의미속성 기반의 개념망을 위한 어휘 연관도 측정

옥은주[0]　　이왕우　　이수동　　옥철영
울 산 대 학 교　　컴 퓨 터 정 보 통 신 공 학 부
{ejock, wwlee, sdlee, okcy}@uou.ulsan.ac.kr

A Measurement of Lexical Relationship for Concept Network Based on Semantic Features

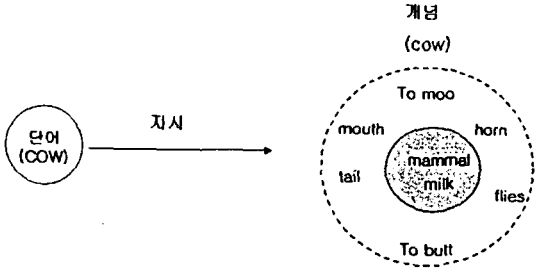
Eun-Joo Ock[0]　Wang-Woo Lee　Soo-Dong Lee　Cheol-Young Ock
Dept. of Computer Engineering and Information Technology, University of Ulsan

요 약

본 논문에서는 개념망 구축을 위해 사전 뜻풀이말에서 추출 가능한 의미속성의 분포 정보를 기반으로 어휘 연관도를 측정하고자 한다. 먼저 172,000여 개의 사전 뜻풀이말을 대상으로 품사 태그와 의미 태그가 부여된 코퍼스에서 의미속성을 추출한다. 추출 가능한 의미속성은 체언류, 부사류, 용언류 등이 있는데 본 논문에서는 일차적으로 명사류와 수식 관계에 있는 용언류 중 관형형 전성어미('∟/은/는')가 부착된 것을 대상으로 한다. 추출된 공기쌍 45,000여 개를 대상으로 정제 작업을 거쳐 정보이론의 상호 정보량(MI)을 이용하여 명사류와 용언류의 연관도를 측정한다. 한편, 자료의 희귀성을 완화하기 위해 수식 관계의 명사류와 용언류는 기초어휘를 중심으로 유사어 집합으로 묶어서 작업을 하였다. 이러한 의미속성의 분포 정보를 통해 측정된 어휘 연관도는 의미속성의 공유 정도를 계산하여 개념들간에 계층구조를 구축하는 데 이용할 수 있다.

1. 서론

개념은 동일 속성을 가진 대상으로부터 추상한 일반화된 관념으로 정의되는데[1], 단어들을 사전에서 찾아보면 핵심적인 단어는 기초 어휘를 형성하는 것이고 주변적인 단어들은 보다 일반적인 지식들을 표현한 것으로써 백과사전에서 언급될만한 것들이다. 이를 바탕으로 'COW' 의 개념 구조를 도식화하면 [그림 1]과 같다.



[그림 1] COW의 개념 구조

[그림 1] 에서 mammal, milk와 같은 것은 COW의 핵심적인 단어이고 to moo, mouth는 주변적인 단어가 된다.

이렇게 하나의 단어가 지니는 개념의 다양성은 어휘 클러스트링이나 계층망 구축에 많은 어려움을 준다. 그러나, 어휘간의 연관도를 측정하면 클러스터들간의 경계에 유연성을 주어 에러율을 낮출 수가 있다. 특히, 동작·상태 등과 같은 추상적인 개념과 수식의 기능을 가지는 용언류의 의미속성의 범주는 세상에 대한 인간의 인식 태도에 따라 결정된다. 또한, 사물의 속성의 표시를 나타내는 형용사는 각 어휘들이 고유의 의미특성을 가지면서 주어지는 상황에 따라 동의관계로 맺어지게 된다.

따라서, 의미속성의 개념은 백과사전적 지식으로 파악할 수 있다. 본 논문에서의 의미속성의 개념은 기존의 논리적 언어학적 관점과 상당히 다르다. 예를 들어, 기존의 논리 언어학적 관점의 의미속성은 '객관적' 인 것으로 간주되어 자동차의 의미는 '비유생적' , '구체적' , '기동적' , '자체추진적' 과 같이 기술된다. 그러나 사전 뜻풀이말에서 추출한 의미속성은 '달리는' , '낡은' ,

'갖춘', '되는', '실은', '지나가는', '빠른' 과 같이 기술된다. 이것은 의미속성이 주위 세계를 지각하고 상호작용하는 언어 사용과 관련된 백과사전적 지식을 반영하고 있다는 것을 나타낸다. 따라서, 사전 뜻풀이말을 기반으로 통계적 정보에 의한 의미속성의 추출은 의미범주 설정에 연구자의 직관성을 줄일 수 있기 때문에 자연어처리에 있어 보다 유용한 결과를 기대할 수 있다.

한편, 하나의 의미범주는 그 경계가 불분명하다는 특징이 있다. 인지범주는 낱말로 표시되며, 낱말은 사전에 기재된다[6]. 따라서 사전에 기재된 항목에서 의미범주에 대한 정보를 얻을 수 있다. 새의 유형에 대한 사전 뜻풀이에 대한 몇 가지 예를 보자.

[표 1] 새의 유형에 대한 사전 뜻풀이말

robin(로빈)	붉은 가슴깃털을 가진, 갈색을 띠는 작은 새(방울새라고도 불린다). (OALD)
parrot(앵무새)	구부러진 부리 및 보통은 보통 밝은 색의 깃털을 가진 열대산 새. 어떤 것은 인간의 말을 모방하도록 가르칠 수 있다. (LDOCE)
ostrich(타조)	타조는 날 수 없는 큰 아프리카 새이다. 그것은 다리와 목이 길고, 머리가 작고, 크고 부드러운 깃털을 가지고 있다. (COBUILD)

[표1]의 사전 뜻풀이말에서 추출된 '새' 유형의 범주에 대한 의미속성의 내부 구조는 다음과 같다.

[표 2] 범주 '새'의 속성들의 분포

속성	범주 구성원				
	>ROBIN<	>SPARROW<	>DOVE<	>PARROT<	>OSTRICH<
알을 낳다	+	+	+	+	+
부리	+	+	+	+	+
두 날개와 두 다리	+	+	+	+	±
깃털	+	+	+	+	+
작고 가볍다	+	+	±	±	-
날 수 있다	+	+	+	+	-
지저귀다/노래하다	+	+	+	±	-
가는/짧은 다리	+	+	+	±	-
짧은 꼬리	+	+	+	±	-
붉은 가슴	+	-	-	-	-

[표2]의 정보를 보면 이들이 속하는 '새' 라는 범주의 공유 의미속성과 각 '새' 의 항목들이 지니는 변별적 의미속성이 있음을 알 수 있다. 따라서 사전 뜻풀이말에서 추출한 의미속성 중 공유하는 속성을 이용하면 다른 개념을 하나의 범주로 클러스트링할 수 있으며, 변별적인 의미속성을 이용하여 개념들간의 계층구조를 설정할 수 있다.

