한국어 분석의 중의성 해소를 위한 하위범주화 사전 구축

이수선*, 박현재*, 우요섭**
인천대학교 정보통신공학과
g9921094@lion.inchon.ac.kr

Development of Subcategorization Dictionary for the Disambiguation Korean Language Analysis

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

요약

자연언어 처리에 있어 문장의 성분 구조를 파악하는 통사적 해석에서는 애매성 있는 결과가 많이 생성된다. 한국어의 경우 어순 등의 통사적 특성뿐 아니라 상황과 의미, 문맥이 문장의 분석에 더 중요한 역할을 하기 때문에 문맥 자유 문법에 의한 접근 방법만으로는 중의적 구조의 해결이 어렵다. 이는 또한 의미 분석시 애매성을 증가시키는 원인이 된다. 이러한 통사적, 의미적 중의성 해결을 위해 용언 중심의 하위범주화 사전을 구축하였다.

본 논문에서는 용언에 따라 제한될 수 있는 하위범주 패턴을 정의하고 패턴에 따라 하위범주 사전을 구축하였다. 하위범주화 사전에는 명사의 시소러스와 정합하여 보어를 선택 제한(Selectional Restriction)할 수 있도록 용언과 명사와의 의미적 연어 관계에 따라 의미마커를 부여했다. 말뭉치를 통해 수집된 용언 12,000여개를 대상으로 25,000여개의 하위범주 패턴을 구축하였고 이렇게 구축한 하위범주화 사전이 120,000여 명사에 대한 의미를 갖고 있는 계층 시소러스 의미 사전과 연동하도록 하였다.

또한 논문에서 구현된 하위범주화 사전이 구문과 어휘의 중의성을 어느 정도 해소하는지 확인하기 위해 반자동적으로 의미 태깅(Sense Tagging)된 2만여 문장의 말뭉치를 통해 검증 작업을 수행하고, 의존관계와 어휘의 의미를 포함하고 있는 말뭉치에 하위범주 패턴이 어느 정도 정합되는지를 분석하여, 하위범주 패턴과 말뭉치의 의존관계만 일치하는 경우와 어휘의 의미까지 일치하는 경우에 대해 평가한다. 이 과정에서 하위범주 패턴에 대한 빈도 정보나, 연어 정보를 수집하여 데이터베이스에 포함시키고, 각 의미역과 용언의 통계적 공기 정보 등을 추출하는 방법도 제시하고자 한다.

1. 서 론

한국어는 상황 중심어이다. 따라서 한 문장은 상황을 전달하는 역할을 한다. 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.

하위범주화 사전의 필요성은 이미 여러 연구에서 지적되어 왔지만 구축에 상당한 노력이 필요하고 대상 용언의 개수가 증가할수록 일관성 등 여러 가지 문제가 발생하기 때문에 제한된 영역에서 실험적으로 구축, 사용되어 온 것이 일반적이다. 따라서 기존 연구의 경우 표층의 격조사만을 대상으로 하여 문장 구성 성분의 의미나 역할을 파악할 수 없는 경우가 많았고, 명사의 계층 의

미 시소러스와 하위범주화 사전을 연계하는 경우보다는 용언에 지배되는 명사의 예[2]를 보이는 정도가 많았다. 본 논문에서는 이러한 하위범주화 사전의 단점을 보안하여 하위범주화 사전을 구축하는 방법론을 제시하고, 한국어 용언 12,000여개에 대해 문형 패턴과 의미역 정보[5]를 기준으로 하위범주화 사전을 구축했다. 또한 이렇게 개발한 하위범주화 사전을 반자동으로 의미 태깅된 2만여 문장의 말뭉치를 통해 검증 작업을 함으로써 하위범주화 사전이 한국어 분석 과정에서 중의성 해소에 얼마나 유용한지를 보인다.

2. 하위범주화 사전의 설계와 구축

본 연구에서는 하위범주화 사전과 연동할 수 있는 명사 시소러스 사전을 고려해 용언과 명사와의 의미적 연어 관계에 따른 구축을 시도하였다. 12,000여개의 용언을 대상으로 했으며 용언에 따라 제한될 수 있는 패턴을 정의하여 반자동적인 방법으로 구축하고 하위범주화 사전의 의미마커 부여는 기구축한 계층적 의미 시소러스[1]를 이용하였다.

