자연어 QA workshop

한국어 의미처리시스템

2015.08.21

KLPLAB 울산대학교 한국어처리연구실

옥 철 영

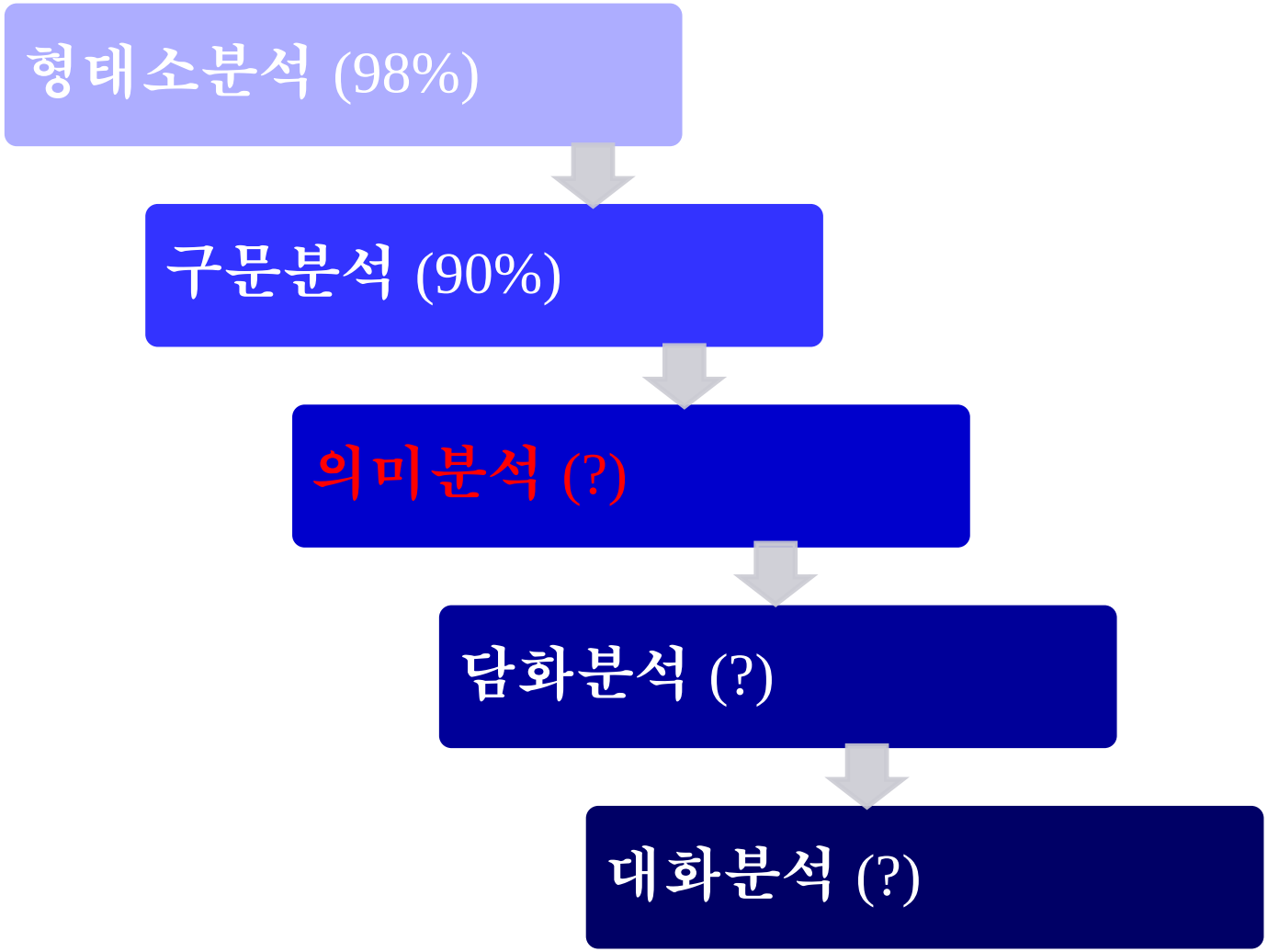
울산대학교

컴퓨터정보통신공학부/국어국문학과

목 차

- ❖ 한국어 의미자원
  - 어휘지도(UWordMap)
  - 의미역 & UPropBank
- ❖ 한국어 의미처리시스템
  - 동형이의어 WSD (UTagger-HM)
  - 의존관계 분석(UTagger-DP)
  - 다의어 WSD (UTagger-PS)
  - 의미역 태깅시스템 (UTagger-SR)
  - 개체명 인식기 (UTagger-NE)
- ❖ 시연(한국어어휘지도, UTagger 시스템)

한국어 분석 단계 및 현 처리수준



의미론(Semantics)

Lexical Semantics
WSD, NER

- 하나의 형태소가 문맥에 따라 여러개의 의미로 해석될 때, 하나의 의미를 결정
- Homograph WSD
- Polysemy WSD
- Named Entity Recognition
- Lexical Semantic Network

Sentence Semantics
Parser, SRL

- 문장 내의 술어-논항(predicate-argument) 관계에 적합한 의미 관계(Semantic Role)를 결정
- Subcategorization
- Semantic Restriction
- Semantic Role Labeling
- Ontology & Inference

한국어 의미자원

- ✓ 한국어 어휘지도
- ✓ 의미역 & UPropBank
- ✓ 용언 의미 군집화 & 계층화

어휘의미망 필요성

- 어휘의미론(Lexical Semantics) 관점
  - 의미관계(상의어, 하의어, 반의어, 유의어, 부분/전체) 정립
  - 의미자질을 통한 개념화

- 문장의미론(Sentence Semantics) 관점
  - 통사구조와 의미구조 분석
  - 논항의 의미제약
  - 의미역(thematic roles) 결정

- Exo Brain의 기반 지식
  - QA에서의 질의유형, 정답유형 판단
  - 어휘의미의 개념화/범주화
  - 논항의 의미제약
  - 의미역 기반 triple 구조의 knowledge 표현
  - 다의어 WSD

한국어 어휘의미망

❑ 부산대 윤애선교수 발표자료 (2008)

명칭	구축방식 (기반)	구축연도	구축기관	의미/개념 vs 어휘 수	구축 품사
KorLex	참조(PWN)	2004-현재	부산대학교	126,653s/143,479w	명,동,형,부, 분류사
한국어 시소러스	참조(PWN)	1997-2000	포항공대	18,362s/ 21,390w	명
다국어 DB	참조(EWN)	2000-2006	고려대 민족문화연구소	5,500w	명
CoreNet	참조(NTT)	1995-2004	KAIST	2,938n/ 62,632w	명,동,형
U-WIN	직접(표준)	2002-현재	울산대학교	57,792s/ 430,000w	모든 품사
세종전자 사전	직접	1998-2007	서울대학교	581n/540,000w	모든 품사

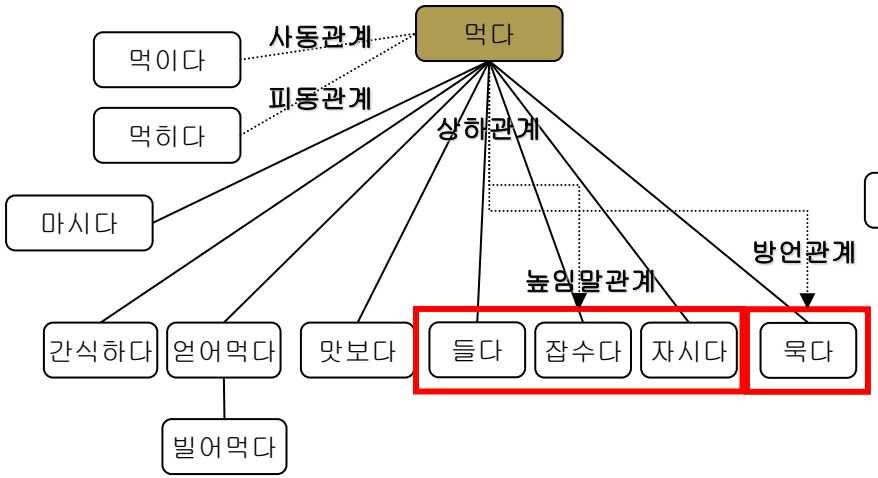
❖ KorLex 규모 (2015. 5. 7 기준, ()는 2012. 3)

단위	Korean WordNet				
	명사	동사	형용사	부사	계
어형	**106,294 (90,909)**	**17,962 (17,957)**	**41,107 (19,694)**	**7,806 (3,032)**	**174,350 (132,877)**
신셋	**101,869 (92,184)**	**17,075 (16,937)**	**18,582 (18,560)**	**3,668 (3,651)**	**142,571 (132,943)**
어의	**121,216 (104,417)**	**20,346 (20,151)**	**51,896 (20,897)**	**9,047 (3,123)**	**203,882 (150,199)**

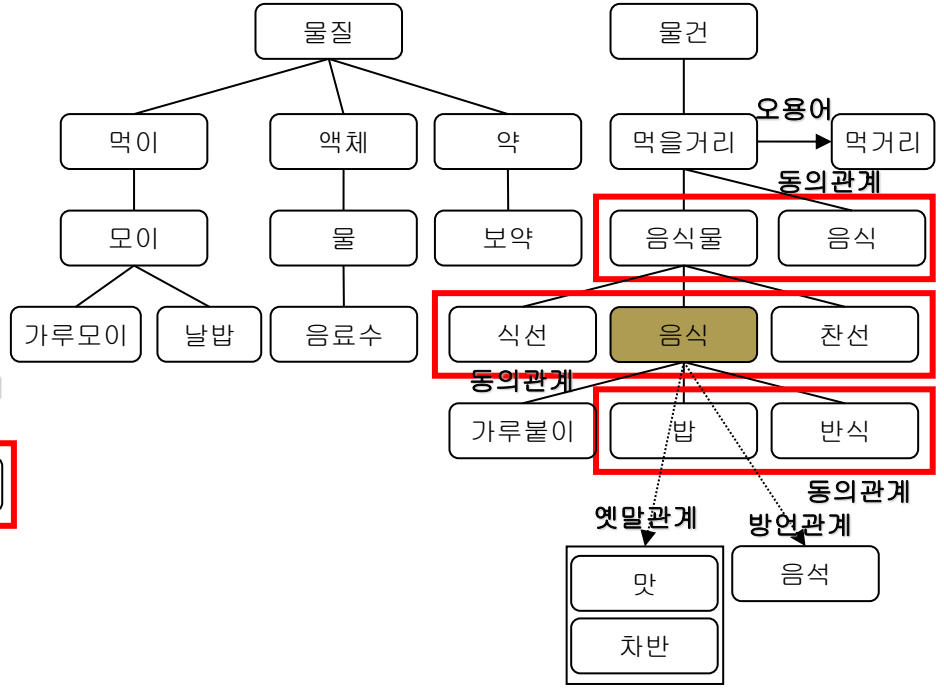
❖ KorLex & 표준국어대사전 mapping

정의문 연동	Korean WordNet 정의문 연동 정보				
	명사	동사	형용사	부사	계
표준국어 대사전	74,446 (67,938)	12,225 (9,635)	19,274 (17,639)	3474 (2,913)	109,419 (99,291)
기타 사전	363	226	2,540	73	3,202

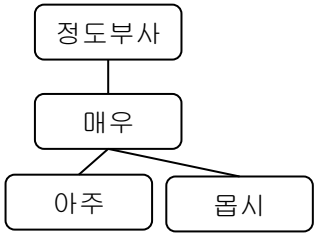
U-WIN의 용언어휘망



U-WIN의 명사어휘망

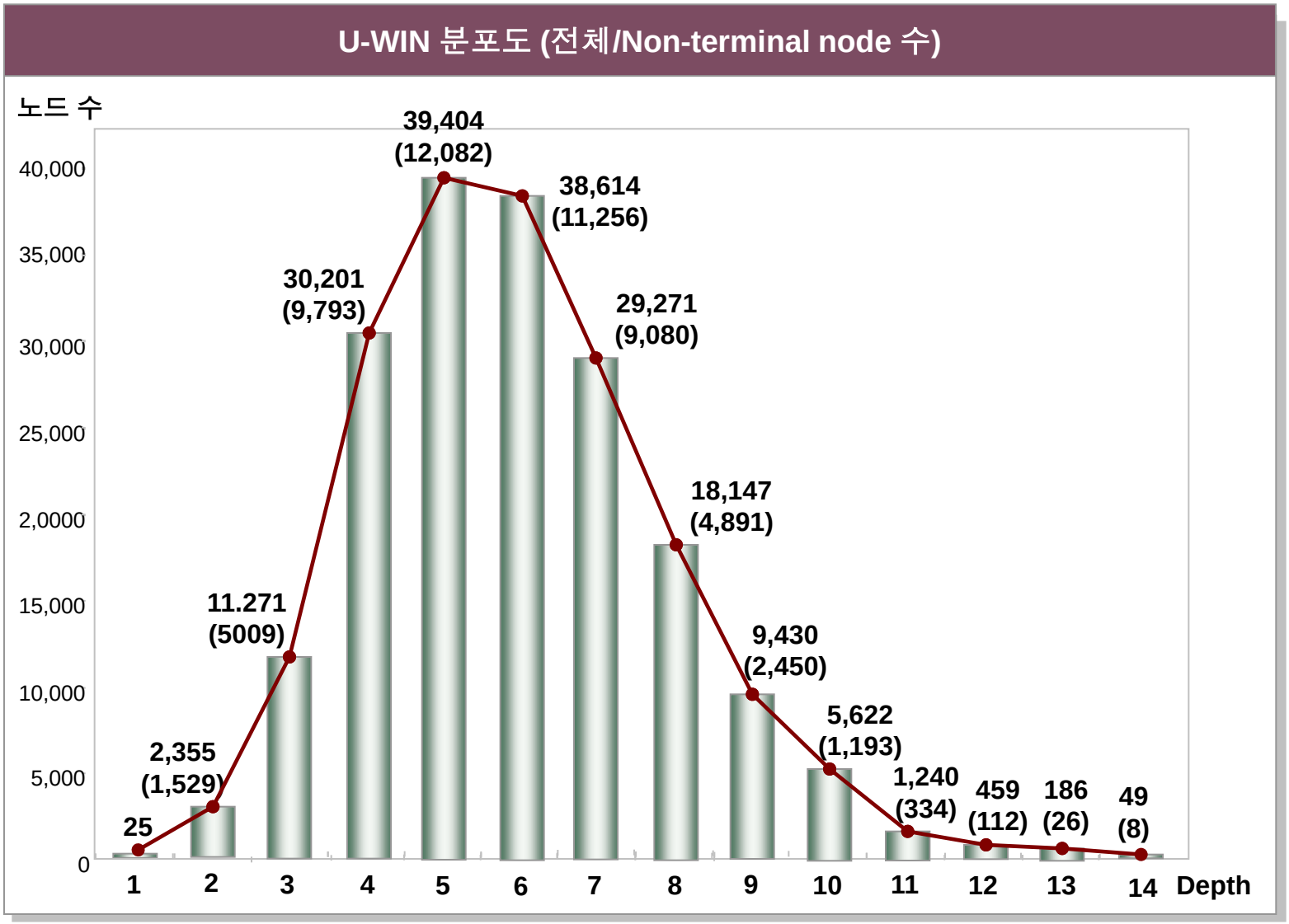


U-WIN의 부사어휘망



- U-WIN ver.1.0(2002 ~ 2007): 160,000 어휘 (수작업)
- 표준국어대사전 뜻풀이/용례 다의어 태깅 (2008 ~2010)
- U-WIN ver.1.9 : 420,000 어휘 (다의어 태깅 사전, 2010)

