정보량과 개념적 밀도를 이용한 단어 의미 중의성 해결

조미영*, 김판구**
*조선대학교 전자계산학과
**조선대학교 컴퓨터공학과
e-mail : irune@chosun.ac.kr

Word Sense Disambiguation using the Information Content and the Conceptual Density

Mi-Young Cho*, Pan-Koo Kim**
*Dept of Computer Science, Chosun University
**Dept of Computer Engineering, Chosun University

요 약

기존의 정보 검색은 단순 키워드 매칭에 의한 패턴 매칭으로 의미적 정보 검색에는 한계가 있다. 이를 해결하기 위한 많은 연구가 이루어졌으나 질의 혹은 문서에 중의적 의미를 가진 단어를 포함하고 있는 경우에 검색시 문제가 되었다. 이에 본 논문에서는 WordNet기반의 단어 빈도수를 고려한 정보량과 단어 영역내 존재하는 노드 수를 고려한 개념적 밀도를 이용한 WSD(Word Sense Disambiguation)를 제안한다. SemCor를 이용하여 테스트한 결과 두 요소를 결합한 방법에 의해 WSD가 약 20% 향상되었다.

1. 서론

기존의 정보 검색은 단순 키워드 매칭에 의한 패턴 매칭으로 검색 단어와 정확한 매칭에 의해 검색한다. 그러나 사용자는 이러한 단순 매칭에 의한 정보 검색이 아닌 의미적 정보 검색을 하고자 한다. 이때, 단어 의미 중의성 해결은 가장 중요한 문제로 이에 대한 많은 연구가 있어 왔다.

단어 의미 중의성 해결 즉 WSD(Word Sense Disambiguation)란 동음이의어 등 중의성을 가진 단어들의 의미를 명확히 하는 것으로, 예를 들어 'bank'라는 단어는 금융 기관이라는 의미와 'snow bank' 혹은 'river bank'의 의미로써 강둑의 의미를 가지고 있다. 정보 검색 시 사용자의 질의로 'bank'라는 단어가 들어왔을 경우 이를 처리하기 위한 WSD 절차가 필수적이다.

논문의 구성은 다음과 같다. 2장에서는 WordNet 및 WordNet 기반의 두 개념간 유사성 측정에 대한 일반적인 내용에 대해 기술한다. 3장에서는 정보량을 이용한 WSD 예제를 보여주고 문제점을 제시한다. 4장에서는 제기한 문제점 해결방안으로 정보량과 개념적 밀도를 이용한 WSD를 제안한다. 5장에서는 두 가지 방법으로 실험하고 이를 비교하며, 마지막 6장에서 결론을 맺는다.

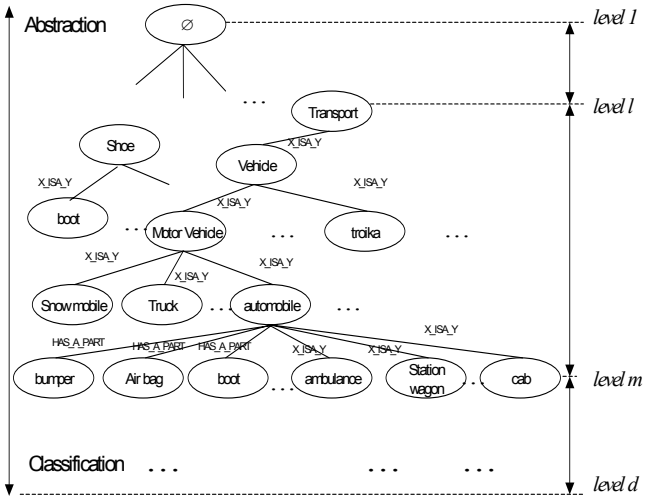
2. 관련연구

2.1 WordNet

Ontology의 일종으로 간주되기도 하는 WordNet은 인간의 어휘지식에 대한 심리언어학 연구의 성과를 토대로 1985년도 중반부터 프린스톤 대학 인지과학연구실이 구축해온 영어어휘 데이터베이스다. WordNet은 인간의 어휘지식을 모방한 만큼 다의성과 동의 관계를 이용하여 의미를 최대한 정확히 표현하고 있고 개념간의 관계 표현 등을 통해 개념을 계층적으로 표현하고 있다.

각 단어들은 synset이라는 동의어의 집합을 기준으로 각 집합 즉 노드들간의 관계를 이용하여 (그림

1)과 같이 정의하고 있다.



