사전에 나타난 인지정보를 이용한 단어 개념의 지식표현

윤덕한O 옥철영
울 산 대 학 교 컴 퓨 터 정 보 통 신 공 학 부
{freshuni1, okcy}@uou.ulsan.ac.kr

Knowledge Representation of Concept Word Using Cognitive Information in Dictionary

Duck-Han YunO Cheol-Young Ock
Dept. of Computer Engineering and Information Technology, University of Ulsan

요 약

인간의 언어지식은 다양한 개념 관계를 가지며 서로 망(network)의 모습으로 연결되어 있다. 인간의 언어지식의 산물 중에서 가장 체계적이며 구조적으로 언어의 모습을 드러내고 있는 결과물이 사전이라고 할 수 있다. 본 논문에서는 이러한 사전 뜻풀이말에서 개념 어휘와 자동적인 지식획득을 통하여 의미 정보를 구조적으로 추출한다. 이러한 의미 정보가 추출되면서 동시에 자동적으로 개념 어휘의 의미 참조 모형이 구축된다. 이러한 것은 사전이 표제어 리스트와 표제어를 기술하는 뜻풀이말로 이루어진 구조의 특성상 가능하다. 먼저 172,000여 개의 사전 뜻풀이말을 대상으로 품사 태그와 의미 태그가 부여된 코퍼스에서 의미 정보를 추출하는 데, 의미분별이 처리 된 결과물을 대상으로 하기 때문에 의미 중의성은 고려하지 않아도 된다. 추출된 의미 정보를 대상으로 정제 작업을 거쳐 정보이론의 상호 정보량(MI)을 이용하여 개념 어휘와 의미 정보간에 연관도를 측정한 후, 개념 어휘간의 유사도(SMC)를 구하여 지식표현의 하나로 연관망을 구축한다.

1. 서 론

인간이 지닌 언어지식이란 어휘들이 갖는 다의적 개념 표상들(representation)과 이들 간의 다양한 개념 관계를 표현한 대규모 어휘 개념 지식베이스이다. 1980년대 이전의 자연언어처리 시스템은 작은 지식베이스와 문법만으로도 인간의 언어지식 처리를 가능하도록 그 영역을 제한하였다. 1,000 단어 미만의 단어와 몇 가지로 제한된 문법에 대해서만 처리가 가능했던 것이다. 따라서 당시의 자연언어처리 시스템의 언어 처리 능력은 매우 자연적이지 못하고 기계적이었다. 그러나 1980년대 접어들면서 자연언어처리 시스템이 실생활에서 사용되는 언어를 처리할 수 있기를 기대하게 된다. 이렇게 기존의 자연언어처리 시스템을 확장하기 위해서는 지식베이스와 문법 등의 확장이 필수적이었고, 따라서 대량의 지식 획득에 대한 문제가 대두되었다. 필요로 하는 지식의 양이 커지면서 수작업에 의한 지식의 획득은 많은 시간과 비용을 요구하게 되었고, 지식 획득의 병목이라는 문제가 생기게 되었다. 따라서 자료의 저장과 검색에 대한 컴퓨터의 기능이 강화되면서 온라인 지식을 이용한 지식의 자동 획득에 대한 관심이 많아지고 있다. 지식 획득에 가장 많이 이용되고 있는 자원은 사전과 코퍼스이다[28].

일반적으로 사전의 역할은 기계번역에서 형태소 해석, 구문 해석, 의미 해석, 개념 구조 및 문맥 해석 등에 필요한 각종 표제어의 등록과 지적 정보를 기록하고, 대응 언어의 구조 변환, 구문 및 형태소 생성 등의 처리에 필요한 각종 정보를 제공해 준다. 또한 사전의 지식정보는 수만에서 수십만 단어의 정보를 대상으로 하기 때문에 수록방법이 간편하고, 신속히 정보를 추출할 수 있는 탐색 기능이 제공된다. 국어 사전은 하나의 어휘에 대해서 그 어휘가 지닌 의미를 뜻풀이말에서 설명하고 있으며, 이러한 의미가 사용되는 용례와 관련어/반의어 등의 정보를 포함하고 있다.

사전은 단어 의미들을 설명하기 위한 목적으로 만들어진 텍스트이므로, 단어 의미에 대한 지식이 가장 명시적으로 기술되어 있다. 그러므로, 단어 의미에 대한 지식을 추출하기 위한 자료로써 사전은 매우 유용한 텍스트베이스이다. 예를 들어, 어원, 유의어, 반어의, 속담, 관용구, 용례, 뜻풀이말과 같은 것들이 사전으로부터 추출될 수가 있다. 이와 같이, 사전으로부터 추출된 단어에 대한 모든 정보를 의미 정보라고 할 수 있다. 특히, 사전 뜻풀이말은 표제어에 대한 정의 및 유의어와 동의

어 개념을 제시하여 단어간의 상하 관계 및 유의 관계에 대한 정보를 쉽게 추출할 수가 있다.

