# 하위범주화 사전의 구축 및 자동 확장

이수선 박현재 우요섭
인천대학교 정보통신공학과
{g9921094, g9921091, yswooo}@lion.inchon.ac.kr

# Development and Automatic Extraction of Subcategorization Dictionary

Su-Seon Lee Hyun-Jae Park Yo-Seop Woo
Dept. of Information and Telecommunication, University of Inchon

요 약

한국어의 통사적, 의미적 중의성 해결을 위해 하위범주화 사전을 구축하였다. 용언에 따라 제한될 수 있는 문형 패턴과 의미역 (semantic roles) 정보의 표준을 정하여 이를 부가하였고 구축한 하위범주화 사전이 명사에 대한 의미를 갖고 있는 계층 시소러스 의미 사전과 연동하도록 용언과 명사와의 의미적 연어 관계에 따라 의미마커를 부여했다. 논문에서 구현된 하위범주화 사전이 구문과 어휘의 중의성을 어느 정도 해소하는지 확인하기 위해 반자동적으로 의미 태깅(Sense Tagging)된 말뭉치와 구문분석된 말뭉치를 통해 검증 작업을 수행했다. 이 과정에서 자동으로 하위범주 패턴에 대한 빈도 정보나, 연어 정보, 각 의미역과 용언의 통계적 공기 정보 등을 추출하여 하위범주화사전에 추가시켰다. 또한 여기서 얻은 정보를 기준으로 하위범주화 사전을 자동으로 확장하는 알고리즘을 적용하여 확장시켰다.

## 1. 서론

한국어 문장이 나타내는 상황 속에는 주된 의미를 갖는 중심어가 있고 이 중심어에 나머지 다른 어휘들이 의미적으로 결속되어 있다. 한국어에서 중심어 역할을 하는 어휘가 바로 용언이고 나머지 어휘들은 이 용언에 의해 역할을 제한받게 되는데, 이 의존 관계를 용언의 하위범주라 한다. 한국어 처리에 있어서 통사적, 의미적 중의성 해결을 위해 하위범주화 사전의 필요성은 여러 연구에서 지적되어 왔고 통사적 수준의 하위범주화 사전이 발표되고 있다.

그러나 지금까지 구축된 사전들은 구문적 특성을 반영한 것이 대부분이며, 구문의 중의성을 파악하는 보조적 자료로서만 이용되어 왔다. 이는 물론 구축의 어려움도 있지만, 함께 연동되어야 할 적절한 의미 사전의 부재에도 그 원인이 있다. 또한 하위범주화 사전에 의미 마커(sematic marker)를 부여할 때, 명사의 의미 사전인 시소러스 (thesaurus)에 기술된 어휘간의 다양한 의미 관계를 고려해서 작성해야 한다는 점이 어려운 문제가 되고 있다. 본 연구에서는 용언에 따라 제한될 수 있는 하위범주 패턴을 정의하고 패턴에 따라 하위범주 사전을 구축하였다. 또한 명사의 시소러스[1]와 정합하여 보어를 선택 제한(Selectional Restriction)할 수 있도록 용언과 명사와의 의미적 연어관계에 따라 의미마커를 부여했다.

이렇게 구축한 하위범주화 사전을 단문분할된 말뭉치와 구문분석된 말뭉치를 통해 매칭 작업을 함으로써 하위범주 패턴에 대한 빈도 정보, 각 의미역과 용언의 통계적 공기 정보, 용언과 명사와의 연어 정보를 수집하여 데이터베이스에 포함시키고자 한다. 또한 여기서 얻은 빈도 정보를 토대로 하위범주화 사전을 자동으로 확장하는 방법을 제안하고자 한다.

## 2. 하위범주화 사전의 설계와 구축

하위범주화 사전과 연동할 수 있는 명사 시소러스 사전을 고려해 용언과 명사와의 의미적 연어 관계에 따른 구축을 시도하였다. 그림 1은 구축된 하위범주화 사전의 예이다. 용언과 연계되어 구문적, 의미적 관계를 형성하게 될 조사는 필수격 보어에 해당하는 것으로 한정하여 대표 조사로 삼고, 대표조사의 대용으로 사용되는 보조사나 복합 조사를 일부 수록하였다.

| ID | 용언 | 품사 | 패턴ID | 피동정보 | 사역정보 | 원형정보 |
|---|---|---|---|---|---|---|
| 1030 | 몰리다 | 동사 | V6 | 1 | 0 | 몰다 |
| 1031 | 몰리다 | 동사 | V9 | 1 | 0 | 몰다 |

| 대표조사 | | 확장조사 | | 의미역 | | 의미마커 | | 예제 |
|---|---|---|---|---|---|---|---|---|
| 이 | 에 | 이/가/은.. | 에/까지 ... | CHD | GOL | R00G00B00U01a,... | R00H,.. | |
| 이 | 에(게)서 | 이/가/은... | 로부터/... | AGT | RCP | R00F00B00D00A | R00G00B00U01a | |

