소 과정을 설명하면 다음과 같다. 먼저 동형이의어와 동형이의어의 의미번호에 따른 맥락 벡터 집합들인 $\{v_{의사_{11}}\}$, $\{v_{의사_{21}}\}$, $\{v_{눈_{11}}, v_{눈_{12}}\}$, $\{v_{눈_{21}}\}$, $\{v_{눈_{31}}\}$를 구성한다. 이후 중의성을 해소하고자 하는 동형이의어 '눈'이 포함된 질의 문장을 양방향 순환 신경망을 이용한 맥락 벡터 계산 모델로 맥락 벡터 v_{query}를 계산한다. 이후 동형이의어 '눈'에 대한 의미별 예문 집합인 $\{v_{눈_{11}}, v_{눈_{12}}\}$, $\{v_{눈_{21}}\}$, $\{v_{눈_{31}}\}$에 포함된 맥락 벡터들과 v_{query}의 맥락 벡터간 코사인 유사도 비교를 진행한다. 계산 결과 v_{query}와 맥락 벡터간 코사인 유사도가 가장 높은 $v_{눈_{11}}$가 가리키는 동형이의어의 '눈'의 의미번호 1을 질의 문장에서 사용된 '눈'의 의미로 반환한다.

IV. Experiments and Analysis

1. Experiment Environment

본 연구의 실험에서는 단어 임베딩 벡터를 계산하고 양방향 순환 신경망을 사용한 맥락 벡터 계산 모델을 학습하기 위해 세 종류의 말뭉치를 사용하였다. 먼저, 국립 국어원에서 제작해 현재 한국어 자연어 처리 연구에 가장 널리 쓰이는 1,000만 어절 규모의 세종 원시 말뭉치[18]를 사용하였다. 세종 원시 말뭉치는 국가 기관에서 제작한 범용 말뭉치로서 사회 각 분야에

Table 1. Examples of Meaning Criteria and Query Sentences

Word	Meaning Number	Sentence Number	Sentence	Vector
의사	1	1	…… '밤에 먹는 사과 한 개는 '의사'를 멀리하게 한다' ……	$v_{의사_{11}}$
	2	1	그리고 반대 '의사'를 밝히거나 묻거나 하지도 못했다 .	$v_{의사_{21}}$
눈	1	1	삽사리는 온몸이 털로 덮여 있어 두 '눈'이 무엇을 보고 있는지 …….	$v_{눈_{11}}$
		2	구기자는 간과 '눈'에 이롭게 작용한다 …….	$v_{눈_{12}}$
	2	1	……물을 얻기가 어려우므로 물 대신 '눈'으로 목마름을 면한다.	$v_{눈_{21}}$
	3	1	…자루나 '눈'이 가는 체로 치면 녹두물이 나오는데…	$v_{눈_{31}}$
Query			그는 충혈된 '눈'으로 강아지를 노려보았다.	v_{query}

서 쓰이는 어휘를 포함하고 있다는 장점[19]이 있다. 하지만 세종 원시 말뭉치는 제작된 지 10년 이상 된 말뭉치로 신조어 정보가 부족하고, 단어 임베딩 벡터를 계산하기 위한 목적으로 쓰기에는 그 규모가 크지 않다는 문제가 있다[20]. 따라서 세종 말뭉치에 부족한 신조어 정보를 추가하고 학습 데이터의 양을 늘리기 위한 추가 말뭉치로 두 종류의 인터넷 사전 즉, 2,400만 어절 규모의 위키피디아 덤프파일[21]과 4,600만 어절 규모의 나무 위키 덤프파일을[22] 학습 말뭉치에 포함해 총 8,000만 어절 규모의 학습용 말뭉치를 단어 임베딩 벡터 계산과 맥락 벡터 계산 모델 학습에 사용하였다. 동형이의어의 의미별 예제 문장으로는 한국어 동형이의어 중의성 해소 모델 성능 평가용 말뭉치인 Senseval-2[23]에서 제공하는 동형이의어 10개에 대한 예문 1,159개를 사용하였고, 실험 문장으로는 Senseval-2에서 제공하는 동형이의어 10개에 대한 예문 428개를 사용하였다. 단어 임베딩 벡터를 구축하기 위한 라이브러리로 Gensim[16]을, 순환 신경망과 양방향 순환 신경망모델을 만들기 위한 딥러닝 라이브러리로 Keras[24]를 사용하였다. 실험을 진행한 컴퓨터에는 i7-6700K, 32GB 램에 GPU로 GTX1080 2개를 사용하였다. 이번 실험에서는 평가 기준으로 중의성 해소 결과의 정확도를 사용하였으며 식 6과 같다.