하위범주를 이루는 보어의 의미적 역할을 정의하는 의미역을 조사 이/가에 대해 8개, 을/를에 대해 7개, 조사 와/과에 대해 1개, 조사 에게/에서/부터/로/으로/으로부터에 대해 16개 총 32개를 정의했다. 그림 1은 하위범주화 사전을 위한 의미역의 예이다.

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

그림 1. 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 반영하는 표준적인 하위범주 패턴이 정의될 필요가 있다. 본 논문에서는 그림 2와 같이 그림 1에서 보여준 의미역들로 구성된 동사 41개, 형용사 17개, 용언화 접사 4개의 표준 패턴을 정의하였다.

[1] 동사의 하위범주 패턴 예

V1. [이 AGT]
예문) 아기가 천천히 걸었다.
V2. [이 AGT] [에서 LOC]
예문) 아이들이 공원에서 논다.

[2] 형용사의 하위범주 패턴의 예

A1. [이 CHD] [에게 RCP]
예문) 그 바지가 나에게 크다.
A2. [이 CHD] [에서/(으)(로)부터 SRC]
예문) 학교가 너희집에서 멀다.
A3. [이 CHD] [에 MGL]
예문) 그 색은 내 눈에 거칠다.

그림 2. 하위범주화 사전 표준패턴의 예

기존의 한국어 하위범주화 관련 연구 결과와 사전, 말뭉치를 활용하여, 12,000 용언에 대해 하위범주 표준 패턴을 부가하고, 패턴의 각 의미역에 해당하는 의미 마커를 시소러스로부터 추출하여 기술하였다. 이렇게 구축한 하위범주화 사전은 말뭉치에 적용하는 검증 작업을 함으로써 의미역이나 표준 패턴에 대해 추가적인 정보를 찾고 이러한 정보들을 다시 하위범주화 사전에 반영하는 작업을 반복적으로 수행하였다.

또한 이 과정에서 하위범주 패턴에 대한 빈도 정보나, 연어 정보를 수집하여 데이터베이스에 포함시키고, 각 의미역과 용언의 통계적 공기 정보 등도 하위범주화 사전에 추가하였다.

하위범주화사전의 구조는 그림 3과 같다. 대상 용언 12,000여개에 대해 구축한 하위범주화 사전의 개수는 25,000여개이다.

그림 3. 확장된 하위범주화사전의 구조

위의 그림 3에서 연어정보, 의미역 빈도 정보, 의미마커 빈도 정보, 패턴 일치 빈도정보는 하위범주화 사전을 말뭉치에 적용하는 검증 과정에서 자동으로 얻어냈다. 예를 들어 "누나가 동생에게 장난감을 주었다"라는 문장이 있을 경우 이 문장이 하위범주화 패턴

[주다] AGT(사람) ACC(물건) RCP(사람, 동물)

과 매칭에 성공한다면 연어정보 필드의 1+에는 '누나'가 2+에는 '선물'이 3+에는 '동생'이 들어간다. 이러한 연어정보는 각 의미역마다 여러개가 올 수 있으므로 '1+', 2+'...로 표시하였다. 한국어의 경우 주어나 다른 성분이 생략되는 경우가 많은데 그러한 경우가 각 용언마다 얼마나 되는지를 보기 위해 의미역 빈도 정보를 두었다. 앞 예문은 AGT, ACC, RCP가 생략되지 않고 모두 있으므로 '의미역 빈도 정보'의 1, 2, 3 필드를 각각 1씩 증가시키면 된다.

위 문장에서는 AGT라는 의미역에 해당하는 의미 마커가 사람밖에 없지만 사람이 아닌 다른 여러 의미들이 들어 있을 경우 문장의 AGT에 해당하는 명사가 하위범주화 사전 속의 의미마커 중 어느 것과 매칭되었는가를 나타낼 필요가 있다. 따라서 각 의미마커마다 그 의미마커의 빈도정보 필드에 그러한 정보를 추가시켰다. 의미역 첫 번째에 해당하는 의미마커가 4개가 있을 경우 이 중 첫 번째 의미마커가 1번, 두 번째 의미마커가 2번, 세 번째 의미마커가 1번, 네 번째 의미마커가 1번 나왔다면 의미마커 빈도정보 '1+'자리에 '1.2.1.1'로 표기하면 된다. 패턴일치 빈도정보는 하위범주 패턴이 말뭉치의 문장과 몇번이나 정합되었는지를 나타낸다. 한번 매칭에 성공할 때마다 1씩 증가시켰다.