그림 1 하위범주화 사전의 예

한국어는 조사가 어휘의 구문적 역할과 의미적 역할을 제한하기 때문에 조사에 따른 의미역의 정의가 필요하다. 기존의 하위범주화 사전은 조사를 표층격으로 하여 해당 보어성분에 의미 마커를 부여하였지만, 의미역을 고려하지 않았기에 각 어휘에 대한 의미적 역할의 파악이 어려운 문제가 있었다. 본 연구에서는 하위범주를 이루는 요소의 개념적 역할을 정의하는 의미역을 조사에 따라 32개를 정의하였다. 의미역의 설정에 있어서는 기존의 의미역에 관한 연구를 참고하여 정의하였다. 의미역을 세밀하고 심층적으로 지정하는 것은 하위범주화 패턴의 수를 크게 증가시키고 작업의 일관성에 영향을 줄 수 있으며, 표층 조사와 연관시키는 것을 불가능하게 만들 수 있다. 따라서 표층 조사와 연계될 수 있는 범위에서 다소 심층적인 의미역을 지향하여 정의하였다. 그림2는 하위범주화 사전을 위한 의미역의 예이다.

| | | |
|---|---|---|
| 조사 이/가 | AGT | AGenT [행위자 주어] |
| | 예문) 영희가 책을 옮겼다. | |
| | CHD | CHaracterizeD [비행위자 주어] |
| | 예문) 그녀는 어리다. | |
| | EXS | EXistent [존재하는 대상 주어] |
| | 귀신은 존재한다. | |
| | TRR | TRansforming Result [변화의 결과] |
| | 철수가 선생님이 되었다. | |
| | TRS | TRansforming Source [변화의 출발] |
| | 준이가 선생님이 되다. | |
| | FCS | FoCuS [형용사의 화제] |
| | 나는 그 소설이 슬프다. | |

그림 2 하위범주화 사전을 위한 의미역의 예

대량의 하위범주화 사전 구축에 있어 가장 큰 문제는 복수의 작업자간에 일관성 있는 하위범주 부가가 어렵다는 점이다. 따라서 작업의 효율성을 위해 문형을 반영하는표준적인 하위범주 패턴이 정의될 필요가 있다. 본 연구에서는 그림3과 같이 그림2에서 보여준 의미역들로 구성된 동사 41개, 형용사 17개, 용언화 접사 4개의 표준 패턴을 정의하였다.

| [1] 동사의 하위범주 패턴 예 |
|---|
| V1. [이 AGT]<br>예문) 아기가 걸었다.<br>V2. [이 AGT] [에서 LOC]<br>예문) 아이들이 놀이터에서 논다. |
| [2] 형용사의 하위범주 패턴의 예 |
| A1. [이 CHD] [에게 RCP]<br>예문) 그 바지가 나에게 크다.<br>A2.[이 CHD] [에서/(으)(로)부터 SRC]<br>예문) 학교가 집에서 멀다.<br>A3.[이 CHD] [에 MGL]<br>예문) 그 색은 내 눈에 거칠다. |

그림 3 하위범주화 사전을 위한 표준패턴의 예

하위범주화 사전의 구축은 기존의 하위범주화 관련 자료 및 작업자의 경험, 국어 사전등의 예문을 하위범주화 표준 패턴에 정합시키는 방법으로 이루어졌다.

하위범주화 사전은 구문분석이나 의미분석시 애매성 해소에 사용된다. 그러기 위해서는 하위범주화 사전의 의미정보와 시소러스의 의미정보와의 정합이 필수적이다. 따라서 하위범주화 사전과 명사 시소러스를 이용한 선택 제약 알고리즘을 위해서는 우선 명사 시소러스의 구조를 하위범주화 사전에 적합하도록 바꿀 필요가 있다. 그림 4는 그러한 시소러스의 예이다. 하위범주화 사전의 의미 마커는 패턴이 이루는 상황하에서 다양한 어휘를 수용하기 위해 가급적 시소러스 레벨에서도 중간 이상의 상위에 속하는 개념을 부여하였다.

| ID | 표제어 | 의미코드 | 상위어 ID |
|---|---|---|---|
| : | : | : | : |
| 16409 | 극 | *0l | 85684 |
| 16410 | 극 | R00H10l00I01S | 1000401 |
| 16411 | 극 | R00I00H05B00V | 69454 |
| 16412 | 극 | R00G00B00T00m00C01k | 108284 |
| : | : | : | : |
| 85613 | 예불 | *3b14q00K00A00Z00A | 6737 |
| : | : | : | : |

그림 4 의미코드가 부여된 시소러스 예

시소러스 계층은 트리 구조가 아니라 상위노드가 복수개 대응될 수 있는 그래프 형태가 일반적이다. 따라서 복수개의 상위 개념을 갖는 노드를 그림 5와 같이 별도의 데이터베이스로 관리하고 의미코드도 '*0l','*0m'과 같

이 '*'로 시작하게 함으로써 다른 명사와 구분지어 정합 때 참조할 수 있도록 하였다.