2. 기존 연구

현재까지 이루어지고 있는 지식획득 방법은 수동 획득과 온라인 자원을 이용한 자동 획득 방법이 있다. 수동 획득과 자동 획득은 그 장점과 단점이 있는데, 오늘날 풍부한 온라인 자원을 바탕으로 자동 획득에 대한 연구가 활기를 띄고 있다. 그러나, 지금까지도 가장 널리 사용되고 있는 지식 획득 방법은 수동 획득(manual acquisition)이다. 가장 노동 집약적인 방법인 수동 획득이 지금까지 널리 이용되고 있는 이유는 다음과 같다. 첫째, 코퍼스나 사전, 분석이나 획득 프로그램이 필요없기 때문에 작업 시작 비용이 가장 적다. 둘째, 작업의 성격이 영역 의존적이어서 시스템이 필요로 하는 단어의 수가 적을 경우, 수동 획득에 의해 매우 정확하고 유용한 지식을 얻을 수 있다. 그러나 모든 작업이 사람에 의해 이루어지므로 지식의 일관성 유지나 확장에 어려움이 있다. 또한 광범위한 영역을 대상으로 하거나 다국어간 자연 언어처리 시스템일 경우, 지식의 수동 획득 방법은 부적당하다. 수작업에 의한 지식 획득과 관련된 가장 대표적인 프로젝트들은 CYC 프로젝트(1986), ONTOS(1985), WordNet(1985) 등이 있다.

그러나, 컴퓨터를 이용한 대량의 자료 저장과 검색이 가능해지면서 지식을 자동으로 획득하려는 연구가 활발하게 진행되고 있다. 지식의 자동 획득은 지식의 양적인 확장이 쉽고 영역에 의존적이지 않다라는 장점이 있는 반면, 지식 획득의 자원의 질적인 문제, 자원의 분석과 획득을 위한 프로그램의 안정성이 보장되어야 한다라는 제한이 있다. 사전과 코퍼스 중심으로 연구가 이루어 지는데 사전을 이용한 지식 획득과 관련된 연구들에서 가장 빈번하게 이용되고 있는 대표적인 표준 사전은 LDOCE(Longman Dictionary of Contemporary English, Procter, 1978)과 COBUILD(Collins Cobuild English Language Dictionary, Sinclair, 1987)이다.

Lisk(1986)은 단어에 대한 사전의 정의가 그 단어의 여러 의미에 대한 좋은 표지가 된다는 생각을 기반으로 연구를 하였다. Yarowsky(1995)는 사전의 정의에 나타나는 단어들을 초기 의미 표지로 하고, 이를 적용하여 의미 분별할 수 있는 결정 목록(decision list)을 코퍼스로부터 자동으로 획득하여 반복적으로 적용하는 부트스트랩핑(bootstrapping) 방법을 제안하였다.

코퍼스에 나타나는 단어 분포를 이용한 지금까지의 주요한 연구는 다음과 같다. Brown(1992)은 단어의 클래스를 결정하기 위해 상호 정보 요구도를 사용하였다. 각 단어를 하나의 클러스터로 하고 그리디 알고리듬(greedy algorithm)을 사용하여 평균 상호 정보 요구도의 손실이 최소가 되게 클러스터들을 합쳐가며 계층적인 클러스터링을 하였다. Yarowsky(1992)는 Brown(1992)와 같이 코퍼스내에서 좌우 50단어 내에 인접해서 나타나는 단어들을 이용하나, 단어 자체가 아니라 단어들의 클래스를 이용하여 의미 중의성을 지닌 단어의 의미를 분별하였다. 그리고, Church & Hanks (1989)는 어휘적 공기관계에 기초하여 단어간의 연관도를 구하기 위해 상호 정보 요구도를 사용하였다. 한편, Schutz(1992)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 Hindle(1990)은 코퍼스내에서 나타나는 술어-논항(predicate-argument) 구조에 따라 명사를 분류하였다. 먼저 구문 분석기를 이용하여 코퍼스내의 문장들에 대한 구문 구조를 얻고, 이로부터 술어-논항(predicate-argument) 관계의 자료를 추출한 뒤, 이들간의 유사 정도에 따라 명사를 분류하였다. Pereira & Tishby & Lee(1994)는 부분 파서를 이용해 코퍼스 상에서 동사-직접목적어 관계의 예를 뽑고, 각 동사가 명사를 직접목적어로 취하는 횟수를 이용해 명사에 대한 의미 벡터를 만들었다. 상대적 엔트로피 분포를 사용하여 명사와 명사, 명사와 클러스트간의 거리를 측정하였고 사용된 클러스트링 알고리즘은 deterministic annealing 기법이다. Grefenstette(1993)는 직접목적어뿐만 아니라 주어, 간접목적어 관계를 모두 이용하여 유사 명사를 추출하는 실험을 하였다. 또한 Hatzivassiloglou(1993)는 코퍼스에서 나타나는 형용사-명사, 형용사-형용사 분포 정보를 이용하여, 의미에 근거한 형용사 분류 실험을 하였다. 그리고, McMahon & Smith(1995)는 평균 클래스 상호 정보 요구도와 locally optimal annealing algorithm를 사용하여 각 단어를 structural tag로 표시하는 클러스트링 기법을 제시하였다.

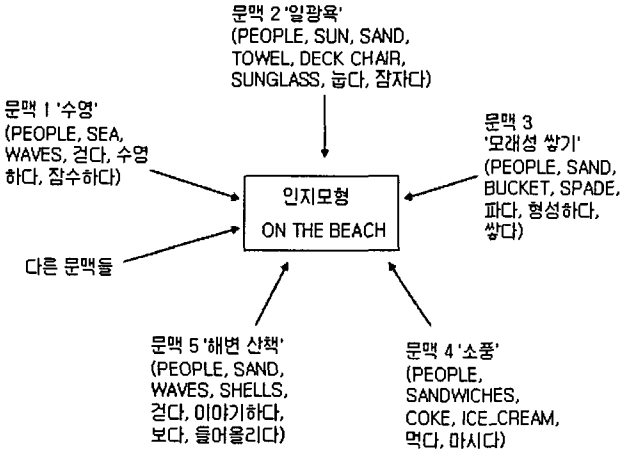
본 연구에서는 사전으로부터 자동으로 지식 획득을 하여, 이를 이용한 지식베이스로 의미망을 구축하는 데 중점을 두고자 한다.

