한국어 어휘지도(UWordMap)와 API 소개[1)]

배영준, 옥철영○
울산대학교 한국어처리연구실
young4862@nate.com, okcy@ulsan.ac.kr

Introduction to the Korean Word Map(UWordMap) and API

Young-Jun Bae, CheolYoung Ock○
Korean Language Processing Lab. University of Ulsan, Korean

요 약

한국어 문장의 의미 분석을 위해서는 어휘 의미들의 상의어, 하의어, 반의어, 유의어 등의 의미관계뿐만 아니라 서술어의 논항이 가지는 의미제약 정보 및 의미역, 서술어와 부사 명사와 부사, 부사와 부사와의 유의미한 결합 정보 등의 다양한 의미 정보가 필요하다. 한국어 어휘지도는 울산대 한국어처리연구실에서 2002년부터 현재까지 구축해 왔으며, 이제 구축된 결과물을 API와 함께 제공한다. 본 논문은 한국어 어휘지도의 대략적인 구조 및 API 등을 소개한다.

주제어: 어휘의미망, 어휘지도, 의미관계, 의미역, 의미제약, 격틀, API

1. 서론

문장의 의미를 분석하기 위해서는 어휘적 의미, 구문적 의미, 담화적 의미를 바탕으로 행위나 현상, 상태 등에 담긴 의미론적·개념론적 특성을 포함하고 있는 의미적 언어자원(semantic language resource)이 필요하다. 이러한 의미적 언어자원을 확보하기 위해 지식베이스(Knowledge Base)와 의미주석된 말뭉치 구축뿐만 아니라 어휘의미망에 대한 연구가 다양하게 이루어지고 있다. 국외에서는 WordNet, Euro WordNet, Cyc, HowNet, Lexical FreeNet, EDR 등이 대표적이며, 국내에서는 카이스트의 CoreNet, ETRI의 어휘 개념망, 부산대의 KorLex 등이 대표적이라 할 수 있다.

울산대에서는 2002년부터 표준국어대사전의 어휘를 기반으로 다의어 수준의 어휘의미망(U-WIN, User Word Intelligence Network)을 구축[1,2]하기 시작하여, 서술어의 필수논항이 가지는 의미제약정보를 명사어휘망의 최소상계노드로 설정하고, 부사와 의미적으로 연결될 수 있는 용언, 부사, 명사 등을 연결하였다. 이렇게 명사, 용언, 부사 어휘들이 상호 유기적으로 연결된 어휘의미망을 **"어휘지도(이하 UWordMap)"**로 명명하였다. UWordMap의 대체적인 구성은 아래 [그림 1]과 같다.

본 논문은 한국어 어휘지도의 대략적인 구축 내용을 소개하고, 실제 의미처리를 위해 제공하는 API를 소개한다.

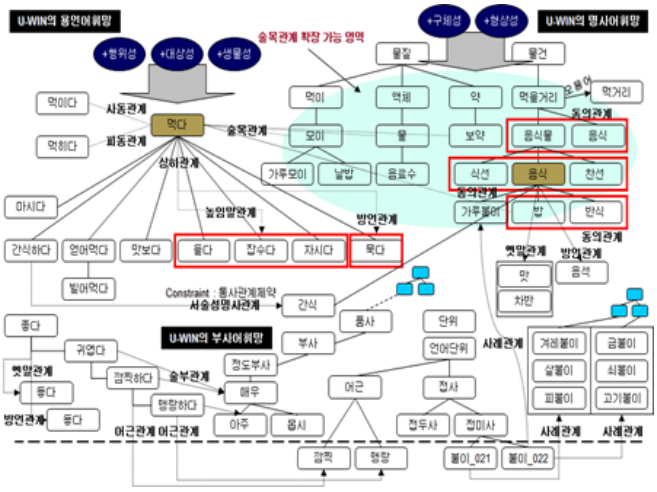

그림 1 한국어 어휘지도(UWordMap) 구조도

2. 명사어휘망

명사 어휘망의 최상위어는 다음 기준으로 설정하였다.

- 사전으로 기반으로 하는 만큼 최상위어는 사전에 등재되어 있는 어휘를 사용
- 의미적으로 명확하게 인지되는 어휘를 사용
- 형태적으로 사람들이 자주 사용하고 인식하는 어휘 사용
- 다른 최상위어와의 개념적 중복성이 적은 어휘 사용
- 하위어의 구성을 고려하여 선택
- 기존 지식베이스에서의 최상위어 중 어휘 군집화를 통하여 공통된 어휘 사용

다음 <표 1>은 명사어휘망에서 설정된 23개의 최상위어를 나타내고 있다.