따라서, 본 논문에서 추출한 의미속성과 공기관련 어휘와의 연관도 측정은 애매한 개념 어휘들간의 클러스트링과 나아가 개념망 구축에 유용한 정보가 될 수 있다.

2. 기존 연구

분포 정보는 코퍼스를 기반으로 하는 방법에서 주로 이용되는 것으로 통계적인 자연어처리에서 매우 유용한 정보이다[9]. 코퍼스내의 단어 분포란 단순히 인접해서 나타나는 단어들에 대해서도 적용이 가능하며 특정 구문 관계로 나타나는 단어들에 대해서도 적용이 가능하다.

단어 분포를 이용한 지금까지의 주요한 연구는 다음과 같다. [16]은 단어의 클래스를 결정하기 위해 상호 정보량을 사용하였다. 각 단어를 하나의 클러스터로 하고 greedy algorithm을 사용하여 평균 상호 정보량의 손실이 최소가 되게 클러스터들을 합쳐가며 계층적인 클러스트링을 하였다. [27]은 [16]과 같이 코퍼스내에서 좌우 50단어 내에 인접해서 나타나는 단어들을 이용하나, 단어 자체가 아니라 단어들의 클래스를 이용하여 의미 중의성을 지닌 단어의 의미를 분별하였다. 그리고, [17]은 어휘적 공기관계에 기초하여 단어간의 연관도를 구하기 위해 상호 정보량을 사용하였다. 한편, [25]는 인접해서 나타나는 단어들만을 이용할 때 나타나는 자료 부족 현상을 보완하기 위해, 인접해서 나타나는 단어와 인접해서 나타나는 단어들의 분포 양상까지 이용하여 단어의 의미를 분별하는 실험을 하였다.

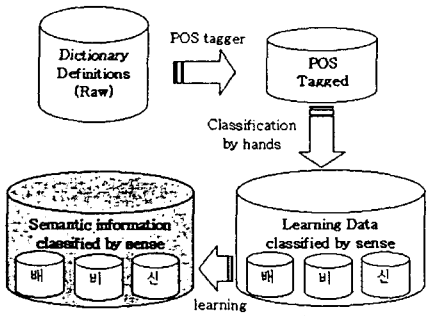
코퍼스에 나타나는 구문 관계를 이용한 연구들로 [20]은 코퍼스내에서 나타나는 술어-논항 구조에 따라 명사를 분류하였다. 먼저 구문 분석기를 이용하여 코퍼스내의 문장들에 대한 구문 구조를 얻고, 이로부터 술어-논항 관계의 자료를 추출한 뒤, 이들간의 유사 정도에 따라 명사를 분류한다. [24]는 부분 파서를 이용해 코퍼스 상에서 동사-직접목적어 관계의 예를 뽑고, 각 동사가 명사를 직접 목적어로 취하는 횟수를 이용해 명사에 대한 의미 벡터를 만들었다. 상대적 엔트로피 분포를 사용하여 명사와 명사, 명사와 클러스트간의 거리를 측정하였고 사용된 클러스트링 알고리즘은 deterministic annealing 기법이다.

[19]는 직접목적어뿐만 아니라 주어, 간접목적어 관계를 모두 이용하여 유사 명사를 추출하는 실험을 하였다. 또한 [26]은 코퍼스에서 나타나는 형용사-명사, 형용사-형용사 분포 정보를 이용하여, 의미에 근거한 형용사 분류 실험을 하였다. 그리고, [23]은 평균 클래스 상호 정보량과 locally optimal annealing algorithm를 사용하여 각 단어를 structural tag로 표시하는 클러스트링 기법을 제시하였다.

3. 사전 뜻풀이말에서 의미속성의 추출

[그림2]는 사전 뜻풀이말에서 의미속성을 추출

하는 과정이다[22]. 의미속성은 품사 부착된 사전(UMRD)의 뜻풀이말로부터 추출한다.



[그림 2] 의미속성의 추출과정

[그림2]의 과정은 사전의 코퍼스로부터 언어 정보를 추출하는 것으로, 품사가 태깅된 한국어 코퍼스로부터 수식 관계에 있는 용언류-명사류 공기쌍을 대상으로 한다. 이렇게 추출된 공기쌍은 의미속성 분포 정보를 구하는 데 이용된다. 그리고, 추출된 명사류와 용언류를 유사어 집합으로 분류한다. 유사어 집합의 정보를 이용하면 자료의 희귀성을 감소시킬 수 있다.

3.1 의미속성 분포 정보

의미속성 분포 정보는 각 용언 수식어구가 그 명사와 공기하는 조건확률로 구할 수 있다. 예를 들어 [표3]의 개념 명사류의 경우를 보자.

[표 3] 수식어구에 따른 하위개념의 분류

	움직이다		뛰어오르다		가지다_1
	나오다_1		구하다_1		느끼다_2
	받다_1		그리다_2		이르다_2
	주다		깨닫다		꾸미다
	읽다		따지다		여기다
	대하다		이르다_2		품다_1
	이루다		대하다		하다
	쓰다_3		말아보다		헤치다
	쏘다		알다		많다
	가볍다		바치다_1		다르다
	어둡다		슬퍼하다		그러하다
느낌	단단하다	마음	여기다	생각_1	좋다_1
	따뜻하다		탐내다		바르다_3
	차갑다		잡다_1		어렵다
	그득하다		지티다		그릇되다
	근지럽다		만들다		그리워하다
	껄끔거리다		싸다_1		기막히다_2
	덥다		남기다		딱하다
	뜨겁다		풀아보다		망령되다
	무겁다		보다_1		못되다
	물렁하다		울다		속되다
	보드랍다		해치다		싫다
	.....		.....		.....

