코어넷을 활용한 비지도 한국어 어의 중의성 해소

한기종○, 남상하, 김지성, 함영균, 최기선

한국과학기술원
{cumgi31, nam.sangha, jiseong, hahmyg, kschoi}@kaist.ac.kr

Unsupervised Korean Word Sense Disambiguation using CoreNet

Kijong Han○, Sangha Nam, Jiseong Kim, YoungGyun Hahm, Key-Sun Choi
KAIST

요 약

본 논문은 한국어 어휘 의미망인 코어넷(CoreNet)을 활용한 비지도학습 방식의 한국어 어의 중의성 해소(Word Sense Dsiambiguation)에 대한 연구이다. 어의 중의성 해소의 실질적인 응용을 위해서는 합리적인 수준으로 의미 후보를 나눌 필요성이 있다. 이를 위해 동형이의어와 코어넷의 개념체계를 활용하여 의미 후보를 나누어서 진행하였으며 이렇게 나눈 것이 실제 활용에서 의미가 있음을 실험을 통해 보였다. 접근 방식으로는 문맥 속에서 서로 영향을 미치는 어휘의 의미들을 동시에 고려하여 중의성 해소를 할 수 있도록 마코프랜덤필드와 의존구조 분석을 바탕으로 한 지식 기반 모델을 사용하였다. 이 과정에서도 코어넷의 개념체계를 활용하였다. 이 방식을 통해 임의의 모든 어휘에 대해 중의성 해소를 하도록 직접 구축한 데이터 셋에 대하여 80.9%의 정확도를 보였다.

주제어: 어의 중의성 해소, 코어넷, 마코프랜덤필드

1. 서론

어휘는 같은 형태에서도 다른 의미를 가질 수 있다. 예를 들면 '배' 라는 어휘는 '배를 타다' 와 '배를 먹다' 라는 두 문장에서 각각 교통수단 배와 과일 배의 의미로 사용된다. 어의 중의성 해소(Word Sense Disambiguation)는 이처럼 문맥 속에서 어휘가 사용된 의미를 파악하는 것이다. 어의 중의성 해소는 기계번역, 정보 추출 등 자연어처리의 여러 문제에서 활용할 수 있는 중요한 문제이다[1].

어의 중의성 해소에서 각 어휘의 의미를 선택하는 후보 대상으로 영어권에서는 어휘의미망인 Princeton WordNet(PWN)에 등재된 의미(Sense)를 기반으로 연구들이 진행되고 있다[1,2]. 본 연구에서는 한국어 어휘의미망인 코어넷(CoreNet)[3]에 등재된 의미를 기반으로 어의 중의성 해소를 진행한다. 코어넷에 대한 자세한 내용은 2장에서 다루었다.

PWN이나 코어넷은 의미가 세분화 된, 잘게 나눔(fine-grained)된 어휘의미망이다. 미묘한 차이를 가지는 의미 후보들이 있어서 사람도 특정 어휘가 어떤 의미로 사용된 것인지 선택하기 어려울 정도이다. 어의 중의성 해소의 실질적인 응용을 위해서는 합리적으로 의미 구분을 할 필요성이 있다[2]. 이에 영어에서는 반자동으로 PWN의 의미를 군집화하여 크게 나눔(coarse-grained)된 어의 중의성 해소를 진행한다[2]. 한국어에서는 동형이의어 수준에서 중의성 해소 연구가 진행된 바 있다[4].

어의 중의성 해소 접근 방식은 크게 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning) 방식이 있다. 지도학습은 의미가 태깅된 말뭉치를 학습하며 비교적 높은 성능을 보인다. 그러나 말뭉치를 구축하는 데에 많은 시간과 비용이 들며 학습 데이터가 없는 어휘에 대해서는 중의성 해소가 어렵다는 단점이 있다. 한국어에서는 의미가 태깅된 약 1000만 어절의 세종말뭉치 학습을 기반으로 하고 어휘의미망을 활용하여 96.5% 수준의 정확도를 나타낸 연구가 있다[4].