1) 어휘지도 API는 울산대 산학협력단과의 기술이전 계약체결 후 UTagger와 함께 DLL형태로 제공(연구용일 경우 무료).

표 1. 명사어휘망에서의 23개 최상위어

최상위 노드	정의
{공간_0502}	물리적으로나 심리적으로 널리 퍼져 있는 범위. 어떤 물질이나 물체가 존재할 수 있거나 어떤 일이 일어날 수 있는 자리가 된다.
{과정_0300}	일이 되어 가는 경로.
{관계_0501; 계관_0101}	둘 이상의 사람, 사물, 현상 따위가 서로 관련을 맺거나 관련이 있음. 또는 그런 관련.
{기호_0100; 심벌_0002}	어떠한 뜻을 나타내기 위하여 쓰이는 부호, 문자, 표지 따위를 통틀어 이른 말.
{단위_0201; 하나치_0000}	길이, 무게, 수효, 시간 따위의 수량을 수치로 나타낼 때 기초가 되는 일정한 기준. 근, 되, 자, 그램, 리터, 미터, 초 따위가 있다.
{대상_1101}	어떤 일의 상대 또는 목표나 목적이 되는 것.
{모양_0201}	겉으로 나타나는 생김새나 모습.
{물건_0001}	일정한 형체를 갖춘 모든 물질적 대상.
{방법_0001}	어떤 일을 해 나가거나 목적을 이루기 위하여 취하는 수단이나 방식.
{범위_0001}	테두리가 정하여진 구역.
{생물_0101; 유생물_0000}	생명을 가지고 스스로 생활 현상을 유지하여 나가는 물체, 영양·운동·생장·증식을 하며, 동물·식물·미생물로 나뉜다.
{성질_0002; 성분_0300}	사물이나 현상이 가지고 있는 고유의 특성.
{시간_0401}	어떤 시각에서 어떤 시각까지의 사이.
{요소_0401}	사물의 성립이나 효력 발생 따위에 꼭 필요한 성분. 또는 근본 조건.
{인지_0803; 인식_0002}	자극을 받아들이고, 저장하고, 인출하는 일련의 정신 과정. 지각, 기억, 상상, 개념, 판단, 추리를 포함하여 무엇을 안다는 것을 나타내는 포괄적인 용어로 쓴다.
{작용_0101}	어떠한 현상을 일으키거나 영향을 미침.
{재료_0101}	물건을 만드는 데 들어가는 감
{정도_1101; 정한_0300}	사물의 성질이나 가치를 양부, 우열 따위로 본 분량이나 수준.
{존재_0001}	현실에 실재로 있음. 또는 그런 대상.
{종류_0201; 종_0902;종속_0400}	사물의 부분을 나누는 갈래.
{집단_0000}	여럿이 모여 이룬 모임.
{행위_0001}	사람이 의지를 가지고 하는 짓.
{힘_0103}	어떤 일을 할 수 있는 능력이나 역량.

명사의휘망에서 의미관계는 상하 관계, 동의 관계, 유의 관계, 반의 관계, 전체-부분 관계, 연관 관계 등 총 6가지 기본 의미 관계를 설정하였으며, 각 의미관계의 설정 기준은 최호섭(2007)[3,4]에서 자세히 서술하였다. 2006년까지 구축된 명사어휘망(버전 1.0)은 약 160,000 어휘 규모로 수작업에 의해 구축되었다. 이후 표준국어대사전 전체의 뜻풀이와 용례를 대상으로 형태·다의어 주석을 하고, 형태·다의어 주석된 표준국어대사전을 이용하여 반자동으로 계층구조를 확장하였다.

3. 용언어휘망

3.1 하위범주화정보 구축

용언어휘망은 기본적으로 상하 관계의 어휘계층 구조를 가지고 있을 뿐만 아니라, 용언의 하위범주화 정보(필수논항의 의미제약)를 명사어휘망의 최소상계노드(LCS, Least Common Subsumer)로 연결하였다. 용언의 필수논항의 LCS를 설정하기 위해서 다의어 수준의 의미 주석된 표준국어대사전의 뜻풀이와 용례에서 <그림 1>과 같이 “먹다”의 목적어로 결합될 수 있는 최소상계노드로 {먹이, 음식물, 물, 음료, 약}이 연결되어 있다. 이 때, 해당 최소상계노드의 하위어 중 “먹다”와 연결될 수 없는 어휘는 “N_OBJ"의 관계를 가지도록 설정하였다. 예를 들어 “먹다”의 목적어로 “약”이 설정된 경우 “약”의 하위어가 “먹을 수 없는 약({바르는약, 붙이는약, 주사약, 해충약 류})”의 어휘들은 ”먹다“와 "N_OBJ"의 관계를 가지도록 설정하였다.