위의 표에서 용언류들은 실제 뜻풀이말에서는 관형형 전성어미('ㄴ,은,는')를 취하여 명사류들의 개념을 보다 구체적으로 설명하는 것으로 개념 '느낌, 마음, 생각' 에 대한 하위개념의 분류기준이 된다. 따라서, 이들을 의미속성(semantic feature)으로 설정할 수 있다. 이와 같이, 뜻풀이말에서 중심어를 수식하는 어절까지를 표제어의 상위개념을 설정하는 기준으로 삼으면 보다 정밀하고 객관적인 하위개념을 분류할 수 있다[11].

상위개념에 대한 하위개념의 의미를 보다 구체적으로 설명하는 용언류들을 의미속성으로 설정하고, 다음 표와 같이 이들 의미속성 다음에 출현하는 어절들을 살펴보자. [표4]는 동형이의어 '쓰다_1, 쓰다_2, 쓰다_3' 가 의미속성으로 쓰이는 경우에 이들과 수식관계의 있는 개념 명사류들을 나열한 것이다.

[표 4] 동일한 의미속성을 지닌 개념

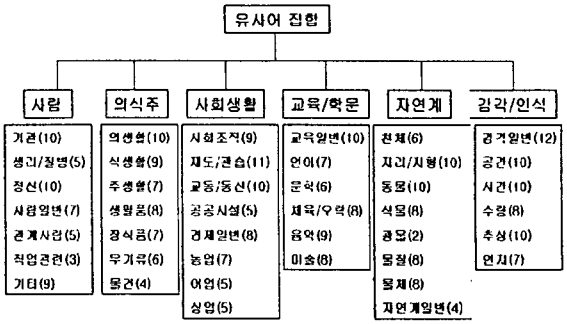
	곡식		갓_1		가공_1
	글_1		갓모자		가공품
	글밭		건_2		가축
	글씨		고깔		간편
	글씨체		관_4		갈퀴
	글자		네모		갈풀
	금빛		대_3		감탄형
	기구_6		뎃개비		값
쓰다_1	기름	쓰다_2	두건	쓰다_3	강철
	기예_1		때_1		개인_1
	끝		마스크		거름
	단자_2		말총		거적_1
	때_1		모양		건축_1
	뜻		모자_5		걸망_1
	맡_6		물건		걸레
	모피_1		방한구		검
	문학		벙거지		격조사
	.....		.....		.....

[표4]에서와 같이, 유사한 상위개념들은 동일한 의미속성을 공유하는 것을 알 수 있다. 이는 다른 개념들간에도 동일한 의미속성에 의해 구조화될 수 있으며, 나아가서는 의미속성의 공유 정도에 따라서 개념들간의 상/하위 개념들을 설정할 수 있는 가능성을 보여준다. 따라서 사전에서 변별 가능한 모든 의미속성을 추출하여 이를 바탕으로 의미체계를 구축할 수 있다면, 보다 정밀한 용언의 의미호응 관계를 설정할 수 있는 장점이 있으며, 추후 새로운 명사가 추가되더라도 전체 의미체계는 큰 변화가 일어나지 않을 것이다[5].

3.2 유사어 집합

자연어처리에서는 단어자체보다 의미적으로 유사한 단어들의 집합을 사용하는 것이 유용한 경우가 많다. 예를 들어, 수식어들은 의미적으로 유사한 명사들을 선택하고, 의미해석에서 사용되는 선택제한도 단어들의 의미 클래스에 의해서 표현되며, 또한 의미 타입은 명사 합성의 가능성을 제한한다[18]. 따라서 본 논문에서는 단어들을 유사어 집합으로 묶어 작업을 하였다. [표5]는 유사어 집합 목록을 나타낸다.

[표 5] 명사류 유사어 집합



본 논문에서 유사어 집합이란 기초어휘를 중심으로 변별되는 것으로 상하위 의미범주를 포괄하는 집합을 의미한다. 기초어휘란 언어생활에서 빈도수가 높고 분포가 넓으며, 이차조어의 근간이 되는 최소한의 필수어를 말한다[7]. 기초어휘를 기준으로 둔 근거는 개념망에서 중간 정도에 위치하는 기본 레벨에 속하는 것으로 공통된 게슈탈트이며 변별되는 전범주적인(category-wide) 의미속성이 가장 많기 때문이다[6]. 명사류를 대상으로 모두 6개의 상위 레벨의 의미범주를 설정할 수 있으며, 이들 상위 레벨의 의미범주로부터 모두 42개의 하위 레벨 의미범주를 둘 수 있다. [표6]은 의미범주 '사람'과 관련된 하위 레벨의 의미범주와 유사어 집합을 나타낸 것이다. 괄호 안의 번호는 유사어 집합의 의미범주 구분번호이다.

[표 6] 의미범주 '사람'의 유사어 집합

신체기관	머리(1100), 팔/다리(1110), 신체내부기관(1120), 뼈/골격(1130), 눈(1140), 신체외부기관(1150), 몸통(1160), 수염/털(1170), 귀/코(1180), 입(1190)
생리/질병	분비물(1200), 신체증세(1210), 질병/질환(1220), 신진대사(1230), 의료/치료(1240)
정신일반	정신일반(1300), 긍정/적극적 감정(1310), 소극/부정적 감정(1320), 생각/사고(1330), 이성작용(1340), 피/슬기(1350), 성격/인격(1360), 느낌/기분(1370)
사람일반	사람일반(1400), 남자(1410), 여자(1420), 아이/청소년(1430), 어른/노인(1440), 속성사람(1450), 자타(1460),
관계사람	형제자매(1500), 부부(1520), 친구(1530), 일반관계사람(1540)
직업관련	직업(1600), 직위/계급(1610), 직책(1620)
기타	나이/연령(1700), 이름/명칭(1710), 성_2(1720), 표정/얼굴(1730), 버릇/습관(1740), 태도/외모(1750), 행동/동작(1760), 건강(1770), 인생(1780)

한편, 용언류들도 그 의미의 속성상 유사어 집합으로 묶일 수가 있다. 용언의 유사어 부류도 명사류와 같이 공통 의미속성을 공유하면서 변별 의미속성을 가지는 경우이다. 다음 [표7]과 같은 예를 들 수 있다[3].

