어휘망(U-WIN)의 구문관계 자동구축

(Automatic Construction of Syntactic Relation in Lexical Network(U-WIN))

임 지 희 † 최 호 섭 †† 옥 철 영 †††
(Jihui Im) (Hoseop Choe) (Cheolyoung Ock)

요 약 본 연구에서는 사용자 어휘지능망(U-WIN)의 어휘 관계 중의 하나인 구문관계를 자동으로 구축하는 방법을 제시하고자 한다. 먼저, 구문관계를 형성할 수 있는 후보명사를 용언의 용례에서 문형 정보를 기준으로 추출함으로써, 용언의 세분화된 의미별로 정확하고 다양한 후보명사를 추출할 수 있다. 그러나 추출된 후보명사는 다양한 의미를 지니고 있으므로, 어휘 간의 명확한 구문관계를 설정하기 위해서는 후보명사의 여러 의미 중에서 정확한 의미로 결정해야 한다. 그래서 본 연구에서는 용례 매칭 규칙, 구문패턴, 의미 유사도, 빈도 정보 등을 이용하여 후보명사의 의미를 분별한다. 또한 구문패턴의 빈도 정보를 이용하여 용례에 나타나지 않지만 구문관계를 형성할 수 있는 명사를 추출하여 구문관계를 확장하고자 하였다. 이러한 연구는 명사 중심의 어휘망이 용언과의 구문관계 구축을 통해 형태소 분석, 구문 분석, 의미 분석 등에 광범위하게 활용할 수 있는 어휘망의 기반을 다지는 작업이 될 수 있을 것이다.

키워드 : 어휘망, U-WIN, 구문관계, 구문패턴, 의미유사도

Abstract An extended form of lexical network is explored by presenting U-WIN, which applies lexical relations that include not only semantic relations but also conceptual relations, morphological relations and syntactic relations, in a way different with existing lexical networks that have been centered around linking structures with semantic relations. So, This study introduces the new methodology for constructing a syntactic relation automatically.

First of all, we extract probable nouns which related to verb based on verb's sentence type. However we should decided the extracted noun's meaning because extracted noun has many meanings. So in this study, we propose that noun's meaning is decided by the example matching rule / syntactic pattern / semantic similarity, frequency information. In addition, syntactic pattern is expanded using nouns which have high frequency in corpora.

Key words : Word Network, User-Word Intelligent Network, Syntactic Relation, Syntactic Pattern, Semantic Similarity

· 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT 연구센터 육성지원사업의 연구결과로 수행되었습니다(IITA-2008-(C1090-0801-0039)).

† 학생회원 : 울산대학교 컴퓨터정보통신공학과
arisu80@ulsan.ac.kr
†† 정 회 원 : 한국과학기술정보연구원 정보기술개발단 선임연구원
hschoe@kisti.re.kr
††† 종신회원 : 울산대학교 컴퓨터정보통신공학부 교수
okcy@ulsan.ac.kr
(Corresponding author임)

논문접수 : 2008년 1월 29일
심사완료 : 2008년 8월 29일

정보과학회논문지: 소프트웨어 및 응용 제35권 제10호(2008.10)

1. 서 론

최근 의미적 언어자원에 대한 관심이 증가하여, 의미부류, 어휘망, 시소러스, 온톨로지 등이 핵심 연구 대상으로 부각되고 있다. 그중에서 어휘망은 한 어휘가 다른 어휘와 가지는 다양한 관계를 망(Network) 형태로 나타내어 데이타베이스화한 것으로, 어휘망의 활용은 언어자원의 효율적인 관리, 의미적 자연언어처리 기술 향상 등의 기대효과가 있다.

2002년부터 개발 중인 한국어 사용자 어휘지능망(User-Word Intelligent Network, 이하 U-WIN)[1]은 한국어의 공통적이고 개별적인 속성을 바탕으로 한국인의 보편적인 인지 체계와 개념 관계를 파악하여 이를 어휘의 의미적·개념적 네트워크로 형성한 대규모 어휘

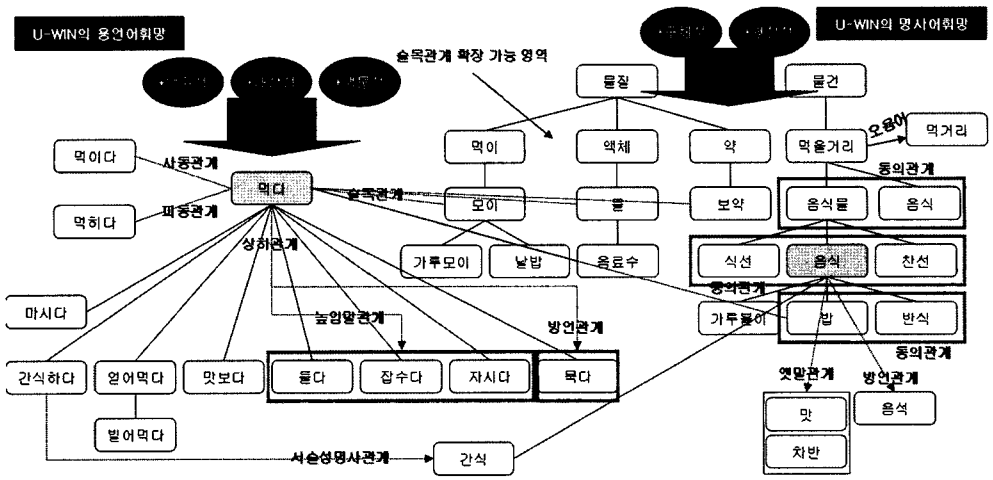


그림 1 U-WIN의 구축 사례

망이라 할 수 있다. 특히 일반적인 어휘망이 의미 관계에 의한 연결 구조를 중심으로 연구 개발된 것과는 달리, U-WIN은 의미관계를 비롯하여 개념 관계, 형태 관계, 구문 관계 등과 같이 의미 관계의 범위를 확장한 어휘 관계를 적용함으로써 어휘망의 확장적 형태를 모색하고자 한다. 예를 들어 그림 1은 동사 '먹다'를 중심으로 상하관계·동의관계의 의미관계, 사동관계·피동관계·방언관계의 형태 관계, 술목관계의 구문관계 등으로 어휘 간의 다양한 관계를 표현하고 있다. 그 중에서 형태관계와 구문관계는 U-WIN을 구성하는 어휘 집합이 모든 품사를 대상으로 함으로써 고려한 어휘 관계로서 형태소 분석 및 구문 분석에 활용할 수 있을 것이다.