상하관계를 중심으로 한 분포



U-WIN에서의 의미관계

❖ 어휘 의미 관계 (다의어 수준)

품사	동의어 (=, ≒)	반대말	비슷한 말	준말	본말	낮춤말	높임말	참고 어휘	합계
명사	114,715	4,513	13,583	496	20	382	378	28,474	162,561
동사	12,702	1,474	10,942	0	141	6	21	3,165	28,451
형용사	1,463	443	1,317	170	46	3	1	2,451	5,894
부사	2,284	68	393	842	117	1	0	11,722	15,427
합계	131,164	6,498	26,235	1,508	324	392	400	45,812	212,333

뜻풀이 및 용례 의미 태깅 (1)

❖ 표준국어대사전 뜻풀이 태깅
- 다의어 수준 뜻풀이 : 587,833
- 다의어/동형이의어 태깅 대상 어휘 : 107,306
- 뜻풀이 태깅 어휘 수 : 3,610,106

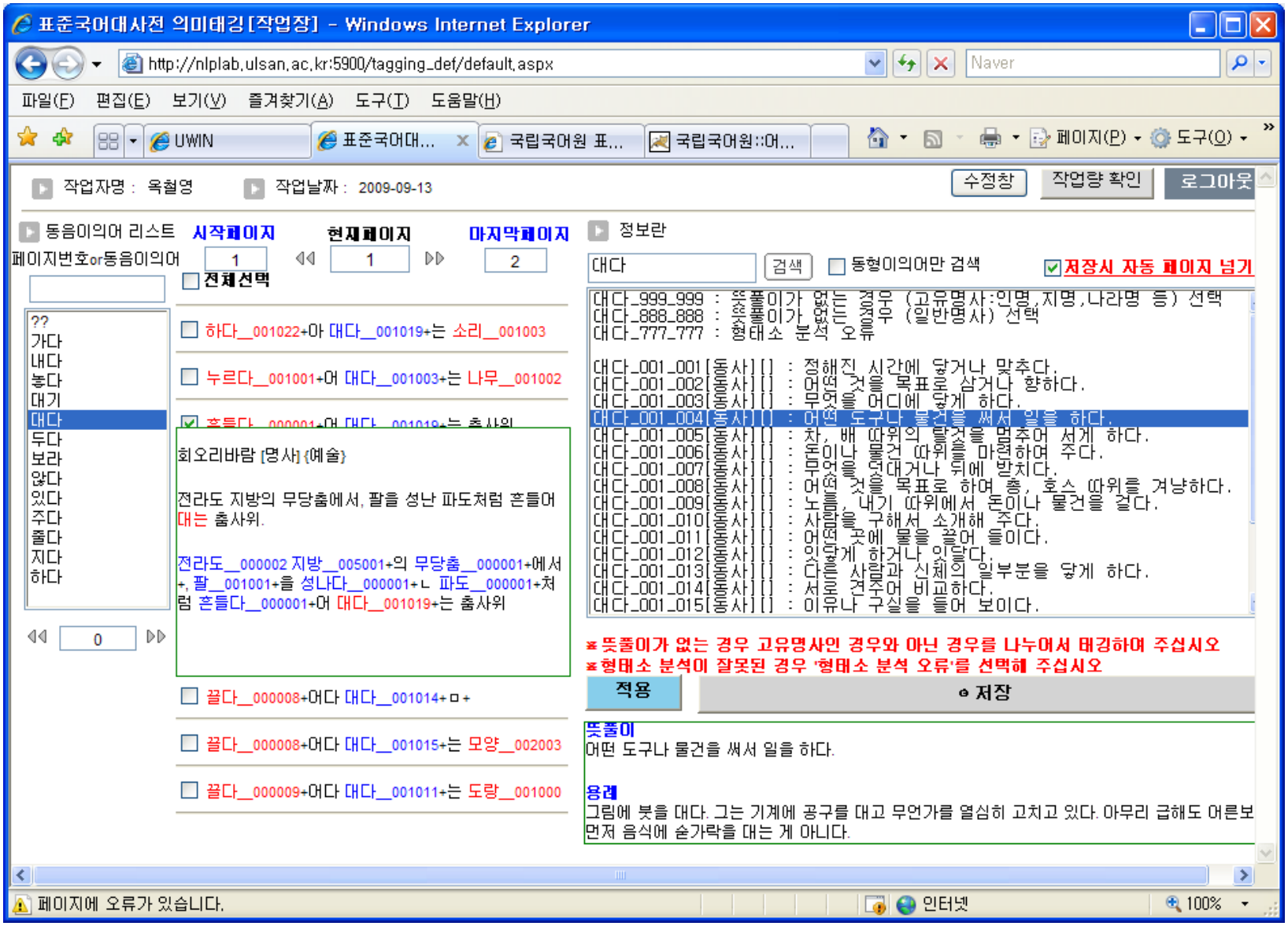
❖ 표준국어대사전 용례 태깅
- 다의어/동형이의어 태깅 대상 어휘 : 78,083
- 용례 태깅 어휘 수 : 1,121,759

❖ 작업기간 : 2008 ~ 2010

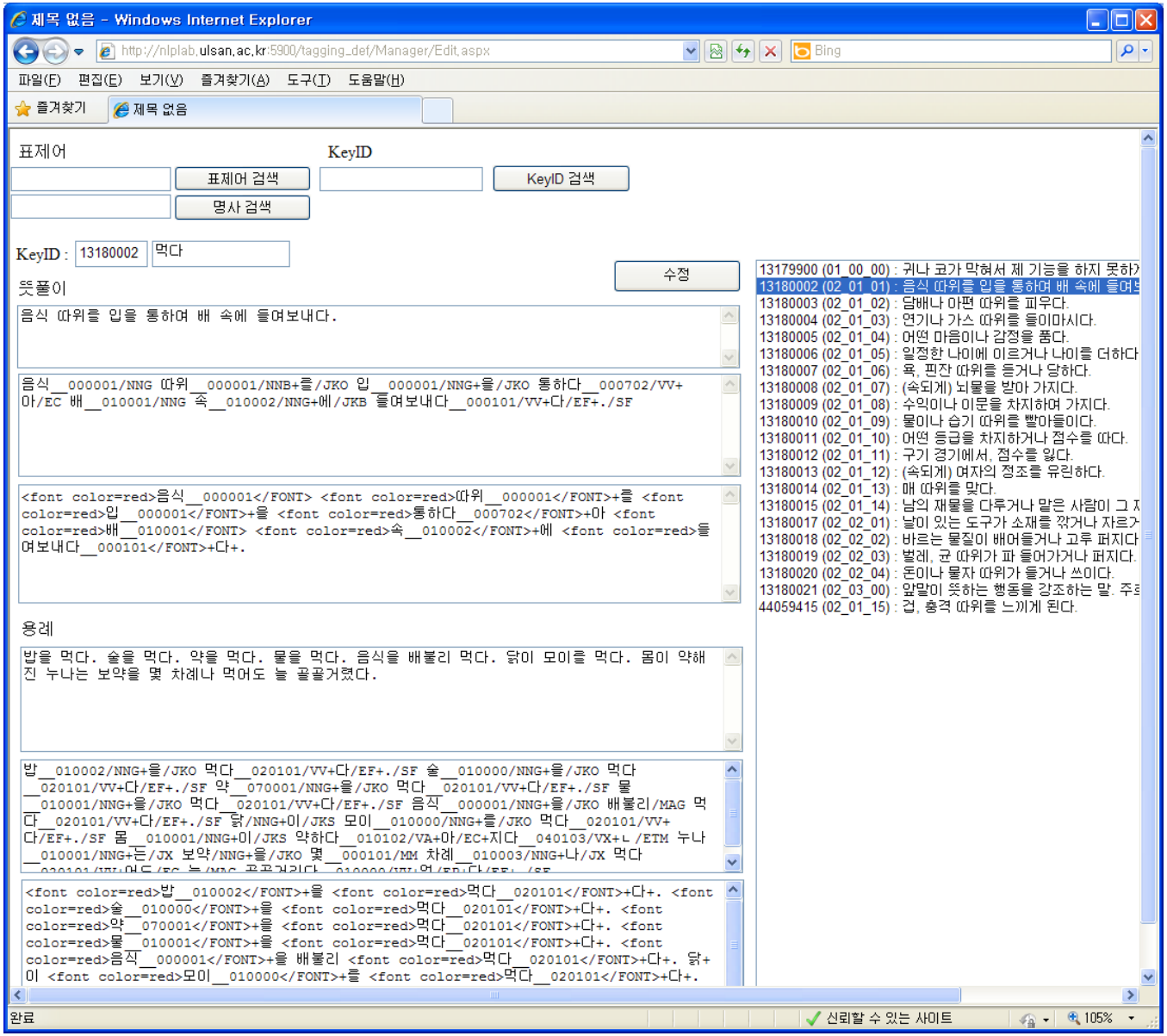
❖ 사전(어휘지식) 해독하는 컴퓨터
- 동형이의어/다의어 WSD
- 의미역 triple 추출
- 개념화 => 어휘지도

38	가격_010000	5	1
39	가격_020000	1	
40	가격_030000	679	154
41	가결_010000	6	2
42	가결_030001	2	
43	가결_030002	1	
45	가결의_000001	2	
46	가결의_000002	2	
48	가경_040002	1	1
49	가경_080000	1	
50	가계_010001	1	
51	가계_010002	1	
52	가계_030000	2	1
53	가계_040000	1	
54	가계_050000	1	
55	가계_060000	23	3
56	가계_080001	21	13
57	가계_080002	1	1
58	가계_100000	1	
59	가곡_010001	101	1
60	가곡_010002	65	8
61	가공_010001	493	15
62	가공_010002	2	2
63	가공_020000	1	
64	가공_040001	12	
65	가공_040002	2	2
70	가공하다_010001	404	13
71	가공하다_030000	1	7
72	가과_010000	1	

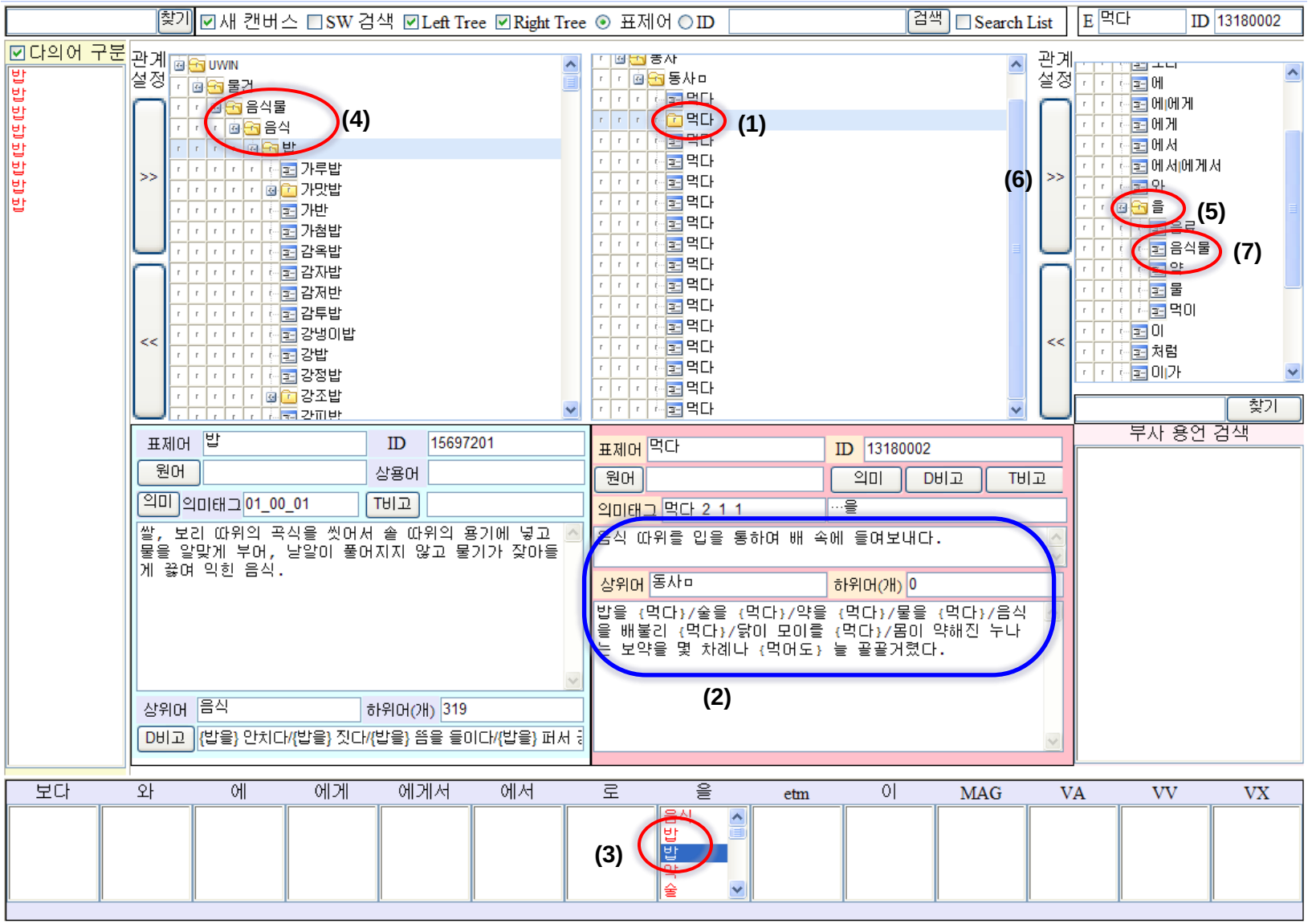
뜻풀이 및 용례 의미 태깅 (2)