[표 7] 용언류 유사어 집합

대상언어	상위언어
줍다/집다	<PICK>
뽑다/빼다	<EXTRACT>
나르다/옮기다	<CARRY>
끌다/당기다	<DRAW>
쏟다/붓다	<POUR>

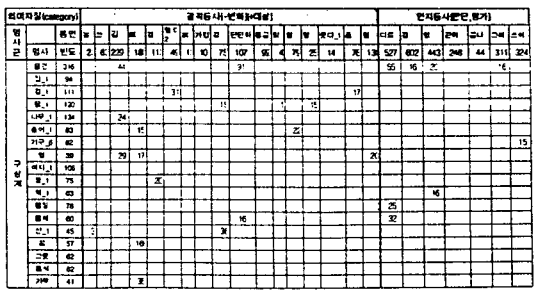
이러한 유사어 집합으로 묶은 후에 의미속성으로 사용된 동사 부류를 95개로 분류했다. [표8]에서 보듯이, 크게 8범주로 나누었으며, 이동동사의 경우는 [+이동성]의 영향권이 주체냐 대상이냐에 따라 주체이동과 대상이동으로 나누었다.

[표 8] 용언류 범주

동사부류		동사 예
이동동사	주체이동	가다류{가다, 나가다, 들어가다, 나아가다, 돌아가다, 달리다. 올라가다, 오르다 등} 오다류{오다, 나오다, 내려오다, 돌아오다 등}
	대상이동	옮기다, 내리다, 나르다, 놓다 등
화행동사		말하다, 이르다, 알리다, 일컫다 등
태도동사		갖추다, 대하다, 따르다, 여기다 등
신체행위동사		앉다, 놀다, 보다, 잡다 등
감각동사		가깝다, 멀다, 가볍다, 무겁다 등
인지동사		강하다, 약하다, 어렵다, 쉽다 등
감정동사		좋다, 나쁘다, 아름답다, 훌륭하다 등
상태변화동사		갈라지다, 나누어지다, 흐르다 등

유사어 집합으로 클러스트링한 후 이들이 공기관계로 나타나는 교차표를 제시하면 [표9]와 같다.

[표 9] 의미범주와 의미속성의 연관 교차표



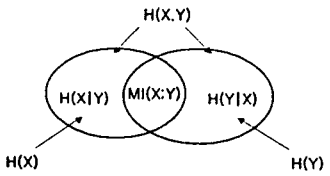
위의 [표9]를 보면 상위 의미범주를 중심으로 의

미속성들이 군집을 이루어 나타난다. 이로부터 하나의 의미범주는 특정 의미속성과 연관성이 있음을 알 수 있다. 따라서, 이들 의미범주들과 의미속성간의 연관도를 계산하면 개념들간의 상하관계를 설정할 수 있다[4].

4. 의미속성 분포 정보를 통한 어휘 연관도 계산

어휘 연관도를 구하기 위해 통계에 정보 이론을 도입하였다. 정보 이론은 Shannon에 의해 1948년에 처음으로 고안되었다[18].

직관적으로 두 확률 x와 y의 사건 분포가 얼마나 관련이 있는가는 함께 나타나는 정도로 파악할 수 있다. 이러한 두 사건의 관련성을 측정하기 위해 상호 정보량(Mutual Information)을 사용하는데, 이것을 두 확률분포 사이에 거리(연관성)를 측정하는 데 사용되는 상대적 엔트로피의 특별한 예이다. 다음 [그림3]은 상호 정보량과 엔트로피 사이의 관계를 나타내는 것이다.



H(X): 사건 x에 대한 불확실성

MI(X;Y):Y가 X에 대해 제공해 줄 수 있는 정보의 양

H(X|Y): Y를 알고 있을 때 X에 대한 불확실성

엔트로피와 상호 정보량 사이의 관계:
H(X)-H(X|Y) = MI(X;Y)

[그림 3] 상호 정보량과 엔트로피

본 논문에서의 의미속성과 명사 클래스들간의 연관도를 계산하기 위해 위의 상호 정보량(MI)을 이용하였다. 상호 정보량을 이용하면 주어진 의미속성과 명사류간의 유사도를 계산할 수 있다. 의미속성과 공기관계에 있는 명사류 의미범주의 상호 정보량은 [수식1]과 같다.

$$MI_{sf}(nv) = \log_2 \frac{\frac{f_{sf}(nv)}{N}}{\frac{f(n)}{N}\frac{f(v)}{N}} \quad \text{(수식 1)}$$

(수식1)에서 $f_{sf}(nv)$ 는 사전 뜻풀이말에서 의미속성이 명사 클래스와 공기하여 나타나는 빈도수이고, f(n)과 f(v)는 각각 뜻풀이말에서 추출한 명사 n과 동사 v의 빈도수이며, N은 코퍼스에서 나타난 수식관계에 있는 어적의 총수이다. 이러한 상호 정보량을 이용하면 단순히 빈도수만을 이용하여 연관도를 측정했을 때보다 정확률을 확보할 수 있다. 예를 들면, 용언 '하다, 되다, 있다' 의 경우는 빈도수가 굉장히 높을 뿐 아니라 거의 모든 명사와 같이 쓰일 수 있으므로 명사와 동사에 나타나는 빈도수도 다른 동사에 비해서 굉장히 높을 것이고, 그러므로 당연히 명사들을 분류하는데 큰 영향을 끼칠 것이다. 그러나, 이러한 동사들은 쓰임이 굉장히 일반적이지만 실제로 분류를 할 때는 오히려 명사와의 관계가 밀접한 동사와의 정보가 중요하다[8,17]. 그래서 빈도수보다 두 개의 사건이 연관된 정보의 양을 나타내는 상호 정보량을 사용하였다.

5.실험 및 분석

5.1 상호 정보량 측정

수식관계에 있는 용언류와 명사류는 모두 45,000개가 추출되었다. 이들을 유사어 집합으로 묶어서 용언류 약 95범주, 명사류 약 349범주로 수작업으로 구축했다. 이러한 과정을 거친 후, 의미속성인 용언류에 대해 명사류 의미범주의 상호 정보량을 측정했는데, 이들 중 몇 가지 예를 보면 [표10]과 같다.