본 연구에서의 구문관계[1]는 'N이(가) P', 'N을(를) P', 'N에(에서, 으로) P' 등의 술주관계, 술목관계, 술부관계를 말하며, 용언 'P'를 중심으로 'N'[2]을 추출하여 자동으로 구문관계를 구축하고자 한다. 그러나 U-WIN은 다의어의 뜻풀이 하나하나를 개별적인 어휘로 구분하여 모든 관계를 설정하였으므로, 'N'의 의미를 다의어 수준으로 분별해야 한다. 예를 들어, 문장 "밥을 먹다"에서 '먹다'를 중심으로 '밥'을 추출하여 술목관계를 구축할 수 있으나, '밥'[3]의 19개 의미 중에서 "음식 따위를 입을 통하여 배 속에 들여보내다"의 의미임을 분별해야 하는 것이다.

그러므로 본 논문에서는 용언을 중심으로 구문관계를 형성할 수 있는 의미적 조응 관계가 성립하는 후보명사를 추출하고, 후보명사의 의미 분별을 통해 자동으로 구문관계를 구축하는 방법을 모색하고자 하였다. 의미적 조응 관계가 성립하는 후보명사는 선택 제약이나 결합 관계 등의 정보를 제공하는 용례와 구문정보를 제공하는 문형정보를 이용하여 추출하였으며, 추출한 후보명사는 용례 매칭 규칙, 구문패턴, 의미 유사도 등을 적용하여 의미를 분별하였다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 어휘망 구축과 의미 유사도 측정방법에 관한 관련 연구를 설명하고, 3장에서는 본 논문에서 제시하는 구문관계 후보명사 추출 방법과 후보명사의 다의어 수준의 의미 분별 및 구문관계 확장 방법에 대해 설명한다. 그리고 4장에서는 이러한 방법으로 구축한 구문관계 결과를 제시·분석하고, 마지막으로 5장에서 결론을 도출한다.

2. 관련 연구

2.1 어휘망 구축 관련 연구

세종 명사 의미부류 체계[2]는 한국어 명사 어휘들의 의미영역을 엄격하고 정밀하게 분할하고, 각각의 의미영역에 대해 이를 공유하는 명사 어휘들과 해당 의미영역의 정의에 근거가 되는 어휘와 적정술어들을 묶어 놓은 위계적 어휘·의미 분류체계이다. 최상위부류는 <구체물>, <집단>, <장소>, <추상적대상>, <사태> 등 5개의 대부류로 나눈 다음, 각각을 분할하여 보다 세밀한 의미영역을 지닌 2~7층위의 하위부류를 포함하고

1) 본 논문에서 말하는 술주관계, 술목관계, 술부관계의 구문 관계는 각각 다음과 같다. 술주관계는 서술어와 주어(명사/대명사...)의 역할을 담당하는 어휘를 연결하는 관계, 술목관계는 서술어(타동사)와 목적어 역할을 담당하는 어휘를 연결하는 관계, 술부관계는 부사어를 요구하는 서술어와 부사어 역할(부사/체언)을 하는 어휘를 연결하는 관계로 정의한다.

2) 'N'은 술주관계, 술목관계, 술부관계 등의 구문관계를 형성할 수 있는 명사로서, 일반적으로 논항이라고 표기한다. 그러나 본 연구에서는 다의어로 의미가 결정되지 않은 'N'은 구문관계를 형성할 수 있는 명사라고 하여, 이하 후보명사라고 한다.

3) '밥'은 2개의 동형이의어로 구성되며, 다시 19개의 다의어로 의미가 세분화되어 있다.

있다. 이때 <사태> 부류는 술어명사의 의미부류이고, 나머지 4개는 논항명사의 의미부류이다.

KAIST CoreNet(코어넷)[3]은 3,000여개의 개념이 한국어-중국어-일본어로 대응되어 동일한 개념체계를 공유한 다국어 어휘의미망이며, 그 개념의 기본은 일본 NTT(Nippon Telegraph Telephone Corporation)의 어휘대계를 바탕으로 하고 있다. 또한 한 개념 체계 아래에 명사, 동사, 형용사의 어휘들이 함께 나타나며, 다시 각 개별 단어는 사전 뜻풀이와 서술어의 경우에는 격틀 정보를 보여주고 있다.

ETRI 명사개념망[4]은 한국어 명사 단어들을 의미관계로 연결시켜 놓은 어휘 데이타베이스이다. 구축된 명사개념망의 규모는 약 4만 9천여 단어로서 31개의 최상위 레벨의 어휘와 깊이 12레벨로 구성되어 있으며, 백과사전 질의응답 시스템으로의 활용을 위해 약 25여 만 고유명사들이 Instance_Of 관계로 연결되어 있다.