$$\text{정확도} = \frac{\text{모델이 정확히 뜻을 분별한 문장의 수}}{\text{전체 질의 문장의 수}} \quad (6)$$

본 연구에서 단어 임베딩 벡터, 양방향 순환 신경망을 구축하기 위해 사용한 파라미터 값은 Table. 2와 같다.

Table 2. Parameters Used for Experiments

Category	Value
Dimension of Word Embedding Vecotr	100
Dropout rate - Input Layer	0.1
Dropout rate - Inner Gate	0.3
Optimizer	SGD(Stochastic Gradient Descent)
Dimension of FNN's Perceptron	350
Dimension of Recurrent Unit's Inner Gate	300
Inner Activate Function	tanh

2. Experiments of Word Sense Disambiguation

본 연구에서 제안하는 양방향 순환 신경망을 사용한 동형이의어 중의성 해소 모델의 성능을 검증하기 위해 단어 출현 순서의 한쪽 방향 정보만 반영하는 순방향 순환 신경망 모델과 역방향 순환 신경망 모델과 제안한 양방향 순환 신경망 모델의 동형이의어 중의성 해소 정확도 비교를 진행하였다. 또한 한국어를 위한 동형이의어 중의성 해소에 적합한 recurrent unit을 선택하기 위해 세 종류의 동형이의어 중의성 해소 모델 별로 recurrent unit이 LSTM일 경우와 GRU인 경우에 대해 동형이의어 중의성 해소 실험을 진행하였으며 그 결과는 Fig. 7과 같다.

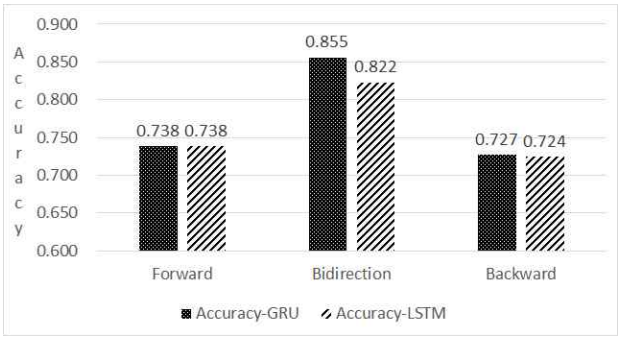


Fig. 7. Result of WSD Test

먼저, recurrent unit에 따른 동형이의어 중의성 해소 실험 결과 맥락 벡터 계산 모델에 사용한 순환 신경망의 종류와 상관없이 LSTM보다 GRU의 성능이 유사하거나(순방향 혹은 역방향 순환 신경망), 0.033만큼 높은 것을(양방향 순환 신경망) 확인 할 수 있었다. 두 recurrent unit에 따라 큰 성능 차이가 있는 것은 아니지만, recurrent unit으로 GRU로 사용하는 모델이 LSTM을 사용하는 모델보다 더 적은 파라메터를 가지고 있다. 따라서 GRU를 사용하는 모델이 LSTM을 사용하는 모델보다 학습시간에 이점이 있다. 유사한 정확도를 보이면서도 학습시간에 더 적은 시간이 소요되므로, 한국어에 대한 동형이의어 중의성 해소 실험의 순환 신경망 모델을 위한 recurrent unit으로는 GRU가 적합한 것으로 나타났다.

