Journal of Korean Institute of Intelligent Systems, Vol. 23, No. 4, August 2013, pp. 325-331
http://dx.doi.org/10.5391/JKIIS.2013.23.4.325

워드넷 기반의 단어 중의성 해소 프레임워크

A Framework for WordNet-based Word Sense Disambiguation

임초람* · 조세형*†
Chulan Ren, Sehyeong Cho†

*명지대학교 컴퓨터공학과
† Department of Computer Engineering, MyongJi University

요 약

본 연구에서는 단어의 의미 중의성을 해소하기 위한 방법을 제안하고 그 결과를 제시한다. 본 연구에서는 워드넷을 두가지 차원에서 활용하였는데, 하나는 사전으로서의 활용이며 다른 하나는 단어간의 개념 계층 구조를 가진 일종의 온톨로지로서 활용하였다. 이 중의성 해소 방식의 장점은 첫째 매우 단순하다는데 있다. 둘째로는 코퍼스를 활용하는 지식 기반/통계 기반 방식이 아니기 때문에 의미 태그 부착된 코퍼스의 부족으로 인한 문제가 발생하지 않는다는 것이다. 현재는 워드넷 온톨로지 중에서 개념 계층 구조, 즉 상위어-하위어 (hypernym-hyponym)의 관계만을 사용하였으나 향후 어렵지 않게 다른 관계들, 즉 유사어(synonym), 반의어(antonym), 부분어(meronym) 등의 관계를 활용하여 확장함으로써 성능의 향상을 기대할 수 있다.

키워드 : 단어 중의성, 시맨틱 웹, 워드넷, 온톨로지, 자연어

Abstract

This paper a framework and method for resolving word sense disambiguation and present the results. In this work, WordNet is used for two different purposes: one as a dictionary and the other as an ontology, containing the hierarchical structure, representing hypernym-hyponym relations. The advantage of this approach is twofold. First, it provides a very simple method that is easily implemented. Second, we do not suffer from the lack of large corpus data which would have been necessary in a statistical method. In the future this can be extended to incorporate other relations, such as synonyms, meronyms, and antonyms.

Key Words : Word Sense Disambiguation, Semantic Web, WordNet, Ontology, Natural Language Processing

1. 서 론

자연언어는 여러 가지 형태의 중의성을(ambiguity) 가지고 있다[1]. 그 중에서도 가장 기본적인 부분은 단어의 중의성이다(lexical ambiguity). 하나의 단어가 어떤 문맥에서 사용되느냐에 따라서 전혀 다른 뜻으로 사용되는 것은 매우 흔한 일이며 거의 모든 단어가 중의성을 가지고 있다. 컴퓨터에 의하여 자연어 처리를 할 때에 이 단어의 중의성 문제는 보기보다 매우 어려운 문제이다. 왜냐하면 완벽한 중의성의 해소를 위해서는 문장 전체, 나아가서는 글 전체의 문맥과 문장의 의미를 정확히 파악해야만 개별 단어의 뜻을 알 수 있기 때문이다. 그러나 컴퓨터에 의한 완전한 자연언어의 이해는 아직도 해결되지 않은 많은 난제들이 남아있는 상태이며 응용에 따라서는 이러한 자연어 처리 기법들을 총동원하기 어려운 경우가 많다. 예를 들어 온라인 외국어 사전을 생각해보자. 웹 서핑을 하던 중 모르는 단어가 등장하면 마우스를 올려놓으면 자동으로 영한사전에서 해당 단어를 찾아주는 소프트웨어를 흔히 만나볼 수 있다. 이러한 소프트웨어는 자동번역과는 또 다른 응용으로서 이 경우 영어의 태깅, 파싱, 의미 분석에 이르는 모든 절차를 거치도록 하는 것은 매우 무거운 일이 된다. 이 경우 사전에 정의된 몇 가지의 의미 중에서 가장 가능성이 많은 해석을 먼저 보여주는 것만으로도 충분히 의미가 있으며 이러한 경우는 신속하게 답을 주는 가벼운 프로그램이 선호되며 또한 영어 공부를 하는 학생의 입장에서는 오히려 해석된 전문을 보는 것 보다 적절한 단어의 정의와 예제를 보여주는 것이 더 도움이 된다.

본 논문에서는 워드넷(WordNet)을[2] 활용하여 주어진 단어의 여러 가지 의미 중에서 가장 가능성이 높은 것을 고를 수 있는 단순한 기법을 제시한다. 워드넷은 1985년 프린스턴 대학에서 개발이 시작되었으며 15만 단어, 11만5천개의 동의어 집합(synset)과 20만여 단어-의미 쌍으로 구성되어있다.