울산대의 U-WIN[1]는 일반적인 어휘망이 가지는 의미 관계를 비롯하여 개념 관계, 형태 관계, 구문 관계 등과 같이 의미 관계의 범위를 확장한 어휘 관계를 적용하고, 의미 정보, 확장 정보 등의 다양한 정보를 포함하고 있는 확장된 어휘망이다. U-WIN은 명사, 동사, 형용사를 중심으로 한국어 어휘의 모든 품사 및 언어 단위를 대상으로 구축하여, 현재 46만 여개의 관계정보가 구축된 상태이다.

U-WIN을 제외한 어휘망에 관한 연구들은 상하관계 및 의미관계에 대해 구축하였으며, 대상 어휘들은 동형이의어 수준에 그치고 있다.

2.2 의미 유사도 측정방법

2.2.1 링크(Link) 기반 측정방법

링크를 기반으로 한 개념의 의미 유사도 측정방법은 개념간의 최단 경로 수를 계산하거나, 계층 구조상에서의 개념의 깊이, 관계 종류 등을 고려할 수 있다.

링크 기반 측정 방법에는 Rada, et. al[5], Leacock and Chodorow[6], Wu and Palmer[10], Hirst and St.Onge[11] 등이 있다.

Rada, et. al[5]는 가장 기본적인 방법으로, 의학용어 시소러스(MeSH)를 대상으로 두 개념 간 최단 경로의 링크(Link) 개수를 기반으로 유사도를 측정하였다.

Leacock and Chodorow[6]는 Rada, et. al.의 방법에 계층 구조의 깊이를 함께 고려하여 식 (1)을 이용하였다.

$$Sim_{lch}(c_1,c_2) = max\left[-log\left(length(c_1,c_2)/(2 \cdot D)\right)\right] \quad (1)$$

length(C1,C2)는 계층 구조에서 두 개념을 연결하는 최단 경로의 링크(Link) 개수이고, D는 분류체계·시소러스·온톨로지 등의 계층구조의 최대 깊이이다.

그리고 Wu and Palmer[10]와 Hirst and St.Onge [11]은 각각 계층구조의 깊이, 관계종류를 기준으로 유사도를 측정하였다.

2.2.2 정보량(Information Content:IC) 기반 측정방법

정보량(Information Content)은 대용량 말뭉치 내 개념의 발생 빈도를 기반으로 MLE(Maximum Likelihood Estimate)방법으로 얻는다. 많은 정보량이 할당된 개념은 특정 주제에 매우 세부적인 개념이고, 적은 정보량이 할당된 개념은 더 일반적인 개념으로 판단할 수 있다.

$$IC(concept) = -log\left(P(concept)\right) \quad (2)$$

의미 주석 말뭉치가 없을 경우에 개념 발생 빈도는 단어별 빈도를 해당 단어의 동형이의어/다의어 개수로 나누어 할당하거나, 단어별 빈도를 해당 단어의 동형이의어/다의어에 그대로 할당하는 방법을 사용한다.

정보량 기반 측정 방법에는 Resnik[9], Jiang and Conrath[7], Lin[8] 등이 있다. Resnik[9]은 정보량을 사용하여 식 (3)에 의해 유사도를 측정한다.

$$Sim_{res}(c_1,c_2) = IC\left(lcs\left(c_1,c_2\right)\right) \quad (3)$$

lcs(c1,c2)는 개념 c1과 c2의 공통 상위어 중에서 가장 하위에 위치한 개념[4](LCS : lowest common subsumer)를 의미한다. 식 (3)에 의해, 부모 노드가 같은 개념들의 유사도는 최소 공통 상위어가 같아서 항상 같은 값을 가진다. 그러나 주로 계층이 큰 덩어리 형태로 이루어진(coarse-grained) 동사 어휘망은 동일한 최소 공통 상위어를 가지는 개념들이 많으므로, Resnik[9]은 가장 좋은 coarse-grained measure로 알려져 있다.

Jiang and Conrath[7]과 Lin[8]는 각각 Resnik 기반의 명사들의 유사도 측정방법과 문서간의 유사도 측정방법 중에 하나인 Dice Coefficient를 이용한 방법이다.

본 논문에서는 식 (1)의 링크 기반 유사도 측정 방법에 최소 공통 상위어의 깊이(Dlcs)를 추가하여 수정하였다. 이것은 계층 구조에서 두 개념을 연결하는 최단 경로의 링크(Link) 개수가 동일하더라도, 상위 계층에 위치하는 어휘들의 링크들보다 하위 계층에 위치하는 어휘들의 링크일수록 어휘간의 유사도가 더 높은 점을 적용하기 위함이다.

3. 구문관계 자동구축 방법

명사와 용언 간의 구문관계를 자동으로 구축하는 개괄적인 과정은 그림 2와 같다.

본 논문에서 사용하는 U-WIN 어휘 사전 데이타베이스의 문형정보는 구문정보를 제공하고, 용례는 선택 제약이나 결합 관계 등의 정보를 추출할 수 있는 말뭉치 역할을 담당하고 있다. 그러므로 용언의 용례에서 문형정보에 해당하는 논항을 추출하여, 해당 용언과 구문관

4) 이후 개념 c1과 c2의 공통 상위어 중에서 가장 하위에 위치한 개념을 '최소 공통 상위어'라고 한다.

