# 딥러닝을 이용한 한국어 어의 중의성 해소

김홍진○, 김학수
강원대학교 컴퓨터정보통신공학과
jin3430@gmail.com, nlpdrkim@kangwon.ac.kr

# A Word Sense Disambiguation for Korean Language Using Deep Learning

Hong-Jin Kim○, Hark-Soo Kim
Kangwon National University Department of Computer and Communications Engineering

**요 약**

어의 중의성 문제는 자연어 분석 과정에서 공통적으로 발생하는 문제로 한 가지의 단어 표현이 여러 의미로 해석될 수 있기 때문에 발생한다. 이를 해결하기 위한 어의 중의성 해소는 입력 문장 중 여러 개의 의미로 해석될 수 있는 단어가 현재 문맥에서 어떤 의미로 사용되었는지 분류하는 기술이다. 어의 중의성 해소는 입력 문장의 의미를 명확하게 해주어 정보검색의 성능을 향상시키는데 중요한 역할을 한다. 본 논문에서는 딥러닝을 이용하여 어의 중의성 해소를 수행하며 기존 모델의 단점을 극복하여 입력 문장에서 중의적 단어를 판별하는 작업과 그 단어의 의미를 분류하는 작업을 동시에 수행하는 모델을 제안한다.

주제어: 어의 중의성 해소, 딥러닝, Stacked BiGRU-CRFs

## 1. 서론

의미적 중의성 문제는 자연어 분석 과정에서 공통적으로 발생하는 문제로 한 가지의 단어 표현이 여러 의미로 해석될 수 있다는 점에서 발생한다. 이러한 문제를 해결하기 위한 어의 중의성 해소(Word Sense Disambiguation)는 입력 문장 중 여러 개의 의미로 해석될 수 있는 단어가 현재 문맥에서 어떤 의미로 사용되었는지 분류하는 기술이다. 어의 중의성 해소는 입력 문장에서 단어가 가질 수 있는 중의적인 의미를 제거하고 의미를 명확하게 해주어 정보 검색이나 정보 추출의 성능 향상에 매우 중요한 역할을 한다[1]. 본 논문에서는 자연어 처리 분야에서 좋은 성능을 보인 딥러닝(Deep Learning)을 이용하여 어의 중의성 해소를 수행한다. 대부분의 딥러닝을 이용한 자연어 처리 연구에서는 입력을 형태소 단위나 음절 단위로 받는다. 형태소 단위를 사용할 경우 신조어와 같은 미등록어의 형태소 분석 오류가 전파된다는 문제가 있고, 음절 단위의 경우 단어가 가지는 의미가 충분히 반영되지 않는 문제가 있다. 본 논문에서는 형태소 분석의 오류 전파 문제를 보완하기 위해 음절 단위로 어의 중의성 해소를 수행하며 음절 품사 임베딩(Embedding)과 중의성을 갖는 단어들의 사전 자질을 사용하는 어의 중의성 해소 모델을 제안한다.

## 2. 관련 연구

어의 중의성 해소를 수행하는 연구는 크게 딥러닝 기법 중 하나인 순환신경망(Recurrent Neural Network)을 이용한 중의성 해소연구[2-6]와 비지도 학습에 속하는 그래프 기반(Graph Based) 중의성 해소 연구[7-8]가 있다. 이 중 순환신경망 기반의 연구들이 자연어 처리 분야에서 좋은 성능을 보이고 있으며, 특히 양방향(Bidirectional) 순환신경망과 CRFs(Conditional Random Fields)를 결합한 BiGRU-CRFs(Bidirectional Gated Recurrent Unit CRFs)[9]가 순차적 레이블링(Sequence Labeling)으로 접근 가능한 개체명 인식 등에 좋은 성능을 보이고 있다. [5]는 GRU 기반의 양방향 순환신경망과 멀티 헤드 어텐션(Multi-head Attention)을 이용하여 어의 중의성 해소를 수행한다. 그러나 [5]에서는 입력 문장에서 중의적 의미를 가진 단어가 2개 이상이면 한 번에 모든 단어의 중의성을 해소 하지 못하고, 같은 문장에 대해 중의적 의미를 가진 단어의 개수만큼 중의성 해소 작업을 반복해야한다. 본 논문에서는 BiGRU-CRFs에 기반하여 음절 단위의 어의 중의성 해소를 순차적 레이블링 방법으로 수행하며 입력 문장에서 중의적 의미를 가진 단어의 개수에 상관없이 중의적 단어를 탐색하는 작업과 그 단어의 의미를 분류하는 작업을 모두 수행하는 어의 중의성 해소 모델을 제안한다.