비지도 방식은 비교적 성능은 낮지만 학습 데이터가 없어도 가능하다. 이 방식에서는 어휘 의미망을 활용하는 지식기반의 알고리즘 연구가 성과를 내고 있다[1,5]. 이 연구들은 어휘의미망의 구조화된 정보를 활용해 그래프 형태의 의미 네트워크를 구성하여 가장 적합한 의미를 찾는 방식이다. 넓은 범위의 어휘를 커버하고 좋은 성능을 내고 있다[6]. 대표적인 최신 연구로 마코프랜덤필드(Markov Random Field) 모델을 기반으로 한 연구가 있다[1]. 문장을 형태소 분석과 의존구조 분석(Dependency parsing)을 통해서 마코프랜덤필드 모델로 구축한 뒤 해당 모델에 최대 사후 확률 추론(Maximum a posteriori inference)을 통해 문장 속의 모든 중의성 후보 대상 어휘의 의미를 찾는 방식이다. 이 모델을 통해 문장 속에서 서로 영향을 주는 어휘의 의미들을 동시에 고려하여 어의 중의성 해소를 할 수 있다.

본 논문은 한국어 어휘 의미망인 코어넷을 활용한 비지도학습 방식의 한국어 어의 중의성 해소에 대한 연구이다. 의미 후보의 크게 나눔을 위해 동형이의어와 코어넷의 개념체계를 활용하였다. 이에 대한 자세한 설명은 2장에서 다루었다. 비지도학습 접근 방식으로는 마코프랜덤필드를 바탕으로 한 지식 기반 모델[1]을 코어넷의 개념체계를 활용하여 한국어에 적용하였다. 이에 대한 내용은 3장에 서술하였다. 4장에서는 이 방식을 평가하기 위해 직접 구축한 데이터에 대하여 설명하였다. 5장에서는 코어넷의 개념체계를 활용하여 의미 후보 구분을 한 것이 실제 응용에서 의미가 있음을 보이는 실험과 본 논문의 접근 방식의 성능에 대해 서술하였다. 6장에서는 결론과 향후 연구에 관해 서술하였다.

2. 코어넷과 개념체계를 통한 어의 중의성 해소

2.1 코어넷

코어넷[3]은 한국어의 명사, 형용사, 동사의 의미와 관계를 나타내며 중국어, 일본어, 영어의 개념과도 연결된 다국어 어휘 의미망이다. 한국어 기본 어휘 체계의 약 90% 이상을 커버한다고 알려져 있다[7]. 총 7만 3000여개의 의미가 등재되어 있으며 각 의미의 정의문, 예문과 같은 부가적인 자원이 포함되어 있다. 각 의미들은 한글 학회 우리말 큰 사전의 동형이의어와 다의어 번호를 통해 구분되어 있다.

2.2 코어넷의 개념체계

코어넷의 각 의미는 좀 더 보편적인 의미를 가지는 코어넷의 개념체계와 연결되어있다. 이 개념체계는 일본 NTT '어휘의미속성체계'를 기반으로 한국어 특징에 맞게 227개를 확장하여 총 2,954개의 개념으로 이루어져 있다. 각 개념이 상/하위 관계를 맺고 있는 최대 깊이 12인 트리 형태의 계층적인 구조이다. 또한 각 개념은 동사, 명사, 형용사를 아우르는 품사 독립적인 체계이다. 이 개념을 통해서 중국어 일본어와 연결되어 있고, 영어 WordNet의 의미와도 연결되어 있다. 그 예시는 그림 1과 같다. 깊이 7인 '경쟁'이란 개념이 있고, 이 개념 아래에 동사인 '겨루다_0_1'과 명사인 '경기_12_1', '결승전_0_0' 등의 의미가 연결된 식이다. 각 의미에서 어휘 뒤에 첫번째 숫자는 동형이의어 번호(vocnum)를 나타내고 두번째 숫자는 다의어 번호(semnum)를 나타낸다.