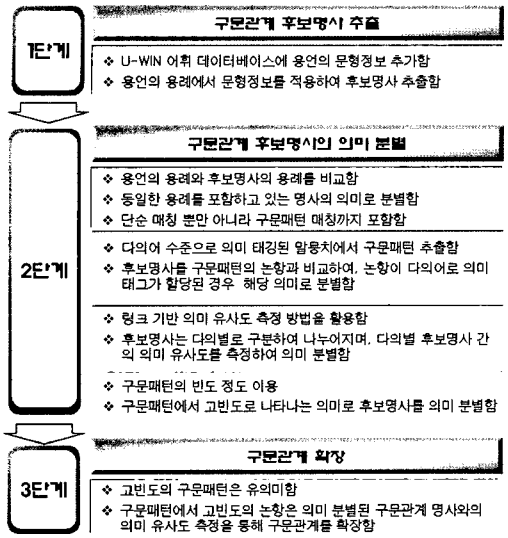


그림 2 명사-용언 구문관계 자동구축 과정

계를 형성할 수 있는 후보명사로 이용함으로써, 용언의 세분화된 의미별로 다양하고 명확한 후보명사를 추출할 수 있다. 그러나 U-WIN은 다의어를 개별 어휘로 판단하고 있어, U-WIN의 구문관계 자동구축을 위해서는 후보명사를 다의어 수준으로 의미 분별을 해야 한다.

그래서 본 연구에서는 용례 규칙, 의미 주석 말뭉치에서 추출한 구문패턴, 의미 유사도 등을 차례로 적용하여 후보명사의 의미를 분별하였으며, 특히 구문패턴은 다의어 수준으로 의미 태그가 부착된 말뭉치에서 추출함으로써 의미 태그가 부착된 명사는 후보 명사의 의미 분별 및 구문관계 확장에 활용하였다.

3.1 구문관계 후보명사 추출

기존의 U-WIN 어휘 사전 데이타베이스는 [표준국어대사전]을 기반으로 표 1과 같이 명사를 중심으로 어휘 내적 정보를 설계 및 구축되어 있다.

그러나 본 연구에서는 후보명사를 추출하는 기준으로 각 용언의 문형정보[5]를 활용하였다. 그러므로 본 연구에 맞게 표 2와 같이 U-WIN 어휘 사전 데이타베이스에 문형정보 항목을 추가하고, [표준국어대사전]에서 용언의 다의어별 문형정보를 추출하여 문형정보 항목에 할당하였다.

그리고 표준국어대사전 편찬 지침[12]를 살펴보면 "용례는 뜻풀이를 이해하거나 실제로 사용하는 데에 도움을 줄 수 있고 가급적 선택 제약이나 결합 관계 등을 보여 주는 전형적인 것이어야 하며, 특히 명사 표제어라

5) 문형정보는 문장을 구성하는 데 있어서, 각 용언이 요구하는 필수적 성분들, 즉 논항의 순서 있는 목록을 말한다.

표 1 U-WIN 어휘 사전 데이타베이스의 어휘 내적 정보 구조

'어휘 내적 정보' 항목	설 명
식별자(Identifier : ID)	각 어휘가 가지는 식별자
의미 코드(Sense Code)	각 어휘가 가지는 의미 관리 코드
어휘소(Lexme)	어휘의 형태
의미 표지(Sense Tag)	동형이의어/다의어 정보
일차 분석 어휘소 (Analyzed lexme)	일차 형태소 분석 결과 및 띄어쓰기 정보
뜻풀이(Definition)	어휘의 뜻풀이 정보
품사(Parts-Of-Speech)	어휘의 품사 정보
한자(Chinese Character)	어휘의 한자 정보
원어(Original Word)	외래어의 원어 정보
대역(Translation)	어휘의 대역 정보
전문분야(Domain)	전문용어 분야 정보
용례(Example)	어휘의 용례 정보
출처(Source)	어휘의 출처 정보

표 2 문형정보 추출 및 할당(예-'먹다')

의미표지	뜻풀이	문형정보
먹다_001000	귀나 코가 막혀서 제 기능을 하지 못하게 되다.	(-을)
먹다_002001	음식 따위를 입을 통하여 배 속에 들어보내다.	-을
먹다_002002	담배나 아편 따위를 피우다.	-을
먹다_002003	연기나 가스 따위를 들이마시다.	-을
먹다_002004	어떤 마음이나 감정을 품다.	-을
먹다_002005	일정한 나이에 이르거나 나이를 더하다.	-을
먹다_002006	욕, 핀잔 따위를 듣거나 당하다.	-을
먹다_002007	(속되게) 뇌물을 받아 가지다.	-을
먹다_002008	수익이나 이문을 차지하여 가지다.	-을
먹다_002009	물이나 습기 따위를 함유하다.	-을
먹다_002010	어떤 등급을 차지하거나 점수를 따다.	-을
먹다_002011	구기 경기에서, 점수를 얻다.	-을
먹다_002012	(속되게) 여자의 정조를 유린하다.	-을
먹다_002013	해 따위를 맞다.	-을
먹다_002014	남의 재물을 다루거나 맡은 사람이 그 재물을 부당하게 자기의 것으로 만들다.	-을
먹다_002015	날이 있는 도구가 소재를 깎거나 자르거나 갉거나 하는 작용을 하다.	-에
먹다_002016	바르는 물질이 배어들거나 고루 퍼지다.	-에
먹다_002017	벌레, 균 따위가 파 들어가거나 퍼지다.	-에
먹다_002018	돈이나 물자 따위가 들거나 쓰이다.	-에

면 함께 쓰이는 서술어의 대표 예가, 동사 표제어라면 함께 쓰이는 명사의 대표 예가 충분히 제시될 필요가 있다"고 기술되어 있다. 그러므로 용례는 논항 정보를 추출할 수 있는 대표적인 말뭉치라고 할 수 있다.

또한 용언 표제어의 용례는 문형정보별로 제시되어 있으므로, 용례에서 문형정보를 기준으로 구문관계를 형성할 수 있는 의미적 조응 관계가 성립하는 후보명사를 추출하였다.