용언의 하위범주화 정보 구축은 다음의 지침에 따라 구축하였다[5].

1 용언의 논항정보는 표준국어대사전의 각 용언의 각 용언의 용례와 의미를 참조하여 구축한다.
2 문형정보 해당 명사는 명사 어휘망의 최소상계노드와 연결한다.
3 의미정보상 최소상계노드 연결이 부적합하면 노드를 조정한다.
- 상위노드가 논항의 동위노드를 아우르지 못하면 자기노드로 연결한다.
- 한 용언의 두 논항이 상하 관계에 있으면 상계 상위노드로 연결한다.
- 논항들이 유사 의미장에 같은 줄기이면 공통 상위노드로 연결한다.
- 두 논항 형태가 같고 상하 관계이면 의미가 넓은 상위노드로 연결한다.

4 논항이 여러 노드에 걸쳐 나타나면 의미정보로 노드를 설정한다.
- 논항의 두 동형이 의미정보에 포함되면 복수 상위노드로 설정한다.
- 논항이 동형이라도 의미정보가 다르면 단수 상위노드로 설정한다.

5 논항이 상위노드를 설정하지 못하는 상태이면 제자리를 유지한다.
- 논항의 상위노드를 설정하면 그 의미가 부적합할 때 자기노드에 둔다.
- 논항이 최상위노드일 때 그 의미가 포괄적이더라도 자기노드에 둔다.

3.2 의미역 및 격틀사전

문장의 의미를 파악하기 위해서는 용언과 공기관계에 있는 필수논항인 보어들의 의미역할을 파악해야 한다. 이를 위해서 표준국어대사전의 격조사의 의미와 용례에 따라 22개의 의미역 결정하고, 문형이 있는 문형이 있는 41,518개의 용언에 대해 22개의 의미역을 부착하여 용언의 다음 <표 2>와 같은 격틀사전(UPropBank)을 구축하여, 용언어휘망에 포함시켰다[6,7].

표 2. 격틀 사전(UPropBank)

디자인하다 000000	[동사]		⇒ 디자인. >>의상, 공업 제품, 건축 따위 실용적인 목적을 가진 조형 작품의 설계나 도안.	이 경기장은 기와집의 처마를 살려 {디자인했다}.
사용하다 030001	[동사]	{X:행동주 Y:대상-을/를 Z:착점-에/에게} {X:행동주 Y:대상-을/를 Z:착점-으로/로}	사용04(1).	어른에게 존댓말을 사용하다. 일생을 두고 모은 재산을 사회 복지 사업에 사용하다. 손이나 도구를 사용하던 인간은 기계와 동력을 사용하게 되었다. 인류가 정확히 언제부터 불을 사용하게 되었는지는 아직 알 수 없다. 비둘기를 통신용으로 사용하다. 지렁이를 미끼로 사용하다. 키 큰 소위의 말로는 국회의원을 지낸 사람의 집을 부대의 임시 숙소로 사용하겠다는 것이었다.
사용하다 030002	[동사]		⇒사용04(2). >>사람을 다루어 이용함. '씀', '부림'으로 순화.	
열다 020101	[동사]	{X:행동주 Y:대상-을/를}	닫히거나 잠긴 것을 트거나 벗기다.	문을 열다. 창문을 열다. 서랍을 열다. 수도꼭지를 열다. 자물쇠를 열다. 가방을 열다. 안전기의 스위치를 열고 퓨즈가 끊어진 것을 확인한다. 나는 만약을 위해 한 번 더 약병의 뚜껑을 열고 수건을 대어 흔들었다.
열다 020102	[동사]	{X:행동주 Y:대상-을/를}	모임이나 회의 따위를 시작하다.	국회를 열다. 총회를 열다. 동창회를 열다. 우리 조상들은 씨뿌리기와 가을걷이가 끝나는 음력 5월과 10월에 큰 모임을 열고 하늘에 축원과 감사의 고사를 지냈다.
열다 020201	[동사]	{X:행동주 Z:착점-에 Y:대상-을/를}	사업이나 경영 따위의 운영을 시작하다.	형은 집에서 가까운 네거리에 가게를 열었다. 아직 교육의 혜택을 제대로 받지 못한 오지에 학교를 열었다. 그들은 함께 힘을 모아 아담한 구둣방을 열게 되는 것이 꿈이었다.
열다 020202	[동사]		새로운 기틀을 마련하다.	이 땅에 새 시대를 {열다}. 한반도에 새 왕조를 {열다}. 사람들이 토지에 정착하여 살 수 있게 됨으로써 인류 역사에 농경 시대를 {열게} 되었다.
열다 020301	[동사]		자기의 마음을 다른 사람에게 터놓거나 다른 사람의 마음을 받아들이다.	자기가 하는 일에 마음을 {열어야} 그 일을 통해 진정한 보람을 느낄 수 있다. 그는 결국에는 아내에게 굳게 닫혔던 마음을 {열었다}. 모든 사람에게 마음을 {열고} 살기 위해서는 무엇보다도 타인에 대한 사랑과 이해가 우선되어야 한다.
열다 020302	[동사]		다른 사람에게 어떤 일에 대하여 터놓거나 이야기를 시작하다.	용의자는 마침내 형사에게 입을 {열었다}. 경민은 잠시 침묵을 지킨 뒤 곧 담담하게 앞을 보고 입을 {열었다}. ≪홍성원, 육이오≫
열다 020400	[동사]		어떤 관계를 맺다.	조선은 청나라와 국교(國交)를 {열었다}. 서로 국가 이념이 다른 두 나라가 경제적인 협력을 위하여 국교를 {열었다}.