3. 하위범주화 사전의 보완

하위범주화 사전이 구문과 어휘의 중의성을 어느 정도 해결하는지 보기 위해 반자동적으로 의미 태깅된 2만여 문장의 말뭉치를 통해 검증 작업을 수행하였다.

검증 작업에 쓰인 의미 태깅된 말뭉치는 먼저 자동 의미 정합의 성능 향상을 위해 명사 시소러스의 의미 코드를 다시 정의하고 이 시소러스와 하위범주화 사전을 이용 자동으로 명사의 의미와 의존 관계를 얻어내는 의미 태깅 모듈을 설계한 다음 알고리즘 상의 오류로 잘못된 태깅이 발생한 문장에 대해서는 다시 수작업을 통해 정정하는 과정을 거쳐 구축했다.

3.1 명사 시소러스에 의미 코드 정의 및 부여

하위범주화 사전과 명사 시소러스를 이용한 선택 제약 알고리즘을 위해서는 우선 명사 시소러스의 구조를 하위범주화 사전에 적합하도록 바꿀 필요가 있다. 하위범주화 사전의 의미 마커는 패턴이 이루는 상황하에서 다양한 어휘를 수용하기 위해 가급적 시소러스 레벨에서도 중간 이상의 상위에 속하는 개념을 부여하였다.

말뭉치의 어휘는 시소러스 내에서 하위 레벨에 속하는 세부적 개념을 가지고 있기 때문에 말뭉치의 어휘문에 대해 하위범주화 사전과 명사 시소러스와의 정합시 말뭉치내의 어휘의 의미로부터 하위범주화 사전의 의미에 속하는지에 관해 의미 추적 작업이 이루어져야 하고 이는 상당한 계산량을 요구하게 된다.

이런 계산상의 부담을 줄이면서 빠른 상하위 추적을 위해 시소러스의 각 의미 마커에 대해 레벨에 따른 문자열 의미 코드를 부여하였다. 이러한 규칙적인 의미 코드 부여는 의미 마커 정합시 의미 코드의 문자열 비교를 통해 각 의미간 상하위 관계의 파악이 용이하도록 한다.

선택제약으로 하위범주화사전과 시소러스를 사용하는 것은 문장에 나타난 명사의 시소러스 개념과 하위범주화 사전에 부여된 의미 마커가 정합되는가를 판단하는 것이다. 여기서 정합은 상하위 관계로 정합되는 것이 일반적이다. 그러나 경우에 따라서는 상하위 관계가 아니라 시소러스의 개념 계층에서 근접한 정도, 즉 개념간의 거리를 이용하여야 하는 경우도 있다. 따라서 이러한 개념 정보들은 상하위 관계나 개념간 거리를 판단하기 쉬운

형태로 기술되어야 하며 본 논문에서는 접두어식 개념 계층 표기 방법으로 정의하였다.

또한 시소러스 계층은 트리 구조가 아니라 상위노드가 복수개 대응될 수 있는 그래프 형태가 일반적이다. 따라서 복수개의 상위 개념을 갖는 노드를 별도의 테이블로 관리하고 상하위 정합 때 테이블을 참조할 수 있도록 하였다.

그 테이블의 예는 아래 그림 4과 같다. 예를 들어 '감동'이라는 어휘는 'A00C00H01d'라는 코드를 가진 명사와 '*i'라는 코드를 가진 명사, '*i01U'라는 코드를 가진 명사 이렇게 3개의 명사를 상의어로 가지고 있다. 따라서 아이디 2319인 '감동'이라는 명사의 상의어를 찾고자 할 때 이 테이블을 참고해 3개의 상위 명사에 대해 추적해 보면 된다.

Index	복수 상위 개념	ID	어휘	Mark
251	/A00C00H01d/*i/*i01U/	2319	감동	*D
378	/A00A00A00B/A00A00A00B00L/	3921	개체	*E
436	/A00C00I00Q/A00C00I00W/	4655	거짓	*F
438	/+X/*H00D/	4667	거처	*G

그림 4. 복수 상위 개념 대응 테이블

이렇게 부여된 코드는 의미간 접두어 정합(prefix matching)을 통해 그 어휘의 속성이 하위범주화 사전의 상위어의 속성을 갖고 있는지 파악할 수 있게 된다.

말뭉치내 어휘가 R00b00A00D00H00D라는 의미코드를 가지고 있고 하위범주화사전의 의미 소성이 R00b00A일 경우 두 어휘가 서로 상하위 관계에 있다고 볼 수 있으므로 개념 거리는 3으로 판정한다.