기구축된 선택제약 정보를 살펴보면, '먹다_2'는 '밥, 술, 음식, 나이, 약'의 선택제약 정보를 가지고 있다. 그러나 동형이의어 '먹다_2'의 의미는 그림 3과 같이 18가지로 세분화되어 있으며, 각 의미별로 구분되는 선택제약 정보를 가지고 있지만, 이러한 점을 반영하지 못하고 있다.

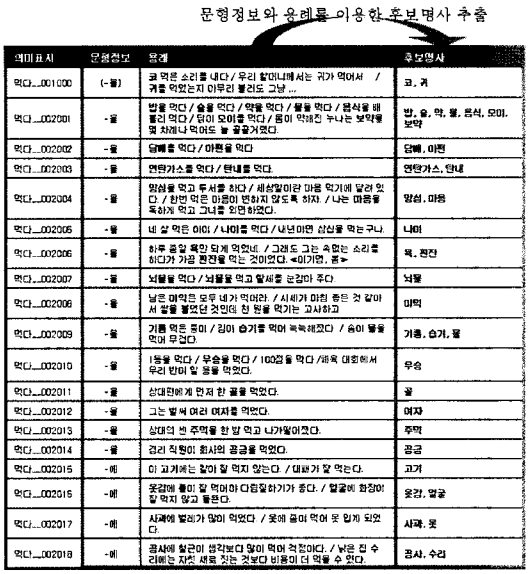


그림 3 문형정보와 용례를 이용한 후보명사 추출 (예-'먹다')

그러므로 본 논문에서는 그림 3과 같이 '먹다_002001'[6]와 '먹다_002005'의 문형정보("...을")를 기준으로 용례에서 각각 {밥, 술, 약, 물, 음식, 모이, 보약}과 {나이}의 의미적 조응 관계가 성립하는 후보명사 목록을 구별하여 추출하였다.

3.2 구문관계 후보명사의 의미 분별

U-WIN은 다의어를 개별 어휘로 판단하고 있어, U-WIN의 구문관계 자동구축을 위해서는 추출한 후보명사를 다의어 수준으로 의미 분별을 해야 한다.

3.2.1 용례 매칭 규칙 적용

명사/용언 표제어의 용례에는 표제어와 함께 쓰이는 대표적인 서술어/명사를 포함하고 있으므로, 이렇게 함께 쓰이는 대표적인 명사와 용언은 각각 같은 용례를 가지고 있는 경우가 많다.

예를 들어 그림 4와 같이, 용례 "약을 먹다"는 '약_007001'과 '먹다_002001'의 용례에 각각 나타난다. 그러므로 동사 '먹다_002001'의 용례와 의미 분별되지 않은 후보명사 '약'의 용례를 차례로 비교함으로써, 동일한 용례 "약을 먹다"를 포함하고 있는 '약_007001'의 의미로 후보명사의 의미를 결정할 수 있다.

또한 명사와 용언이 동일한 용례를 포함하는지 비교하는 과정은 용례 전체를 비교하는 단순 매칭뿐만 아니라, 용례에서 추출한 구문패턴의 매칭도 함께 고려해야 한다. 용언 표제어의 용례는 용언의 활용형이 포함되거나 또는 "숟가락으로 밥을 떠 먹다"의 문장의 '숟가락으로' 등의 수의적 성분을 포함하는 경우도 있기 때문이다.

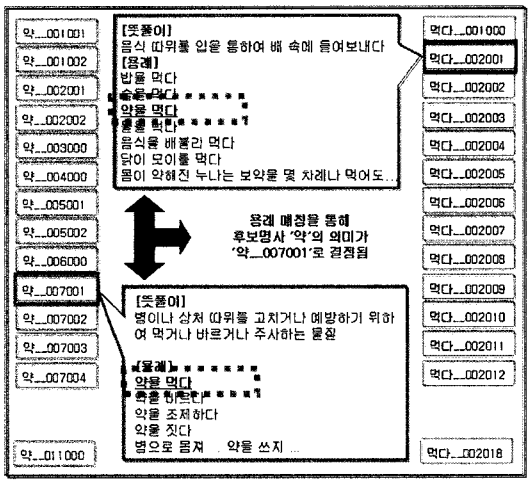


그림 4 용례 매칭을 통한 의미 분별의 예 ('약_007001'의 의미 분별)

그러므로 용례 매칭 규칙을 통한 의미 분별 과정은 우선 후보명사를 추출한 용언의 용례('약을 먹다_002001')를 기준으로 후보명사('약')가 가지는 다양한 용례를 단순매칭, 구문패턴의 매칭을 단계적으로 수행하여 이루어진다.

이 과정에서는 하나의 의미를 가지는 어휘를 추출하여 분별 대상어휘를 줄임으로써, 다음 단계에서의 의미 분별 과정을 용이하게 하도록 한다.

3.2.2 구문패턴 적용

구문패턴은 U-WIN 어휘 사전 데이타베이스의 의미 주석된 뜻풀이 말뭉치[7]에서 추출하였으며, 기존의 동형이의어 수준이 아닌 다의어 수준으로 의미 태그가 부착되어 있어 활용성이 높다.

구문패턴 추출 방법은 중심어 후행의 원리에 의해, 본용언과 본용언 사이로 추출 범위를 지정하여 용언의 인접어절 정보를 중심으로 추출하였다. 용언의 선행 요소 중에서 주격조사(JKS), 부사격조사(JKB), 목적격조사(JKO)를 취하는 어절만을 추출하고, 용언의 후행 요소 중에서 어미와 다음어절을 추출하였다. 선행요소가 보조

6) 본 논문에서는 동형이의어 및 다의어 수준까지 의미가 세분화된 표제어를 나타내기 위해 '[표제어]_[동형이의어번호]{다의어번호}'의 형식에 따라 '먹다_002001'로 표기하기로 한다.