4. 부사어휘망

부사는 일반적으로 용언과 결합하나, 경우에 따라서는 의미적으로 특정한 용언이나 체언, 부사, 관형사에만 한정적으로 결합하는 경우도 있다.

또한 부사는 형태 구성의 관점에서도 품사가 ① 원래 부사인 것, ② 파생부사, ③ 합성부사, ④ 조사가 결합한 형태, ⑤ 어미가 결합한 형태 등 다양한 형태가 있다.

① 품사가 부사 : 매우, 좀, 아주, 퍽, ...
② 파생부사 : 자연히, 없이, 같이, 많이, 달리, 높이, 빨리, ...
③ 합성부사 : 밤낮, 요즈음, 잘못, 좀더, ...
④ 조사가 결합한 형태 : 각중에, 순식간에, 정말로, 별도로, 되는대로, 그나마, 그만큼, ...
⑤ 어미가 결합한 형태 : 가볍게, 격렬하게, 되도록, 어쩌면, 끝없이, 각설하고, 못해도, ...

또한, 특정 부사(예, '결코')는 의미적으로 부정형태 혹은 보조용언과도 결합하는 등 다양한 결합 형태를 보인다.

부사어휘망도 형태·다의어 주석된 표준국어대사전의 뜻풀이와 용례에서 개별 부사의 결합관계를 추출하였으며, 다음 지침에 따라 구축하였다.

① 사전에서 용례를 확인할 수 없는 경우 따로 처리하지 않는다.
② 문장 부사는 따로 처리하지 않는다.
③ 'N+이다' 서술어는 처리하지 않는다.
④ '용언+-이'의 형태 중 부사로 처리하기 힘든 경우는 처리하지 않는다.
⑤ '구 구성, 절'을 수식하는 경우는 처리하지 않는다.
⑥ 사전에서의 용법을 살핀 후 처리할 수 없는 것은 처리하지 않는다.
⑦ 용례가 잘못되어 부사로 보기 어려운 것들은 처리하지 않는다.

5. 어휘지도 구축 현황

현재(2014.03 기준)까지 구축된 UWordMap은 다음 <표 3>과 같으며, 2014년 12월까지 용언어휘망의 완성을 목표로 구축하고 있다.

표 3 어휘지도 구축현황

품사	표준국어 대사전 어휘수	U-WIN (계층관계)	어휘지도 (하위범주화)
명사	377,281	362,726	LCS 72,030
동사	90,237	73,612	29,345 (용례없음: 10,682)
형용사	21,618	16,724	4,653 (용례없음: 413)
부사	25,178	17,697	6,186
합계	514,314	470,759	112,214

U-WIN의 계층별 분포는 다음 <그림 2>와 같다.