워드넷은 각 단어에 대하여 각 의미의 설명 및 예문과 아

접수일자: 2013년 5월 13일
심사(수정)일자: 2013년 6월 18일
게재확정일자 : 2013년 6월 18일
† Corresponding author

울러 빈도수를 표시하고 있어, 가장 빈도가 높은 의미를 먼저 참조하는 것 만으로도 약간의 도움을 받을 수는 있다. 그러나 이러한 빈도는 문맥에 무관하게 채집된 정보이기 때문에 특정한 문맥이 주어진 경우에는 이보다 더 많은 정보를 활용할 수 있다.

예를 들어 다음과 같은 문장을 생각해보자.

"A cat was chasing a mouse in my backyard."

사람들은 이 경우 "mouse"라는 단어가 동물인 쥐를 뜻하는 것을 쉽게 알 수 있다. 반면에

"People prefer a mouse over trackballs; a mouse is probably the best pointing device ever."

라는 문장에서라면 동일한 단어인 "mouse"가 컴퓨터의 주변 장치인 마우스를 뜻한다는 것을 알 수 있다. 이것은 상식 추론에 기반한다. 말하자면 고양이가 쥐를 좇아 다니는 것은 매우 자연스러운 일이지만 고양이가 컴퓨터 장치를 좇아 다닌다는 것은 상식적으로 생각하기가 (불가능하지는 않지만)어렵다. 또 둘째 문장의 경우 트랙볼 보다 쥐를 더 좋아한다는 것은 상식적으로 이해하기 어려운 말이지만 트랙볼 보다 마우스를 좋아한다는 말은 이해가 된다. 그러나 이러한 판단은 문장을 완전히 이해하지 않더라도 상당 부분 추론이 가능하다.

위의 문장에서 "mouse"를 제외한 명사들을 각각 추출하여보자.

A: {cat, backyard} B: {people, trackball, device}

A와 B를 각각 문맥이라고 부르기로 한다. 문장에 대한 구조적인 분석이나 의미 분석이 따르지 않더라도 위의 문맥만을 보더라도 비록 100% 확신할 수는 없지만 거의 직감적으로 문맥 A에서는 쥐를, 문맥 B에서는 컴퓨터 마우스를 지칭한다고 추정할 수 있다. 물론 인위적으로 반대의 예를 만들어 낼 수 없는 것은 아니다. 예를 들어:

"A cat was playing with my broken mouse in the backyard." 또는 "People were discussing trackball device, when the mouse crawled in."

같은 문장들이 그 예이다. 그러지만 대부분의 경우에는 자연스럽게 연관성 있는 단어들이 같은 문장에 출현할 가능성이 크다는 것은 매우 직관적인 일이다. 우리는 이러한 직관을 휴리스틱한 방법론으로 바꾸어 확률적으로 우수한 판단을 하는 단어 중의성 해소 알고리즘을 제시할 것이다.

이제 워드넷을 이용하여 어떻게 중의성을 해소할 수 있는지 살펴보기로 하자.

아래 내용은 워드넷에서 "mouse"를 검색한 결과이다.

(14)S: (n) mouse (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)
S: (n) shiner, black eye, mouse (a swollen bruise caused by a blow to the eye)
S: (n) mouse (person who is quiet or timid)
S: (n) mouse, computer mouse (a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad) "a mouse takes much more room than a trackball"

이 결과는 네 가지의 각기 다른 단어-의미 쌍을 보여주고 있다. 각 쌍에서는 이 단어에 대한 설명이 있는데 이 부분은 gloss(주해)라고 하며 괄호 안에 표시되어 있는 부분이다. 또한 많은 경우에 예문을 보여주고 있다. 이 부분은 단어의 의미를 추측할 수 있는 많은 힌트를 가지고 있다. 다음의 결과는 각 주해에서 추출한 명사들의 집합이다.

1: {rodent, rat, snout, ear, body, tail}
2: {bruise, blow, eye}
3: {person}
4: {computer, hand, device, cursor, screen, pad, bottom, ball surface, pad, room, trackball}

위의 1~4 집합과 분석 대상인 문장에서 추출한 명사의 집합을 비교하여보자.

A: {cat, backyard} B: {people, trackball, device}

집합 A 와 집합 1은 공통의 단어는 하나도 없다. 그러나 이들 중에는 어떤 의미에서 연관성이 큰 단어들의 쌍을 볼 수 있다. 예를 들어 cat-rodent 는 둘 다 동물이라는 특성이 있고 cat-ear는 전체-부분의 관계가 있다. 반면에 A와 4는 공통점이 거의 없다. 한 가지 있다면 backyard-room 정도인데 이들은 집의 일부라는 공통점은 있으나 사실 우연히 등장했을 뿐이다.

문제는 이러한 직관적인 데이터를 어떻게 형식화하여 프로그램에 활용할 수 있는가 하는 것이고 이것이 본 논문에서 다루게 될 주제가 될 것이다.