7) 본 논문에서 사용하는 의미 주석 말뭉치(410만 어절)는 어휘 사전 데이타베이스의 뜻풀이에 수작업으로 다의어 수준의 의미 태그를 부착하여 구축하였다. 의미 주석 말뭉치는 1차적으로 말뭉치에 나타난 41,000여개의 동형이의어(가다, 오다, ...)를 대상으로 의미태그를 부착하여 구축하였고, 2차적으로 말뭉치에 나타난 34,000여개의 다의어(사람, 인물, ...)를 대상으로 의미태그를 부착하여 구축하고 있다. 현재 2차 구축 작업이 95%정도 진행되었으며, 향후 구축이 완료된 의미 주석 말뭉치는 다양하게 활용 가능할 것이다.

사(JX)를 취하는 어절은 그 격의 모호성 때문에 추출대상에서 제외하였다.

위와 같은 방법으로 추출한 구문패턴 중에서 동사 '먹다_002001'는 {음식, 약_007001, 고기_001001, 밥_001002, 먹이, 밥_001001, 아침밥, 과일_00100, 국수_001001....} 등의 명사들이 목적격 조사(JKO)로 구문패턴을 구성하고 있다. 특히, 구문패턴에서 의미가 결정된 명사 목록은 후보명사와 비교함으로써, 일부 후보명사의 의미를 결정짓도록 한다. 예를 들어 후보명사 중에서 '술, 아편, 마음, 나이, 물'이 구문패턴의 명사 목록에 의해 '술_001000, 아편_001000, 마음_001002, 나이_001000, 물_001001'로 그 의미가 결정되었다.

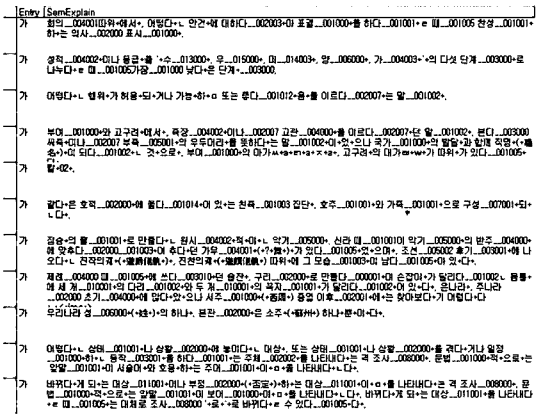


그림 5 의미 주석 뜻풀이 말뭉치의 일부

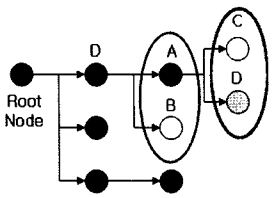
3.2.3 의미 유사도 적용

후보명사는 선택 제약이나 결합 관계 등을 보여 주는 전형적인 용례에서 추출하였으므로, 용언의 의미별로 나누어진 후보명사들은 공통 특성을 가지는 경우가 많다. 이 공통 특성으로 인해 후보명사들이 U-WIN 계층구조 상에서 동일한 부모 노드를 가지거나 가까운 거리에 나타나기도 한다. 예를 들어, '먹다_002009'의 후보명사인 '기름'과 '물'은 "액체"라는 공통된 특성을 가지며, U-WIN의 계층구조에서 '액체_000000'의 하위어로 함께 나타나 있음을 알 수 있다.

그러므로 본 연구에서는 링크기반 의미 유사도를 적용하여 후보명사의 의미를 결정한다. 본 논문에서는 식 (1)의 링크 기반 유사도 측정 방법을 식 (4)와 같이 최소 공통 상위어의 깊이(Dlcs)를 추가하여 수정하였다.

$$similarity(c_1,c_2)=\max\left[-\log\left(\frac{length(c_1,c_2)}{2*D}*\frac{1}{D_{lcs}}\right)\right] \quad (4)$$

이것은 계층 구조에서 두 개념을 연결하는 최단 경로의 링크(Link) 개수가 동일하더라도, 상위 계층에 위치하는 어휘들의 링크들보다 하위 계층에 위치하는 어휘들의 링크일수록 어휘간의 유사도가 더 높은 점을 적용하기 위함이다. 예를 들어, 그림 6과 같은 계층구조에서 식 (1)을 통한 similarity(A,B)와 similarity(C,D)는 동일한 값을 가지게 된다. 그러나 식 (4)를 통해서는 similarity(C,D)가 similarity(A,B)보다 높게 나타남을 알 수 있다.



length(C,D) = 2, length(A,B) =2
Dlcs(C,D) = 3 , Dlcs(A,B) = 2

➔ *similarity(C,D) > similarity(A,B)*

그림 6 의미유사도 측정 예

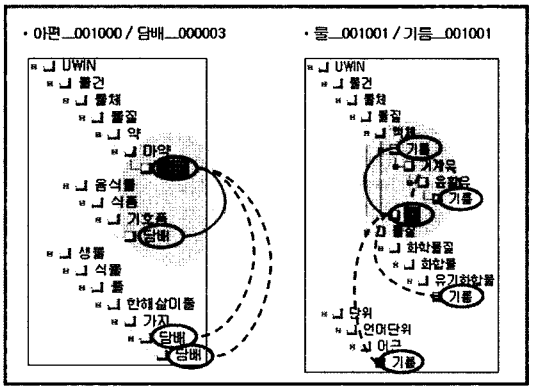


그림 7 의미 유사도에 의한 후보명사의 의미 분별

그리고 1,2단계 의미 분별 과정을 거치며 의미가 결정된 '아편_001000', '물_001001' 은 의미 유사도 측정을 할 때, 기준점으로 작용하여 의미 분별을 용이하게 한다. 예를 들어, '먹다_002009'의 후보명사 {'기름', '습기', '물'}은 2단계 의미 분별 과정을 거쳐 {'기름', '습기', '물_001001'}이 되었다. 그리고 '기름'은 '물_0001001'과 "U-WIN>물건>물체>물질>액체>"의 최소 공통 상위어를 가지며, 가장 큰 값의 의미 유사도를 가지는 '기름_001001'로 의미를 결정할 수 있다.