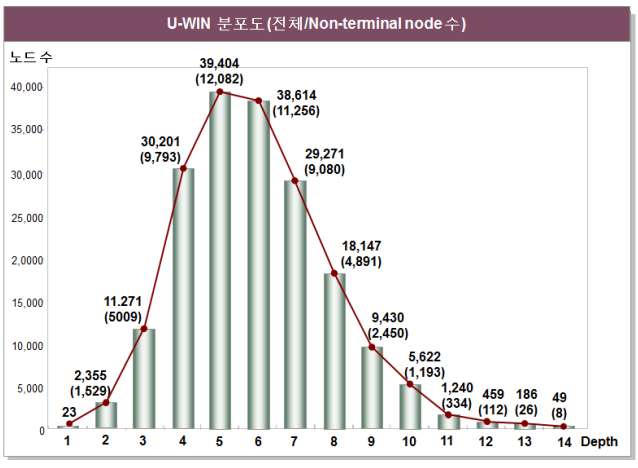

그림 2 U-WIN의 계층별 분포도

6. 어휘지도 API

UWordMap API(Application Programming Interface)는 어휘지도를 사용하기 위한 함수들의 집합이다. 어휘지도는 다의어 단위로 구축되어 있기 때문에 API 사용 시 의미 단위(단어, 동형이의어, 다의어)를 고려하여야 한다. 아래 <표 4>은 API와 관련된 정보이다.

표 4. 어휘지도 API 정보

API 정보	
언어	C / C++
배포	DLL로 배포
반환값	모든 API의 반환값의 타입: int 0: 정상 종료, 1: 오류 또는 없는 단어
마지막 인자	함수 결과 값. 모두 다의어 단위로 저장됨
의미 단위	단어, 동형이의어, 다의어 단위로 구분. 예) 배 [단어] : 배 [동형이의어] : 배__01, 배__02 [다의어] : 배__010001, 배__010002 ...
약어	[동형이의어] -> [동형] [다의어] -> [다의]

함수를 <표 5>와 같이 정리하였다.

표 5. 어휘지도 API 함수 종류

함수 종류	
함수명	GetPS
설명	(단어 or 동형)의 다의어 받아오기
인자	1. string* pStrWord : 단어 2. vecotr<string>* pVPS : 다의 리스트
함수명	GetHyperWord
설명	단어의 1레벨 위의 상위어 받아오기
인자	1. string* pStrWord : 단어or동형or다의 2. vector<string>* pVHyper : 상위어
함수명	GetHyperAllWord
설명	다의어의 상위어 전체 받아오기 (루트까지)
인자	1. string* pStrWord : 다의 2. vector<string>* pVAllHyper : 상위어
함수명	GetHypoWord
설명	다의어의 1레벨 아래의 하위어 받아오기
인자	1. string* pStrWord : 단어 2. vector<string>* pVHypo : 하위어
함수명	GetNRelV
설명	해당 용언과 논항과 관련된 명사 받아오기
인자	1. string* pStrWord : 다의(용언) 2. string strRel : 논항명 2. vector<string>* pVRWord : 다의(명사)
함수명	GetVRelN
설명	해당 명사와 논항과 관련된 용언 받아오기
인자	1. string* pStrWord : 다의(명사) 2. string strRel : 논항명 3. vector<string>* pVRWord : 다의(용언)
함수명	GetSynSet
설명	다의어의 동의어 받아오기
인자	1. string* pStrWord : 다의 2. vector<string>* pVSyn : 동의어
함수명	GetAntSet
설명	다의어의 반의어 받아오기
인자	1. string* pStrWord : 다의 2. vector<string>* pVAnt : 반의어
함수명	GetDistance
설명	다의어1과 다의어2의 거리 받아오기
인자	1. string* pStrWord1 : 다의1(명사) 2. string* pStrWord2 : 다의2(명사) 3. int* pIDist : 거리
함수명	GetRelSubCt
설명	용언과 명사의 논항명 받아오기
인자	1. string* pStrWord1 : 다의(용언) 2. string* pStrWord2 : 다의(명사) 3. vector<string>* pVAg : 논항명

<표 6>에서 UWordMap API를 사용하는 간단한 프로그램 예제 및 출력결과를 볼 수 있다.