(그림 1) WordNet의 구조

2.2 단어간 유사성 측정

단어의 개념적 접근을 통한 WSD에 대해 논하기 전에 유사성 측정시 고려될 수 있는 특징들에 대해 간략히 살펴보겠다. 크게 3 가지로 분류되는 각각의 특징들의 정의와 유사성 측정시 미치는 영향은 다음과 같다.

1) 거리(path) - 두 개념간 최단거리로 작은 값이 반환될수록 더 가깝다.
2) 깊이(depth) - WordNet상에서 개념의 위치로 깊을수록 더 세부적이고 구체적인 개념이다. 즉, 측정하고자하는 개념을 포함한 노드의 깊이가 더 깊을수록 유사성 측정시 더 큰 값을 반환한다.
3) 정보량(information content) - 두 개념간 Least Common Subsumer(LCS)를 찾고 만약 LCS에 대해 다수의 노드가 존재한다면, 가장 큰 정보량을 선택한다.

위 3가지 특징을 요약하면, 거리가 짧을수록 깊이가 깊을수록 큰 정보량을 가지며, 큰 정보량을 가질수록 두 개념간의 유사성은 더 크다고 볼 수 있다. 다음 3장에서는 위의 요소를 고려한 단어간 유사성 측정을 통한 WSD를 기술하도록 하겠다.

3. 정보량을 이용한 WSD

Resnik에 의한 정보량은 큰 말뭉치에서 단어 의미들간의 의미적 연관성을 계산하기 위한 개념의 빈도수로부터 계산된다. Resnik이 정의한 WordNet을 기반으로 한 정보량은 다음과 같다[3].

$$P(c)=\frac{freq(c)}{N} \quad \text{----------} \quad (1)$$

먼저 수식 1에서 $P(c)$ 는 개념 c와 마주칠 확률로 N은 개념의 총 수를 의미하고 계층적 구조의 경우 $freq(c)$는 개념 c에 포함된 모든 하위 개념들의 합을 의미한다. 만약 c_1 IS-A c_2 라면 $p(c_1) \le p(c_2)$가 된다. 그러므로 WordNet에서 유일한 top node ϕ의 확률은 1이 된다.

정보 이론에 따르면 개념 c의 정보량은 $-\log P(c)$로 확률이 증가하면 정보량은 감소하므로 더 추상적인 상위 개념은 낮은 정보량을 가진다. 즉, ϕ의 정보량은 0이 된다.

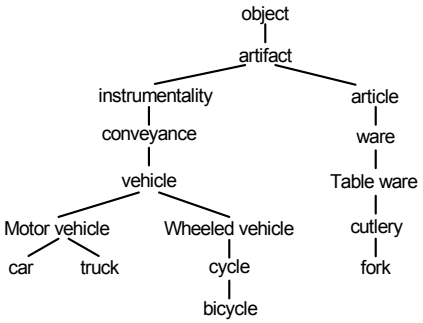
$$H(c)=-\log P(c) \quad \text{----------} \quad (2)$$

측정하고자 하는 두 개념이 공유한 정보는 WordNet에서 두 개념들을 포함하고 있는 개념의 정보량으로 표현할 수 있다. 따라서 두 개념간의 노드 기반 유사성 측정은 다음과 같이 나타낼 수 있다.

$$\mathrm{sim}(c_1, c_2)=\max_{c \in S(c_1, c_2)} [-\log p(c)] \quad (3)$$

수식 3에서 $S(c_1, c_2)$는 c_1과 c_2 두 개념을 포함하고 있는 상위 개념 c를 의미한다.

두 개념이 공유하고 있는 상위 개념이 많을수록 두 개념은 더 유사하다고 할 수 있다. 따라서 개념간 유사성은 c_1과 c_2 를 포함하는 상위 개념 중 가장 큰 H를 가지는 개념의 H이다.



(그림 2) WordNet상에서 car와 bicycle, fork간 상하위 관계

예를 들어 (그림 2)에서 car와 truck 그리고 car와 bicycle간의 유사성을 측정한다고 하면 car와 truck이 공유한 개념은 car와 bicycle이 공유한 개념보다 더 많으므로 더 유사하다고 할 수 있다. 그러므로 sim(car, truck)은 H(motor vehicle)이고 sim(car,

bicycle)은 H(vehicle)이다.

```
<instance id="plant.1000099" docsrc = "BNC/B0S">
<answer instance="plant.1000099" senseid="plant%factory"/>
<context>
(i) the acquisition of land (including buildings thereon), (ii) the acquisition of vehicles, <head>plant</head> , machinery, etc.,
</context>
</instance>
```