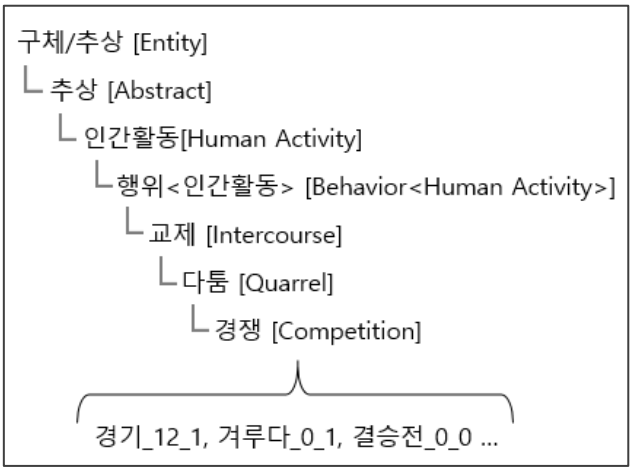

그림 1 코어넷의 개념체계 및 의미

2.3 어의 중의성 해소에 코어넷 개념체계 활용

어의 중의성 해소의 실질적인 응용을 위해서는 합리적인 수준에서 의미 구분을 할 필요성이 있다[2]. 본 연구에서는 분류 대상을 다른 한국어 연구처럼 동형이의어를 사용하되, 코어넷의 개념체계를 활용하였다. 코어넷 상에서 일반적으로 동형이의어 번호가 다른 의미들은 대부분 서로 다른 개념들에 연결되어 있다. 그러나 몇 가지 예외가 있다. '사과_1_0 : 사과참외의 준말', '사과_3_0 : 사과나무의 열매'는 서로 다른 동형이의어 번호를 가지지만 둘 다 '과일'이라는 개념에 연결되어 있다. 이런 경우 같은 후보로 군집화하여 진행하였다. 이렇게 같은 개념에 연결된 의미는 같은 후보로 판단하여, 후보 판별 시 개념을 활용할 수 있도록 하였다. 즉, '사과'라는 어휘에 대해서 표 1과 같은 식으로 의미 후보가 분류가 되고 각 후보에 대해서 적합성을 판단 할 때, '사죄', '과일' 등의 코어넷 개념체계를 활용할 수 있다.

이 결과 표 2와 같이 후보 대상 모호성이 평균 2.94개에서 2.66개로 줄어들었다. 모호성이란 전체 데이터의 의미 후보 개수의 총합을 전체 데이터 개수로 나눈 것이다. 4장에서 서술된 우리가 구축한 데이터에 대하여 측정하였다. 또한 이렇게 후보를 나는 것이 실제 활용에서 의미가 있음을 5장의 실험결과에서 보였다.

표 1 동형이의어와 코어넷 개념체계를 사용했을 때 '사과'의 의미 후보 분류 예시

후보	개념	(동형이의어, 다의어) 번호	정의문
0	사죄	(8,0)	- 잘못을 인정하고 용서를 빎
1	학문분야/학과	(6,1) (6,2)	- 도를 닦는 네 가지 과정 - 유학의 네 가지 학과
2	과일	(1,0) (3,0)	- '사과참외'의 준말 - 사과나무의 열매

표 2 후보 선택 대상별 선택 모호성

분류 대상	다의어	동형이의어	동형이의어+개념
모호성	5.00	2.94	2.66

3. 접근 방식

본 논문의 주 접근 방식은 3.2장에서 서술한 마코프랜덤필드 기반 방식이다. 다른 접근 방식으로 3.1장에 서술한 TF-IDF 벡터 유사도 방식을 활용하였다. 이 방식은 마코프랜덤필드 기반 방식의 모델을 구현하는 데 필요한 개념에 대한 빈도수 값을 구할 때 활용하였다. 이에 대한 자세한 내용은 3.2장에 서술하였다. 각 접근방식의 상세내용은 다음과 같다.

3.1 TF-IDF 벡터 유사도

각각의 후보 의미와 연결된 개념 아래에 있는 모든 의미의 정의문+예문의 TF-IDF 벡터와, 어의 중의성 해소를 하고자 하는 어휘가 포함된 문장의 TF-IDF 벡터의 코사인 유사도(Cosine Similarity)를 측정한다. 이때 가장 큰 값을 가지는 후보를 선택한다. 의미 후보 중 여러 개의 개념이 연결된 것이 있으면 해당 개념 중 하나만 맞추면 맞는 것으로 하였다.