표 6. 프로그램 예제 및 결과

```
//UWordMap 객체생성
UWM uwm;
//변수선언
string strWord=""; // 입력단어(단위: 단어, 동형, 다의)
string strRel="";  // 논항명(예: 을, 로, 에)
string strRWord=""; // 관련단어
vector<string> vec;// 결과저장

// (단어 or 동형) 다의어 출력
strWord="나무"; //(동형)"나무__01";
uwm.GetPS(&strWord, &vec);
//(출력)나무__010001/나무__010002/나무__010003
  /나무__020000/나무__030000/나무__040000/나무__050000
vec.clear();

// 1단계 위 상위어 출력
strWord="나무"; //(동형)"나무__01";//(다의)"나무__010001";
uwm.GetHyperWord(&strWord, &vec);
//(출력)여러해살이식물__000000/재목__010001
  /땔감__000000/엽전__000001/수필__040000
  /염불__020001/춤__010000
//(입력)나무__01: //(출력)여러해살이식물__000000
  /재목__ 010001/땔감__000000
//(입력)나무__010001: //(출력)여러해살이식물__000000
vec.clear();


// 전체 상위어 출력
```

```
strWord = "나무__010001";
uwm.GetAllHyperWord(&strWord, &vec);
//(출력) 나무__010001/식물__020000/생물__010001/UWIN
vec.clear();

// 하위어 출력
strWord = "나무__010001";
uwm.GetHypoWord(&strWord, &vec);
//(출력)
가로수__000000/가목__030000/가시나무__000001/갈잎나무__
000001/거목__010001/고목__010000/고목__040000/고목__050
000/고목__070000/...
vec.clear();

// 용언-격조사 => 가능한 명사 출력
//(입력)용언, 논항
strWord = "가결하다__010000";
strRel = "을";
uwm.GetNRelV(&strWord, &strRel, &vec);
//(출력) 문안__020002/사항__020000
vec.clear();

// 명사-격조사 => 결합가능한 용언 출력
//(입력)명사, 논항
strWord = "표현__000001";
strRel = "을";
uwm.GetVRelN(&strWord, &strRel, &vec);
//(출력)
겁내다__000000/과장하다__020000/꺼리다__000100/내뱉다__
000002/머뭇대다__000000/못하다__000100/부인하다__000000
/빌리다__000202/실수하다__000001/억누르다__000002/...
vec.clear();

//=> 동의어 출력
strWord = "명주실__000000";
uwm.GetSynSet(&strWord, &vec);
//(출력) 면주실__000000/명사__090000/주사__200000
vec.clear();

//=> 반의어 출력
strWord = "개방성__000000";
uwm.GetAntSet(&strWord, &vec);
//(출력) 폐쇄성__000000
vec.clear();

//=> 두 단어 사이 거리
strWord1 = "음식__000001";
strWord2 = "가락국수__000000";
int iDist = 0;
uwm.GetDistance(&strWord1, &strWord2, &iDist);
//(출력) 2

//=> 용언과 명사의 결합이 가능한 격조사 출력
//(입력)용언, 명사
strWord1 = "먹다__020101";
strWord2 = "석류__010002";
uwm.GetRelSubCt(&strWord1, &strWord2, &vec);
//(출력) 을
vec.clear();
```

감사의 글

"본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [10044508, 비기호적 기법 기반 인간모사형 자가학습 지능 원천기술 개발]"

참고문헌

[1] 최호섭 외(2002), "사전을 기반을 한 한국어 의미망 구축과 활용", 제29회 한국정보과학회 춘계학술발표회 논문집, 한국정보과학회.

[2] 최호섭·옥철영(2002), "한국어 의미망 구축과 활용", 『한국어학』 17, 한국어학회.

[3] 최호섭 외(2006), "대규모 우리말 어휘지능망 구축 방법", 『한글』 273, 한글학회. p125-151

[4] 최호섭(2007), "대규모 사용자 어휘지능망 구축과 활용", 울산대학교 대학원 컴퓨터정보통신공학부 박사학위논문

[5] 옥철영(2009), 『어휘의미 체계 기반 입체적 국어사전 확장』, 국립국어연구원 11-1371028-000118-01. 연구보고서

[6] 김윤정, 옥철영, "한국어 서술어와 논항들 사이의 의미역", 제 26회 한글및한국어정보처리 학술대회, 2014 (제출)

[7] 김완수, 옥철영, "한국어 격틀사전 기반 의미역 반자동 부착 도구", 제 26회 한글및한국어정보처리 학술대회, 2014 (제출)