3.2.4 빈도 정보 이용

의미 유사도에 의한 후보명사의 의미 분별은 다음과 같은 경우에는 적용할 수 없다. 후보명사가 U-WIN의 계층구조로 표현되어 있지 않거나, 다의어별 용례에서 추출한 후보명사가 '뇌물', '우승', '이익', '골', '여자', '주먹' 등과 같이 하나만 존재하는 경우이다. 또한 '기름(물

건)', '습기(작용)'과 같이 두 후보명사의 의미 유사도가 작은 경우에도 후보명사의 의미를 결정하지 못한다.

이와 같은 경우 앞서 추출한 구문패턴에서의 고빈도 의미를 선택함으로써, 의미가 결정되지 않은 나머지 후보명사의 의미를 분별하였다. 예를 들어, '먹다_002006'의 후보명사 '욕'과 '핀잔' 중에서 '욕'은 1단계 의미 분별 과정에서 '욕_002001'의 의미로 결정되었지만, '핀잔'은 '먹다'의 구문패턴에 나타나지 않고, U-WIN의 계층 구조로 표현되어 있지 않아 의미 유사도를 구할 수 없다. 이러한 경우 전체 구문패턴에서 고빈도로 나타난 '핀잔_001000'으로 의미를 결정하도록 하였다.

3.3 구문관계 확장

말뭉치에서 출현한 특정 어휘가 용례에 나타나지 않지만 해당 용언과 빈번하게 사용되는 경우가 있다. 이러한 정보는 사전에서 추출할 수 없는 어휘의 실제 사용 실태를 표현하므로 의미가 있다고 할 수 있다. 그러므로 구문패턴에서 고빈도로 나타난 명사를 해당 용언과 구문관계를 설정함으로써, 용례에서 추출할 수 없는 구문관계를 추가적으로 구축할 수 있다. 예를 들어 구문패턴 명사 목록에서 고빈도 명사 '돈_001001', '겁_005000'은 각각 '먹다_002014', '먹다_002004'와 구문관계를 설정할 수 있다.

4. 구문관계 자동 구축 실험 및 결과

구문관계 구축 실험을 위한 대상 어휘는 한국어 사용 빈도 조사 결과를 바탕으로 하여 의미가 세분화되어 있으며, 문형정보를 가지고 있는 동사를 중심으로 5개('가다', '먹다', '만들다', '받다', '보다')를 선정하였다.

각각의 실험 어휘의 후보명사는 '가다'(18개), '먹다'(26개), '만들다'(22개), '받다'(47개), '보다'(45개)로 추출하였으며, 용언의 한 의미당 평균적으로 2~3개이다. 그리고 후보명사의 의미 분별 실험 결과, 정확하게 의미를 분별한 확률은 각각 '가다'(83%), '먹다'(92%), '만들다'(90%), '받다'(80%), '보다'(88%)로 나타났다.

특히, 본 연구에서는 많은 어휘 내적 정보를 포함하는 어휘를 실험 어휘로 선정함으로써, 용례 매칭 규칙을 적용하는 1단계에서 의미가 결정되는 경우가 많았다. 이것은 U-WIN 어휘 사전 데이타베이스가 [표준국어대사전]을 기반으로 하여 어휘 내적 정보를 설계 및 구축되었기 때문이다. 그리고 어휘 내적 정보가 충분하지 못한 어휘인 경우에는 2~4 단계에서 의미가 결정되는 경우가 많을 것으로 예상된다.

또한 의미 분별에 실패하는 후보명사는 사전에 등재되어 있지 않은 단어이거나, 의미 분별 과정의 오류로 인해 나타났다.

실험 어휘 중에서 '먹다'의 단계적 의미 분별 과정을 살펴보면 다음과 같다. '먹다'의 14개의 의미('먹다_002001'~'먹다_002014')에서 26개의 후보명사를 추출한 다음, 그림 9와 같이 4단계의 후보명사 의미 분별을 통해 구문관계를 설정하고 고빈도 어휘를 이용하여 구문관계를 확장하였다. 그 결과 26개 후보명사 중 '골_004005'와 '공금_001001'을 제외한 24개의 명사가 정확

표 3 구문관계 구축 결과

어휘	1단계 (용례매칭)	2단계 (구문패턴)	3단계 (의미유사도)	4단계 (빈도)	정확률 (%)
'가다'	7	3	1	4	83
'먹다'	8	6	2	8	92
'만들다'	6	5	3	6	90
'받다'	13	11	5	9	80
'보다'	15	9	7	9	88

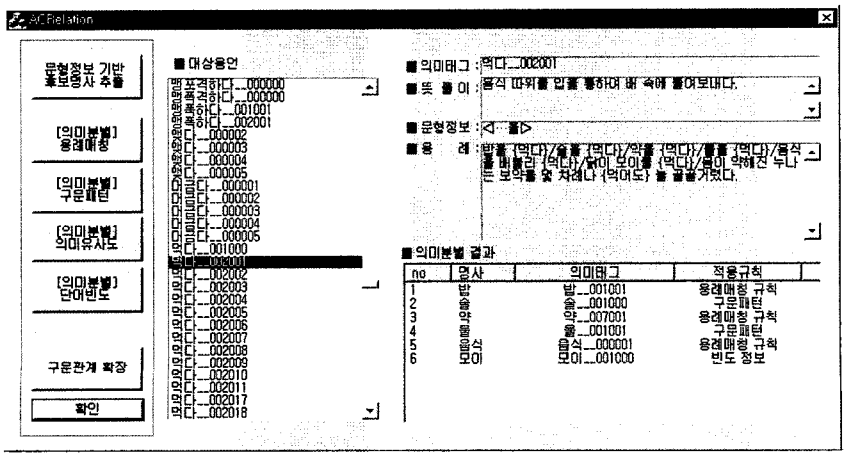


그림 8 구문관계 자동구축 도구

그림 9 후보명사의 의미 분별을 통한 구문관계 자동 구축 및 확장의 예 ('먹다')