예를들면 다음과 같다. '결승전 [경기]에서 연장전 끝에 한화가 이겼다.'라는 문장에서 '경기'라는 어휘의 중의성 해소를 할 때 '체육 운동으로 승부를 겨룸' 등의 뜻을 가지는 의미 후보가 있다. 이 의미 후보는 코어넷에서 그림 1과 같이 '경쟁'이라는 개념과 연결되어 있다. '경쟁' 개념 하위에 있는 '경기_12_1' 뿐만 아니라 '겨루다_0_1', '결승전_0_0' 등의 의미의 정

의문과 예문까지 활용하여 TF-IDF 벡터를 생성하고 입력 문장과 비교하는 방식이다.

TF-IDF 벡터의 차원은 전체 문서 집합에 나타난 서로 다른 어휘의 개수와 같고 벡터의 원소는 각각의 어휘에 대한 값을 나타낸다. 그 값은 다음과 같이 표시된다.

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times idf(t, D)$$

D는 전체 문서 집합을 나타내고 본 연구에서는 한글학회 우리말 큰 사전에 등재된 각각의 의미 하나의 정의문과 예문을 문서 하나로 설정하였다. 단어 빈도 $\text{tf}(t, d)$는 어휘 t가 특정 문서 또는 문장 d에 나타난 회수이다. 역문서 빈도 $idf(t, D)$는 문서 집합 D 에 어휘 t가 출현한 문서의 수의 역수 값에 로그를 취한 값이다. 두 문장 TF-IDF 벡터의 코사인 유사도는 같은 어휘가 서로 많이 등장할수록 높아진다는 원리를 활용한 것이다.

3.2 마코프랜덤필드 기반 방식

마코프랜덤필드(Markov Random Field)란 확률변수들의 집합으로 이루어진 비방향성 그래프 모델이다. 그래프에서 정점은 각각의 확률변수를 나타낸다. 각 확률변수는 간선으로 연결된 다른 정점을 나타내는 확률변수에만 의존적이다. 이 모델은 자연어처리의 여러 문제를 해결하는 데 사용되고 있다[1,8]

본 논문에서는 어의 중의성 해소 문제를 마코프랜덤필드를 통해 접근한 [1] 연구를 코어넷의 개념체계를 활용하여 한국어에 적용하였다. 이 방식에서 문장 속의 중의성 후보 대상인 어휘를 해당 모델의 정점, 즉 확률변수로 택한다. 간선은 의존구조 분석결과 직접 연결된 두 어휘에만 생성한다. 이렇게 구성된 마코프랜덤필드에 최대 사후 확률 추론(Maximum a posterior inference)을 통해 결합 추론을 하여 문장 속에 모든 중의성 후보 대상 어휘의 적합한 의미를 결정한다. 자세한 원리와 방식은 다음과 같으며 [1]을 기반으로 한국어에 어떻게 적용하였는지를 설명하였다.

3.2.1 모델의 작동 원리

먼저 그래프 모델을 사용하여 문장 속의 모든 중의성 해소 대상 어휘에 대해 동시에 의미를 추론하는 이유는 다음과 같다. 어휘의 의미는 기본적으로 같은 문맥 속에 다른 어휘의 의미에 영향을 받는다.

'대상 수상'

이라는 문장에서 '대상'은 1) 최고 상, 2) 객관의 사물 등의 의미가 있고 '수상'은 a) 상을 받음, b) 내각의 우두머리 등의 의미가 있다. '대상' 이 1)의 의미로 쓰인 것을 알기 위해서는 '수상'이란 어휘만으로는 알 수 없고 '수상'이 a)의 의미로 쓰인 것을 알아야 파악할 수 있다. 따라서 문장 속 모든 어휘의 의미를 한 번에 결합 추론하는 그래프 모델이 사용되었다.