3.2 어휘의 의미 태깅 모듈 설계

자동으로 의미 태깅된 결과를 얻어낼 말뭉치는 마침표 등의 문장 구분 표시('.', '?', etc)를 기준으로 먼저 문장 단위로 분리된 형태소 태깅된 문장을 대상으로 한다. 주요 태그가 명사(NN, NU, NX, NP), 조사(JO), 동사(VV), 형용사(VJ) 에 대해서 고려한다. 하위범주화사전이 술어와 필수격 조사에 따른 보어 성분간의 의존 관계로 이루어졌기 때문이다.

의미 태깅 시스템의 알고리즘은 기본적으로 하위범주화 사전과 시소러스간의 개념 정합에 의한 선택 제약에 의한 방법이다. 여기에는 문장에 대해 술어를 중심으로 한 의존 관계를 파악하여 절로 구분하는 과정이 필수적이기 때문에, 그림 5과 같은 간단한 의존 구조 해석기를 설계하였다.

이 알고리즘은 한국어에 대한 문법 규칙에 기반하기보다는 경험적인 지식을 활용한 것이다. 먼저 문장의 형태소 태깅 결과로부터 술어와 명사 성분들을 추출하고, 이를 술어를 중심으로 재배열하는 방법이다. 재배열은 우선 해당 술어의 하위범주 패턴들을 하위범주화 사전에서 추출하고, 표층 조사를 기준으로 술어와 명사간의 가능한 의존 관계를 찾아내는 것이다. 표층 조사는 대표 조사만이 아니라 확장 조사 리스트를 활용한다. 한 패턴에는 2~4개 정도의 하위범주 성분이 있고, 각 하위범주 성분에 정합 가능한 명사들은 복수개가 있을 수 있으며, 이들 중 한 명사의 개념을 시소러스에서 탐색하게 되면 복수개의 개념에 대응하는 경우도 발생한다. 시소러스에서 탐색된 개념을 패턴의 대응하는 하위범주의 의미 마커와 상하위 관계로 정합시키게 된다.

문장내의 모든 술어에 대해 가능한 의존 관계들을 찾아내게 되면, 이들 중에서 가장 적절한 의존 관계들만을 선택하는 경험적인 여과 (filtering) 과정이 수행된다. 즉 명사가 여러 개의 술어에 걸리는 문제가 발생하거나 하위범주 성분을 갖지 못하는 술어가 가급적 없도록 문장 전체의 의존 구조를 결정하는 것이다.

이렇게 얻어진 문장 전체에 대한 술어들과 하위범주 성분들 간의 의존 구조가 최종적으로도 복수개가 있다면 그 중 하나만을 작업자에게 제시하여야 하므로 순위를 결정하는 과정이 필요하다. 순위 결정은 경험적으로 각 술어가 갖는 하위범주 성분의 개수가 비교적 편차가 없이 균일한 것을 우선적으로 하였으며, 하위범주 패턴의 의미 마커와 시소러스 사전의 의미 마커간의 평균 개념 거리가 가까운 것을 그 다음으로 하였다. 평균 개념 거리는 시소러스 계층 구조에서 상호간의 깊이 차이를 말하는 것으로 접두어식 코드이므로 쉽게 계산된다.

만일 보어 성분이 될 만한 표층 조사를 가진 명사구가 어떤 술어에도 하위범주로 할당되지 못하는 경우가 생기면, 표층 조사에 부합하는 하위범주나 술어의 미정합된 하위범주를 대상으로 다시 정합 여부를 검사한다. 이때의 정합은 상하위 정합이 이미 실패한 상황이므로, 개념 거리를 기준으로 한 정합이다.

이러한 방법을 통해 말뭉치에 수록된 보어 성분의 명사에 적절한 개념 후보를 제시해 준다. 그러나 이 후보가 항상 정확한 것은 아니므로 시소러스에 수록된 그 명