(그림 3) SemCor 포맷

정보량을 이용한 WSD를 실험하기 위하여 본 논문에서는 중의성을 띈 단어에 의미가 태깅되어 있는 말뭉치 데이터 집합인 SemCor를 사용하였다. (그림 3)은 'plant'에 대한 SemCor의 일부를 보여주고 있다. 실험시 문맥에서 명사만을 추출하여 중의성을 가진 단어와 정보량을 측정한다. 다음 <표 1>은 (그림 3)에서 plant의 의미 중의성 해결을 위해 문맥내 다른 명사들과 정보량을 측정한 결과이다.

<표 1> 정보량 측정 결과

Word pairs		The information content value
acquisition	plant	2.828
land	plant	3.056
machinery	plant	2.558
building	plant	3.826
vehicle	plant	2.558

<표 1>에서 보듯이 plant와 building간의 정보량이 3.826으로 가장 크며, trace결과 plant의 첫 번째 의미인 factory와 building의 첫 번째 의미 사이의 정보량이 가장 컸다. 이로써 의미적 중의성을 해결할 수 있다. 다음은 또 다른 예로 같은 방식으로 factory의 의미를 가지는 plant를 찾는다.

```
<instance id="plant.1000001" docsrc = "BNC/A0C">
<answer instance="plant.1000001" senseid="plant%factory"/>
<context>
Bodfari Foods is a processor and supplier of liquid milk. Last year the company saw pre-tax profits of 1.9m. JUS-ROL INVESTS POTATO product manufacturer Jus-Rol has invested 1.5m in its potato manufacturing <head>plant</head> at Amble, Northumberland. …… The company has also achieved a grade A accreditation from the plant evaluation committee of the National Association of Catering Butchers.
</context>
</instance>
```

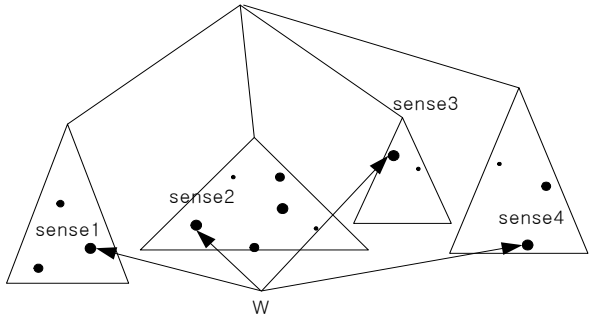
(그림 4) SemCor내 'plant'의 다른 예

위 예제에서 plant와 문맥내 다른 단어들 사이 정보량 측정 후 potato#n#1와 plant#n#2간 가장 큰 정보량인 6.243을 얻었다. 그러나 이 문맥상에서 plant는 (그림 4)에서 보다시피 factory로 정보량만을 이용한 WSD에서는 다음과 같은 오류가 발생할 수 있다. 따라서 본 논문에서는 다음 장에서 개념적 밀도를 이용한 새로운 방법을 제안하고자 한다.

4. 개념적 밀도

개념적 밀도란 중의성을 가지고 있는 단어의 각 의미별 영역을 설정한 후 그 영역의 개념적 밀도로 영역내 노드의 수를 카운팅함으로써 구할 수 있다 [1]. 즉, 영역내 노드수가 많을수록 높은 밀도값을 가지며 노드수가 적을수록 상대적으로 적은 밀도값을 가진다.

(그림 5)는 개념적 밀도를 통해 어떻게 단어의 모호성을 해결하는지 보여주고 있다. 예를 들어, 단어 W는 4가지의 의미를 가지고 있으며 각각은 WordNet의 subhierarchy에 포함되며 이는 각 의미(sense)별 영역이 된다. 영역내의 각 점은 문맥상 단어들과 단어 W의 의미이다. 그리고 영역내 존재하는 노드의 수를 카운팅하는 방법을 이용하여 개념적 밀도를 구한다. 개념적 밀도가 클수록 즉 영역이 클수록 각 문맥에서 정확한 의미를 표현할 가능성이 크다. (그림 5)의 경우 sense 2가 찾고자 하는 문맥에 가장 알맞은 의미이다.



(그림 5) 개념적 밀도

제안하고자 하는 정보량과 개념적 밀도를 이용한 WSD는 다음과 같은 같다.

1) 문맥내 단어들(즉, 명사)와 중의적 의미를 가진 단어 W간 정보량 측정을 한다.