두 번째로, 문장 속에 모든 어휘가 서로 직접 영향을 미치지는 않는다. 그 예시는 그림 2와 같다. '배를 먹은 뒤 배가 아팠다' 라는 문장에서 첫 번째 '배' 라는 어휘에는 '먹다' 라는 어휘만 영향을 미치고 두번째 배는 '아프다' 라는 어휘만 영향을 미친다고 볼 수 있다. 영향을 미치는 어휘를 선택하는 방식에는 여러 가지가 있을 수 있지만 이 모델에서는 그림 2처럼 의존구조 분석 결과 직접 연결이 되어있는 어휘를 서로 영향을 미치

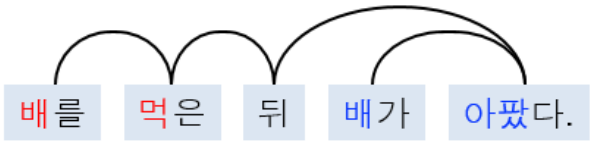

그림 2 문장의 의존구조 분석 결과

는 대상으로 선택하여 이 경우에만 마코프랜덤필드의 간선을 생성해 주었다.

3.2.2 모델 상세

먼저 문장이 주어지면 문장에서 모든 일반명사(NNG), 동사(VV), 형용사(VA) 중 코어넷에 등재된 어휘를 중의성 해소 대상 어휘로 선택한다. 이 과정에서 ETRI 언어분석기를 활용하였다. , 즉 마코프랜덤필드의 정점들을 $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ 로 표현한다. x_i 값들은 각각 m_i 개의 개념 값을 가질 수 있다. x_i가 가질 수 있는 후보 개념들을 $s_i^1, s_i^2, \ldots, s_i^{m_i}$ 로 표현한다. 예를 들면 그림 2의 문장 같은 경우 x_1= '배' 가 가질 수 있는 개념을 나타내는 확률변수, x_2= '먹다' 가 가질 수 있는 개념을 나타내는 확률변수를 의미한다. x_1 의 후보 s_1^1=탈 것(본체(물)), s_1^2=과일 등이 되는 식이다.

이 방식도 3.1 방식과 마찬가지로 의미 후보가 여러 개의 개념과 연결되어 있으면 그 중 하나만 맞추면 알맞게 중의성 해소를 한 것으로 보았다.

이 마코프랜덤필드 모델의 포텐셜 함수 값으로 정점의 포텐셜 함수 $\psi(x_i)$ 와 간선의 포텐셜 함수 $\psi(x_i, x_j)$가 있다.

$$\psi(x_i = s_i^a) \propto \log(frequency(s_i^a) + e)$$

먼저 어휘 자체가 어떤 개념을 가질지 나타내는 정점의 포텐셜 함수는 위와 같다. $frequency(s_i^a)$ 는 해당 개념이 얼마나 자주 나타나는지를 의미한다. 이 값은 3.1장에서 서술한 TF-ID 유사도 방식에서 역치값을 0.14로 설정하여 우리가 구축한 데이터에 대해 정밀도(Precision) 0.951, 재현율(Recall) 0.287인 성능으로 위키피디아 전문의 약 10%에 해당하는 170만여 어절에 대해서 측정하였다. 0인 경우에 대비해 e 를 더하고 최종 모델에서 정점의 포텐셜이 너무 큰 영향을 미치는 것을 방지하기 위하여 log 를 적용하였다.

$$\psi(x_i = s_i^a, x_j = s_j^b) \propto Relatedness(s_i^a, s_j^b)$$

연관이 있는 두 어휘가 동시에 특정 개념들을 가질 경우를 나타내는 간선 포텐셜 함수는 위와 같이 두 개념간의 관련성에 비례한다. 이 모델에서 간선은 ETRI 언어분석기를 활용해 의존구조 분석결과 직접적으로 연결된 경우만 간선을 생성해 주었다. 이 간선 집합을 E 라 할 때 $\{x_i, x_j\} \in E$ 이다. 관련성은 아래 두가지 방식으로 측정하여 각각 실험하였다.

$$(1)\ Relatedness(s_i^a, s_j^b) = 1/(shortestpath(s_i^a, s_j^b) + 1)$$
$$(2)\ Relatedness(s_i^a, s_j^b) = \log(frequency(s_i^a, s_j^b) + e)$$

(1) 의 경우는 코어넷의 개념들간의 관계를 활용하여 두 개념간의 최단 경로의 역수를 취하였다. 이 방식은 [1]에서 사용한 방식과 같다.(2)의 경우는 $frequency(s_i^a)$ 를 구할 때와 같은 방식으로 두 개념이 한 문장에서 동시에 등장하는 횟수를 세었다.