| ID | 표제어 | 의미코드 | 상의어 의미코드 |
|---|---|---|---|
| : | : | : | : |
| 16409 | 극 | *0l | /R00J00H00D00S/*3b14w00i00B/ |
| 16558 | 극장 | *0m | /R00H10H/*0L00K/ |
| 17347 | 금액 | *0n | /R00G00A00B07g00A00I/R00I00H05B00U00y00X/ |
| : | : | : | : |

그림5 중복 상의어를 가진 명사 테이블 예

### 3. 하위범주화 사전의 자동확장

이렇게 구축한 하위범주화 사전과 단문분할된 말뭉치와 구문분석된 말뭉치와의 매칭시키는 과정에서 그림 6과 같은 유용한 정보들을 구해 이를 사전에 추가시켰다. 하위범주 패턴에 대한 빈도 정보, 각 의미역과 용언의 통계적 공기 정보, 용언과 명사와의 연어 정보등을 하위범주 데이터베이스에 포함시켰다. 또한 여기서 얻은 정보를 토대로 하위범주화 사전을 자동으로 확장하는 알고리즘에 적용시켜 반자동적으로 구축한 하위범주화 사전과는 별도의 하위범주화 사전을 구축했다. 자동으로 구축한 하위범주화 사전은 정확률 면에서 반자동으로 구축한 사전에 미치지 못하지만 한국어 분석에 이용시 반자동적으로 구축한 하위범주화 사전을 이용하고 여기서 해결하지 못한 애매성을 자동으로 구축한 하위범주화 사전을 이용해 해결함으로써 분석의 성능을 향상시키는데 도움을 줄 수 있다.

| 연어정보 | | | | 의미역 빈도정보 | | | | 의미마커 빈도정보 | | | | 패턴일치 빈도정보 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1+ | 2+ | 3+ | 4+ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | |

그림 6 확장된 하위범주화 사전 구조

### 4. 실험 및 평가

단문분할된 말뭉치[5] 42,323문장과 구문분석된 말뭉치에서 추출해낸 문장 23,141문장을 가지고 하위범주화 사전과의 매칭을 수행했다.

용언에 의존하는 명사별로 하위범주화 사전의 의미와 매칭된 수를 찾아보면 전체 대상명사는 193,692개이고 이중 하위범주화사전의 의미와 정합된 수는 145,082개이다.

의미 매칭이 성공한 명사를 DB에 포함시킴으로써 하위범주화 사전에 용언과 명사와의 연어 정보를 추가시켰다. 또한 각 용언마다 의미역들이 몇 번 나왔는지 정보를 의미역 빈도 정보에 추가시키고 패턴이 일치하는 경우 패턴 일치 빈도 정보의 값을 증가시켰다. 또한 실험 문장의 명사들이 하위범주화 사전의 명사들의 의미 마커 중 어느 것과 매칭됐는지 그 빈도정보를 포함시켰다.

매칭이 실패한 경우의 정보는 따로 수집하고 하위범주 자동 확장 알고리즘을 통해 새로운 하위범주사전을 구축했다. 이 과정에서 구축한 하위범주사전의 개수는 234개이다.

### 5. 결론

본 연구에서는 한국어 용언 중심의 하위범주화 사전을 설계하고 구축하였다. 구축된 하위범주화 사전을 단문분할된 말뭉치와 구문분석된 말뭉치와 매칭시켜 명사의 연어 정보와 하위범주 패턴의 공기정보 등을 하위범주화 사전 DB에 포함시킴으로써 하위범주화 사전의 성능을 향상시키고 매칭에 실패한 문장을 확장 알고리즘에 적용하여 하위범주 사전을 자동으로 확장할 수 있음을 보였다. 자동확장된 사전의 수가 적은 것은 실험에 이용한 단문분할된 말뭉치와 구문분석된 말뭉치의 규모가 작기 때문이다.

이렇게 구축한 하위범주화 사전은 의미 사전과의 연동으로 구문 분석 후보의 수를 줄이며, 동시에 어휘의 의미를 결정할 수 있기 때문에 대량의 말뭉치에 적용시켜, 술어와 보어간 의존 관계와 의미 정보가 태깅된 말뭉치 구축에 활용할 수도 있을 것이라 생각된다.

이 논문은 정보통신부 대학 기초 연구 지원 사업의 연구비 지원에 의해 수행되었음

참 고 문 헌

[1] 우요섭, "토큰 기반 한국어 분석기 개발 - 한국어 의미 분석 사전 및 하위범주화사전 구축", 한국전자통신연구원,1997

[2] 홍재성 외, "현대 한국어 동사 구문 사전", 두산동아, 1997

[3] 추교남, "개념 기반 정보 검색을 위한 한국어 어휘의 의미 분석", 인천대학교 석사학위논문, 12.1998

[4] 조정미, "한국어 의미 해석시 중의성 해소에 관한 연구", 정보과학회지, 1997.6

[5] 박현재, "의미 개념을 이용한 이단계 단문 분할 알고리즘", 한글 및 한국어 정보처리 학술대회, 1999.10

[6] 옥철영, "우리말 개념망 명사 데이터 구축", 한국전자통신연구원 보고서, 1997