3. 사전으로부터 지식 획득

3. 1 의미망

의미망(semantic networks)은 언어학 및 컴퓨터 과학에서 다양한 관점과 방법으로 연구되어 온 한 연구 분야이다. 의미망의 개념의 기원은 개념들의 층위적인 구조와 그 특성에 대한 의문들이 여러 가지 주제에 대해 제기 되었다. 여러 실험들에서 중요한 결과는 몇몇 개념들에 결합되어 있는 정보들이 하위 개념들에 전이될 수 있지만 이 하위 개념들에 직접적으로 저장되어 있지 않다는 것이다. 의미망은 바로 사람들이 그러하듯이 연상에 의한 기억의 모형으로 이루어진다. 우리 인간들의 기억이 의미망의 형태로 저장되는지는 확실치 않지만, 의미망은 그래프 형태로 나타내면 편리하다. 이 그래프는 각 개념을 나타내는 마디와 서로 연관이 있는 마디를 이어주는 선으로 구성된다. 이 개념이라는 기본 단위는 다른 개념들과 연결시켜 주는 관계에 의해서만 의미를 가질 수 있다. 그래프의 선들은 개념들 간의 관계를 표상한다. 각 마디에는 개념을 나타내는 표지가, 각 선에는 개념간의 관계를 나타내는 표지가 달려있다[21]. 이러한 의미망의 기본적인 개념은 인간의 인지 작용에 그 기반을 두고 있다.

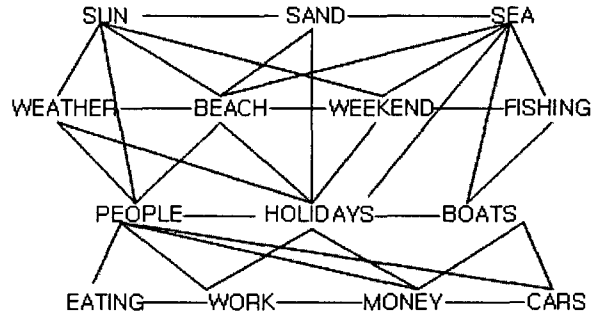
인간은 인지 작용의 하나로 일상에서 만나는 모든 종류의 현상에 대해, 상호 관련된 문맥은 체험하고 저장한다. 인지범주는 직접적인 문맥에 의존할 뿐만 아니라, 그 문맥과 연상되는 한 묶음의 전체 문맥들에도 의존한다. 이와 같이, 어떤 한 개념과 관련된 모든 인지적 표상들을 인지모형(cognitive model)이라고 한다[24].



[그림 1] 인지모형 ON THE BEACH에 대한 도식

[그림 1]은 인지모형 'ON THE BEACH'에 대한 도식을 나타내는데 이것은 주요 범주들과 범주 서로간 상호작용 하는 방식을 나타낸다.

이러한 인지모형은 다른 의미 모델에 비해 다음과 같은 특성을 가진다. 첫째, 다른 개념에 대해 개방적이다. 다시 말해, 개념이란 몇 가지의 일관된 의미 속성으로 정의 내릴 수 없으며 학자들의 주관이나 시간과 공간에 따라 달라진다는 말이다. 인지모형이 개방적이란 말은 인지모형을 기술하는 것은 어렵지만, 항상 선택적이고 변화하는 자연 언어의 특성을 잘 반영하고 있음을 의미한다. 둘째, 인지 모형은 고립된 개념이 아니라 상호관련 된다. 위의 [그림 1]에서 범주 PEOPLE, SEA, SAND 와 같은 다른 범주들은 인지모형 'ON THE BEACH'와 관련된 다양한 문맥에서 빈번하게 나타난다. 다시 말해, PEOPLE, SEA, SAND의 인지모형은 모형 'ON THE BEACH'와 밀접하게 관련된다. 따라서 이러한 인지모형의 연결은 인지망 혹은 의미망을 형성하기 위해 결합하게 되는 것이다. 다음 [그림 2]는 망의 개념을 제시하는데, 망은 복합적 연결을 통해 상호 관련된 다양한 인지모형을 구성한다.



[그림 2] 인지모형의 전형적인 망

셋째, 이러한 인지모형은 우리의 일상생활에 편재한다는 것이다. 이러한 인지모형을 구조적으로 잘 드러내고 있는 중요한 데이터베이스가 사전이라고 할 수 있다.

3.2 사전뜻풀이말을 통한 지식획득

자동으로 지식을 획득하는 방법은 크게 사전을 이용하는 방법과 코퍼스를 이용하는 방법이 있다. 지식획득으로부터 얻어지는 결과물이 의미 정보인데, 코퍼스로부터 의미 정보를 추출하는 기존의 연구는 선택 제한 지식을 이용한 규칙 기반의 의미 정보를 획득하는 방법과 공기 정보를 이용한 통계 정보 기반의 의미 정보를 획득하는 방법이 있는데, 선택 제한 지식을 이용한 방법은 구문 지식에 해당하는 정보이므로, 의미 정보 추출을 위해서는 구문 구조가 부착된 코퍼스가 필요하다. 공기 정보를 이용한 의미 정보는 단어들의 공기 정보에 따른 빈도 및 확률값을 의미 정보로 추출할 수 있다.