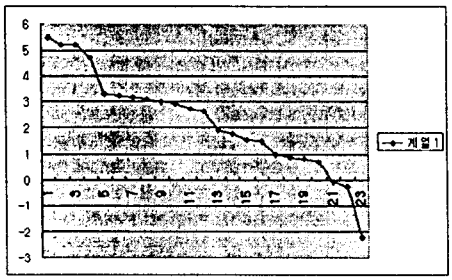
[표 10] '심다, 피다_1'와 의미범주의 상호 정보량

(심다, 명사군)	MI(v,n)	f(v, n)	f(v)	f(n)
산/숲	5.32	2	119	86
담배	4.59	2	119	19
풀/식물	4.57	30	119	289
농작물	4.03	2	119	28
땅/터	3.99	27	119	369
물체기타	3.48	2	119	41
때/시간	3.35	7	119	320
나무류	3.27	9	119	214
어익/손해	2.82	11	119	357
운동종류	2.8	1	119	33
과일/열매	2.7	3	119	105
날	1.82	1	119	65
액체	1.77	1	119	67
관계시간	1.55	2	119	157
간격/틈	1.48	1	119	82
식기류	1.2	2	119	199
재물/도랑	1.18	1	119	101
거리/무게	0.782	1	119	133
장소/굴	0.66	2	119	280
용기류	0.5	1	119	161
도구	-0.004	4	119	918
건물일부	-0.23	1	119	268
건설/거짓	-0.68	1	119	366
모양/형상	-0.7	1	119	373
속성사람	-1.18	1	119	519

(피다, 명사군)	MI(v,n)	f(v, n)	f(v)	f(n)
꽃	6.39	33	88	109
밤낮	5.92	3	88	15
잎	4.19	3	88	51
단위	4.06	2	88	37
시제	3.88	2	88	42
재물/도랑	3.53	8	88	214
산/숲	3.43	3	88	85
고형물	3.24	2	88	67
정치/체경	3.23	1	88	33
때/시간	3.12	9	88	320
풀/식물류	3.07	4	88	289
달	3.066	1	88	37
가간	2.06	1	88	74
모양/형상	2.05	5	88	373
관계시간	1.98	2	88	157
금속	1.84	2	88	173
의사소통행위	1.157	1	88	139
식물류부분	0.989	1	88	156
가치/용도	0.76	2	88	366
감정/마음	0.135	1	88	282
재넘/대상	0.07	1	88	295

위의 [표10]에서 알 수 있는 것은 다음과 같다. 첫째, 출현 빈도수와 측정된 상호 정보량을 비교해 보면, 출현 빈도수가 많다고 해서 그 의미속성과 관련하여 중요도가 높은 것은 아니다. 예를 들어, 의미속성 '심는' 과 공기하는 명사류를 보면 '땅/터' 가 빈도수 27로 '농작물' 2보다 훨씬

많지만 상호 정보량은 '농작물' 이 오히려 높음을 알 수 있다. 둘째, 각 의미속성이 동일한 명사부류와 공기하더라도 상호정보량은 다름을 알 수 있다. '개울/도랑' 의 경우 '심는' 에서는 상호정보량이 1.18로 비교적 낮은 편이지만 '피는' 과 공기하는 경우에는 3.53으로 상대적으로 높다. 셋째, 의미속성들로 사용된 용언류들은 사전 뜻풀이말에서 다의적인데 상호 정보량을 통해서 중심의미와 부수적 의미를 설정해 줄 수 있다. 다음 [그림4]를 보면 쉽게 확인할 수 있다.



[그림 4] '피다_1'와 의미범주의 상호 정보량

[그림4]는 '피다_1'에 대한 상호 정보량을 나타낸 것이다. 세로축은 상호 정보량 수치를 나타내며, 가로축의 번호는 명사류 의미범주이다. 이를 통해 의미속성 '피다'와 연관도가 높은 명사류는 '꽃(번호1), 밤낮(번호2), 잎(번호3), 단위(번호4)'와 같은 것임을 알 수 있다. 따라서, 다음의 뜻풀이말 리스트에서 '피다_1'의 여러 의미 중 의미 '꽃봉오리 따위가 벌어지다'가 중심적 개념의미가 된다고 할 수 있다. 다음 [표11]은 명사류 '느낌, 마음, 생각_1' 과 연관된 의미속성을 나타낸다.

[표 11] '느낌, 마음, 생각_1'과 연관 의미속성

느낌	고프	5.647304944	마음	겸연쩍	5.743931779	생각_1	품다_1	6.98230901
	그득하	5.647304944		너그럽	5.743931779		원하	6.713941746
	근지럽	5.647304944		두렵	5.743931779		감추어지	6.394519345
	쌀쌀거리	5.647304944		성스럽	5.743931779		아깝	6.394519345
	당황하	5.647304944		안타깝	5.743931779		딱하	5.701372165
	따갑	5.647304944		조마조마하	5.743931779		망령되	5.701372165
	트이다_2	5.647304944		고요하	5.050784599		잡히다_2	4.785081433
	서운하	5.647304944		굳세	5.050784599		그리워하	4.602759876
	섭섭하	5.647304944		튀어오르	5.050784599		속되	4.602759876
	괜찮하	5.647304944		분하다_3	5.050784599		그릇되	4.448609196
	...	...		당네	5.050784599		기막히다_2	4.448609196
					...		...	...

[표11]을 보면, '마음'은 '겸연쩍다, 너그럽다, 두렵다, 성스럽다, 안타깝다, 조마조마하다 등'의 의미속성과 연관도가 높은 반면, '생각_1'은 '품다_1, 원하다, 감추어지다, 아깝다, 딱하다, 망령되다 등'과 연관도가 높다는 것을 알 수 있다. 이것은 추상적이고 애매한 개념 어휘도 연관된 의미속성에 따라 변별될 수 있음을 나타낸다.

'쓰다' 의 경우 [표12]에서와 같이 동형이의어로 나타나는데, 이들과 연관관계에 있는 명사류을 보면, 의미속성에 의해 서로 다른 클래스로 클러스트링됨을 알 수 있다. '쓰다_1' 은 '글자를 이루다, 글을 짓다' , '쓰다_2' 는 '모자 따위를 머리에 얹다' , '쓰다_3' 은 '사용하다,이용하다' 가 주된 개념의미를 이루는데, 이에 따라 연관된 명사류가 클러스트링되어 나타남을 알 수 있다.