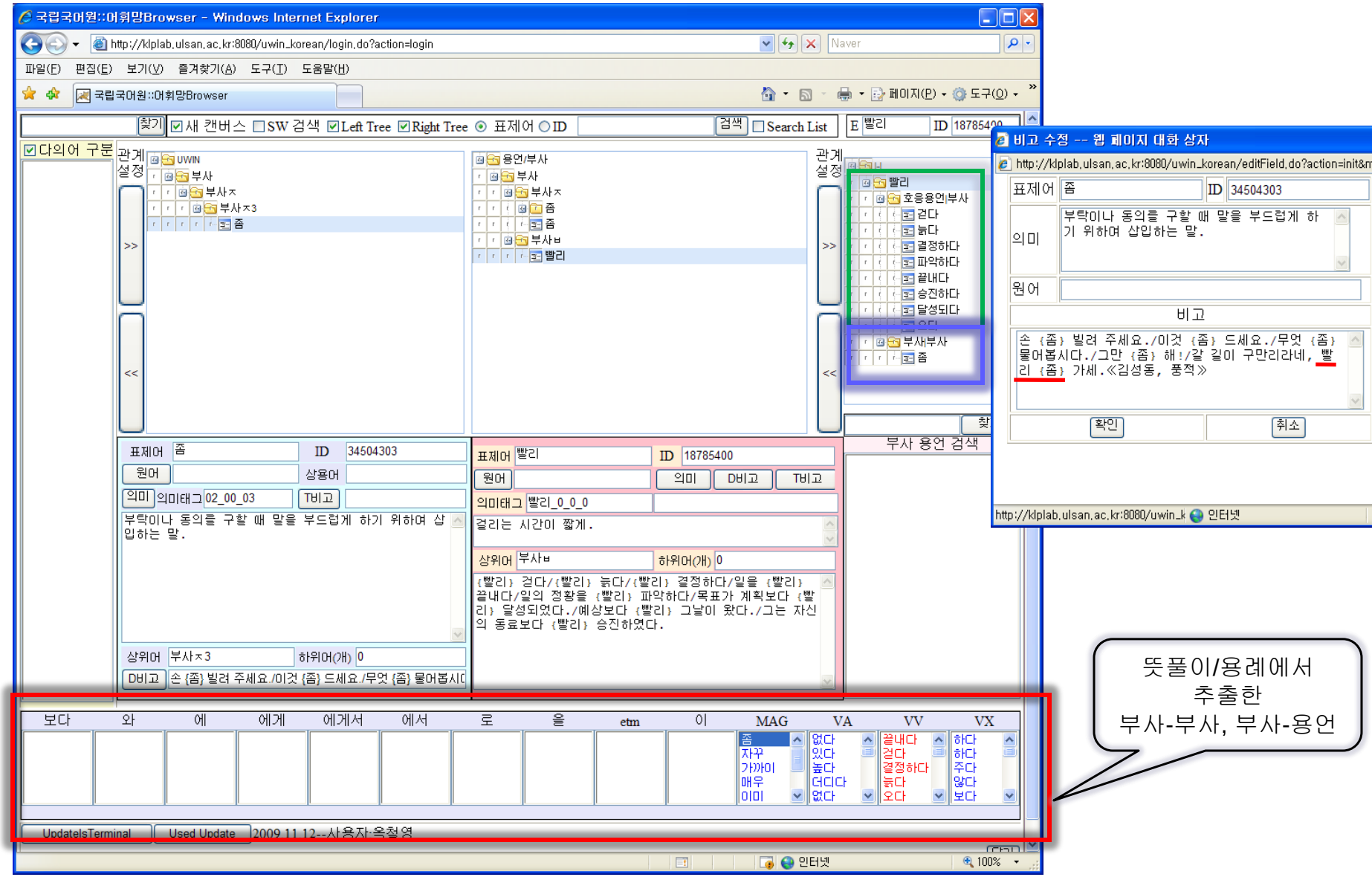
뜻풀이 및 용례 의미 태깅 (3)



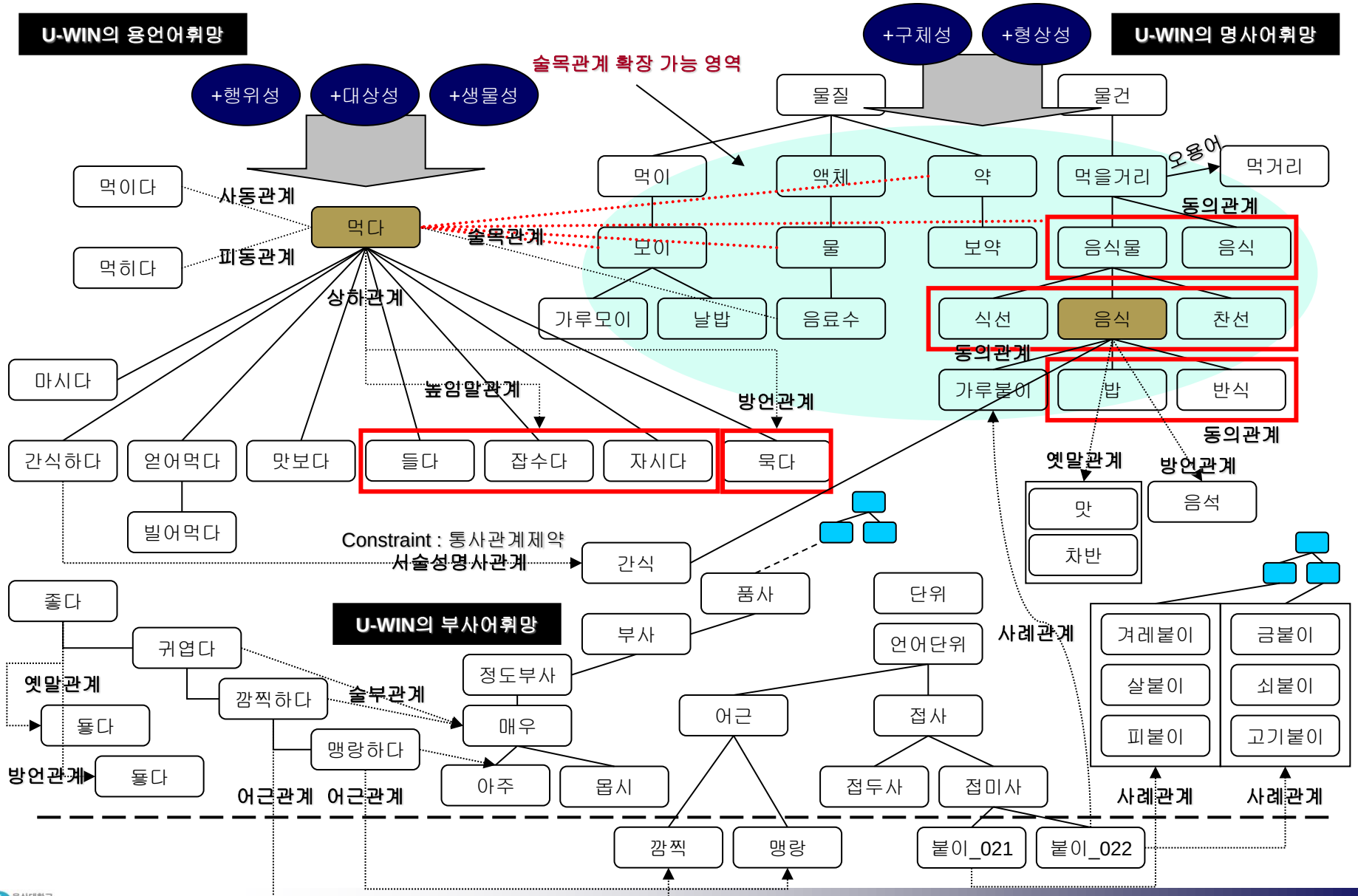
용언어휘망: 하위범주화 정보 구축



부사 어휘망: 부사, 용언, 명사 결합정보



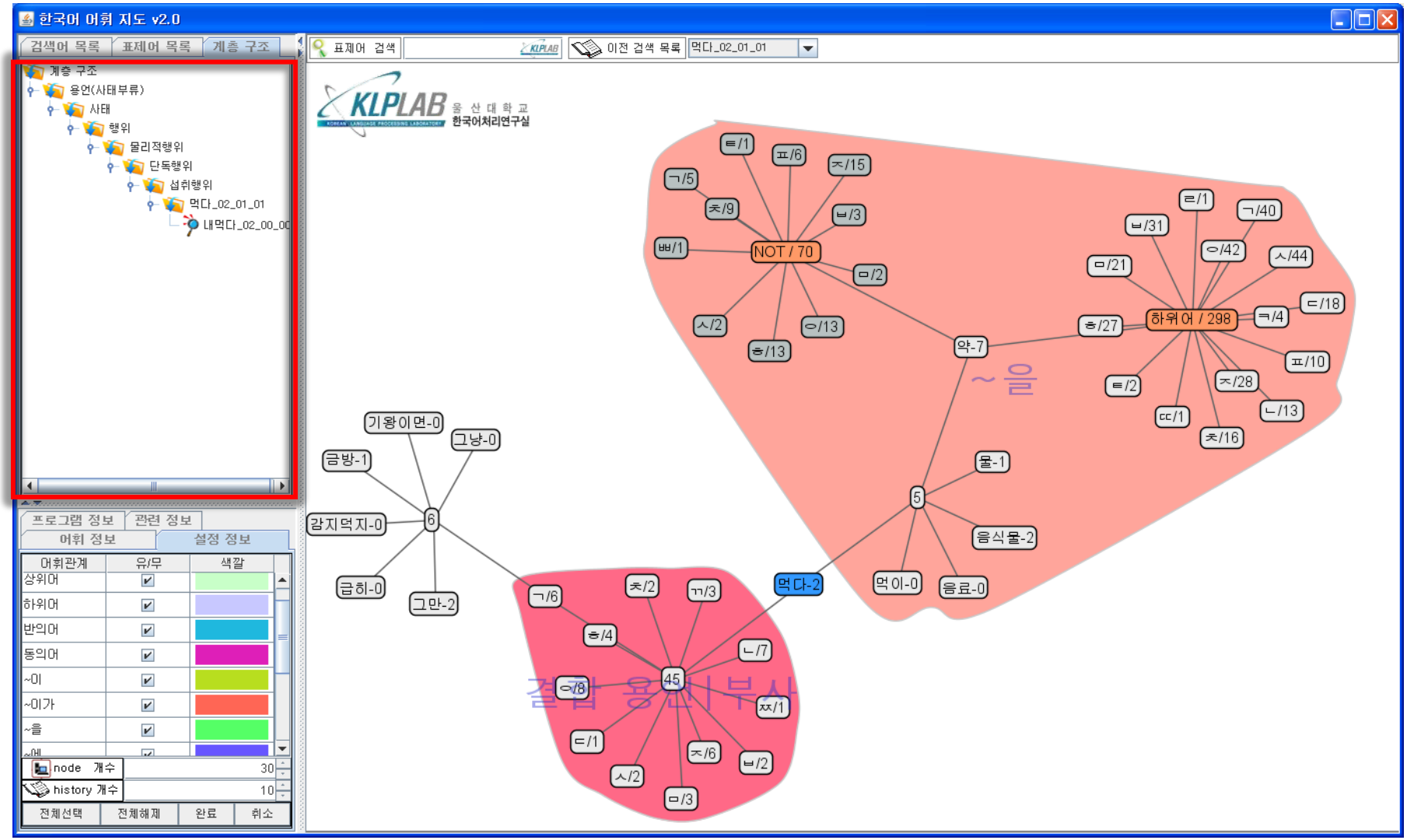
한국어 어휘지도(UWordMap) (1)