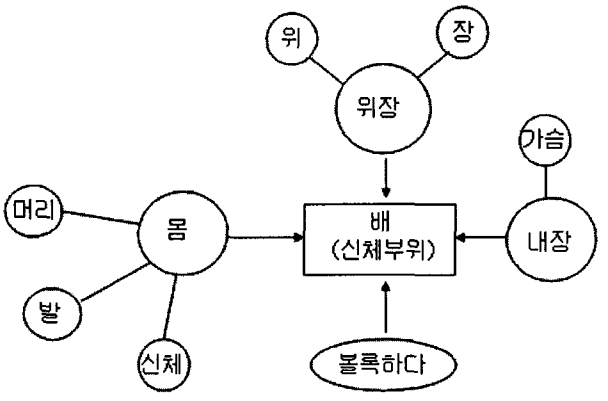
한편, 사전으로부터 의미 정보를 획득할 수가 있는데,

이러한 사전은 어떤 단어와 그 단어의 의미를 기술하기 위해 사용된 단어들 간에는 의미적인 연관 관계가 있다. 일반적으로 사전의 한 의미 항목은 의미 기술 문장과 몇 개의 예문으로 구성된다. 이러한 사전의 특성을 이용하여 네트워크를 자동으로 구축하여 단어들간의 의미 유사 관계를 측정한 연구들이 있다. 그러나 이들 연구에서는 의미 기술 단어들의 의미 중의성을 해결하지 않고 네트워크를 구축하였기 때문에 단어 의미들 간의 연관 관계를 정확하게 나타내지 못한다. 일반적으로 사전의 의미 기술 문장에 사용된 많은 단어들이 여러 개의 의미를 가지며, 이 의미들의 대부분은 정의 되는 의미와 연관 관계가 없다. 그러나, 본 연구에서는 의미 분별된 사전 자료를 대상으로 하기 때문에 의미 중의성 문제는 발생하지 않으며 논외로 한다. 다음 〈표 1〉은 사전 뜻풀이말에 나타난 의미 정보의 예를 보여준다.

〈표 1〉 개념어 '배'에 대한 1차 의미 정보

배(신체부위)
표제어 : 뜻풀이말
배(신체부위) : 사람이나 동물의 **몸**에서 **위장** 따위의 **내장**이 들어 있는 부분. 긴 몸건 **가운데**의 **볼록한** 부분 **몸**: 사람이나 동물의 **머리**로부터 **발**까지 거기에 **딸린** 모든 것의 총칭. **신체** **위장** : 위와 장. **내장** : **가슴**과 **배속**에 있는 여러 **기관**의 총칭. **해부학**상으로 **호흡기**, **소화기**, **비뇨생식기** 및 **내분비선**으로 나눔.
배(운송수단)
표제어 : 뜻풀이말
배(운송기관) : 사람 몸건을 **싣고** **물위**로 **떠다니는** 몸건. **선박** **물위**: 물의 겉면, 수면. 물이 **흘러내리**는 위쪽 부분. **선박**: **배**.
배(과일)
표제어 : 뜻풀이말
배(과일) : **배나무**의 **열매** **열매** : **식물**이 **수정**하여 **씨방**이 자라서 된 것. **과실** **배나무** : **능금 나무**과의 **활엽 교목**. 봄에 흰 꽃이 피며 가을에 **열매**가 **익는데** ,맛이 **달고** **물**이 많음.

위의 〈표 1〉은 동형이의어 '배'에 대한 표제어와 뜻풀이말을 보여주고 있다. '배'는 '신체부위, 운송수단, 과일' 등의 의미 영역에 속하는데 각 영역에서 그 개념을 가장 핵심적으로 기술하고 있다는 가정을 두고 의미 정보를 추출한다. 또한, 우리가 의미를 잘 모르거나 애매한 어휘가 나타날 때, 그 어휘를 기술한 정의문을 사전에서 찾아서 참조하듯이 그 어휘를 기술한 정의문의 의미 정보를 참조하게 된다. 위의 〈표 1〉에서 신체부위'배'를 기술하는 뜻풀이말에서 관련 의미 정보'사람, 동물, 몸, 내장, 위장, 가운데, 볼록하다'가 있다. 이 중에 '사람, 동물'은 기본어휘 선정에서 낮은 변별력을 지니게 되므로 제거된다. 따라서'배'(신체부위) 의 의미 정보로는'몸, 내장, 위장, 가운데, 볼록하다'로 추출된다. 그리고, 이들을 기술하는 뜻풀이말을 참조하게 되어'머리, 발, 신체, 가슴, 배속, 기관'등도 관련 의미 정보로 추출되게 된다. 추출된 의미 정보는 다음과 같은 의미 참조 모형을 나타내게 된다.