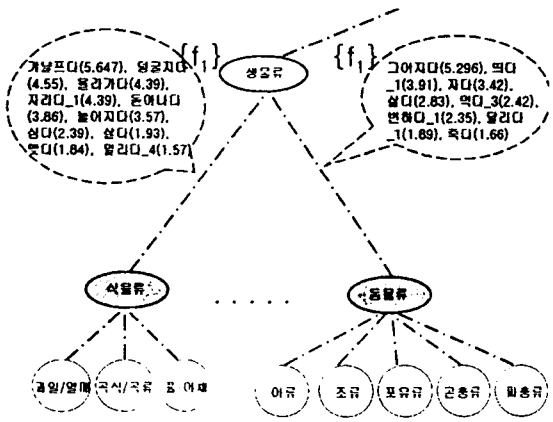
[표 12] 동형이의어 '쓰다'와 연관 의미범주

쓰다_1	소설/설화	3.111	쓰다_2	모자/갓	5.291	쓰다_3	기호/부호	2.351
	글	2.659		의류기타	5.118		농기구	1.928
	시/시조	2.461		포유류부분	3.444		칼/창	1.866
	역사/문화	2.461		가지	3.039		가공품	1.801
	기둥	2.1		가전제품	1.751		거름	1.801
	언어요소	2.026		금속	1.525		화장품	1.801
	기술_1/기교	2.012		생활도구	1.504		가공할품	1.696
	종이류	1.985		여자	1.342		고무류	1.647
	곡류/곡식	1.767		격식/도리/관습	1.326		국면	1.556
	사타	1.469		풀/식물류	0.641		무기류부분	1.513
	말/언어	1.371		방향/쪽	0.589		생활도구	1.49
	직업	1.195		모양/형상/모습	0.174		문방구	1.483
	과목/학문	1.189		도형	0.119		총/대포	1.342
	흠_4	1.153		물건/사물	6.20E-02		놀이_1	1.29
	유희/오락	1.074		때/시간/시각	-0.853		금속	1.253
	의식/예식	1.042					식기류	1.226
	언어단위	0.914					포유류부분	1.226
	속성사람	0.644		...	...		널이	1.195
	...	...					소리기구	1.195
							...	...

따라서, 이들 상호 정보량은 의미속성의 공유정도를 계산하여 개념들간에 계층구조를 구축하는 데 이용할 수 있다.

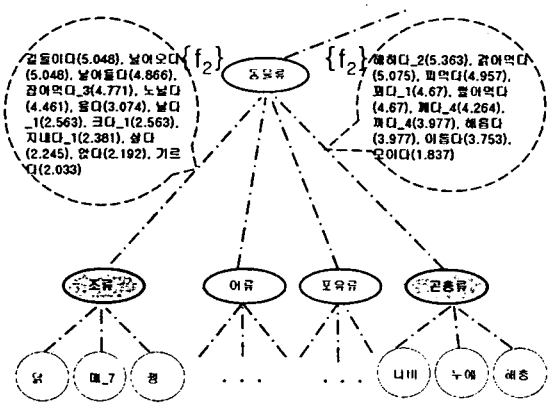
5.2 계층망 구축 예

다음에는 의미속성에 의해 계층망이 구축되는 과정을 보자. 먼저, [그림5]는 계층망 구조의 일부로 '생물류' 노드에서 그 하위범주로 클러스트링되는 과정이다.



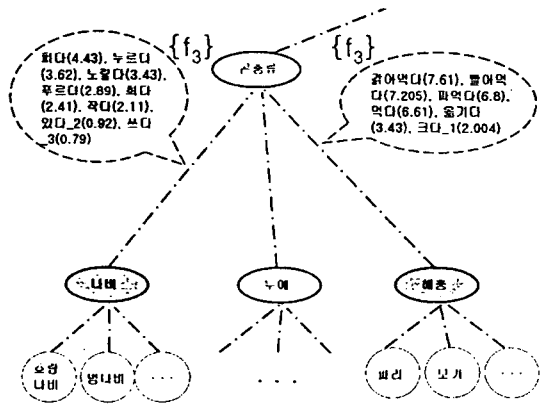
[그림 5] 생물류의 클러스트링과 의미속성 집합

[그림5]에서 생물류는 의미속성 집합 $\{f_1\}$에 의해 식물류와 동물류로 클러스트링된다. 이 때, 괄호 안의 수치는 생물류의 하위 의미범주와 각 의미속성들 간의 상호 정보량을 나타낸다. 다음 [그림6]은 그 다음 하위 레벨인 동물류 범주에서 의미속성 집합 $\{f_2\}$에 의해 클러스트링되는 것을 나타낸다. 현 레벨에서 의미범주들은 의미속성 집합 $\{f_1\}$을 상속받게 된다. 한편, 의미범주 '조류'와 '곤충류'는 '닭, 매_7, 꿩, 나비, 누에, 해충 등'과 같은 기본 레벨의 개념 어휘를 포함하는 유사어 집합이다. 이 레벨에서 '날다, 울다, 살다, 해롭다, 먹다, 기르다 등'과 같이 변별되는 '전범주적' 의미속성이 가장 많음을 알 수 있다.



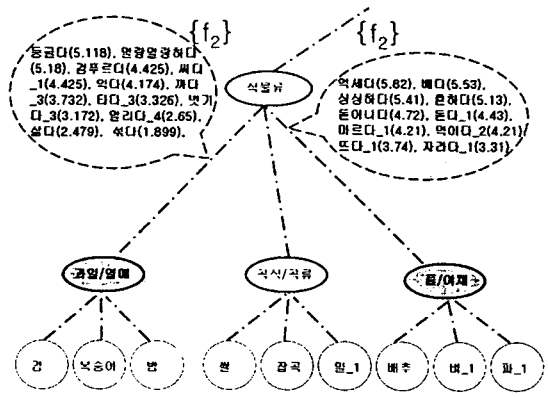
[그림 6] 동물류의 클러스트링과 의미속성 집합

다음 [그림7]은 [그림6]의 하위 레벨인 곤충류에서 의미속성 집합 $\{f_3\}$에 의해 클러스트링되는 의미범주들이다. 현 레벨에서는 모든 상위 레벨의 의미속성 집합 $\{f_1\}$과 $\{f_2\}$ 모두를 상속받게 된다. 그리고 현 레벨에서 '현저하고 특정한' 의미속성들이 그 분류기준이 됨을 알 수 있다.