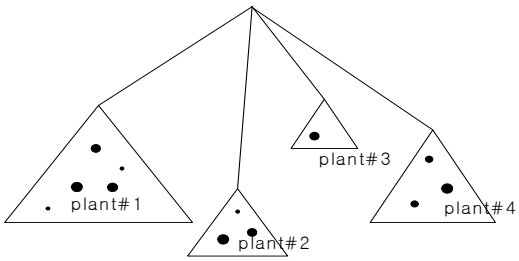
$$S(w, c_i) = \max_{c \in S(w,\ c_i)} [-\log p(c)] \text{-------} (4)$$

여기서 c_i는 W와 유사성 측정을 위한 문맥내 단어들로 이를 통해 WordNet상에서 W와 c_i의 의미(즉, 위치)를 결정한다.

2) WordNet상에서 각 W_s의 의미를 따라 영역 R을 생성한다. 그리고 각 의미별 R_s를 기반으로 한 D_s는 다음과 같다.

$$D_s = \sum_{i=1}^{n} N_i \text{----------} \quad (5)$$

여기서 n은 region 내 노드의 총 수이다. 위에서 언급했듯이 밀도는 R에 비례한다. (그림 6)는 예제 2의 개념적 밀도 측정 결과이다. 밀도 측정 결과 plant#1 즉 factory의 의미로 정확한 의미를 찾아내었다.

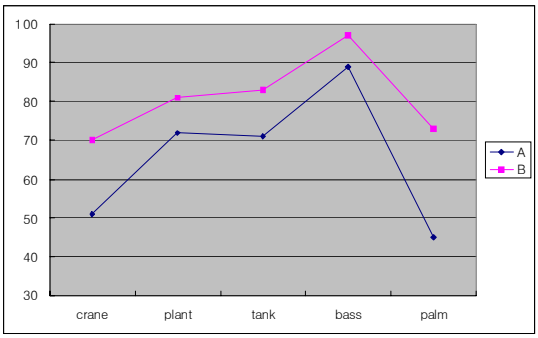


(그림 6) 그림 4 예제의 개념적 밀도

5. 실험

제안한 정보량과 개념적 밀도를 이용한 WSD 실험을 위해 WordNet 2.0버전을 이용하였으며, 대표적인 동음이의어에 대해 의미를 주석 처리해 둔 171개의 문맥을 포함한 SemCor를 이용하였다. 본 논문에서 다음과 같은 두 가지 방식으로 테스트하였다.

A. 정보량만을 이용한 WSD
B. 정보량과 개념적 밀도를 이용한 WSD



(그림 7) A와 B 방법을 통한 WSD 결과

중의성을 가진 단어인 'crane', 'plant', 'tank', 'bass', 'palm' 총 5가지 단어로 실험하였다. (그림 7)에서 보듯이 전반적으로 A방법에 비해 B방법이 좋은 결과를 보였으며, 특히 'crane'과 'palm'은 A 방법에 비해 월등히 좋은 결과를 보였다.

6. 결론

본 논문에서는 의미적 정보 검색시 문제가 되는 단어 의미 중의성 해결을 위해 WordNet을 기반으로 정보량과 개념적 밀도를 이용한 WSD를 제안하였다. SemCor라는 대용량의 말뭉치 데이터베이스를 이용하여 실험한 결과 정보량만을 적용한 경우에 비해 정보량과 개념적 밀도를 같이 고려한 경우 WSD가 20% 향상되었다. 제안한 방법은 사용자에게 의미적 정보 검색이 가능하도록 도와줄 것이다.

Acknowledgement

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음. (IITA-2005-C1090-0502-0009)

참고문헌

[1] Eneko Agirre, German Rigau, "Word Sense Disambiguation using Conceptual Density", International Conference On Computational Linguistics, 1996

[2] Sang-Gum Kim, Hee-Cheol Seo, Hae-Chang Rim, "Information retrieval using Word Senses: Root Sense tagging Approach", Annual ACM Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom, 2004

[3] Philip Resnik, " Semantic Similarity in Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research, 1999

[4] William D. Lewis "Measuring Conceptual Distance Using WordNet: The Design of a Metric for Measuring Semantic Similarity", The University of Arizona Working Papers in Linguistics, 2002

[5] Peter D. Turney, " Word Sense Disambiguation by Web Mining for Word Co-occurrence Probabilities", Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3), Barcelona, Spain, 239-242, July, 2004

[6] Philip Resnik, David Yarowsky, "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation", Natural Language Engineering, 1999