[그림 3] 개념어 '배' 의 의미 참조 모형

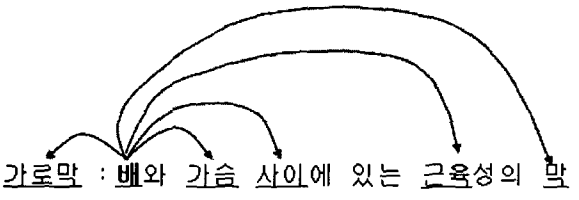
위의 모형에서'몸, 내장, 위장, 볼록하다'와 같은 의미 정보는 1차 의미 정보를 거쳐'머리, 발, 신체, 위, 장, 가슴'과 같은 2차 의미 정보를 참조하게 된다. 그러나, 2차 참조 의미 정보를 기술하는 뜻풀이말을 참조하는 것은 개념'배'(신체부위)와 연관도가 깊지 않다고 판단하여 재참조를 하지 않는다. 이렇게 네트워크로 참조하다 보면 의미들이 물고 물리는 순환 망 구조가 되어버리기 때문에 유용하지 않다. 따라서, 본 연구에서는 2차 의미 정보까지 참조하여 의미 참조 모형을 지식 베이스로 구축하게 된다.

한편, 사전 뜻풀이말 구조를 이용한 의미 정보 추출 외에 개념 어휘가 뜻풀이말에서 함께 공기하여 나타나는 어휘들을 의미 정보로 추출할 수 있다. 이것은, 공기하여 나타나는 어휘들은 연관성을 지닌다는 가정을 근거로 의미 정보를 추출하게 된다. 다음 〈표 2〉는 사전 표제어와 뜻풀이말을 나타낸다. 뜻풀이말상에 나타난 어휘와 그 어휘들과 공기하여 나타나는 어휘들간의 공기 정보를 의미 정보로 추출할 수가 있다.

〈표 2〉 개념어 '배'에 대한 2차 의미 정보

배(신체부위)
표제어 : 뜻풀이말
가로막 : 배와 가슴 사이에 있는 근육성의 막.
개복수술 : 배 안에 있는 기관의 수술 또는 이물을 없애기 위하여 복벽을 갈라내는 수술.
똠배 : 품품하게 불러 나온 배.
물배 : 물만 먹고 부른 배.
배불뚝이 : 배가 불뚝하게 나온 사람.
뱃살 : 배를 싸고 있는 살이나 가죽.
젖배 : 젖을 먹는 어린아이의 배.
헛배 : (소화 불량 등으로) 음식을 먹지 아니하고도 부른 배.
배(운송수단)
표제어 : 뜻풀이말
강나루 : 강가의 배가 건너다니는 일정한 곳.
강선 : 강철로 만든 배.
객선 : 손님의 태우는 배.
계류선 : 부두나 바닷가에 매어놓은 배.
고주 : 외로이 떠 있는 배.
귀항 : 배가 출발하였던 항구로 돌아오거나 돌아가는 항해.
난파선 : 항해 중 폭풍우나 그 밖의 장애로 파괴된 배.
닻 : 배를 한 곳에 머물게 하기 위하여, 줄을 매어 물 밑 바닥으로 가라앉히는, 갈고리가 달린 제구.

위의 사전 표제어와 뜻풀이말에서 의미 정보는 함께 공기하여 나타나는 어휘 중 명사류와 용언류와의 연관도를 측정하게 된다. 아래 도식은 어휘 '배'(신체부위)와 연관 관계에 있는 의미 정보들을 나타낸다.



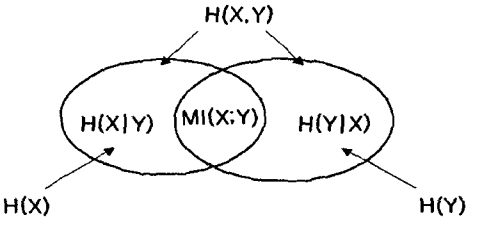
[그림 4] 개념어 '배' 의 공기 의미 정보

4. 개념 어휘의 연관도 측정과 연관망의 구축

4.1. 분포 정보를 통한 연관도 계산

추출된 의미 정보를 이용하여 어휘의 연관도를 구하여 이를 바탕으로 의미망을 구축하게 된다. 어휘 연관도를 구하기 위해 통계이론 바탕의 가장 보편적으로 사용되고 있는 상호 정보량(*Mutual Information*)을 이용하는데, 이는 정보 이론에 기반을 두고 있다. 정보 이론은 Shannon에 의해 1948년에 처음으로 고안되었다[17].

직관적으로 두 확률 x와 y의 사건 분포가 얼마나 관련이 있는가는 함께 나타나는 정도로 파악할 수 있다. 이러한 두 사건의 관련성을 측정하기 위해 상호 정보량을 사용하는데, 이것을 두 확률분포 사이에 거리(연관성)를 측정하는 데 사용되는 상대적 엔트로피의 특별한 예이다. 다음 [그림 5]는 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 5] 상호 정보량과 엔트로피

본 연구에서는 의미 정보와 개념 어휘들간의 연관도를 구하기 위해 이용하는 상호정보량(MI)은 다음과 같다.

$$MI = \log_2 \frac{\frac{f(n,m)}{N}}{\frac{f(n)}{N}\frac{f(m)}{N}}$$

(수식 1) 상호정보량

(수식1)에서 f(n,m)은 사전뜻풀이말에서 의미속성이 어휘 클래스와 공기하여 나타나는 빈도수이고, f(n)과 f(m)은 각각 뜻풀이말에서 추출한 어휘의 빈도수이며, N은 코퍼스에서 나타난 수식관계에 있는 어절의 총수이다. 이러한 상호 정보량을 이용하면 단순히 빈도수만을 이용하여 연관도를 측정했을 때보다 정확률을 확보할 수 있다. 그래서 빈도수보다 두 개의 사건이 연관된 정보의 양을 나타내는 상호 정보량을 사용하였다.