하게 의미가 결정되었고, 구문패턴에서 추출한 고빈도 어휘 중에서 '아침밥', '겁_005000', '돈_001001'이 추가적으로 구문관계를 설정하였다.

5. 결론 및 향후 연구방향

본 연구는 U-WIN의 어휘 관계 중의 하나인 구문관계를 자동 구축하는 방법을 제시하였다. 즉, 용언의 의미별 용례와 문형정보를 이용하여 다양한 후보명사를 추출한 다음, 용례 규칙, 구문패턴, 의미 유사도를 이용한 명사 의미 분별 과정을 수행함으로써 명사와 용언 간의 구문관계를 자동으로 구축하는 방법을 제시하였다. 이러한 자동 구축방법은 향후 확장된 형태의 어휘망을 구축하는 기반을 다지는 작업이라고 할 수 있다.

본 연구는 다양한 구문관계 중에서 명사와 용언의 관계가 가장 밀접한 술목관계를 대상으로 제시한 연구방법의 실효성을 검증함으로써, 술주관계, 술부관계 등의 다양한 구문관계를 자동 구축하기 위한 기반 작업이라 할 수 있다. 그러므로 본 연구의 실험결과를 통해, 제시한 연구방법은 술목관계, 술주관계, 술부관계 등의 다양한 구문관계의 자동 구축을 가능하도록 할 것이다.

이렇게 구축한 구문관계는 기존의 선택제약 정보와 달리 용언과 명사가 모두 다의어 수준으로 의미가 결정되어 있으므로, WSD, 격틀사전 구축, 정보검색, 클러스터링, 구문분석, 의미분석 등의 다양한 자연언어처리 분야에서의 활용을 기대해 볼 수 있을 것이다.

참 고 문 헌

[1] 최호섭, "대규모 사용자 어휘지능망 구축과 활용", 울산대 박사학위논문, 2007.

[2] 홍재성 외, "21세기 세종 계획 <전자사전 개발> 연구 보고서", 국립국어원, 2005.

[3] 전문용어언어공학센터[KORTERM], 『다국어 어휘망』 총3권, KAIST Press, 2005.

[4] 최호섭, "한국어 명사 개념망 구축-경제용어를 중심으로", ETRI 지식정보검색연구팀 경제개념망 구축결과 보고서, 2001.

[5] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric on semantic nets, IEEE Transactions on Systems, Man and Cybernetics 19 (1) 17-30, 1989.

[6] C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, in: C. Fellbaum (Ed.), WordNet: An electronic lexical database, MIT Press, pp. 265-283, 1998.

[7] J. Jiang, D. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: Proceedings on International Conference on Research in Computational Linguistics, Taiwan, pp. 19-33, 1997.

[8] D. Lin, Using syntactic dependency as a local context to resolve word sense ambiguity, in: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, pp. 64-71, 1997.

[9] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the 14th International Joint Conference onArtificial Intelligence, Montreal, pp. 448-453, 1995.

[10] Wu, Z., Palmer, Verb semantics and lexical selection, 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico State University, LasCruces, New Mexico, 1994.

[11] Hirst, G. and D. St.Onge. Lexical Chains as Represenations of Context for the Detection and Correction of Malapropisms. WordNet. C. Fellbaum. Cambridge, MA, The Mit Press, 1995.

[12] 국립국어연구원, "<표준국어대사전> 편찬 지침 I-II", 국립국어연구원, 2000.

임 지 희

2003년 울산대학교 컴퓨터정보통신공학부(학사). 2005년 울산대학교 컴퓨터정보통신공학과(석사). 2007년 울산대학교 컴퓨터정보통신공학과 박사과정 수료. 2005년~2007년 울산대학교 외래강사. 2007년~현재 (주)오픈베이스 연구소 선임연구원. 관심분야는 한국어정보처리, 온톨로지, 정보검색

최 호 섭

1998년 경남대학교 국어국문학과(학사) 2000년 경남대학교 국어국문학과(석사) 2007년 울산대학교 컴퓨터정보통신공학과(박사). 2000년~2001년 한국전자통신연구원 지식정보검색연구팀 파견연구원 2002년~2006년 (주)시소러스 선임연구원, 감사. 2004년~2006년 울산대학교 외래강사. 2006년~현재 한국과학기술정보연구원 정보시스템개발팀 선임연구원. 관심분야는 한국어정보처리, 지식베이스, 어휘망, 온톨로지, 지식처리

옥 철 영

1982년 서울대학교 컴퓨터공학과(학사) 1984년 서울대학교 컴퓨터공학과(석사) 1993년 서울대학교 컴퓨터공학과(박사) 1994년 러시아 TOMSK 공과대학 교환교수. 1996년 영국 GLASGOW 대학교 객원교수. 2007년~현재 한국정보과학회 언어공학연구회 위원장. 2008년~현재 국립국어원 객원교수 1984년~현재 울산대학교 컴퓨터정보통신공학부 교수. 관심분야는 한국어정보처리, 의미분별, 온톨로지, 지식베이스, 기계학습, 문서분류