2장에서는 연관된 과거 연구 결과들을 살펴볼 것이다. 3장에서는 중의성 해소를 위한 기본적인 방법론을 제시함과 아울러 이를 위하여 필요한 단어간의 연관성을 정량적으로 측정할 수 있는 여러 가지 기준(metric)을 제시하여 비교 분석할 것이며 이를 활용한 알고리즘의 시행 결과를 제시할 것이다. 4장에서는 결과를 정리하고 앞으로의 연구 방향을 논의하도록 한다.

2. 관련 연구

단어 중의성 해소 문제(WSD: word sense disambiguation) AI-complete 문제로서 매우 어려운 문제이다[3]. 이 문제에 대한 접근법은 크게 세 가지가 있는데 이는 지도학습(supervised learning), 자율 학습(unsupervised learning), 그리고 지식에 기반한 방법(knowledge-based approach)이다[4]. 지도학습 방식은 단어들에 대한 의미 태그가 붙은 말뭉치를 사용하여 통계적인 정보를 추출하고 이를 기반으로 판단하는 방식이다. 이 방식은 방대한 양의 훈련 말뭉치를 필요로 하기 때문에 현실적인 제약이 뒤따른다[5,6,7,8,9,10]. 자율 학습은 의미 태그되어 있지 않은 말뭉치를 이용하여 학습을 하는데 이는 자료의 준비가 쉬운 반면에 정확도에 있어서 지도학습에 비해 좋은 성능을 내기가 어렵다[11][12]. 지식 기반 방법론은 사전이나 시소러스에 의존하며 말뭉치를 활용하지 않는다. 이러한 방법은 사전이라는 잘 정제된 양질의 정보를 사용한다는 장점이 있는 반면에 지도학습의 경우처럼 문맥에서 통계적인 정보를 끄집어내기는 어렵다는 단점이 있다. 이러한 이유로 성능 면에서는 지도학습 방식에 못 미치는 것으로 보고되어 있다[4]. 반면에 사전의 특성 때문에 활용 범위가 넓다는 장점을 가

지고 있다. 워드넷은 가장 많이 활용되고 있는 지식자원이다.

지식 기반 방식은 다시 세 가지 유형으로 나눌 수 있다. 첫째는 주해의 중첩(gloss overlap)을 이용하는 방법이고 [13, 14], 둘째는 선택 제약 방식(selectional restriction), 셋째는 구조적인 방식이다. 선택 제약이란[15] 단어의 역할에 있어서 특정 단어는 특정한 대상을 취한다는 데에 착안한 방법이다. 예를 들어 "dish"라는 단어의 용례 중에서 "The dish was delicious."라는 문장이 있다면 delicious라는 형용사는 음식을 주어로 한다는 선택 제약이 있으므로 이 dish는 음식을 뜻하는 dish일 것으로 추정하는 방법이다. 반면에 "the dish was broken"의 경우에는 break의 대상은 깨지기 쉬운 물체이므로 그릇을 뜻하는 것으로 추정한다.

2.1 중첩에 의한 방식

Lesk[13] 알고리즘은 단어 의미를 정의한 주해 중첩(gloss overlap)을 이용한다. 예를 들어 단어 w_1의 정의를 S_1이라 하고 w_2의 정의 S_2가 있을 때 두 의미의 연관도를 나타내는 점수 score는 다음과 같이 정의된다.

$$score_{Lesk}(S_1, S_2) = |gloss(S_1) \cap gloss(S_2)| \quad ...1)$$

여기서 gloss(s)는 의미 정의 부분에서 gloss에 해당하는 부분에 있는 단어의 집합이다. 특정 문맥상의 단어 w에 대한 가장 가능성 있는 정의를 골라내기 위해서는 문맥에 있는 단어와 사전의 의미 정의에 있는 단어의 중복을 계산한다.

$$score_{LeskV}(w, S) = |context(w) \cap gloss(S)| \quad ...2)$$

이 방법은 단순하다는 장점이 있지만 정확도가 매우 떨어진다. (50~70% 정도로 보고되고 있다.) 그 이유는 일반적으로 gloss가 매우 작기 때문에 중첩이 없을 가능성이 너무 높기 때문이다. 그러기 때문에 단어 하나가 있느냐 없느냐에 따라서 결과가 완전히 달라지게 된다.

Lesk 알고리즘을 확장한 Banerjee와 Pedersen [16]의 알고리즘은 워드넷의 단어 간 연관 관계를 이용하는 방식으로 확장을 하였다.

$$score_{exLesk}(s) = \sum_{s \equiv s'} |context(w) \cap gloss(s')| \cdots 3)$$

여기서 ≡ 기호는 두 센스가 워드넷의 관계를 가지고 있음을 의미한다. (예: synonym) 이 방법으로 많은 성능 향상을 가져왔으나 지식 기반 방법에 비하면 보잘것없는 성능을 보이고 있다.