다음 〈표 3〉은 의미부류'동물류'중 추출된 의미 정보 예이다.

〈표 3〉 '동물류'의 의미 정보

동물류(NNG)	동물류(VV)
사람(82), 몸(81), 모양(27), 반응(27), 에너지(25), 강장(24), 가죽(21), 때(21), 개체(20), 일(20), 조직(20), 곳(19), 뜻(19), 뼈(19), 일부(19), 입(19), 개념(18), 고등(18), 때(18), 먹이(18), 새끼(18), 물질(17), 소리(10), 전쟁(8), 짐(8), 양쪽(8), 물건(7), 음식(7), 끝(7), 점(7), 줄(6), 비유(5), 나무(5), ...	살다(45), 있다(40), 되다(38), 다르다(36), 하다(33), 만들다(33), 가지다(29), 속하다(23), 같다(22), 잡다(21), 위하다(20), 이루다(19), 먹다(18), 생기다(18), 딸리다(17), 작다(14), 나다(13), 사납다(13), 많다(12), 움직이다(12), 이르다(10), 나오다(9), 얻다(9), 희다(9), ...

위의 〈표 3〉은'동물류'에 대해 추출된 다양한 의미 정보를 보여주고 있다. NNG는 명사류 의미 정보이며, VV는 동사와 형용사 의미 정보에 대한 모임이다. 괄호는 해당 어휘와 함께 나타난 공기 빈도수 f(n,m)이다. 그러나, 빈도수만으로는 해당 어휘의 의미 정보를 변별하기에는 부족하다.'동물류'와 함께 나타난 의미 정보 중에 고빈도로 나타난'사람(82), 몸(81), 모양(27), 반응(27), 에너지(25)'와 같은 어휘는 다른 많은 어휘와도 공기하여 나타나므로 어휘들을 분류하는 데 많은 영향을 끼치게 된다. 용언류(VV)의 경우도'살다(45), 있다(40), 되다(38), 다르다(38), 하다(33)'같은 의미 정보가 고빈도로 나타나고 있다.

그러나, 이들 고빈도로 나타난 의미 정보가 다른 어휘와의 공기빈도와 이들 의미 정보의 출현 빈도수가 높다면 해당 어휘의 핵심 의미 정보로서 변별력이 약화된다. 그러나 이들 어휘와 의미 정보와의 관련을 상호정보량(MI)을 이용하면 어휘와 의미 정보 사이의 관련성을 정제할 수 있다. 위의 〈표 3〉의 의미 정보를 기반을 상호정보량(MI)을 이용하여 연관도를 구하면 다음과 〈표 4〉와 같다.

〈표 4〉 '동물류'의 의미 정보 연관도

동물류(NNG)	동물류(VV)
삼배엽성(2.219), 육식(1.183), 후생(1.048), 낭배(0.986), 신경판(0.986), 포배(0.986), 초식(0.616), 신경관(0.616), 등마루뼈(0.581), 두색류(0.493), 미색류(0.493), 연등(0.493), 다세포(0.443), 성체(0.443), 편형(0.410), 윤형(0.410), 연체(0.399), 절지(0.396), 분류학(0.394), 생태학(0.328), 혈관계(0.328), 선형(0.317), ...	빨아먹다(0.246), 울부짖다(0.232), 기어다니다(0.116), 뛰리다(0.098), 배다(0.083), 길들이다(0.075), 기다(0.072), 할퀴다(0.061), 성내다(0.058), 잡아먹다(0.052), 사납다(0.049), 달려들다(0.044), 부르짖다(0.025), 기어가다(0.022), 살리다(0.022), 물다(0.020), 지르다(0.020), 날아다니다(0.019), 쫓다(0.019), 가두다(0.017), ...

위의 〈표 4〉에서 괄호안의 수치는'동물류'어휘와 해당 의미 정보들 사이의 연관된 정보량을 나타내는 상호정보량(MI)이다. 어휘'동물류'에서 명사(NNG) 경우'상배엽성, 후식, 후생, 낭배, 신경관, 포배'등의 의미 정보가 높은 수치로 측정되었고, 용언류(VV)의 경우도'빨아먹다, 울부짖다, 기어다니다, 물리다'등과 같은 의미 정보가 변별력이 높은 의미 정보로 측정되었다. 이들은 사전 뜻풀이상에 출현한 빈도수와 공기 빈도수는 적지만, 상호정보량은 높게 측정되었다. 따라서, 쓰임이 매우 일반적이지만 실제로 분류를 할 때는 오히려 해당 어휘와의 관계가 낮은 의미 정보들을 걸러낼 수 있게 된다.

4.2 개념 어휘 연관망 구축

앞에서 구한 개념 어휘와 의미 정보간의 연관도를 속성 벡터 정보로 이용하여 개념어들간의 거리를 측정하여 망을 구축하는 정보로 이용한다.