## 3. 음절 단위 어의 중의성 해소 모델

### 3.1 Stacked BiGRU-CRFs

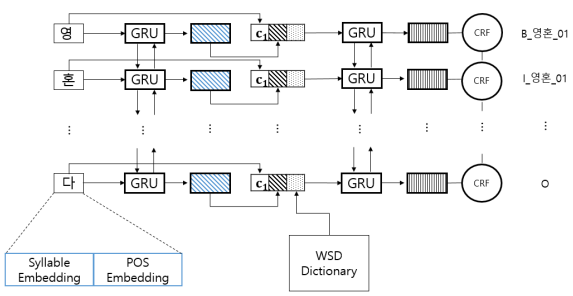


그림 1 음절 단위 어의 중의성 해소 모델 구조도

그림 1은 음절 단위 어의 중의성 해소 모델의 전체 구조도이며 2개의 계층(Layer)으로 구성되어 있다. 첫 번째 계층은 정방향(Forward)과 역방향(Backward) GRU로 구성된 양방향 GRU이다. 첫 번째 계층에서는 각 음절의 임베딩 벡터, 품사 임베딩 벡터, 그리고 중의적 단어 음절 확률 분포 벡터를 연결(Concatenation)하여 입력으로 사용하여 양방향 음절 정보를 담은 벡터를 출력한다. 두 번째 계층은 양방향 GRU-CRFs이다. 두 번째 계층에서는 첫 번째 계층에서 사용한 입력 벡터, 첫 번째 계층의 출력, 그리고 중의적 단어 사전 자질 벡터를 연결하여 입력으로 사용한다. CRFs에서는 두 번째 계층의 각 은닉 단계의 결과를 받아, 인접한 결과 간의 전이 확률을 고려하여 최종적으로 각 음절에 맞는 단어의 의미 태그를 부착한다.

### 3.2 임베딩

양방향 순환신경망 모델은 단어 임베딩에 따라 성능이 좌우된다. 음절 임베딩만으로는 단어를 표현하기에 부족하기 때문에 품사 임베딩을 이용하여 입력 표상을 확장하였다. 품사 중의성은 형태소 분석 단계에서 발생하며 입력된 단어를 형태소 단위로 나누고 복원하는 과정에서 그 형태소에 해당하는 품사가 다수일 때 발생한다[1]. 따라서 품사 중의성을 해소하기 위한 품사 정보가 중요하다. 즉, 입력 단어의 품사 정보를 자질로 주면 어의 중의성 해소 성능을 향상 시킬 수 있다. 본 논문에서는 무작위로 초기화(Random Initialize)한 음절 품사 임베딩을 사용한다. 음절이 형태소의 시작 음절이면 “B” 태그를 부착하였고, 그 외에는 “I” 태그를 부착하였다. “홍길동씨는”이라는 어절에 대응되는 품사 태그는 표 1과 같다.

표 1 음절 단위 품사 태그 예시

| | 홍 | 길 | 동 | 씨 | 는 |
|---|---|---|---|---|---|
| 품사 태그 | B_NNP | I_NNP | I_NNP | B_NNB | B_JX |

### 3.3 음절 단위 중의적 단어 의미 태그

표 2 음절 단위 중의적 단어 의미 태그 예시

| | 씨_01 | 씨_02 | ... | 영혼_01 | 영혼_02 |
|---|---|---|---|---|---|
| B | B_씨_01 | B_씨_02 | ... | B_영혼_01(영) | B_영혼_02(영) |
| I | - | - | ... | I_영혼_01(혼) | I_영혼_02(혼) |

표 2는 본 논문에서 사용한 음절 단위 중의적 단어 의미 태그 예시이다. 음절 단위 중의적 단어 의미 태그는 중의적 의미를 갖는 단어를 각 음절로 나누어 단어의 시작 부분이면 “B” 태그를 부착하였고, 그 외에는 “I” 태그를 부착하였다. 중의적 의미를 갖는 단어마다 가질 수 있는 의미의 개수가 다르기 때문에 단어 별 의미 태그 개수가 다르다. 또한 단어를 이루고 있는 음절의 개수도 다르기 때문에 단어별 음절 태그 개수도 다르다. 각 중의적 의미를 가지는 단어에서 생성되는 의미 태그의 개수는 (단어가 가지는 의미의 개수 * 단어를 이루고 있는 음절 개수)이다. 표 2에서와 같이 “영혼”이라는 단어가 가지는 의미가 2개이고, “영혼”을 이루는 음절이 “영”과 “혼” 2개이므로 생성되는 단어 의미 태그 개수는 4(2*2)개이다.

### 3.4 중의적 단어 사전 자질

본 논문에서는 어의 중의성 해소의 성능 향상을 위해 중의적 단어 사전 자질을 사용한다. 중의적 단어 사전은 중의성을 갖는 단어들을 사전 형식으로 구축한 것이다. 중의적 단어 사전 자질을 생성한 방법은 다음과 같다. 현재 입력 음절로 이루어지는 한 음절(Unigram) 단어나 현재 입력을 포함한 인접한 음절들과 조합으로 이루어지는 2 음절(Bigram) 단어부터 5 음절(Pentagram) 단어가 중의적 단어 사전에 존재하면 “1”, 그렇지 않으면 “0”으로 한다. “홍길동씨는”의 어절에서 현재 입력이 “씨”이라면 중의적 단어 사전 자질은 표 3과 같다. (“<SP>” 태그는 어절의 경계를 나타내는 띄어쓰기이다.)