구조적인 방법은 의미론적인 거리를 계산하는 방식이거나[17] 어휘 연쇄(lexical chain) 개념을 이용한다[18,19].

2.2 유사도의 활용

유사도를 활용하는 구조적인 방법은 유사도를 점수로 대응시키는 함수를 사용한다. 즉,

$$score : Senses_D \times Senses_D \rightarrow [0,1] \quad \cdots 4)$$

여기서 $Senses_D$란 사전에 있는 모든 단어 의미의 집합을 말한다. 주어진 텍스트 $T_1 = (w_1, w_2, ...w_n)$의 의미 다음과 같은 식을 만족하는 $\hat{S}$를 선택하는 것이다.

$$\hat{S} = \arg max_{S \in Sense(w_i)} \sum_{w_j \in T, i \neq j} \max_{S' \in Senses_D(w_j)} score(S, S') \quad ... 5)$$

계층 구조상의 위치를 이용하는 방법으로서 가장 단순한 것은Rada[20]의 유사도 척도로서 단순히 계층 구조상에서의 거리, 즉 number of hops 를 사용하였다.

$$score_{Rada}(S_w, S_{w'}) = d(S_w, S_{w'}) \quad \cdots 6)$$

이 방식은 지나치게 단순하여 실제 계층 구조에서 우리가 얻을 수 있는 정보를 충분히 활용하지 못하기 때문에 성능이 좋지 않다. Sussna[21]의 방식은 계층 구조에서 깊을수록 같은 거리에도 연관 관계가 깊다는 점에 착안하여 (즉, car와 limousine의 관계는 location과 entity의 관계보다 가깝다는 것) 다음과 같은 척도를 제안하였다.

$$score_{Sussna}(s_w, s_{w'}) = \frac{w_R(s_w - s_{w'}) + w_{R^{-1}}(s_w, s_{w'})}{2D} \quad ... 7)$$

여기서 R은 w를 중심으로 한 계층 relation이며 R^{-1}은 R의 역 관계이다. D는 전체 계층의 깊이이며 각 에지(edge)는 다음과 같이 weight를 준다.

$$w_R(S_w, S_{w'}) = \max_R - \frac{\max_R - \min_R}{n_R(S_w)} \quad \cdots 8)$$

여기서 $n_R(S_w)$는 S_w에서의 분기 수이며 $\max_R$, $\min_R$은 각각 이 분기에서 S_w의 하위개념들에게 할당하고자 하는 최대 및 최소값이다.

Qun Liu의 방식은 [22] Rada의 방식에서 정규화를 함으로써 유사도 값이 0 ~ 1 사이에 위치하도록 하였으며 그 식은 다음 9와 같다.

$$score_{QLiu}(w_1, w_2) = \frac{\lambda}{d(w_1, w_2) + \lambda} \quad \cdots 9)$$

Leacock와 Chodoro는[23] Rada의 거리 기반 방식에 기초하여 점수를 부여하는 다른 방법을 고안하였다. 그들은 경로 길이를 전체의 깊이D로 규모 조정을 하였다.

$$score_{Lch}(s_w, s_{w'}) = -\log \frac{d(s_w, s_{w'})}{2D} \cdots 10)$$

계층 구조에서 거리를 기반으로 하는 방법은 다소의 정보는 제공을 하지만 그 거리의 성질에 따라서 실제 유사성은 매우 다른 양상을 보인다는 문제가 있다. 예를 들면 계층 구조의 상위 층에서의 거리와 하위 층에서의 거리는 그 중요성이 다르다고 볼 수 있다. 예를 들어 car와 motor vehicle은 거리가 1이지만 매우 구체적인 개념들이다. 따라서 많은 정보를 가지고 있는 개념이며 유사성, 즉 공통적인 정보 혹은 속성이 많다고 볼 수 있다. 그러나 상위에 있는 entity와 physical entity는 매우 추상적인 개념이고 속성이 별로 없는 개념이다. 이 경우는 둘 사이에 유사성은 상대적으로 적다고 볼 수 있다. 이러한 개념은 Feng Li[24]의 접근법에서 찾아볼 수 있다. Li의 경우는 다음과 같은 점수 계산 방식을 사용하였다.

$$score_{FLi}(s_1,s_2)=\frac{\lambda \min(depth(s_1),depth(s_2))}{d(s_1,d_2)+\lambda \min(depth(s_1),depth(s_2))} \quad ...11)$$

Li의 방식은 두 개념 중에서 얕은 쪽의 깊이에 정규화 상수를 곱하여 분모와 분자에 더해줌으로써 깊이가 깊은 개념들에게 더 후한 점수를 부여하는 방법을 제공한다.