개념 어휘간의 유사도 측정은 상대 엔트로피(Kullback-Leiber distance : DKL)를 이용하여 구하는데, 상대 엔트로피는 확률 분포의 각 시점에서 확률 분포 p(x)와 q(x) 사이의 평균적인 차이를 측정하는 것이다. 그 수식은 다음과 같이 정의 된다.

$$D(p\|q) = -\sum_x p(x)[\log\frac{1}{q(x)} - \log\frac{1}{p(x)}]$$

(수식 2)

실제 발생한 상호정보량을 전체 공기쌍에 대한 상호정보량으로 나눈 것을 개념어와 의미 정보 사이의 연관계수라고 정의하고 이를 개념어에 대한 속성으로 정의하면 다음과 같다.

$$p(m^i|n) = \frac{MI(m^i, n)}{MI(\sum_{j=1}^{k} m^i, n)}$$

(수식 3)

개념 어휘(n_1)과 (n_2)에 대한 분포 정보 $d(n_1)$과 $d(n_2)$가 주어질 때, 상대 엔트로피에 의한 두 분포간의 유사도는 다음과 같다.

$$D_1(d(n_1)\|d(n_2)) = -\sum_{i=1} kp(m_i|n_1)\log\frac{p(m_i|n_1)}{p(m_i|n_2)}$$

(수식 4)

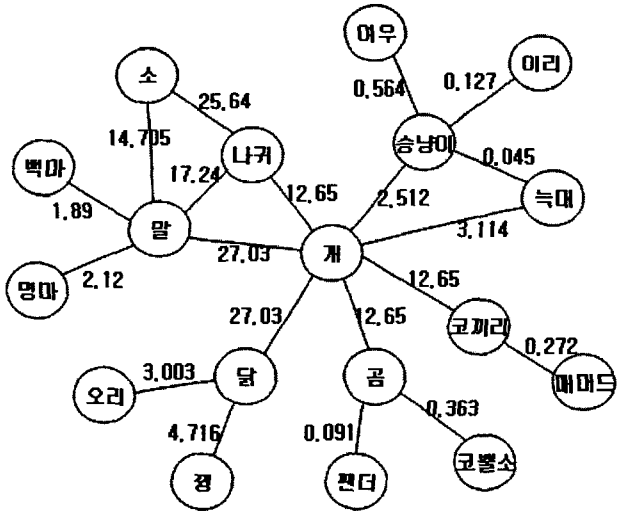
(수식 4)에서 n_1과 n_2는 서로 다른 개념 어휘를 나타내며, m은 의미 속성 정보이다. 따라서, 동일한 의미 속성 정보에 대해 상이한 개념 어휘들간의 유사도를 측정하게 되는 것이다. 유사도는 거리 값의 의미를 가지게 되므로 그 값이 작을수록 어휘간에 유사성이 높음을 나타낸다. 다음은 측정된 개념 어휘들간의 유사도 거리

중에 일부를 나타낸 것이다.

〈표 5〉 '동물' 부류의 개념어간의 유사도 거리

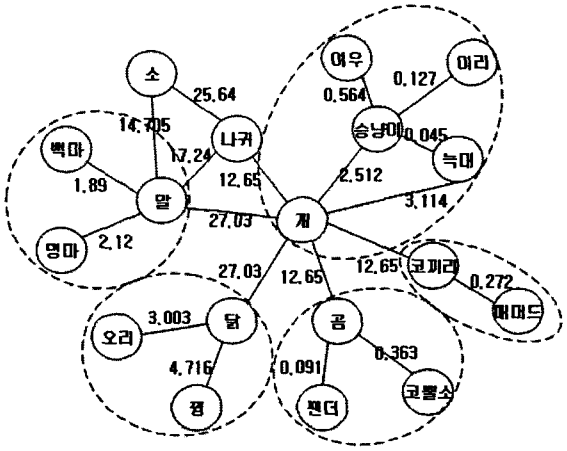
승냥이	늑대(0.045)-이리(0.127)-여우(0.564)-개(2.512)
코끼리	매머드(0.272)-코뿔소(0.363)-곰(1.821)-범(2.638)-고양이(3.095)-쥐(6.666)-소(37.03)
닭	참새(2.403)-오리(3.003)-꿩(4.716)-암탉(8.403)-돼지(8.711)-참새(10.0)-병아리(10.204)-제비(12.048)-학(13.88)-호랑이(20.0)-거북(21.739)-토끼(27.027)-소(34.482)
원숭이	긴꼬리원숭이(0.109)-명주원숭이(0.109)-늙보원숭이(0.218)-고릴라(0.327)-유인원(0.982)-다람쥐(1.091)-말(1000)
말	백마(1.923)-명마(2.123)-경마(4.672)-꼰소(12.82)-낙타(17.24)-당나귀(17.24)-양(26.315)-돼지(27.77)-길마(37.037)-노루(38.461)

위의 〈표 5〉에서 알 수 있듯이, 서로 연관 관계가 있는 개념 어휘들의 거리가 가깝게 나타난다. 상호정보량의 단점은 빈도수가 빈약한 정보를 확대 해석하는 오류를 범하지만, 이러한 상호정보량의 최대값과 최소값의 비율로 나타내면 보다 오류를 줄 일 수 있다. 다음[그림 5]는 이들 개념 어휘들의 거리 관계를 연관망으로 도식화한 것이다.



[그림 5] '동물' 부류 개념어들의 연관망

위의 연관망을 보면, 어휘들간의 유사도에 따라 분류되고 있음을 알 수 있다. 즉, 대부분의 개념 어휘들이 표시된 것처럼 {승냥이-여우-이리-늑대-개}, {백마-명마-말}, {닭-오리-꿩}, {곰-팬더-코뿔소}, {코끼리-매머드} 의 묶음으로 분류되어 나타난다. 즉, 이들 무리는 서로간의 유사도 거리 값이 작은 것끼리 클러스터를 이루고 있다.