또한 recurrent unit으로 GRU를 적용한 양방향 순환 신경망 모델은 동형이의어 중의성 해소의 정확도가 0.855로 이는 순방향 모델보다는 0.117, 역방향 모델보다는 0.128만큼 높은 정확도를 기록하는 것으로 나타났다. 비록 양방향 순환 신경망 모델이 더 많은 내부 파라메터를 가지고 있어 순방향 혹은 역방향 순환 신경망 모델보다 모델을 학습하는데 더 많은 시간을 필요로 하지만, 더 정확하게 중의성을 해소함을 알 수 있다.

V. Conclusions

본 연구에서는 기존 순환 신경망을 이용한 동형이의어 중의성 해소 모델이 단어순서의 순방향 정보만 사용해 뒤에 존재하는 단어가 앞에 등장하는 단어의 뜻에 영향을 미칠 수 있는 자연어의 특징을 제대로 반영하지 못하는 문제점을 해결하기 위해 입력의 순방향 정보와 역방향 정보를 반영 할 수 있는 양방향 순환 신경망을 맥락 벡터의 계산에 사용한 동형이의어 중의성 해소 모델을 제안하였다. 본 연구에서 제안하는 양방향 순환 신경망 모델을 사용한 동형이의어 중의성 해소 모델은 단어 출현 순서의 순방향 정보와 역방향 정보를 동시에 고려해 맥락 벡터를 계산함으로써 단어 출현 순서의 한쪽 방향 정보만 반영한 순환 신경망 모델보다 더 많은 정보를 맥락 벡터 계산 과정에 반영하였다. 본 연구에서는 실험을 통해 한국어를 위한 동형

이의어 중의성 해소에 최적화 된 순환 신경망 내부의 recurrent unit으로 중의성 해소 정확도가 유사하거나 조금 더 우수하게 나온 GRU를 선정하였다. 또한 단어 출현 순서의 순방향 정보와 역방향 정보를 동시에 반영하는 양방향 순환 신경망 모델이 동형이의어 중의성 해소의 정확도 면에서 단어 출현의 순서의 순방향 혹은 역방향 정보만 반영하는 순환 신경망 모델보다 0.117과 0.128 만큼 더 정확한 중의성 해소가 가능함을 실험을 통해 확인하였다. 이러한 실험 결과를 통해 단어 출현 순서의 양방향 정보를 반영하면 더 정확한 동형이의어 중의성 해소가 가능함을 보였다.

향후 연구로는 더 많은 학습 데이터로 맥락 벡터 계산 모델을 학습하고 더 다양한 동형이의어 실험 문장을 사용한 평가를 진행해 실제 어플리케이션에 적용할 수 있는 중의성 해소 모델을 고도화 할 예정이다. 또한 제안한 모델을 표절 탐색 시스템, 키워드 추출과 같은 다른 자연어 처리 연구에 적용해 중의성 해소에 따른 다른 자연어 처리 연구 결과의 성능 개선 여부를 확인 할 계획이다.

REFERENCES

[1] M. Sundermeyer, and R. Schlüter, “LSTM Neural Networks for Language Modeling” Interspeech, pp. 194-197, Sep. 2012.

[2] mykang, bgkim and jslee, “Word Sense Disambiguation using Word2Vec”, Proceeding of 27th Conference on Human & Cognitive Language Technology, pp. 81-84, Oct. 2015

[3] jcshin, and, cyock, “Homograph Word Sense Disambiguation using Korean Lexical Semantic Map(UWordMap) and Word-Embedding”, Proceeding of Korean Computer Congress 2016, pp. 702-704, Jun. 2016

[4] D. Yuan, J. Richardson, R. Doherty, C. Evans and E. Altendorf, “Word Sense Disambiguation with Neural Language Models”, arXivpreprint arXiv:1603.07012, 2016.

[5] jhmin, jwjeon, khsong, and yskim, “Study on Word Sense Disambiguation Using Recurrent Neural Network for Korean”, Proceeding of Winter Conference on Korean Association of Computer Education 2017, pp. 93-96, Jan. 2017.