Dekang Lin[25]은 유사한 개념을 전혀 다른 차원에서 분석하였다. Lin은 세가지 차원에서 유사도를 분석할 것을 제안한다. 첫째는 공통점에서 기인하는 정보량을 반영해야 한다는 것이며 A와 B의 공통점 정보는 I(common(A,B))로 표기한다. 둘째는 두 개념의 차이점을 각 개념의 정보량의 합에서 공통 정보량을 뺀 것으로 모델링 한다. 즉, I(description(A)+description(B)-common(A,B)) 가 되며 여기서 description(A)는 A라는 개념을 설명하기 위한 proposition의 합으로 본다. 마지막으로 유사도는 common(A,B)의 정보량과 description(A),description(B)의 함수가 되어야 한다는 개념이다. 이 Lin의 개념은 여러 가지로 분화될 수 있을 것으로 보이며 본 논문에서도 이러한 개념에 바탕을 두고 방법론을 개발하였다.

3. 제안된 중의성 해소 방식

개요에서 소개한 바와 같이 우리는 대상 단어를 둘러싼 문맥을 활용하여 중의성을 해소하고자 한다. 기본적으로 우리는 중의성 해소를 하기 위한 단어에 대해서 그 단어를 둘러싼 문맥과 주어진 단어에 대해 워드넷이 제공하는 각 의미의 정의, 즉 주해(gloss)와 예문을 비교하여 가장 연관성이 많은 의미를 선택하게 될 것이다.

워드넷은 디지털화된 사전으로서 단순히 사전으로서의 기능뿐 아니라 명사, 동사, 형용사 및 부사들이 유사어의 집합이라 할 수 있는 신셋(synset)으로 정리되어 있다. 하나의 신셋은 같은 의미를 가진 단어들의 집합으로서 그 집합 자체로서 하나의 의미를 제공한다. 예를 들면 car라는 단어는 다음과 같은 신셋으로서 나타내진다.

{ car, auto, automobile, machine, motorcar }

이렇게 유사어의 집합으로 나타내어 짐으로써 모호하지 않고 명확한 의미를 파악할 수 있다. 또한 나아가서 이 개념은 아래의 그림과 같이 상위 개념들을 가지고 있다.

{ car, auto, automobile, machine, motorcar }
=>motor vehicle, automotive vehicle
=>self-propelled vehicle
=>wheeled vehicle
=>vehicle
=>conveyance, transport
=>instrumentality, instrumentation
=>artifact, artifact
=>whole, unit
=>object, physical object
=>physical entity
=>entity

워드넷의 각 단어에 대한 의미 항목은 각각의 신셋을 형성하며 이들은 단어가 아닌 단어-의미의 쌍이기 때문에 여기에는 중의성이 없이 하나의 의미로 정의가 된다. 워드넷은 또한 반의어 관계(antonymy), 부분어 관계(meronymy), 관련어 관계(pertainymy), 포함어 관계(holonymy), 그리고 수반 관계(entailment)등을 제공한다[26]. 그러나 본 논문에서는 hypernym-hyponym의 관계, 즉 계층 구조만을 활용할 것이다.

앞서 본 바와 같이 단어의 중의성 해소 task는 주어진 텍스트 $T_1=(w_1,w_2,...w_n)$에서의 중의성 해소 대상 w_i에 대하여 다음과 같은 식을 만족하는 의미 $\hat{S}$를 선택하는 것으로 보았다.

$$\hat{S}= \arg max_{S\in Sense(w_i)} \sum_{w_j\in T, i\neq j} \max_{S'\in Senses_D(w_j)} score(S,S') \quad ... 5)$$

그러나 식 5에서와 같이 최대치의 합으로 단순 계산하는 것은 문제가 있다. 왜냐하면 우리의 경우 워드넷에 있는 주해와 예문으로 한정하여 찾기 때문에 어떤 의미 해설은 단어가 많고 어떤 것은 단어가 적다. 이 경우 단어가 많을수록 채택될 가능성이 많다는 문제점으로 불공평하게 된다. 이러한 이유로 우리는 식 5를 변경하여 5′으로 사용하기로 하였다.

$$\hat{S}= \arg max_{S\in Sense(w_i)} \frac{\sum_{w_j\in T, i\neq j} \max_{S'\in Senses_D(w_j)} score(S,S')}{Size_{gx}(Senses(w_i))} \quad \cdots 5')$$

여기서 size_gx는 해당 단어 의미의 gloss와 예제에 있는 명사의 개수이다.