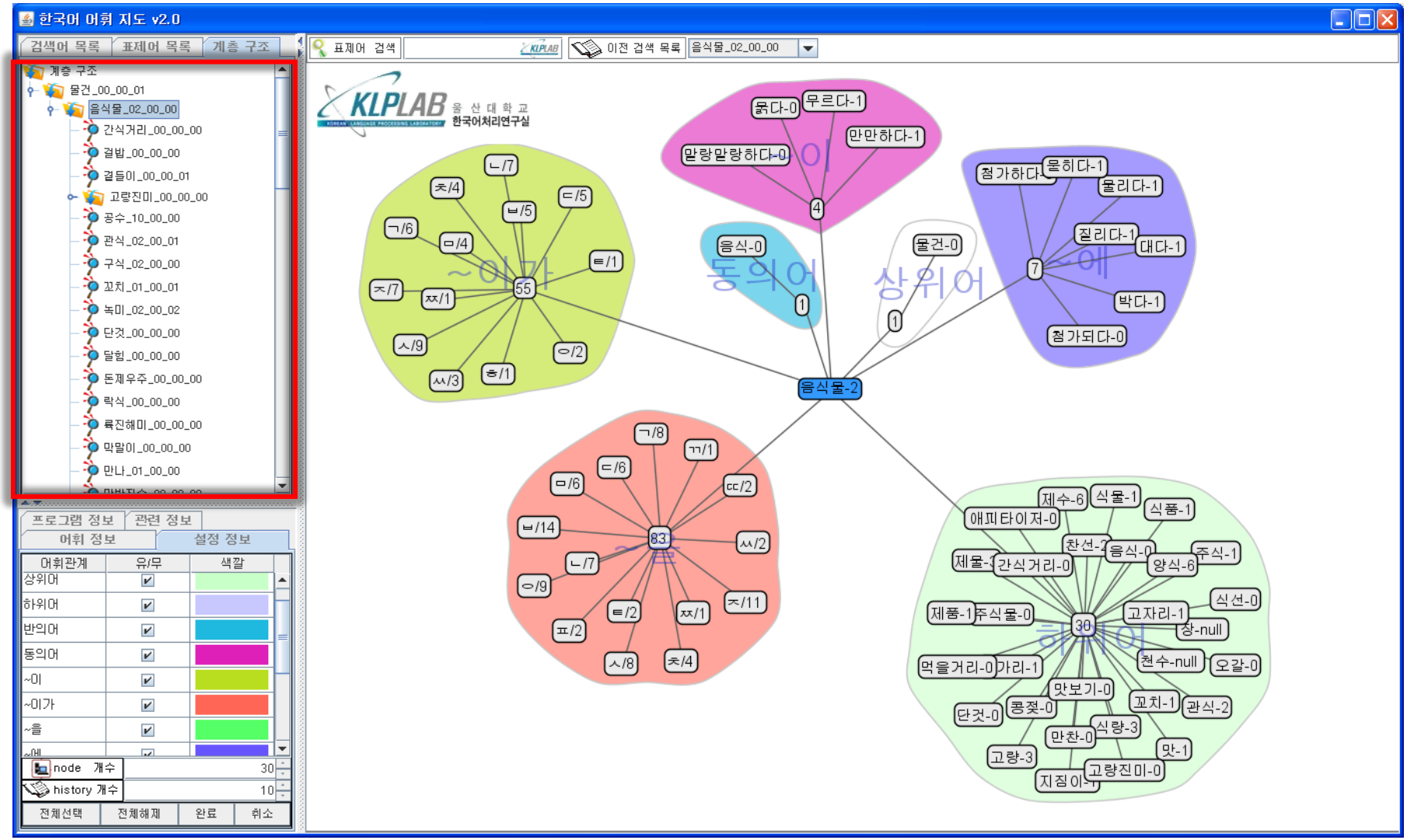
한국어 어휘지도(UWordMap) (2)

품사	표준국어 대사전	U-WIN (계층관계)	UWordMap (Exo Brain Project)		
			1차년도 (14.03.24)	2차년도 (15.01.20)	3차년도 (15.06.19)
명사	377,281	365,774	LCS 72,020	97,410	98,264
동사	90,237	73,694	29,345	46,410	51,336
형용사	21,618	16,853	4,653	7,709	12,438
부사	25,178	17,697	6,186	6,187	6,187
합계	514,314	474,018	123,823	157,718	168,225

한국어 어휘지도 Browser (1)



한국어 어휘지도 Browser (2)



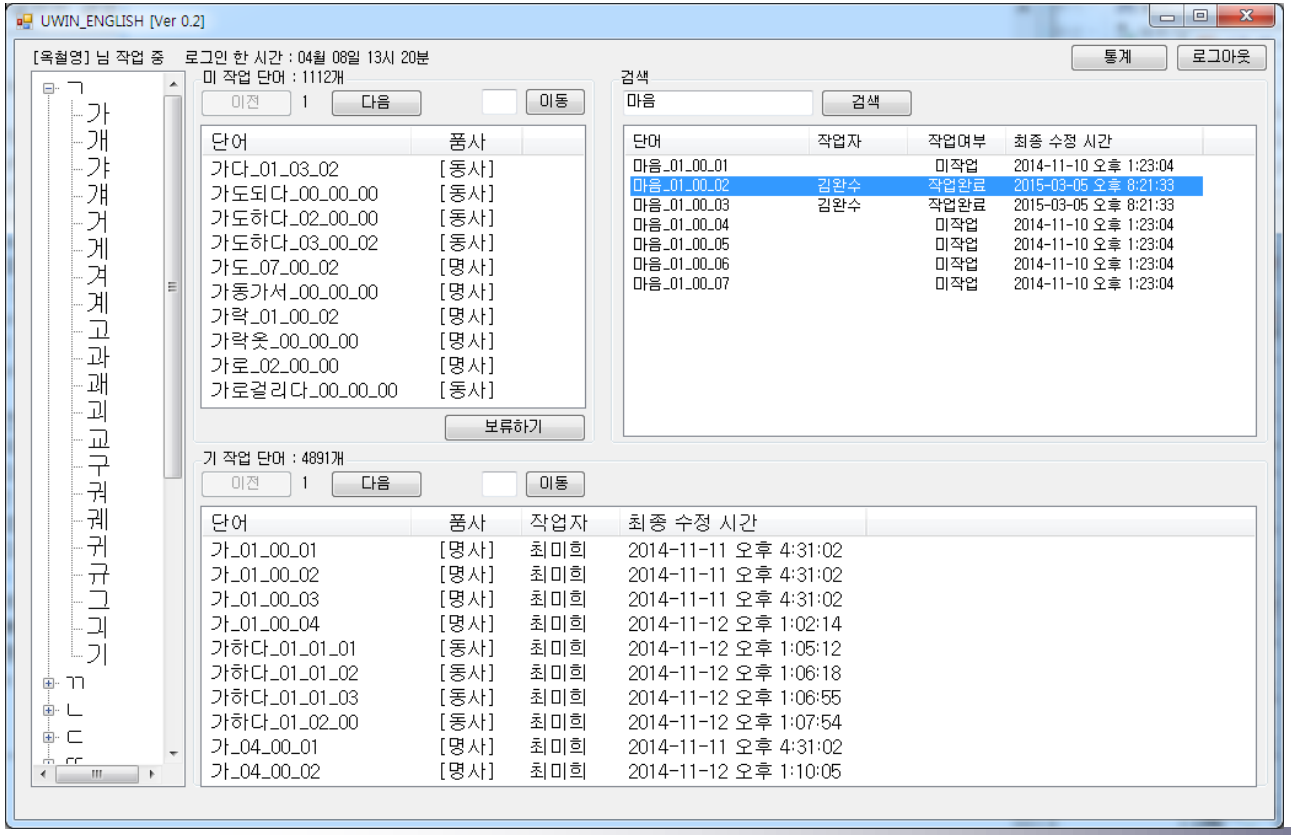
한국어 어휘지도 API

함수명	설명
GetPS	(단어 or 동형)의 다의어 받아오기
GetHyperWord	단어의 1레벨 위의 상위어 받아오기
GetHyperAllWord	다의어의 모든 상위어 받아오기
GetHypoWord	다의어의 1레벨 아래의 하위어 받아오기
GetNRelV	해당 용언과 논항(격조사)으로 관련된 명사 받아오기
GetVRelN	해당 명사을 논항(격조사)으로 가지는 용언 받아오기
GetRelSubCt	용언과 명사가 연결된 논항(격조사) 받아오기
GetSynSet	다의어의 동의어 받아오기
GetAntSet	다의어의 반의어 받아오기
GetDistance	다의어1과 다의어2의 거리 받아오기

- ❖ 부사 기준 API 추가 예정(2015)
- ❖ 어휘간 유사도 측정 (다양한 option 제공)

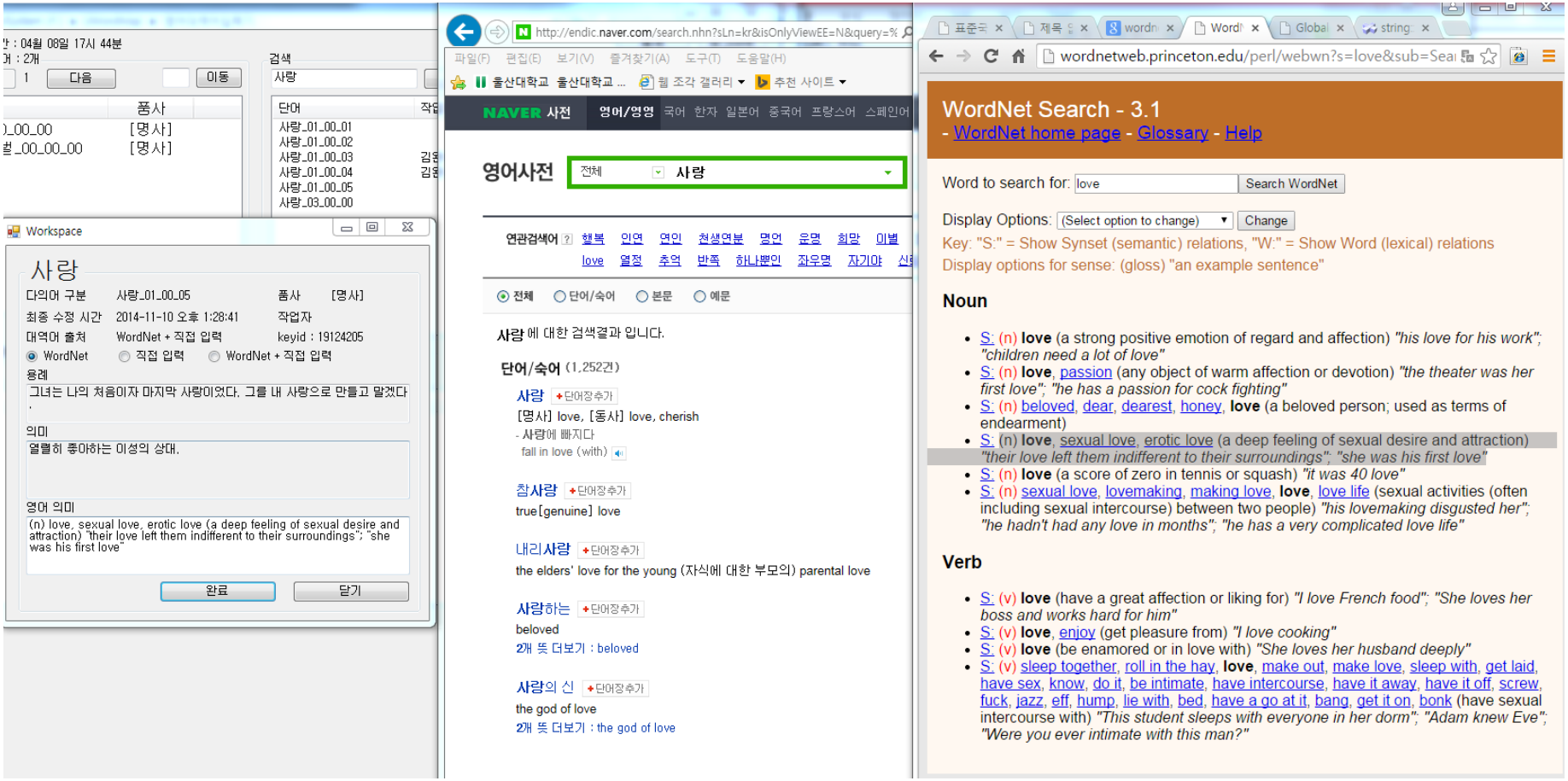
Mapping UWordMap to WordNet(1)

- 표준국어대사전 다의어 의미를 WordNet Synset으로의 mapping
  - 전문용어, 방언 제외한 239,000 다의어 대상
  - KorLex-표준국어대사전 mapping 정보 활용(약 100,000 entry)
  - 다의어별 의미에 맞는 대역어(NAVER 사전) – WordNet synset
- Mapping 도구



Mapping UWordMap to WordNet(2)

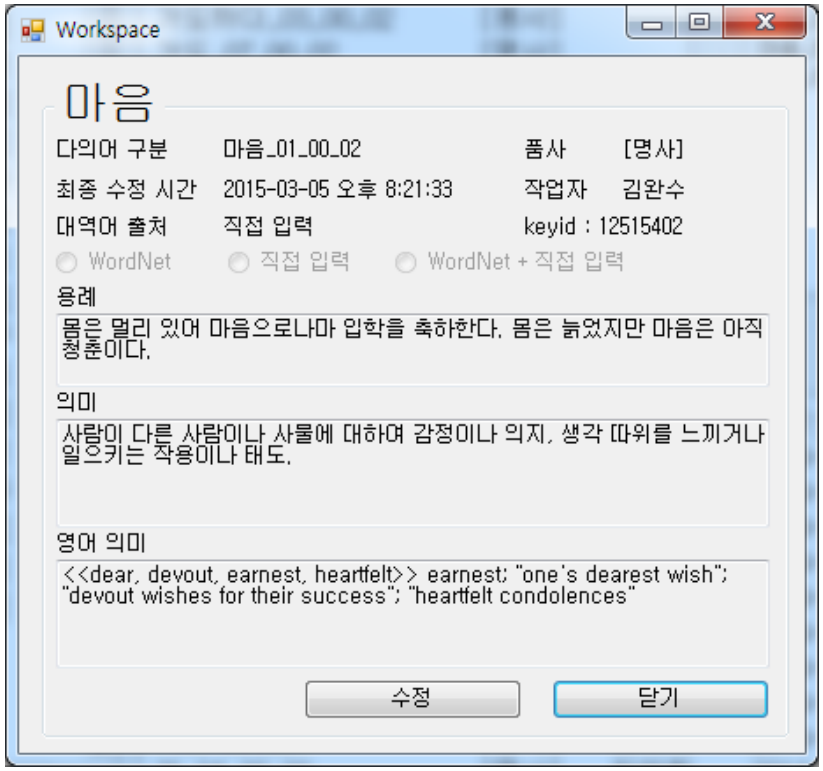
❖ 다의어별 의미에 맞는 대역어(NAVER 사전) – WordNet synset

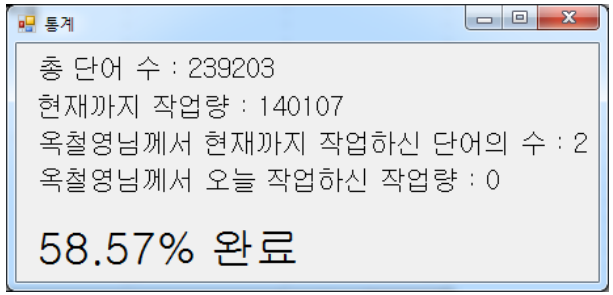


Mapping UWordMap to WordNet(3)