[6] jsbae, and cklee, “End-to-end Learning of Korean Semantic Role Labeling Using Bidirectional LSTM CRF”, Proceeding of 42th Winter Conference on Korean Institute of Information Scientists and Engineers, pp. 566-568, Dec. 2015.

[7] C. Irsoy, and C. Cardie, “Opinion Mining with Deep Recurrent Neural Networks”, Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing, pp. 720-728, 2014.

[8] H. Sak, A. Senior, and F. Beaufays, “ Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition”, arXiv preprint arXiv:1402.1128, 2014.

[9] Junyoung Chung, C. Gulcehre, KyungHyun Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling”, arXiv preprint arXiv:1412.3555, 2014.

[10] shjung, “Inference of Context-Free Grammars using Binary Third-order Recurrent Neural Networks with Genetic Algorithms”, Journal of The Korea Society of Computer and Information, Vol. 17, No. 3, pp. 11-25, Mar. 2012

[11] hmkim, jmyoon, jhan, kmbae, and yjko, “Syllable-based Korean POS Tagging using POS Distribution and Bidirectional LSTM CRFs”, Proceeding of 28th Conference on Human & Cognitive Language Technology, pp. 3-8, Oct. 2016

[12] smhan, “Deep Learning Architectures and Applications”, Journal of Intelligence Information System, Vol. 22, No. 2, pp. 127-142, Jun. 2016

[13] Y. Bengio, P. Simard, and P. Frasconi, “Learning Long-Term Dependencies with Gradient Descent is Difficult”, IEEE Transactions on Neural Networks, Vol. 5, No. 2, Mar. 1994

[14] bmkang, “Text Context and Word Meaning: Latent Semantic Analysis”, Journal of the Linguistic Society of Korea, Vol. 68, pp. 3-34, Apr. 2014.

[15] Gensim, https://radimrehurek.com/gensim/models/word2vec.html

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space”. arXivpreprint arXiv: 1301.3781. 2013.

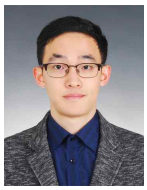
[17] mbchung, “Color matching application which can help color blind people based on smart phone”, Journal of The Korea Society of Computer and Information, Vol.20, No. 5, pp. 65-72, May. 2015

[18] National Institute of Korean Language, http://ithub.korean.go.kr,

[19] hgkim, mbkang, and jhhong, “21st Century Sejong Modern Korean Corpora: Results and Expectations”, Proceeding of 19th Conference on Human & Cognitive

Language Technology, pp. 311-316, Oct. 2007
[20] shchoi, jsseol and sglee, "On Word Embedding Models and Parameters Optimized for Korean", Proceeding of 28th Conference on Human & Cognitive Language Technology, pp. 252-256, Oct. 2016
[21] Korean Wikipedia, https://ko.wikipedia.org/
[22] Namuwiki, https://namu.wiki/
[23] SENSEAVAL-2, http://www.hipposmond.com/senseval2/
[24] Keras, http://keras.io

Authors



Jihong Min received the B.S., and M.S. degrees in Information and Communication Engineering from Inha University, Korea in 2015 and 2017, respectively. Mr. Min joined the student of the Department of Information Communication Engineering at Inha University, Incheon, Korea, in 2008. He is interested in text mining.

Joon-Woo Jeon received the B.S. degrees in Information and Communication Engineering from Inha University, Korea in 2016. Mr. Jeon is currently a graduated student in the Department of Information and Communication Engineering, Inha University. He is interested in text mining, clustering, and ontology.

Kwang-Ho Song received the B.S., and M.S. degrees in Information and Communication Engineering from Inha University, Korea in 2015 and 2017, respectively. Mr. Song is currently a Ph D student in the Department of Information and Communication Engineering, Inha University. He is interested in database, and text mining.

Yoo-Sung Kim received the B.S. in Computer Science from Inha University in 1986. He recevied M.S. and Ph.D. degrees in Computer Science from KAIST, Korea in 1988 and 1992, respectively .

Dr. Kim is currently a professor in the Department of Information and Communication Engineering, Inha University. He is interested in multimedia information mining, big data, and intelligent video surveillance system.