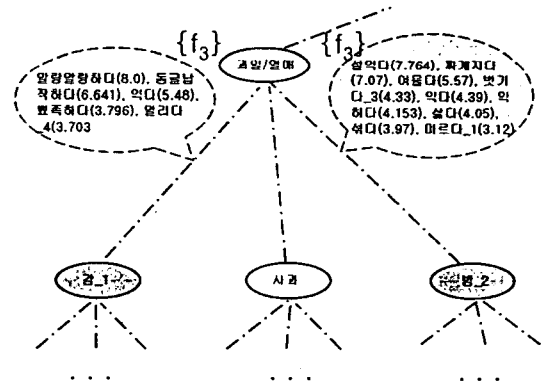
[그림 7] 곤충류의 클러스트링과 의미속성 집합

다음 [그림8]과 [그림9]는 '생물류-식물류-과일/열매-감_1'의 순서로 상위 레벨에서 하위 레벨로 의미속성 집합 $\{f_1\}$-$\{f_2\}$-$\{f_3\}$에 의해 클러스트링되는 과정을 나타낸 것이다. 역시 이 과정 중에도 각 레벨은 상위레벨의 의미속성 집합을 상속받게 된다.



[그림8] 식물류의 클러스트링과 의미속성 집합

마찬가지로, 의미범주 '과일/열매' 노드는 개념 어휘를 포함하는 유사어 집합으로 변별되는 전범주적 의미속성이 가장 많음을 다음 [그림9]에서 알 수 있다.



[그림9] 과일류의 클러스트링과 의미속성 집합

한편, 자연어 처리의 많은 예가 말뭉치에서 존재하지 않기 때문에 단어 자체에 대한 통계값에 전적으로 의존하는 것은 불가능하다. 따라서 단어 기반 모델은 매우 큰 확률 변수가 필요하게 되어 적절한 학습 크기를 결정하는데 어려움이 있고 실제 문제 해결에 적용시에도 자료 희귀성의 문제가 따른다[2]. 마찬가지로, 본 연구의 명사류 의미범주와 관련된 의미속성들도 모두가 항상 관찰대상으로 나타나는 것은 아니다. 다음 [표13]은 '과일/열매' 의 하위 의미범주들인데, 관련 의미속성이 결여됨을 알 수

있다.

[표 13] 의미속성이 결여된 '과일/열매' 하위범주

FI	감_1	대추_1	배실_2	밤_2	배_4	복숭아	사과_3	살구	잣	참외	호박
동그랗										5.8349	
말리다_4	4.5142		5.2074								
둥글										4.4222	4.54
물리다_2		6.1494							5.4562		
붉										2.5027	3.3136
익						5.9891					
빨갛										5.8349	
조리							8.8794				
싸_1					8.1863						
아름다_1				4.6797							
달어지											
빠다_1								6.8425			
빗기다_3				6.4227							
적어지											
살				5.7295							
떫					5.5713						
보죽하	5.2959										
검다_2										2.616	
섞				5.1497							
얼						4.495					
푸르											4.3722

그러나, 본 논문에서 제안한 모델에서는 의미속성이 상위 레벨에서 하위 레벨로 상속되기 때문에 결여된 의미속성 정보를 바로 위 상위 레벨의 의미범주로부터 구할 수가 있다. 예를 들어, 의미범주 '감_1' 의 의미속성 정보를 구하는 과정을 나타내면 다음과 같다.

$F_{감\_1} = \{f_1\} + \{f_2\} + \{f_3\}$

$\{f_1\}$: $\{f_{식물류}\}$ = {심다, 맺다, 가냘프다, 덩굴지다, …}

$\{f_2\}$: $\{f_{열매류}\}$ = {둥글다, 싸다_1, 익다, …}

$\{f_3\}$: $\{f_{감\_1}\}$ = {말랑말랑하다, 말리다_4, …}

이와 같은 과정에 의해 구해진 '과일/열매' 범주의 의미속성 정보는 다음 [표14]와 같다.

[표 14] 범주 '과일/열매'의 의미속성 정보

FI	감_1	대추_1	배실_2	밤_2	배_4	복숭아	사과_3	살구	잣	참외	호박
동그랗	4.559	4.559	4.559	4.559	4.559	4.559	4.559	4.559	4.559	5.835	4.559
말리다_4	4.514	2.65	5.2074	2.65	2.65	2.65	2.65	2.65	2.65	2.65	2.65
둥글	3.705	3.705	3.705	3.705	3.705	3.705	3.705	3.705	3.705	4.422	4.54
물리다_2	2.388	6.149	2.388	2.388	2.388	2.388	2.388	2.388	5.456	2.388	2.388
붉	1.92	1.92	1.92	1.92	1.92	1.92	1.92	1.92	1.92	2.503	3.314
익	4.174	4.174	4.174	4.174	4.174	5.989	4.174	4.174	4.174	4.174	4.174
빨갛	3.865	3.865	3.865	3.865	3.865	3.865	3.865	3.865	3.865	5.835	3.865
조리	4.02	4.02	4.02	4.02	4.02	4.02	8.879	4.02	4.02	4.02	4.02
싸_1	4.425	4.425	4.425	4.425	4.425	8.186	4.425	4.425	4.425	4.425	4.425
아름다_1	2.122	2.122	2.122	4.68	2.122	2.122	2.122	2.122	2.122	2.122	2.122
달어지	2.074	2.074	2.074	2.074	2.074	2.074	2.074	2.074	2.074	2.074	2.074
빠다_1	1.983	1.983	1.983	1.983	1.983	1.983	1.983	6.843	1.983	1.983	1.983
빗기다_3	3.172	3.172	3.172	6.423	3.172	3.172	3.172	3.172	3.172	3.172	3.172
적어지	4.425	4.425	4.425	4.425	4.425	4.425	4.425	4.425	4.425	4.425	4.425
살	2.479	2.479	2.479	5.73	2.479	2.479	2.479	2.479	2.479	2.479	2.479
떫	0.712	0.712	0.712	0.712	5.571	0.712	0.712	0.712	0.712	0.712	0.712
보죽하	5.296	1.822	1.822	1.822	1.822	1.822	1.822	1.822	1.822	1.822	1.822
검다_2	1.052	1.052	1.052	1.052	1.052	1.052	1.052	1.052	1.052	2.616	1.052
섞	1.899	1.899	1.899	5.15	1.899	1.899	1.899	1.899	1.899	1.899	1.899
얼	0.724	0.724	0.724	0.724	0.724	4.485	0.724	0.724	0.724	0.724	0.724
푸르	0.899	0.899	0.899	0.899	0.899	0.899	0.899	0.899	0.899	0.899	4.372

이러한 접근법은 단어 기반 모델과 클래스 기반 모델을 모두 고려한 통계 모델로 두 모델의 약점을 어느 정도 극복할 수 있다. [표14]와 같은 의미속성 정보는 '과일/열매' 와 같은 하나의 의미범주에서 속하는 개념 어휘들의 유사도(Similarity)를 측정하는 데 사용된다. 이것은 향후 연구 과제로 남겨두고자 한다.