[그림 6] '동물' 부류 개념어들의 클러스터링

5. 결 론

본 논문에서는 사전 뜻풀이말을 기반으로 의미 정보를 추출하고 이를 이용하여 개념 어휘의 유사도를 측정하였다. 사전의 구조적인 특성을 이용하여 개념 어휘의 의미 참조 모형을 자동으로 추출할 수 있었고, 이 의미 참조 모형의 의미 정보와 공기하여 나타나는 어휘들을 의미 정보로 추출하면 개념 어휘에 대해 많은 수의 의미 정보가 구축될 수 있었다. 기존에 어절 단위로 공기하는 어휘나 구문상의 관련성을 가지고 공기하는 어휘들만을 다루었을 때는 자료의 희귀성으로 인해 개념 어휘의 특성이 잘 드러나지 않았는데, 사전의 표제어와 뜻풀이말을 이용하면 뜻풀이말 문장 자체가 공기 정보의 단위가 되므로 보다 의미 정보가 확장되어 자료의 희귀성을 어느 정도 극복할 수 있었다. 또한, 어떤 어휘는 코퍼스에서 한 번 사용된 의미 정보와 공기하여 나타나기도 하는데 이렇게 빈도수가 극도로 적은 경우에 기존 연구에서는 제외되거나 변별력을 갖지를 못했는데, 본 연구에서는 보다 잘 드러났다.

본 논문은 약간의 개선해야 할 점이 있지만 개념 어휘들을 명사와 동사를 막론하고 사전 뜻풀이말을 단위로 의미 정보를 추출하고 그 정보를 이용하여 유사도를 구했을 경우 보다 관련 있는 모습으로 개념어휘들이 분류될 수 있음을 보여주고 있다.

앞으로 본 연구에서 자동으로 구축된 의미 정보를 실용적인 시스템에서 이용할 수 있게 보다 체계적이고 확장된 연구가 필요하다. 또한, 본 연구는 사전 뜻풀이말이 하나의 표제어를 완벽하게 기술하고 있다는 가정을 두고 실험을 하였기 때문에, 실생활에서 사용되는 모습과는 차이가 있을 수 있다. 따라서, 사전의 예문이나 다

양한 코퍼스자료를 이용한다면 보다 정교한 언어 지식을 획득하고 지식표현을 구축할 수 있다고 보고 이러한 자료들도 연구 대상으로 확장하려고 한다.

6. 참고 문헌

[1] Alpha k, Luk. 1995. "Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions", 33rd Annual Meeting of the ACL

[2] 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.

[3] 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.

[4] Hindle,D. 1990. Noun Classification From Predicate-Argument Structures. In Proceedings of the 28st Meeting of the Association for Computational Linquistics.

[5] Fano,R.1961.Transmission of Information.In Cambridge, Mass : MIT Pres.

[6] Grefenstette,G.1993.Evaluation techniques for automatic semantic extraction : comparing syntactic and window-based approaches. Technical Report, Department of Computer Science, University of Pittsburgh

[7] Ido Dagan. 1995.「 Contextual Word Similarity and Estimation form Sparse Data」.

[8] J. Hur, C.-Y. Ock. 2001.「A Homonym Disambiguation System based on Semantic Information extracted from Definition in Dictionary」. 19th ICCPOL 2001.

[9] McMahon, J. and Smith, F.,Improving statistical langualge model performance with automaticallygenerated word hierarchies, Computational Lingyistics.

[10] Pereira,F.,Tishby,N. and Lee,L.,Dstributional clustering of English words, Proceeding of the 31st Annual Meeting of ACL.

[11] Schutze,H.1992.Word Sense disambiguation with suble-xical representations.In Workshop Notes, Statistically-Based NLP Techniques.AAAI

[12] 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.

[13] Yarowsky,D.1992.Word-sense disambiguation using statistical model of Roget's categorties trained on large corpora. In Proceedings of COLING-92.

[14] 김영택,「자연언어처리」, 교학사, 1994.

[15] 김병희, 2002, "정보이론 기반의 병합식 클러스터링 기법 연구"

[16] 김봉주, 개념학, 한신문화사. 1992

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

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

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

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

[21] 송도규, 인지언어학과 자연언어 자동처리, 홍릉과학출판사, 1996.

[22] 옥은주, 2003, 사전의 의미 정보와 격틀을 이용한 단어의 연관도 측정과 의미 클러스터링.

[23] 이기동.「언어와 인지」, 한산문화사, 1992

[24] 임지룡, 김동환. 1998.「인지언어학개론」.한신문화사.

[25] 임지룡. 1997.「인지의미론」.탑출판사.

[26] 정연수, 조정미, 김길창. 1995.「개념분류기법을 적용한 한국어 명사분류」. 제7회 한글 및 한국어 정보처리 학술대회.

[27] 조정미, 김길창. 1995,「분포정보를 이용한 의미 중의성을 지닌 한국어 동사의 의미 분별」. 제7회 한글 및 한국어 정보처리 학술대회.

[28] 조정미. 1998. 코퍼스와 사전을 이용한 동사 의미 분별. 한국과학기술원 박사학위논문.

[29] 조평옥, 옥철영. 1999.「사전 뜻풀이말에서 구축한 한국어 명사 의미계층구조」. 한국인지과학회 논문지 제10권 제4호.

[30] 허정. 2000. 사전 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템. 울산대학교 석사학위논문.