유사성의 척도

앞서 언급한 바와 같이 여러 가지 유형의 유사도가 여러 연구에서 사용된 바 있다. 우리는 유사도가 단순히 근거 없는 휴리스틱에 기반하기 보다는 어떤 원칙에 근거한 척도가 되는 것이 바람직하다고 보았으며 Lin의 정보량 개념을 도입하였다.

관찰1. 개념 계층도 상에 있는 두 개념은 공통의 조상이 많으면 많을수록 유사도가 커질 것이다.

관찰 2. 개념 계층도 상에 있는 두 개념은 동일한 공통점이 있다면 두 개념이 가진 정보량의 합이 클수록 유사도는 작아질 것이다.

관찰 3. 개념 계층도에서 hypernym과 그의 hyponym 간에는 최소한의 정보량의 차이가 있다.

이러한 관찰로부터 우리는 다음과 같은 유사도 Sm을 정의하였다.

$$Sm(s_1,s_2)=\frac{I(s_1)\cap I(s_2)}{I(s_1)\cup I(s_2)} \quad \cdots 12)$$

여기서 관찰 3에 의해 계층 하나를 내려갈 때 늘어나는 정보량을 단위 정보량으로 단순화하여 보면 12는 13과 같은 식이 된다.

$$Sm(s_1,s_2)=\frac{OL(s_1,s_2)}{depth(s_1)+depth(s_2)+OL(s_1,s_2)} \quad \cdots 13)$$

여기서 OL은 overlap을 뜻하며 두 개념의 공통 조상의 개수이다.

아래의 그림을 보자. Entity로부터 아래로 내려갈수록 한 단계씩 더 구체적인 개념이 된다. 한 단계를 내려올 때마다 하나씩 정보가 더 추가된다고 볼 수가 있다. 그렇다면 location와 object 가 가진 총 정보량은 3이다. 그런데 공통의 정보는 entity가 가진 정보뿐이므로 1로 볼 수 있다. (depth(location)=1, depth(object)=1, OL(location, object)=1) 따라서 유사도는 1/3이 된다.

life와 organism의 경우 공통 정보는 3이며 전체 정보는 5가 된다. 따라서 유사도는 3/5가 된다. 주목할 것은 이 두 경우가 단순한 거리 기반에서는 동일한 거리 2로 계산되었다는 것이다.

여기서 한 가지 더 추가할 정보가 있다. 동일한 개념의 두 하위 개념 즉, 형제에 해당하는 경우와 하나의 개념과 그의 상위 개념, 즉 할아버지에 해당하는 개념의 경우를 비교해보자. 전자의 예는 life와 organism이고 후자의 경우는 life와 object이다. 형제에 해당하는 개념은 여러모로 연관도가 크다. 예를 들면 하나의 문맥에서 두 가지 종류가 대비되어 등장하는 예는 매우 흔하다. 예를 들면 "car and truck"이 그러하다. 다른 측면에서 보자면 이러한 경우는 하나의 정보가 다른 정보로 대치되는 경우가 많은 것이다. 그러므로 이러한 경우의 차별성을 두기 위해 lso 항을 추가하였다. Lso는 lowest super-ordinate의 뜻으로 다음과 같은 식을 가진다.

$$lso(s_1,s_2)=\begin{cases}\beta, one\ is\ the\ lso\ of\ the\ other\\ 0, otherwise\end{cases} \quad \dots\ 14)$$

식 13을 15로 대체한다.

$$Sm(s_1,s_2)=\frac{OL(s_1,s_2)}{depth(s_1)+depth(s_2)-OL(s_1,s_2)+lso(s_1,s_2)} \cdots 15)$$

워드넷의 계층 구조 중에서 일부를 발췌한 아래의 다이어그램에서 우리는 4가지의 유사성 척도에 의하여 4개의 개념에 대한 유사도를 계산하여 보았다.4 가지 모두 거리, 즉 최단거리 엣지의 수는 2로서 동일하다. 계산 방법에 따라 유사도는 각기 다르게 나오는 것을 볼 수 있다. 직관적인 판단으로는 우리는 표의 아래쪽으로 갈수록 유사도가 높게 나오기를 선호한다. 이 표는 우리의 직관과 일단 일치하는 것을 볼 수 있다.