- Mapping 도구 기능
  - Mapping **방법**: WordNet, **직접입력**, WordNet+**직접입력**
  - **작업자별 통계**
  - **보류**
  - **검색**

(2015.08.18 기준)





의미역(Semantic Role) (1)

- 세종 의미역 15개

- PropBank
  - ArgN 최대 6개
  - ArgM 13개

- ExoBrain 22개(2013 결정)
  - 격조사 의미분석
  - 남승호 기준(16개) + 추가 (6개)
    - 시간격
    - MANNER
    - 목적
    - 내용
    - 정도
    - 재료

세종전자사전	PropBank	Exo-Brain
행위주	Agent	행동주
경험주		경험주
심리경험주		경험주
동반주		동반주
대상격	Theme	대상
처소격	LOC	처소
방향격	DIR	방향
도착격		착점
결과상태		착점
출발격		기점
도구격	INS	도구
영향주	Patient	피동주
기준치		비교기준
목적		목적
내용		내용
…	…	…
		자격
	TMP	시간
	MNR	방법
15개		22개

의미역(Semantic Role) (2)

의미역	이,가, 는,은	께서, (이)라서, 서	을, 를, ㄹ	더러, 보고	에	에게, 한테	에게로, 에게를, 한테로	에서	에게서, 에서부터, (으)로부터, 한테서	에다가	에를, 에의, 에야	(으)로	(으) 로서	(으) 로써	보다, 처럼, 마따나	같이	와,과, 하고, (이)랑	고, (이)라 고
행동주 AGT	o	o			o	o		o				o						
대상 THM	o		o		o													
경험주 EXP	o	o	o															
피동주 PAT	o	o	o	o		o												
자극 STM	o		o		o							o						
처소 LOC	o		o		o			o		o								
기점 SRC	o		o		o	o		o	o	o			o					
착점 GOL	o		o	o	o	o	o				o	o						
원인 CAU	o				o			o				o						
수혜자 BEN	o		o	o	o	o	o											
경로 ROU			o									o						
방향 DIR			o									o						
목적 PUR			o															
정도 DGR			o															
재료 MAT			o									o		o				
도구 INS			o		o							o		o				
시간 TMP					o							o		o		o		
비교기준 CRT					o			o							o	o	o	
동반주 COM												o					o	
내용 CNT												o						o
자격 ROL					o							o	o					
방법 MNR												o						

의미역(Semantic Role) (3)

❖ UPropBank : 90,090개(다의어, 동사, 형용사)의 필수논항에 의미역 부여

- 문형이 있는 용언: 41,518 (2013년)
- 문형이 없는 용언(자동사, 형용사) : 48,572 (2014년)

표제어	문형	뜻풀이	용례	의미역 부착 격틀
감금되다	◁ …에▷	감금.	감옥에 감금되다. 독방에 감금되다. 저녁이 되자 두 죄인은 좁고 어두운 방에 감금되었다.	X:행동주 Z:처소-에
감금하다	◁ …을 …에▷	감금.	자식 놈을 곳간에 감금해 놓고 아비 꼴이 어떤가 …. 사장은 … 운전수를 이상한 창고 같은 방 속에 감금해 버리고는 혼자 그 집 안으로 사라져 버린다.	X:행동주 Y:대상-을 Z:처소-에
감급되다	◁ …으로▷	감급.		X:대상 Z:정도-으로, 착점
감급하다	◁ …을 …으로▷	감급.	직원들 봉급을 100만 원에서 90만 원으로 감급했다.	X:행동주 Y:대상-을 Z:착점-으로
감기다 020101	◁ …에▷	'감다03[1]⑴'의 피동사.	왕거미의 거미줄에 풍뎅이가 감기듯이 그가 그 검은 노끈에 감기고 있었다. 아이들은 뛰어가면서 발목에 감기는 개울물의 감촉을 미리부터 즐긴다.	X:피동주 Z:도구-에
감기다 020102	◁ …에▷	옷 따위가 몸을 친친 감듯 달라붙다.	젖은 치맛자락이 맨살에 감기는 듯하다. 이불 속에 감겨 있던 몸이 홀쩍 문밖으로 나서니 추위는 살을 에는 듯하였다.	X:대상 Z:착점-에
감기다 020103	◁ …에▷	음식 따위가 감칠맛이 있게 착착 달라붙다.	며칠 같이 유하면서 송도 특유의 맛깔스러운 음식과 집에서 담근 혀에 감기는 약주술로 송별을 겸해 회포를 풀 기회를 갖고 싶어서 청해 본 거였다. 큰아들이 혀에 착착 감기는 조청이라면 작은아들은 목구멍에 걸린 가시였다.	X:대상 Z:착점-에
감기다 020104	◁ …에▷	사람이나 동물이 달라붙어서 떠나지 아니하다.	강아지가 내 옆에 감겨서 꼼짝도 않는다. 손자가 할머니 다리에 감겨 떨어질 줄을 모른다.	X:대상 Z:착점-에
감기다 020105	◁ …에▷	음식을 너무 먹어 몸을 가누지 못하다.	유복이가 여러 날 변변히 먹지 못하고 굶주린 끝에 배불리 먹고 음식에 감기어서 길 갈 기운이 없어졌다.	X:피동주 Z:원인-에
감기다 030000	◁ …을▷	'감다01⑴'의 사동사.	눈을 감기다.	X:행동주 Y:대상-을
감기다 040000	◁ …을▷	'감다02'의 사동사.	머리를 감기다. 어머니는 할머니의 손톱과 발톱도 깎아 드리고 머리도 감겨 드렸다. 어머니는 창포물에 언니의 머리를 감기면서, 여자란 머릿결이 고와야 용모가 아름다운 것이라고 말씀하였다.	X:행동주 Y:대상-을
감기다 050000	◁ …에게 …을▷	'감다03[1]'의 사동사.	어머니는 아들에게 헝클어진 실을 감겼다.	X:행동주 Z:피동주 -에게 Y:대상-을

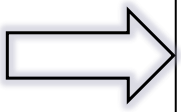
한국어 의미처리시스템(UTagger)

- ✓ 동형이의어 WSD(UTagger-HG)
- ✓ 의존관계 분석(UTagger-DP)
- ✓ 다의어 WSD(UTagger-PS)
- ✓ 의미역 태깅시스템(UTagger-SR)
- ✓ 개체명 인식기(UTagger-NE)

의미처리(어휘WSD) 필요성

❖ 형태소 분석 오류 : 문맥 반영 못함
- 과도한 음주로 인해 간이 좋지 않다/ 간이 식당에서 요기를 했다
- 논이 걸어서 벼가 잘 자란다/ 옷을 걸어서 보관하다/ 걸어서 하늘까지
- 그는 칠 주야를 걸어 서울에 왔다/ 헤엄을 칠 줄 아는 …

❖ 구문/의존관계 분석 시 Case frame (용언) 제공
- 주한 미군 철수 땐 일 재무장 초래
- 중국과 깊이 관계
- 너 기저귀 차고 놀 때 나는 공 차고 놀았다



차다02 전체 보기 〔차, 차니〕
「동사」
【…을】
「1」 발로 내어 지르거나 발아 올리다.
「2」 발을 힘껏 뻗어 사람을 치다.
「3」 혀끝을 입천장 앞쪽에 붙였다가 떼어 소리를 내다.
「4」 발로 힘 있게 밀어젖히다.
「5」 (속되게) 주로 남녀 관계에서 일방적으로 관계를 끊다.
「6」 날쌔게 빼앗거나 움켜 가지다.

차다03 전체 보기 〔차, 차니〕
「동사」
[1] 【…에 …을】
「1」 물건을 몸의 한 부분에 달아매거나 끼워서 지니다.
「2」 수갑이나 차꼬 따위를 팔목이나 발목에 끼우다.
[2] 【…을】
(속되게) 애인으로 삼아 데리고 다니다.

❖ 전문용어/개체명, 합성어, 파생어의 의미 분석
- "한국전자통신연구원", "한국 전자통신 연구원"
- 전기기록 : 前期+기록, 電氣+기록, 轉機+기록
- 내각제(制), 추모제(祭), 미국제(製), 위염제(劑)

❖ 표준국어대사전 표제어 507,100 개 중 125,152 개 동형이의어(25%)

UTagger 특징(1)

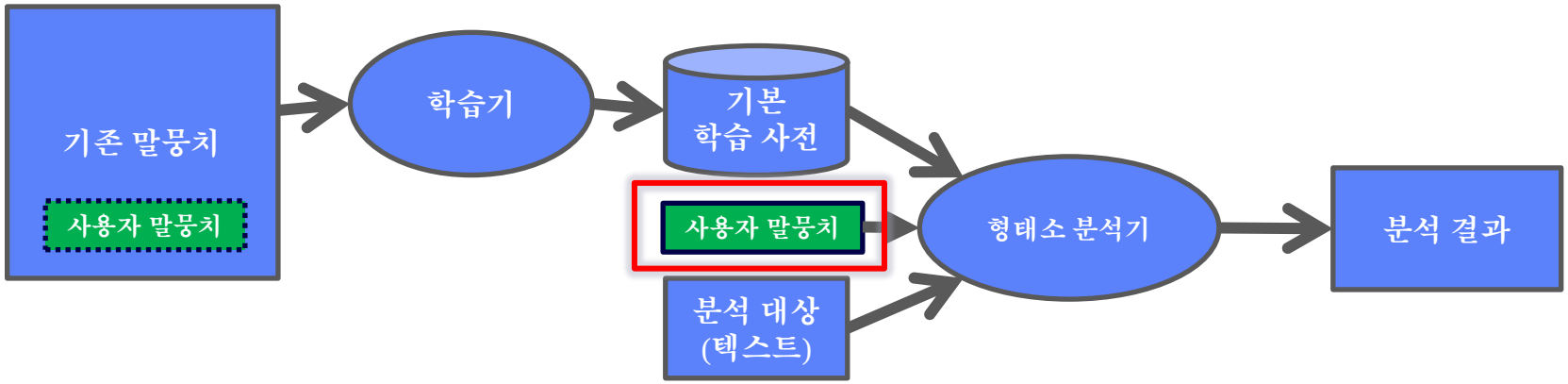
❖ 학습말뭉치 기반
- 세종형태의미부착 말뭉치 : 11,116,320 어절(339파일)
- 세종원시말뭉치 태깅 : 52,338,450 어절(1,770 파일)

❖ 학습사전
- 자체 개발 파일DB 사용
- CKMA : 형태소분석용, 인접형태소/품사 전이확률
- 품사/동형이의어 태깅용, 인접 두 어절 간의 전이 : biAF, biEF, biFF

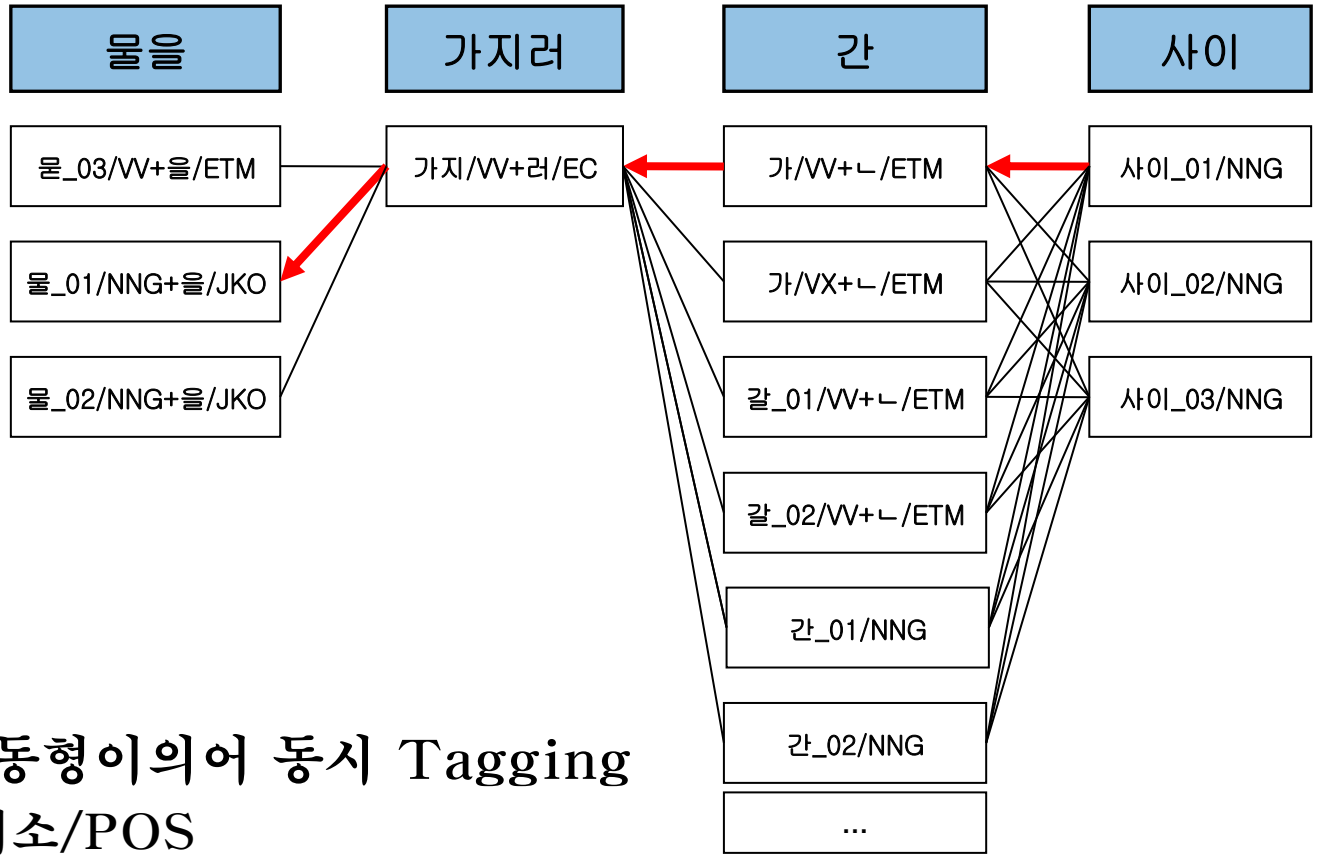
❖ CKMA (Corpus-based Korean Morphology Analysis)
- 형태소분석 : 기분석 어절의 형태소간 전이 확률 이용 (재현율 99.9%)
- 불규칙용언에 대한 분석/변형/복원 규칙 없음
- 체언류/용언류 형태소분석 루틴 불필요
- 띄어쓰기 오류 어절 분석(주민들사이에, 그리기시작했으며, 박전대표위원)