최종적으로 이 모델의 포텐셜 함수는 다음과 같다.

$$\psi(\mathrm{X}) = \prod_{x_i \in X} \psi(x_i) \prod_{\{x_i, x_j\} \in E} \psi(x_i, x_j)$$

S를 각각의 어휘에 대해서 선택된 개념들 이라고 할 때, 위와 같은 포텐셜 함수를 가지는 모델에 아래와 같은 최대 사후 확률 추론을 통해서 중의성 해소를 한다. 이를 구현하기 위하여 [9]을 사용하였다.

$$\arg\max_S \psi(\mathrm{X} = \mathrm{S})$$

3.2.3 모델의 작동 예시

그림 3에 나타나 있는 문장은 우리의 데이터 중 하나이다. '자고산 서쪽 [사면]은 낙동강의 동쪽 강변과 닿아 있다' 라는 문장에서 '사면' 에 대해 중의성 해소를 하는 경우이다. 이 문장에서 '사면' 은 의존구조 분석결과 '서쪽' 과 '닿다' 에 영향을 받는다. '서쪽' 과 '닿다' 는 해당 모델에서 각각 '방위' 와 '연결' 이라는 개념을 가지는 의미 후보로 선택되었다. 이 때 사면의 의미 후보 중 '비스듬한 바닥' 이란 뜻의 후보는 '면(面)' 이라는 개념과 연결되어있다. 면(面) 이란 개념이 서쪽-'방위' 와 닿다-'연결' 이라는 개념과 가지는 관련성이 다른 후보들에 비해 높기 때문에 이 의미 후보가 선택되었고 알맞게 선택된 경우이다. 빈도수와 비례하는 정점 포테셜 값은 '죄를 용서하여 형별을 면제함' 의 뜻을 가진 후보가 높지만 관련성까지 고려하여 알맞은 후보를 선택되었다. 그림 3의 관련성 포텐셜 함수 값은 MRF co-occur 기준으로 기재되었다.

4. 데이터

먼저 위키피디아 알찬글 문서에서 임의로 문장을 추출하였다. 각 문장에서 일반명사, 동사, 형용사 중 하나이면서 코어넷에 등재된 어휘 중 임의로 하나를 선택하는 식으로 구축하였다. 이렇게 선택된 하나의 문장과 문장 속의 어휘로 이루어진 데이터에 대해서 3명의 사람이 알맞은 의미 후보를 선택하였다. 이 방식으로 총 470개의 데이터가 구축되었으며 구체적인 통계는 표 3과 같다.

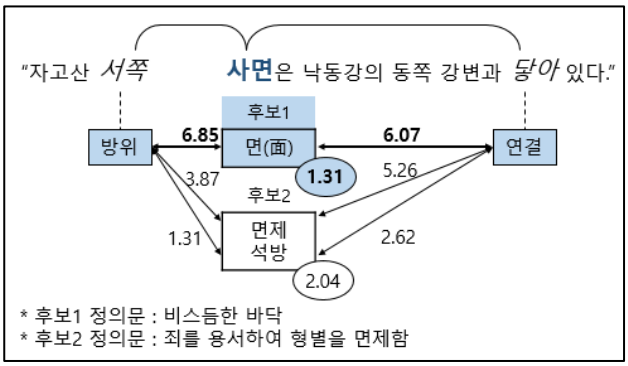

그림 3 마코프랜덤필드 모델의 작동 예시

구축한 데이터 중 중의성이 있는 어휘, 즉 후보가 2개 이상인 데이터는 총 215개 였다. 표 3에서 '# 어휘' 는 데이터에서 중의성 해소 대상이 되는 서로 다른 어휘의 개수를 의미한다. '# 의미' 는 데이터에서 어휘는 같아도 서로 다른 후보 의미로 쓰였으면 따로 세어서 계산한 것이다.