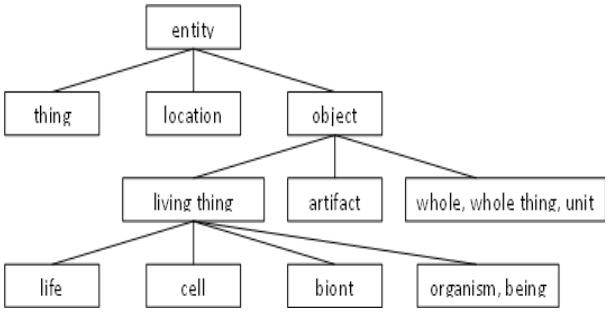


그림 1. 워드넷 계층도의 일부
Fig. 1. Part of WordNet hierarchy

표 1. 여러 가지 척도에 의한 개념간 유사도의 비교
Table 1. Similarities among concepts by each metric

D=2	Our method	Qun Liu's method	Feng Li's method	Dekang Lin's method
object & thing	0.33	0.44	0.44	0.33
object & life	0.44	0.44	0.44	0.50
living thing & artifact	0.50	0.44	0.62	0.50
life & cell	0.60	0.44	0.71	0.60

우리는 이 척도들을 이용하여 COCA[27] 말뭉치에서 최초 200개의 명사를 추출하여 실제 텍스트의 문맥에 있는 단어와 그 단어의 주해 및 예문에 있는 단어들 간의 연관성을 식 5'에 의해 계산하여 가장 유사도가 높은 의미를 선택하는 방식으로 실험을 하였다. 각 실험에서 신셋 간의 유사도를 정의하는 척도를 위 4가지 방법으로 각각 실행하여 본 결과는 아래의 표와 같았으며 제안된 유사도에 의한 방식이 가장 높은 정확도를 보이는 것을 확인하였다.

표 2. 각 유사도를 이용한 단어 의미 찾기의 정확도
Table 2. The accuracy of disambiguation by each similarity metrics

	Number of Correct guess	Accuracy of first candidate
Qun Liu's method	82	0.3886
Feng Li's method	144	0.6825
Dekang Lin's method	179	0.8483
Proposed method	195	0.9242

4. 결론 및 향후의 연구 방향

본 연구에서는 단어의 의미 중의성을 해소하기 위한 방법을 제안하였다. 연구 결과는 크게 두 가지로 요약할 수 있다. 그 하나는 신셋 간의 유사도를 측정하는 방식으로서 계층 구조를 활용하였으며 단순 거리 방식에서 탈피하여 개념이 가지고 있는 정보량을 활용함으로써 추상적인 개념간의 거리보다 구체적인 개념 사이의 거리를 더 가깝게 계산

할 수 있게 하였다. 다른 한가지의 결과는 별도의 말뭉치를 사용하지 않고 워드넷 상에 있는 주해와 예문만을 활용하여 유사도를 측정하게 하였다는 것이다.

이 방식의 장점은 워드넷에서 제공하는 데이터를 사용하였기 때문에 별도의 가공된 말뭉치를 사용하지 않는다는 것이다. 이러한 이유로 알고리즘이 매우 단순하고 프로그램이 가벼워 짐으로써 다양한 응용에 사용될 수 있다. 그 중 한 가지 응용은 영어를 모국어로 하지 않는 사람들이 사전을 검색할 때에 가장 가능성이 많은 의미를 먼저 선택하도록 순서를 결정지어 주게 된다. 여기서 워드넷을 두 가지 차원에서 활용하였는데, 하나는 사전으로서의 활용이며 다른 하나는 단어간의 개념 계층 구조를 가진 일종의 온톨로지로서 활용하였다. 현재는 워드넷 온톨로지 중에서 개념 계층 구조, 즉 상위어-하위어 (hypernym-hyponym)의 관계만을 사용하였으나 향후 어렵지 않게 다른 관계들, 즉 유사어(synonym), 반의어(antonym), 부분어(meronym) 등의 관계를 활용하여 확장함으로써 성능의 향상을 기대할 수 있을 것으로 보인다.

References

[1] Daniel Jurafsky and James H. Martin, *Speech and Language Processing,* 2nd edition, Pearson 2009

[2] Christiane Fellbaum(ed.), *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. 1998

[3] MALLERY, J. C. *Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers*. Ph.D. dissertation. MIT Political Science Department, Cambridge, MA. 1988.

[4] Roberto Navigli. "Word Sense Disambiguation: A Survey," *ACM Computing Surveys*, 41(2), 2009, pp. 1-69.

[5] A. Novischi, M. Srikanth, and A. Bennett, "Lcc-wsd: System description for English coarse grained all words task at semeval 2007," in *Proc. of the 4th International Workshop on Semantic Evaluations*, pp. 223-226, Prague, Czech Republic, 2007.

[6] M. Ciaramita and Y. Altun, "Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger," in *Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, pp. 594-602, 2006.

[7] L. M'arquez, G. Escudero, D. Martinez, and G. Rigau, "Supervised corpus-based methods for WSD," in *Word Sense Disambiguation: Algorithms and Applications*, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, pp. 167-216, 2007.

[8] R Mihalcea and E. Faruque, "Senseleamer: Minimally supervised word sense disambiguation for all words in open text," in *Proc. of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text* (Senseval-3), Barcelona, Spain, pp. 155-158, 2004.