UTagger 특징(2)

- 품사/동형이의어 태깅
  - 형태의미 주석말뭉치에서 bigram 간의 전이 확률 학습(단계별 적용)
  - HMM 모델(2012)
  - SCP(부분어절 조건부확률) 모델(2013)
  - 복합명사/접미사 동형이의어 태깅
- 미등록어/신조어 추정
  - CKMA 기반
  - 사용자사전 지원
- 사용자말뭉치 문맥학습(incremental learning)

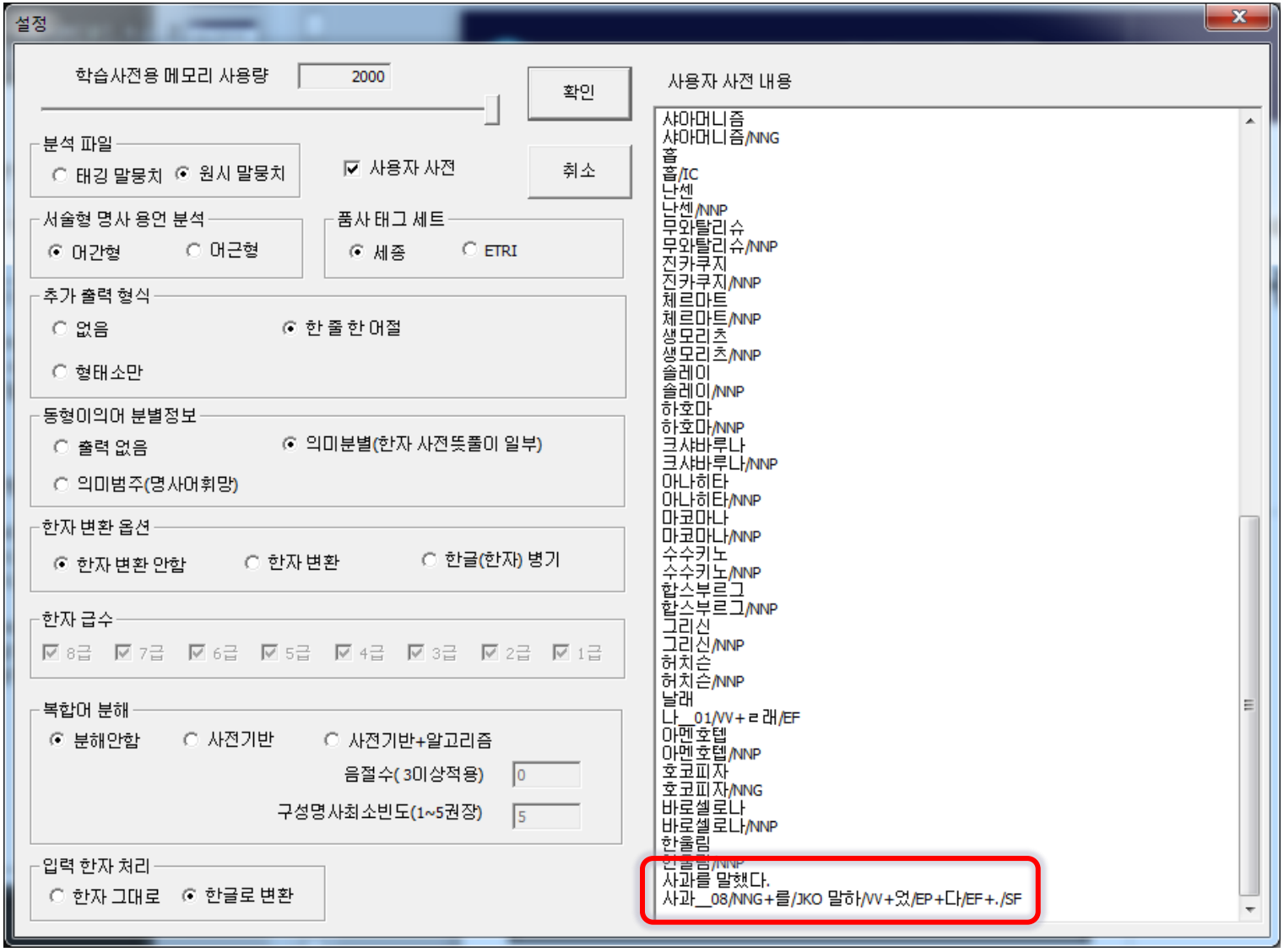


단계별 전이 모델 & Viterbi



- ❖ POS/동형이의어 동시 Tagging
  - 형태소/POS
  - biAA : 앞어절 전체 형태소, 뒷어절 전체 형태소
  - biAD : 앞어절 전체 형태소, 뒷어절 첫 형태소
  - biLD : 앞어절 마지막 형태소, 뒷어절 첫 형태소
  - 논문 참조 : 정보과학회논문지 39/5, 39/11 (2012)

UTagger option (대화형사용자용)



UTagger option (분석 option)


```
MEMORY 2000                  // 학습사전을 load할 memory 양(KByte), 많을수록 빠름
HLX_DIR ..\학습사전\          // 학습사전 폴더
cache 100000                 // 캐쉬 사용량 조절. 윈도우는 100000 권장. 리눅스 유닉스는 0 권장
hanja_to_hangul 0            // 입력 한자 처리 한자그대로(0), 한글로 변환(1)
depen                        // 의존관계 분석시 용언의 의미제약(UWordMap) 사용
tag_poly_uwm 1               // 다의어 태깅

Noun_Attribute 0             // 복합명사 분해시 속성정보사용 여부 미사용(0), 사용(1)
analyzeMore 0                // 고유명사 분해할 최소 음절(3) , 고유명사 추가분해 안함(0)
analyzeMoreNNG 1             // 고유명사 분해 시, 일반명사도 적용(분해할 최소 음절 적용)
analyzeMoreMiniFreq 5        // 복명명사 분해할 경우, 분해된 구성명사의 최소출현빈도
separate_compound 0          // 사전등재 복합명사(A^B, A-B 유형) 추가 분해 여부

Light 2                      // CKMA 분석방법(정확률-속도): 가장느림(0), 약간느림(1), 빠름(2)
tagging 2                    // 형태소분석(CKMA)만(0), HMM태깅(1), SCP태깅(2)
useAD 0                      // HMM태깅: 앞 어절 전체형태소, 뒤 어절의 첫 형태소
useLD 0                      // HMM태깅: 앞 어절 마지막형태소, 뒤 어절의 첫 형태소
useAF 1                      // SCP태깅: 앞 어절전체, 뒤 어절의 첫 2개 음절
useEF 1                      // SCP태깅: 앞 어절의 마지막 2개음절, 뒤 어절의 첫 2개 음절
recursive 1                  // 띄어쓰기 오류 재귀분석
probability_equation 0       // 전이확률 계산시 확률식으로 계산 (정확률 조금 낮아짐)
```

UTagger option (출력 option)


```
TAG_STYLE 0                          // 울산대/세종 태그(0), ETRI 태그(1)
hadaVerb 0                           // 서술형 명사 용언  어간형(0), 어근형(1)
print_sense_num 1                    // 태깅 결과 출력 시 어깨번호 출력 안함(0), 동형이의어(1)
hangul_to_hanja 0                    // 한자변환 안함(0), 변환(1), 병기(2)
hanjaLevel 0 1 2 3 4 5 6 7 8         // 한자 변환/병기 시 출력할 한자능력검정 급수
one_length_hanja_word_no_trans 1             // 1음절 한자어 변환여부 변환(0), 미변환(1)
ucs2le 1                             // 한자 출력시 유니코드 출력 안함(0), 출력(1)
hanja_UCS2 1                         // ucs2le=1 일 때 나라별 한자 한국ANSI(0), 한국(1), 대만(2), 중국(3), 일본(4)
CATE 1                               // 의미매핑정보출력(대화식) 없음(0), 한자-뜻풀이(1), 의미범주/상위어(2)

print_end_empty_line 1               // 줄 단위로 모든 출력이 끝나면 마지막에 빈 줄을 출력한다.
print_original_sentence 0            // 입력 문장을 출력 여부. 안함(0), 출력(1)
print_one_line_sentence 1            // 태깅 결과를 한 줄로 출력. 안함(0), 출력(1)
print_ex 3                           // 태깅결과 출력 안함(0), 한줄에 한어절(3)
print_depen 1                        // 의존관계 출력 여부 안함(0), 출력(1), 규칙포함(2)
print_guess_line 0                   // 미학습어절 별도 출력여부 안함(0), 출력(1)
preserve_spliter 1                   // 입력 문장에서 어절 사이에 띄어쓰기 모양 유지 여부 안함(0), 유지(1)
preserve_newLine 1                   // 입력 빈줄 출력 여부 안함(0), 유지(1)
debug_msg 1                          // 콘솔로 실행시 각종 디버깅용 메세지를 출력. 0안함. 1사용
```

UTagger 성능 평가 (1)

	2010년	2012년	2013년
형태소분석	규칙기반 (icma)	학습기반 (CKMA)	학습기반 (CKMA)
tagging	HMM	HMM (AD,LD)	SCP (AF,EF,FF)
정확률	88.93%	96.49%	96.37% (2013) 96.53%(2015)
속도(full option)	180sec	42.6sec	25.7sec
공유메모리	지원안함	지원	지원
띄어쓰기 오류 처리	처리 못함	기분석어절포함 된 경우 분석	기분석어절포함 된 경우 분석
복합명사 의미분석	못함	분석(접사 포함)	분석(접사 포함)
Code (한자)	KS5601	KS5601	Unicode

❖ 학습사전 : 세종형태의미말뭉치 1천만 어절 대상, 약 1G

UTagger-HM 성능 평가 (2)

단어	동형이의어/POS	학습사전 출현빈도	10% 정답	UTagger 태깅	정답률
눈	눈__01/NNG	12,455	1,236	1,244	100.56%
	눈__04/NNG	1,379	141	133	94.33%
	소계	13,834	1,377	1,377	100.0%
손	손__01/NNG	9,209	925	928	100.32%
	손__05/NNB	3	1	0	0.00%
	손__08/NNP	188	18	18	100.0%
	손__09/NNG	27	1	2	20.0%
	소계	9,427	945	948	100.32%
말	말__01/NNG	34,136	3,337	3,337	100.0%
	말__03/NNB	118	7	6	85.71%
	말__05/NNG	584	57	58	101.75%
	말__07/NNG	2	1	1	100.0%
	말__11/NNB	2,723	219	262	119.63%
	소계	37,563	3,621	3664	101.19%
거리	거리__01/NNG	2,056	220	210	99.55%
	거리__02/NNB	339	31	31	100.0%
	거리__04/NNB	14	1	1	100.0%
	거리__08/NNG	1,491	152	153	100.66%
	소계	3,900	404	395	97.77%
바람	바람__01/NNB	1,222	128	125	97.66%
	바람__01/NNG	3,468	357	363	101.68%
	바람__02/NNG	120	10	7	70.0%
	소계	4,810	495	495	100.0%
자리	자리__01/NNG	7,589	717	723	100.84%
	자리__02/NNG	94	12	6	50.0%
	소계	7,683	729	729	100.0%
의사	의사__02/NNG	1,018	108	107	99.07%
	의사__03/NNG	66	6	6	100.0%
	의사__04/NNG	1	1	1	100.0%
	의사__09/NNG	3	1	1	100.0%
	의사__11/NNG	6	2	2	100.0%
	의사__12/NNG	1,754	183	184	100.55%
	의사__14/NNG	24	1	1	100.0%
	소계	2,872	302	302	100.0%
점	점__02/NNG	7	1	1	100.0%
	점__03/NNG	86	9	9	100.0%
	점__10/NNB	1,557	237	154	64.98%
	점__10/NNG	9,861	942	1,025	108.81%
	소계	11,491	1,189	1,189	100.0%
밤	밤__01/NNG	5,388	494	494	100.0%
	밤__02/NNG	124	17	17	100.0%
	소계	5,512	511	511	100.0%
목	목__01/NNG	1,968	188	189	100.53%
	목__09/NNG	3	1	1	100.0%
	목__10/NNG	65	7	7	100.0%
	목__12/NNG	3	1	1	100.0%
	목__13/NNG	9	1	1	100.0%
	목__14/NNG	13	2	0	0.0%
	소계	2,061	200	199	99.5%

UTagger-DP(의존관계 분석) (1)

- 문장 구성성분의 의존 관계 분석
- UTagger-HM의 분석결과 이용 : 동형이의어 분별
  - 동형이의어 용언의 문형 정보 활용
    - 차__01/VV : ◁…에▷◁…으로▷ "독에 물이 차다"
    - 차__02/VV : ◁…을▷ "공을 차다/혀를 차다"
    - 차__03/VV : ◁…에 …을▷ "허리에 칼을 차다"
    - 차__04/VA : "성격이 차고 매섭다 / 바람이 차다"
  - 서술성 명사
    - 중국과 깊이 관계한 협상은 …