표 3 중의적 단어 사전 자질 예시

| 단어 | 사전 포함 여부 |
|---|---|
| “씨” | 1 |
| “동씨” | 0 |
| “씨는” | 0 |
| “길동씨” | 0 |
| “동씨는” | 0 |
| “씨는<SP>” | 0 |
| “길동씨는” | 0 |
| “동씨는<SP>” | 0 |
| “길동씨는<SP>” | 0 |

## 4. 실험 및 결과

### 4.1 실험 환경

본 논문에서는 어의 중의성 해소 모델의 학습 및 평가를 위해 ‘UCorpus-HG 말뭉치’를 사용하였다. 데이터에 포함된 어절은 약 1,887만개이며 전체 데이터를 8:1:1로 나누어 각각 학습, 검증, 평가에 사용하였다. 데이터에서 단어의 의미 개수가 1 개만 등장한 단어는 제외하였다. 음절 임베딩은 50 차원, 음절 품사 임베딩은 16차원으로 실험하였다. 중의적 단어 사전은 학습데이터에서 나타난 중의적 단어들로 구축하였다.

### 4.2 실험 결과

표 4 어의 중의성 해소 성능 비교

| | precision | recall | F1 score |
|---|---|---|---|
| BiGRU-Attention[5] | 0.9652 | - | - |
| 제안 모델 | 0.9646 | 0.9349 | 0.9495 |

표 4는 기존 모델과 제안 모델의 평가 데이터 성능이다. 성능 척도는 micro average precision, recall, F1 score로 측정하였다. [5]는 GRU 기반의 양방향 순환신경망과 멀티헤드 어텐션을 이용한 모델이다. precision은 [5]가 더 높지만 입력 문장에서 중의적 단어를 탐색하는 작업이 없고, 한 문장에 중의성을 가진 단어가 2개 이상이면 그 개수만큼 같은 문장으로 중의성 해소 작업을 반복해야한다. 그러나 제안 모델은 중의적 단어를 탐색하는 작업과 단어의 의미를 분류하는 작업을 동시에 수행하므로 입력 문장에 중의성을 가진 단어의 개수에 상관없이 중의성 해소 작업을 한 번만 수행한다.

## 5. 결론

본 논문에서는 두 계층의 양방향 GRU를 구성하여 어의 중의성 해소를 수행하였다. 입력으로 음절 임베딩과 음절 품사 임베딩을 사용하여 미등록어가 입력되었을 때 전파되는 오류를 보완하였다. 또한 중의적 단어 사전 자질을 사용하여 입력 문장에서 중의적 단어를 탐색하는 작업과 그 단어의 의미를 분류하는 작업을 한 번에 수행할 수 있도록 하였다. 향후 연구에서는 중의적 단어의 음절 확률 분포를 생성한 후 CNN(Convolution Neural Network)을 이용하여 확률 분포에서 지역적인 정보를 추출하여 자질로 사용하는 연구를 할 계획이다.

## 감사의 글

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2016R1A2B4007732)

## 참고문헌

[1] 김민호, 권혁철, “한국어 어휘의미망의 의미 관계를 이용한 어의 중의성 해소”, 정보과학회논문지 : 소프트웨어 및 응용, 제38권, 제10호, pp. 554-564, 2011.

[2] Marvin, Rebecca, and P. Koehn. “Exploring word sense disambiguation abilities of neural machine translation systems (non-archival extended abstract)”, *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, Vol. 1, 2018.

[3] Ahmed, Mahtab, M. R. Samee and R. Mercer, “A novel neural sequence model with multiple attentions for word sense disambiguation”, *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2018.

[4] Amrami, Asaf, and Y. Goldberg, “Word Sense Induction with Neural biLM and Symmetric Patterns”, *arXiv preprint arXiv:1808.08518*, 2018.

[5] 김민호, 조상현, 권혁철, “양방향 순환신경망과 멀티 헤드 어텐션 기반 한국어 어의 중의성 해소 모형”, *한국정보과학회 2019 한국컴퓨터종합학술대회 논문집*, pp. 593-595, 2019.

[6] J. Min, J. Jeon, K. Song, and Y. Kim, “A Study on Word Sense Disambiguation Using Bidirectional Recurrent Neural Network for Korean Language” *한국컴퓨터정보학회논문지*, 제22권, 제4호, pp. 41-49, 2017.

[7] Chaplot, D. Singh, and R, Salakhutdinov, “Knowledge-based word sense disambiguation using topic models”, *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] Edilson A. Correa Jr, A. Lopes, D. R. Amancio, “Word Sense Disambiguation: A Complex Network Approach”, *Information Sciences 442*, pp. 103-113, 2018.

[9] 김선우, 최성필, "Bidirectional GRU-CRF 기반의 한국어 개체명 인식을 위한 어휘 사전 자질 적용 네트워크 토폴로지 연구", *정보과학회 컴퓨팅의 실제 논문지*, 제25권, 제2호, pp. 99-105, 2019.