[9] S. Tratz, A. Sanfilippo, M. Ggregory, A. Chappell, C. Posse, and P. Whitney, "PNNL: A supervised maximum entropy approach to word sense disambiguation," in P*roc. of the 4th International Workshop on Semantic Evaluations* (SemEval), Prague, Czech Republic, pp. 264-267, 2007.

[10] M'ARQUEZ, L., ESCUDERO, G., MART'INEZ, D., AND RIGAU, G., "Supervised corpus-based methods for WSD," in *Word Sense Disambiguation: Algorithms and Applications*, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 167-216. 2006.

[11] PEDERSEN, T. "Unsupervised corpus-based methods for WSD," in *Word Sense Disambiguation: Algorithms and Applications*, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 133-166. 2006.

[12] R Mihalcea, "Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling," in *Proc. Of HLT/EMNLP*, Vancouver, BC, Canada, pp. 411-418, 2005.

[13] LESK, M., "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proceedings of the 5th SIGDOC* (New York, NY). Pp.24-26. 1986.

[14] PEDERSEN, T., PATWARDHAN, S., AND MICHELIZZI, J. "WordNet::Similarity-measuring the relatedness of concepts," in *Proceedings of the 19th National Conference on Artificial Intelligence* (AAAI, San Jose, CA) pp.144-152. 2004.

[15] MCCARTHY, D. AND CARROLL, J. "Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences," *Computational Linguistics* 29-4, pp. 639-654. 2003.

[16] BANERJEE, S. AND PEDERSEN, T., "Extended gloss overlaps as a measure of semantic relatedness," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. 805-810. 2003.

[17] PEDERSEN, T., BANERJEE, S., AND PATWARDHAN, S., "Maximizing semantic relatedness to perform word sense disambiguation," *Res. rep. UMSI 2005/25*. University of Minnesota Supercomputing Institute, Minneapolis, MN. 2005.

[18] NAVIGLI, R, "Consistent validation of manual and automatic sense annotations with the aid of semantic graphs," *Computational Linguistics,* 32- 2, pp.273-281. 2006.

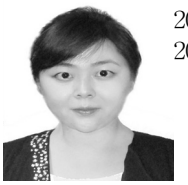
[19] NAVIGLI, R. "Experiments on the validation of sense annotations assisted by lexical chains," in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* 129-136. 2006.

[20] RADA, R., MILI, H., BICKNELL, E., AND

BLETTNER, M. "Development and application of a metric on semantic nets," *IEEE Trans. Syst. Man Cybernet*. 19, 1, 17-30. 1989.
[21] SUSSNA, M. "Word sense disambiguation for free-text indexing using a massive semantic network," in *Proceedings of the 2nd International Conference on Information and Knowledge Base Management*, 67-74., 1993
[22] Qun Liu, Sujian Li, "Word Similarity Computing Based on How-net," *Computational Linguistics and Chinese Language Processing*, Vol.7, No.2, pp.59-76. , August 2002
[23] LEACOCK, C., CHODOROW, M., AND MILLER, G., "Using corpus statistics and WordNet relations for sense identification," *Computational. Linguistics,* 24, 1, 147-166. 1998.
[24] Feng Li, Fang Li, "an new approach measuring semantic similarity in Hownet 2000," *Journal of Chinese Information Processing,* vol.21, No.3, May 2007.
[25] Dekang Lin, "An information-theoretic definition of similarity," in *Proceedings of ICML*, pages 296-304. 1998.
[26] Vaclav Snael, Pavel Moravec, Jaroslav Pokorny. "WordNet Ontology Based Model for Web Retrieval," *International Workshop on Challenges in Web Information Retrieval and Integration* (WIRI'05), 0-7695-2414-1/05.
[27] Brigham Young Universiy, *Corpus of Contemporary American English,* Available: http://www.americancorpus.org/, 2013 [Accessed August, 19, 2013]

저 자 소 개

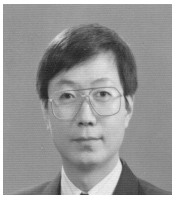
임초람(Chulan Ren)



2005년 : 중국 동북대학교 공학사
2008년: 명지대학교 공학석사
2008～현재 : 명지대학교 박사과정

관심분야 : Ontology, Natural Language Processing, Semantic Web
E-mail : renchulan@gmail.com

조세형(Sehyeong Cho)



1981년: 서울대학교 공학사
1983년: 서울대학교 이학사, 계산통계학
1992년: 펜실베니아 대학 이학박사
1984-2000: 한국전자통신연구원 책임연구원
2000년~현재:명지대학교
컴퓨터공학과교수

관심분야 : Ontology, Natural Language Processing,
Phone : +82-31-330-6779
E-mail : shcho@mju.ac.kr