      중국__01/NNP+과/JKB 깊이__02/MAG 관계하/VV+ㄴ/ETM 협상__01/NNG+은/JX

    - 중국과 깊이 관계 …

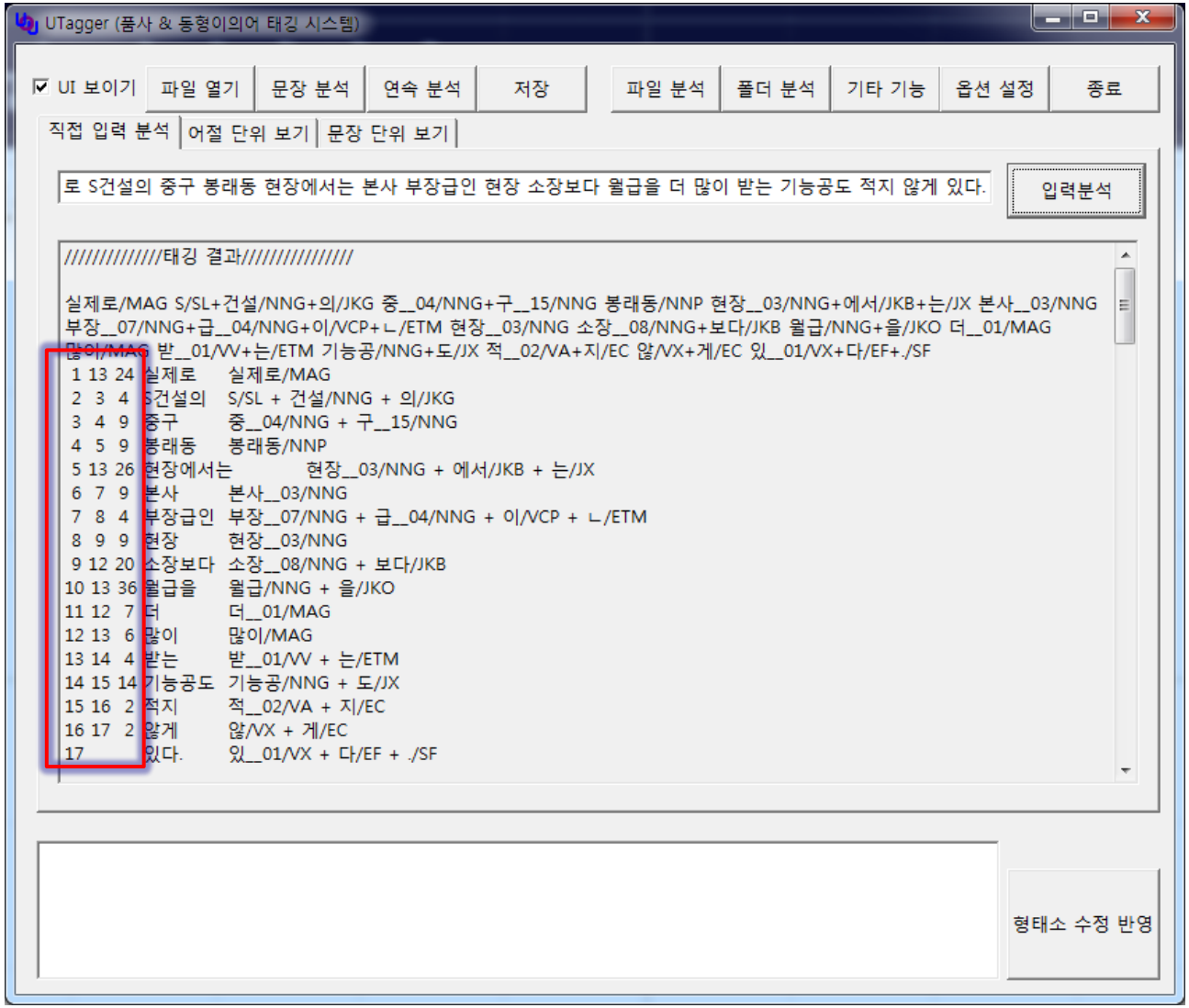
      중국__01/NNP+과/JKB 깊이__02/MAG 관계__05/NNG

      ? ?

- 규칙기반 정확률 : 85.53% (의존관계로 변환된 세종구문분석말뭉치 대상)
- 기계학습기반 의존관계분석 (2015) : 정확률 ??.??% > 87%

UTagger-DP(의존관계 분석) (2)

❖ UTagger-DP (ver. 0.9)



UTagger-DP(의존관계 분석) (3)

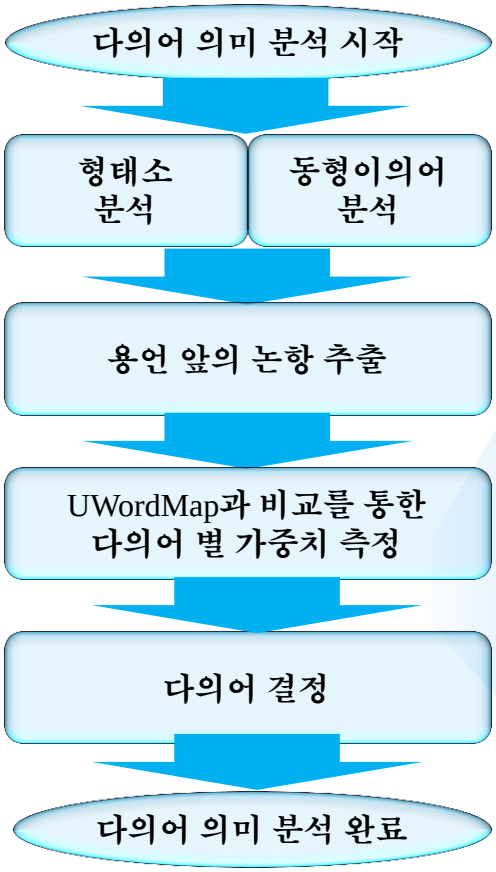
- 규칙기반 정확률 : 85.53% (의존관계로 변환된 세종구문분석말뭉치)

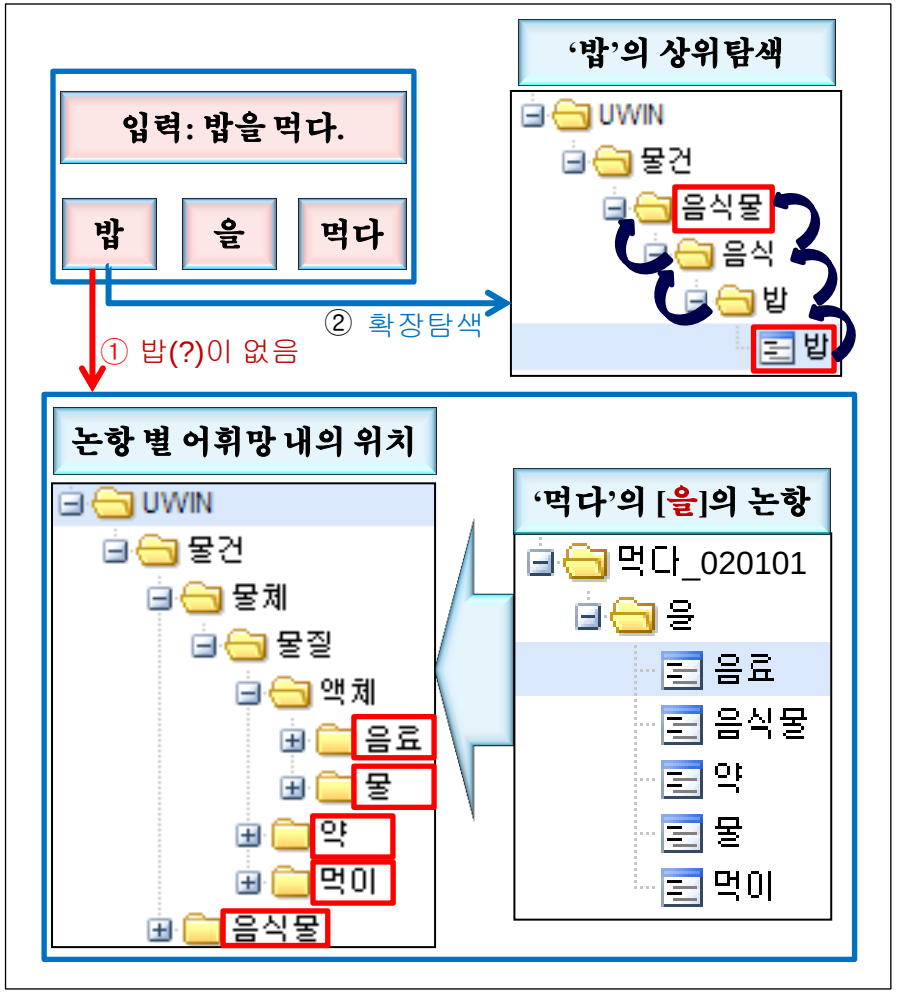
규칙번호	정답	오답	규칙적용	전체비율	정확률
rule_1	3,801	1	3,802	1.02	99.97
rule_2	46,134	380	46,514	12.49	99.18
rule_3	2,816	516	3,332	0.89	84.51
rule_4	55,565	10,330	65,895	17.70	84.32
rule_5	7,685	2,360	10,045	2.70	76.51
rule_6	8,832	192	9,024	2.42	97.87
rule_7	392	4	396	0.11	98.99
rule_8	454	21	475	0.13	95.58
rule_9	34,008	5,106	39,114	10.51	86.95
rule_10	518	0	518	0.14	100.00
rule_11	2,124	131	2,255	0.61	94.19
rule_12	494	497	991	0.27	49.85
rule_13	47,941	579	48,520	13.03	98.81
rule_14	16,089	2,876	18,965	5.09	84.84
rule_15	29,330	2,392	31,722	8.52	92.46
rule_16	6,435	892	7,327	1.97	87.83
rule_17	211	2	213	0.06	99.06
rule_18	2,597	472	3,069	0.82	84.62
rule_19	0	0			
rule_20	402	200	602	0.16	66.78
rule_21	2,455	691	3,146	0.84	78.04

규칙번호	정답	오답	규칙적용	전체비율	정확률
rule_22	3,203	1,575	4,778	1.28	67.04
rule_23	695	107	802	0.22	86.66
rule_24	783	742	1,525	0.41	51.34
rule_25	468	152	620	0.17	75.48
rule_26	1,236	496	1,732	0.47	71.36
rule_27	10,145	3,919	14,064	3.78	72.13
rule_28	11	2	13	0.00	84.62
rule_29	8,334	7,380	15,714	4.22	53.04
rule_30	2,470	577	3,047	0.82	81.06
rule_31	510	108	618	0.17	82.52
rule_32	1,599	1,021	2,620	0.70	61.03
rule_33	2,048	569	2,617	0.70	78.26
rule_34	508	348	856	0.23	59.35
rule_35	691	167	858	0.23	80.54
rule_36	3,001	1,301	4,302	1.16	69.76
rule_37	0	0			
rule_38	82	47	129	0.03	63.57
rule_39	491	101	592	0.16	82.94
rule_40	13,904	7,602	21,506	5.78	64.65
total	318,462	53,856	372,318	100.00	85.53
UTagger 오류	3,605				

UTagger-PS(다의어 WSD) (1)

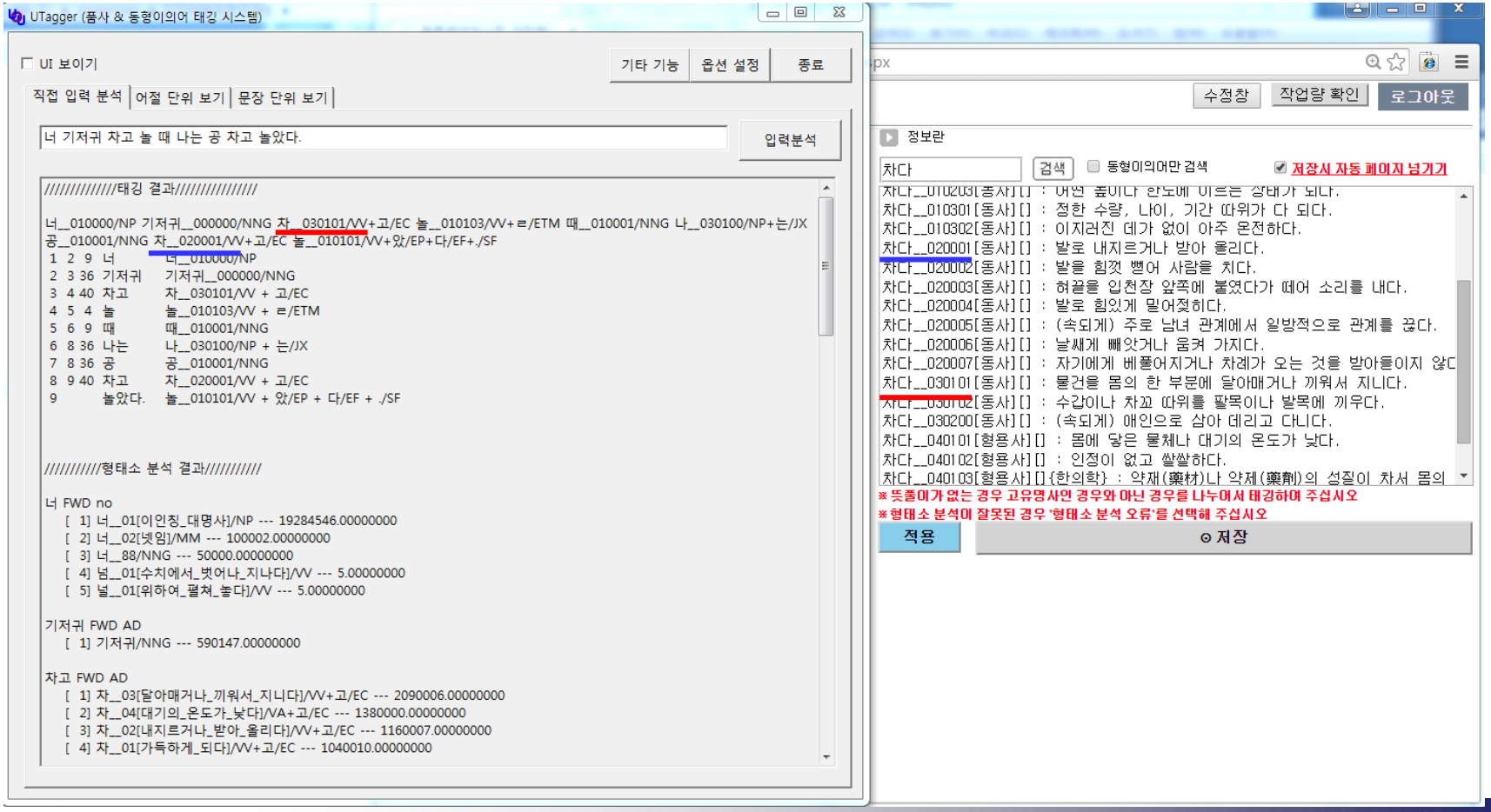
❖ 의미 분석 과정





UTagger-PS(다의어 WSD) (2)