표 3 데이터 통계

	# 데이터	# 어휘	# 의미
모든 데이터	470	322	328
후보 2개 이상인 데이터	215	144	150

5. 실험 및 결과

5.1 성능

표 4 접근 방식별 정확도(Accuracy)

방식	RANDOM	TF-IDF	MRF SP	MRF co-occur
모든 데이터 성능	73.6	88.5	91.3	91.1
후보 2개 이상인 데이터 성능	42.2	74.9	80.9	80.5

우리가 구축한 데이터에 대하여 정확도를 측정한 결과는 위와 같다. 정확도는 다음과 같이 측정하였다.

$$\text{정확도} = (\text{시스템이 맞춘 개수})/(\text{데이터 개수}) * 100$$

RANDOM은 후보 중 임의로 하나를 선택하는 방식이며 5회 시행하여 평균을 내었다. TF-IDF는 3.1장에서 서술한 TF-IDF 벡터 유사도를 기반으로 한 방식이다. MRF SP와 MRF co-occur은 3.2장에서 서술한 마코프랜덤필드 기반의 방식이다. 차이는 두 개념간의 관련성 포텐셜함수 값을 다른 방식으로 구한 것인데, MRF SP는 3.2장의 (1)방식인 개념간의 최단 경로의 역수, MRF co-occur은 (2)방식인 두 개념의 동시 출현 빈도수로 설정한 경우이다.

마코프랜덤필드 기반 방식은 우리 데이터의 중의성이 있는 어휘에 대해서 80.9%의 정확도를 보였다. 임의로 선택한 베이스라인보다 많이 높은 정확도이고 정의문 및 예문의 유사도만을 비교한 TF-IDF 방식보다 약 6%의 성능향상을 보여주었다. 이를 통해 문장 속에 서로 영향받는 어휘의 의미들을 동시에 고려하여 추론하는 것이 의미가 있음을 보였다.

5.2 코어넷을 활용한 어의 중의성 해소 적용 결과

문장에서 관계추출(Relation Extraction)을 하는 문제에 본 연구의 어의 중의성 해소를 적용하였을 때 성능향상을 보였다. 이를 통해 코어넷의 개념체계를 활용하여 의미 구분을 한 후 어의 중의성 해소를 한 것이 실제 활용에서 의미가 있음을 알 수 있다.

구체적인 적용방식은 다음과 같다. Convolutional Neural Network 기반의 관계추출 기법[10]을 본 연구팀에서 한국어를 대상으로 구현한 모델에 어의 중의성 해소를 적용하였다. 관계추출 대상이 되는 문장을 토큰화한 후 각 토큰에 대한 임베딩 벡터가 이 모델의 입력으로 들어간다. 임베딩 벡터는 의미적 연관성이 높은 토큰들은 유사한 실수 벡터값으로 생성한다는 장점이 있다. 그런데 단어를 토큰화 할 때 형태소 단위로만 토큰화를 하면 중의성이 있는 단어들은 구분하지 못한다. 어의 중의성 해소를 통해 이 각각의 토큰들에 의미를 태깅하여 형태는 같아도 서로 다른 의미를 가지는 어휘가 임베딩 시 구분될 수 있도록 하였다.

표 5에서 형태소만으로 임베딩 하였을 때 '시장' 과 가까운 단어는 '물건을 사고 파는 장소' 라는 뜻과 관련된 단어들만 나오는 반면 의미를 태깅한 후에는 '시장' 의 두 가지 의미와 각각 관련된 단어들이 나오도록 임베딩이 된 것을 확인할 수 있다. 또한 표 6 에서 보듯이 의미 태깅 후에 임베딩을 한 것이 형태소 단위로만 토큰화 해서 임베딩 하는 것보다 관계 추출의 F1-score 성능이 약 7% 향상되었다.