6. 결론

본 논문에서는 사전 뜻풀이말을 기반으로 의미속성을 추출하고 이와 공기관계에 있는 명사류를 분포 정보의 대상으로 어휘 연관도를 측정하였다. 측정 방법은 정보 이론의 상호 정보량을 이용하였는데, 자료의 희귀성을 완화하기 위해 단어들을 유사어 집합으로 묶어서 작업을 하였다. 어휘 연관도의 측정을 통해 알 수 있었던 것은 다음과 같다. 첫째, 출현 빈도수가 많다고 해서 그 의미속성과 관련해서 중요도가 높은 것은 아니다. 둘째, 각 의미속성이 동일한 명사부류와 공기하더라도 상호 정보량은 다름을 알 수 있다. 셋째, 의미속성들로 사용된 용언류들은 사전 뜻풀이말에서 다의적인데 상호 정보량을 통해서 중심적 개념의미와 부수적 개념의미를 설정해 줄 수 있다.

한편, 의미속성과 의미범주의 상호 정보량에 의해 개념망을 구축할 수가 있었다. 상위 레벨의 의미속성은 하위 레벨로 상속되기 때문에, 하위 레벨에서 결여된 의미속성 정보는 상위레벨에서 구할 수가 있었다. 이렇게 사전 뜻풀이말에서 추출한 의미속성 정보를 이용한 개념망은 Top-down 접근과 Bottom-up 접근의 단점을 극복하고 보다 정밀하고 객관적으로 구축될 수가 있다.

본 논문에서는 사전 뜻풀이말에서 추출가능한 의미속성 중 관형형 전성어미를 가지고 수식어 역할을 하는 용언류만을 고려하였다. 그러나, 체언류, 부사류, 문법 표지와 같은 정보들도 의미속성으로 추출해서 개념망 구축의 정보로 이용하는 작업이 요구된다. 또한, 사전 뜻풀이말에서 변별가능한 다양한 의미속성의 정보를 이용하여 보다 확장된 개념망을 구축하는 작업으로 이어져야 할 것이다.

7. 참고 문헌

[1] 김봉주, 1992. 개념학. 한신문화사.

[2] 김선호. 1996. 통계 정보를 기반으로 한 어휘 관계 예측. 연세대학교 석사학위논문.

[3] 김준기. 2000.「 한국어타동사 유의어 연구 」. 한국문화사.

[4] 김준수, 옥은주, 옥철영. 2001.「 사전의 뜻풀이말에서 추출한 개념어휘 및 의미자질 」, ASIALEX 2001 PROCEEDINGS.

[5] 박영자. 1997. 사전을 이용한 단어 의미 자동 클러스트링: 유전자 알고리즘 접근법. 연세대학교 박사학위논문.

[6] 임지룡, 김동환. 1998.「 인지언어학개론 」.한신

문화사.
[7] 임지룡. 1997.「인지의미론」.탑출판사.
[8] 정연수, 조정미, 김길창. 1995.「개념분류기법을 적용한 한국어 명사분류」. 제7회 한글 및 한국어 정보처리 학술대회.
[9] 조정미, 김길창. 1995.「분포정보를 이용한 의미 중의성을 지닌 한국어 동사의 의미 분별」. 제7회 한글 및 한국어 정보처리 학술대회.
[10] 조정미. 1998. 코퍼스와 사전을 이용한 동사 의미 분별. 한국과학기술원 박사학위논문.
[11] 조평옥, 옥철영 . 1999.「사전 뜻풀이말에서 구축한 한국어 명사 의미계층구조」. 한국인지과학회 논문지 제10권 제 4호.
[12] 한영균. 1994.「명사류 의미망 구축을 위한 사전 뜻풀이의 어 휘구조 분석」. 제4회 한국어 정보처리 학술대회 발표논문집. 한국정보과학회.
[13] 한영균. 1998.「<한국어 기초 빈도용례 사전>의 편찬을 위한 기초적 연구」.
[14] 허정. 2000. 사전 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템. 울산대학교 석사학위논문.
[15]Alpha k, Luk. 1995. "Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions" , 33rd Annual Meeting of the ACL
[16]Brown,P.F.,Pietra,V.J.D.,DeSouza,P.V.,Lai,J.C., and Mercer,R.L.1992.Class-based n-gram models of natural language. In Computational Liguistics.
[17]Church.K.,Hanks.P. 1989. Word association norms, mutual information, and lexicography. In Proceedings of the 27th Meeting of the Association for Computational Lin guistics. Vancouver, B.C.
[18]Fano,R.1961.Transmission of Information.In Cambridge, Mass:MIT Pres.
[19]Grefenstette,G.1993.Evaluation techniques for automatic semantic extraction: comparing syntactic and window-based approaches. Technical Report. Department of Computer Science, University of Pittsburgh
[20]Hindle,D. 1990. Noun Classification From Predicate-Argument Structures. In Proceedings of the 28st Meeting of the Association for Computational Linquistics.
[21]Ido Dagan. 1995.「Contextual Word Similarity and Estimation form Sparse Data」.
[22]J. Hur, C.-Y. Ock. 2001.「A Homonym Disambiguation System based on Semantic Information extracted from Definition in Dictionary」. 19th ICCPOL 2001.
[23]McMahon, J. and Smith, F.,Improving statistical langualge model performance with automaticallygenerated word hierarchies, Computational Lingyistics.
[24]Pereira,F.,Tishby,N. and Lee,L.,Dstributional clustering of English words, Proceeding of the 31st Annual Meeting of ACL.
[25]Schutze,H.1992.Word Sense disambiguation with sublexical representations.In Workshop Notes, Statistically-Based NLP Techniques.AAAI
[26]Vasileios Hatzivassiloglou,and Kathleen R.McKeown.1993. Towards The Automatic Identification Of Adjectival Scales: Clustering Adjectives According To Meaning.In Proceedings of the 31st Meeting of the Association for Computational Linguistics. Columbus.
[27]Yarowsky,D.1992.Word-sense disambiguation using statistical model of Roget's categorties trained on large corpora. In Proceedings of COLING-92.