- UTagger-PS (ver. 0.5)
  - 동형이의어분별 후 어휘지도 상에서 "용언-명사" 발견
  - 의존관계 사용하지 않음
  - 용언의 왼쪽, 오른쪽 어절에 대해 다른 가중치 적용



UTagger-PS(다의어 WSD) (3)

❖ 전체 다의어 대상 실험 (2015.03.05)

- 표준국어대사전에서 용언을 포함하는 용례 대상 : 210,426문장
- 현재의 어휘지도 구축 현황 (18쪽 참고)

	명사	부사	동사	형용사	합계
정답	572,182	74,292	337,613	88,567	1,072,654
Base line (첫번째다의어)	430,660 (75.27%)	48,170 (64.84%)	175,440 (51.96%)	40,307 (45.51%)	694,577 (64.75%)
어휘지도를 이용한 다의어 WSD	418,873 (73.21%)	47,586 (64.05%)	191,799 (56.81%)	43,640 (49.27%)	710,898 (65.44%)

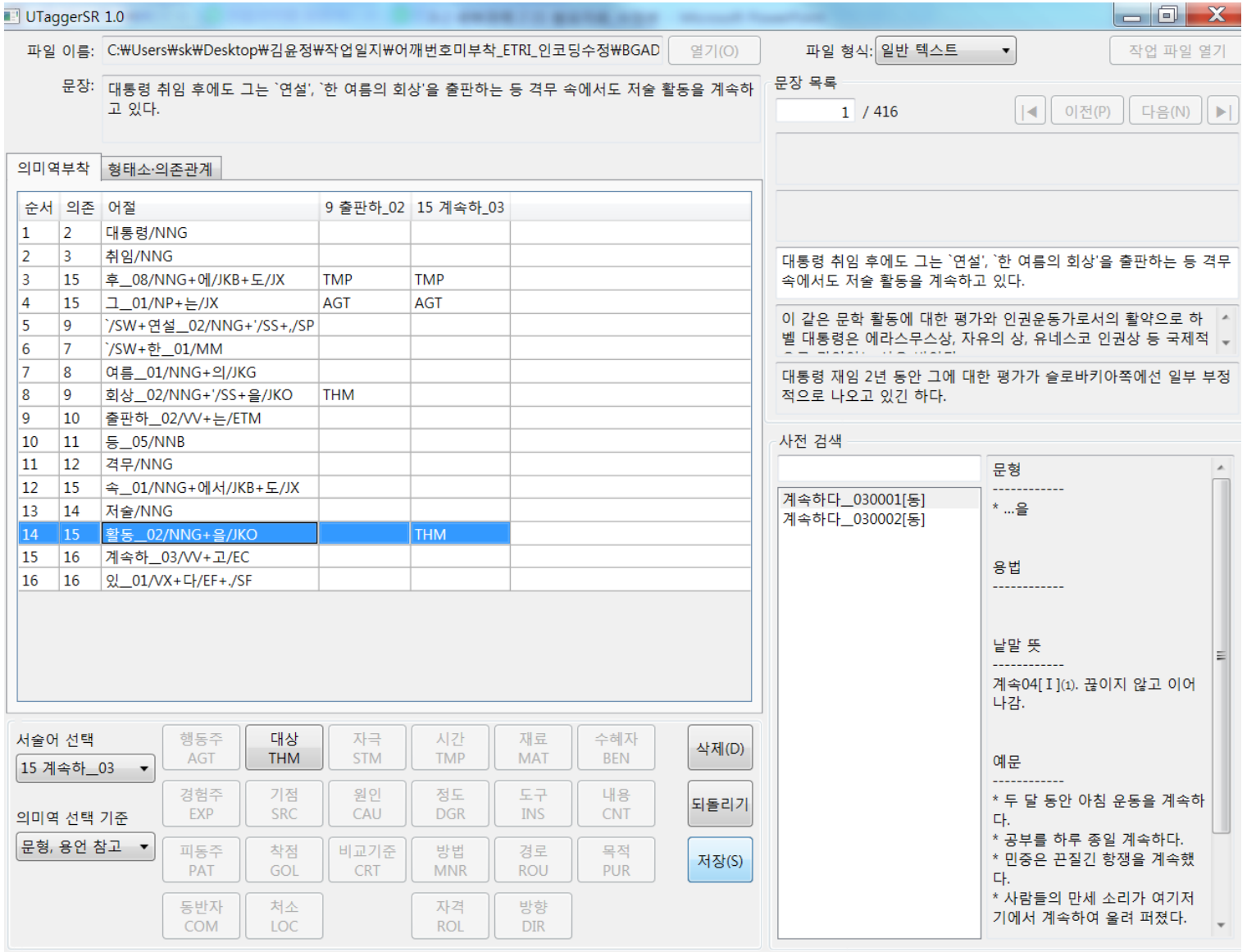
UTagger-SR (UPropBank 기반 의미역 부착 도구) (1)

❖ 의미역 부착 말뭉치 구축(31개 파일, 2014)
- 세종구구조부착말뭉치를 의존구조로 변경
- 수작업으로 의미역 부착
- 개별 의미역 통계
- 용언별로 격조사-의미역 통계

❖ UPropBank
- 다의어 수준, 90,090 용언
- 동형이의어 수준으로 의미역 통합

전체 서술어 개수			240,747
의미역 부착 서술어 개수			181,404
행동주(AGT)	48,150	경험주(EXP)	3,085
피동주(PAT)	5,095	동반자(COM)	3,073
대상(THM)	129,919	기점(SRC)	8,200
착점(GOL)	20,718	처소(LOC)	12,197
자극(STM)	412	원인(CAU)	3,519
비교기준(CRT)	6,404	시간(TMP)	6,946
정도(DGR)	3,583	방법(MNR)	7,652
자격(ROL)	1,741	재료(MAT)	341
도구(INS)	2,975	경로(ROU)	250
방향(DIR)	2,291	수혜자(BEN)	1,289
내용(CNT)	7,910	목적(PUR)	548

UTagger-SR (UPropBank 기반 의미역 부착 도구) (2)



❖ 실험 결과

- 65,529개 문장
- 59,257개 문장 격조사-용언별 의미역 통계
- 6,272개 문장 실험

방법	태깅 정답수	의미역 수	정확률(%)
UPropBank & 후보 1개	5,548	7,106	78.08
후보 2개 이상 이거나 없을 때, 격조사별 의미역 빈도	7,390	10,668	69.27
후보 2개 이상 이거나 없을 때, 격조사-용언별 의미역 빈도	7,891	11,191	70.51
후보 2개 이상 이거나 없을 때, 뒤 서술어에 대해서 격조사-용언별 의미역 빈도	7,882	10,901	72.31

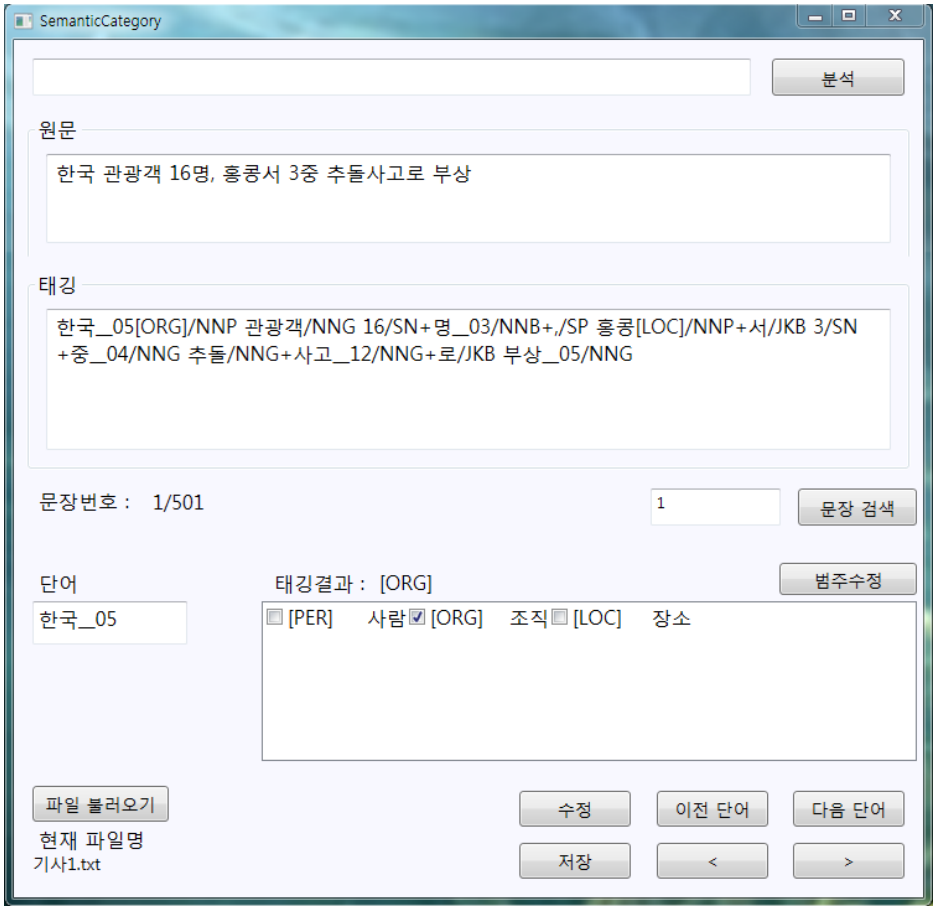
UTagger-NE (개체명 인식기) (1)

❖ 의미범주 : ETRI 대분류 + UWordMap의 상위 노드 참조 (13개 범주)

범주명	의미	범주명	의미
[PER]	사람	[ACC]	정도
[KNW]	지식	[ACT]	활동
[MAT]	물질	[ANM]	동물
[CHR]	성질	[PNT]	식물
[ORG]	조직	[REL]	관계
[SPC]	공간	[ETC]	기타
[TME]	시간		

UTagger-NE (개체명 인식기) (2)

- UTagger-NE (ver. 1.0)
  - 사용자 의미범주 정의 기능
  - 신경회로망 기반 학습
  - 자동 태깅 및 수정



❖ 실험(PLO) 결과

문서	PER	ORG	LOC	소계	정확률(%)
신문1	172/195	135/194	125/150	432/539	80.15
신문2	42/52	134/142	349/372	525/566	92.76
신문3	109/133	146/165	522/591	777/889	87.40
백과사전1	111/122	206/236	221/261	538/619	86.91
백과사전2	134/138	412/417	126/133	672/688	97.67
소설1	225/237	4/5	59/81	288/323	89.16
소설2	33/34	2/2	81/89	116/125	92.80
소설3	65/67	7/8	21/35	93/110	84.55
합계	891/978	1,046/1,169	1,504/1,712	3,441/3,859	
정확률(%)	91.10	89.48	87.85	89.17	

추가 연구(2015)

- 향상된 의존관계파서를 적용
  - 용언과 논항의 관계를 보다 정확히 파악

- 어휘지도 활용
  - 미학습 동형이의어 태깅
  - 의존관계 분석 시 의미제약
  - 의미역 태깅
  - 의미범주

- 다의어 분석 대상 확대
  - (동사, 형용사) + 명사
  - 수의 논항에 의한 WSD 방법

한국어처리시스템 U Series

UTagger-HM (96.6%) ~ 2013

동형이의어 WSD

UTagger-DP (90%) ~ 2015

다의어 WSD/SRL

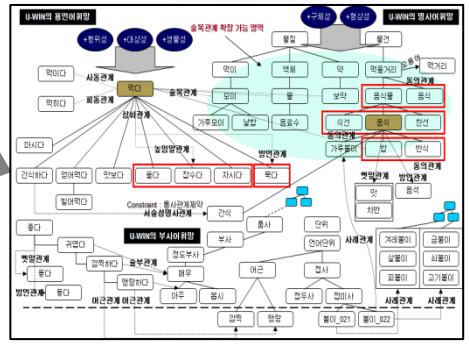
UTagger-SR (85%) ~ 2016

NER

UKnowl (100M) ~2018

UCorpus

- 품사주석 (1억 어절)
- 동형이의어 주석 (6천)
- 의존관계 주석 (1천)
- 개체명/범주 주석
- 의미역 주석 (1천)



UWordMap

현재의 정확률을 2% 개선시키기 위해서는 …

Lexical Semantics
WSD, NER

- 하나의 형태소가 문맥에 따라 여러개의 의미로 해석될 때, 하나의 의미를 결정
- Homograph WSD
- Polysemy WSD
- Named Entity Recognition
- Lexical Semantic Network

Sentence Semantics
Parser, SRL

- 문장 내의 술어-논항(predicate-argument) 관계에 적합한 의미 관계(Semantic Role)를 결정
- Subcategorization
- Semantic Restriction
- Semantic Role Labeling
- Ontology & Inference

- 연구실 홈페이지 : http://nlplab.ulsan.ac.kr DEMO
  - WordsMap Browser
    - http://klplab.ulsan.ac.kr:8080/uwin_nrel.jnlp : UWIN전체 대상
    - http://youtu.be/bE3GtuG6gN8
  - 형태소/동형이의어 태깅시스템(UTagger-HM)
    - http://nlplab.ulsan.ac.kr:8080/KMAClient/KMAClient.jsp (WEB용)
    - 한자자동변환시스템 http://hanjaro.juntong.or.kr/
    - 다의어 WSD http://youtu.be/glV2DNq_xt8
    - 아래한글에 UTagger-HM addon http://youtu.be/Ibhy7okBO98
  - UTagger-SR (격틀사전 기반 의미역 부착도구)
    - https://youtu.be/8AgQnrE71n4
  - UTagger-NE (개체명인식기)
    - https://www.youtube.com/watch?v=BgxoxN_a4n8&feature=youtu.be

Q & A

okcy@ulsan.ac.kr

052-259-2222 (010-2561-5830)

울산대학교 한국어처리연구실/지능형컴퓨터연구실

http://nlplab.ulsan.ac.kr