표 5 임베딩 토큰 단위에 따른 유사한 토큰들

토큰 단위	토큰	유사한 토큰들
형태소	시장	투자, 유통, 수익, 수출, 자산, 대기업,
형태소 + 의미 태깅	시장-0 (물건을 사고 파는 장소)	시장, 산업-0, 업계-0, 경쟁력, 중소기업-0, 사업-4
	시장-1 (지방 자치 단체 장)	교육감-0, 기초자치단체장, 새누리당, 박순자, 고진화, 박영선_(1960년)

표 6 입력 임베딩 단위에 따른 관계추출 F1-score

임베딩 토큰 단위	형태소	형태소+의미 태깅
관계추출 F1-score	0.474	0.544

6. 결론 및 향후 연구

본 연구에서는 한국어 어휘 의미망인 코어넷의 개념체계를 활용하여 비지도학습 방식의 어의 중의성 해소를 진행하였다. 비지도학습이기 때문에 의미가 부착된 말뭉치 학습 데이터가 없어도 되며 마코프랜덤필드 모델과 의존구조 분석을 통해 문맥 속의 다른 어휘의 의미까지 고려하여 중의성 해소를 진행한다는 특징을 가지고 있다. 또한, 이러한 접근 방식과 의미 후보 분류에 있어서 보편적이며 계층적인 의미를 가지는 코어넷의 개념체계를 활용하였다는 데에서 의미가 있다. 이 방식을 통해 우리가 구축한 데이터에서 중의성이 있는 어휘에 대하여 80.9%의 정확도를 보였고, 코어넷의 개념체계를 활용하여 의미 후보 구분을 한 후 어의 중의성 해소를 하는 것이 실제 응용에서 의미가 있음을 보였다.

코어넷을 대상으로는 의미가 태깅된 말뭉치가 없어서 해당 모델의 함수 값 중 하나인 개념의 빈도수를 구할 때 TF-IDF 베이스라인 모델에서 재현율을 줄이고 정밀도를 높인 설정으로 구하였다. 정밀도를 0.951수준까지 높인 상태로 진행하긴 했지만 470개라는 비교적 적은 데이터에 대해서 측정된 것이고 좀 더 정확한 데이터를 활용할 필요성이 있다. 이와 관련해 한국어에는 약 1000만 어절 의미가 태깅된 세종말뭉치가 구축되어 있다. 이는 표준 국어 대사전의 의미 번호를 바탕으로 구성되어 있다. 반면 코어넷은 한글 학회 우리말 큰 사전의 의미 번호를 사용하여 의미가 구분되어 있다. 향후에는 코어넷에서 세종말뭉치를 잘 활용할 수 있는 방안을 연구하여 준지도학습(Semi-Supervised Learning) 등의 방향으로 성능향상을 이룰 수 있을 것으로 기대한다.

사사

이 논문은 2017년도 과학기술정보통신부의 재원으로 한국연구재단 바이오의료기술개발사업의 지원을 받아 수행된 연구임(NRF-2015M3A9A7029725)

참고문헌

[1] Chaplot, Devendra Singh, Pushpak Bhattacharyya, and Ashwin Paranjape. "Unsupervised Word Sense Disambiguation Using Markov Random Field and Dependency Parser." AAAI. 2015

[2] Navigli, Roberto, Kenneth C. Litkowski, and Orin Hargraves. "Semeval-2007 task 07: Coarse-grained english all-words task." Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics, 2007.

[3] Key-Sun Choi, et al. "Korean-Chinese-Japanese Multilingual Wordnet with Shared Semantic Hierarchy." LREC. 2004

[4] 신준철 and 옥철영. "한국어 어휘의미망 (UWordMap)을 이용한 동형이의어 분별 개선." 정보과학회논문지

제43권, 제1호, pp.71-79, 2016
[5] Agirre, Eneko, Oier Lopez de Lacalle, and Aitor Soroa. "Random walks for knowledge-based word sense disambiguation." Computational Linguistics 40.1, pp.57-84, 2014
[6] Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. "Embeddings for Word Sense Disambiguation: An Evaluation Study." ACL (1). 2016.
[7] 한국과학기술원 전문용어언어공학연구센터, “다국어 어휘의미망 제1권: 어휘의미망 구축론”, KAIST Press, 2005.
[8] Jung, Sung-Young, et al. "Markov random field based English part-of-speech tagging system." Proceedings of the 16th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1996.
[9] Ankan, Ankur, and Abinash Panda. "pgmpy: Probabilistic Graphical Models using Python." Proceedings of the 14th Python in Science Conference (SCIPY 2015). 2015.
[10] Zeng, D., et al. "Relation Classification via Convolutional Deep Neural Network." In Proceedings of COLING, pages 2335-2344. 2014.