사의 동음이의 관계의 개념들을 그 다음에 나열하고, 또한 하위범주 패턴의 해당 보어 슬롯에 기술된 의미 마커들을 이후에 추가하여 사용자에 제시하게 된다. 시소러스에 포함된 개념이 하위범주의 의미 마커보다 개념 계층의 하위 노드인 경우가 일반적이므로, 이러한 순서는 정합을 통한 개념을 1순위, 보다 구체적인 개념이 2순위, 광범위한 상위 개념이 3순위의 형태를 갖고 제시하게 되는 것이다.

```
Step 1. Initializing Process
        Input(sentence[max_word]);
        posTaggedSentence[] ← MorphologicalAnalysis(sentence[]);
        predIndex[max_verb] ← PredicateExtraction(posTaggedSentence[]);
        nounIndex[max_noun] ← ComplementCandidateExtraction(posTaggedSentence[]);
    Step 2. Matching Process
        while (each i in predIndex[i])
           subcat[i][max_subcat] ← FetchSubcatDict(predIndex[i]);
           while (each j in subcat[i][j])
                 while (each slot)
                    while (each semanticMarker in slot)
                       leftLimit ← NounRangeForMatching(i)
                       for (k = leftLimit; k < i; k++)
                         matchedQueue[i][k] ← Matching(semanticMarker,
                                                 ThesaurusExtraction(sentence[k]);
    Setp 3. Filtering Process
        matchedQueue[][]
           ← MultipleDependentNounProcess(posTaggedSentence[], matchedQueue[][]);
        while (nonMachedNoun in nounIndex[])
           if (have complemantable postfix)
              MatchingConceptualDistanceProcess(nonMachedNoun,
                                          posTaggedSentence[], matchedQueue[][]);
        while (matchedQueue[][]) {
           FindAveageMachedSlots();
           FindAverageMatchedConceptualDistance();
           OutputFiles ← (SelectFinalCandidate(), posTaggedSentence[],
                                  subcat[][], ThesaurusDB)
        }
    Step 4. Display Process
```

그림 5. 말뭉치와의 선택 제약을 위한 알고리즘 개요

3.3 의존관계와 어휘의 의미를 포함하는 말뭉치 구축

하위범주화 사전과 어휘의 계층적 의미 관계를 나타낸 명사 시소러스를 데이터로 하고 앞에서 제시한 선택 제약 알고리즘을 통해 자동적으로 어휘 의미 태깅된 말뭉치를 구축했다.

자동으로 말뭉치를 태깅한 경우 "철수는 점심을 먹고 산책을 갔다"와 같은 문장에서는 "먹다"와 "가다"의 주어가 철수인 것을 올바르게 찾을 수 있지만 "철수는 수레를 끌고 있는 당나귀를 보았다"와 같은 문장에서는 "끌다"와 "보다"의 주어가 모두 "철수"인 것으로 판단하게 되는 오류가 생길 수 있다. 시소러스 개념과 하위범주 의미 마커 정합에 의해 중복 의존의 형태로 선택해 주는 것을 우선하도록 하였으나 자동적인 개념 선택 과정에서 오류의 발생이 빈번하므로 수작업으로 확인하는 과정을 거쳤다. 의미 정보를 수작업에 의해 부가하는 것은 작업자 개인의 경험에 따라 많은 편차가 있을 것으로 생각되며, 따라서 일관성을 유지하기 위해 적절한 후보들을 제시해 주는 도구 시스템의 개발이 필수적이다. 본 논문에서는 말뭉치의 의미 태깅을 위한 태거 시스템을 설계, 구현하였다.

그림 6은 의미 태거의 작업 화면 일례이다. 그림에서와 같이 말뭉치를 문장 단위로 화면에 제시하도록 하였으며, 그 문장에 대해 자동으로 의미 태깅된 결과가 화면에 보여진다. 수작업자는 태깅된 결과 중 잘못된 부분을 수정하면 된다. 작업의 편의를 위해 형태소 태깅된 결과를 에디트 상자에 출력해 주었다. 동사를 클릭하면 하위범주화 사전을 통해 가능한 하위범주 패턴들이 하단에 출력되고, 이에 따라 관련된 보어 성분의 후보들의 열이 색깔로 구분되어 출력된다. 이들 중 실제 보어에 해당하는 부분들을 클릭하여 개념 정보등을 부여하면 앞서 클릭된 술어와의 의존 관계가 '의존관계' 필드에 출력된다. 술어가 피동형 등이어서 하위범주 패턴에 변형 규칙이 적용된다면 그 정보가 '문형' 필드에 나타난다. '의미역' 필드는 하위범주 패턴이 최종 결정될 때, 하위범주 사전을 통해 부여되는 격 정보이다.

수정 작업이 끝나면, 자동 태깅된 결과가 올바를 경우 그대로 그 정보가 남아있고 잘못되었을 경우 올바른 보어-술어 의존 구조가 파악되고, 각 보어 성분의 의미적

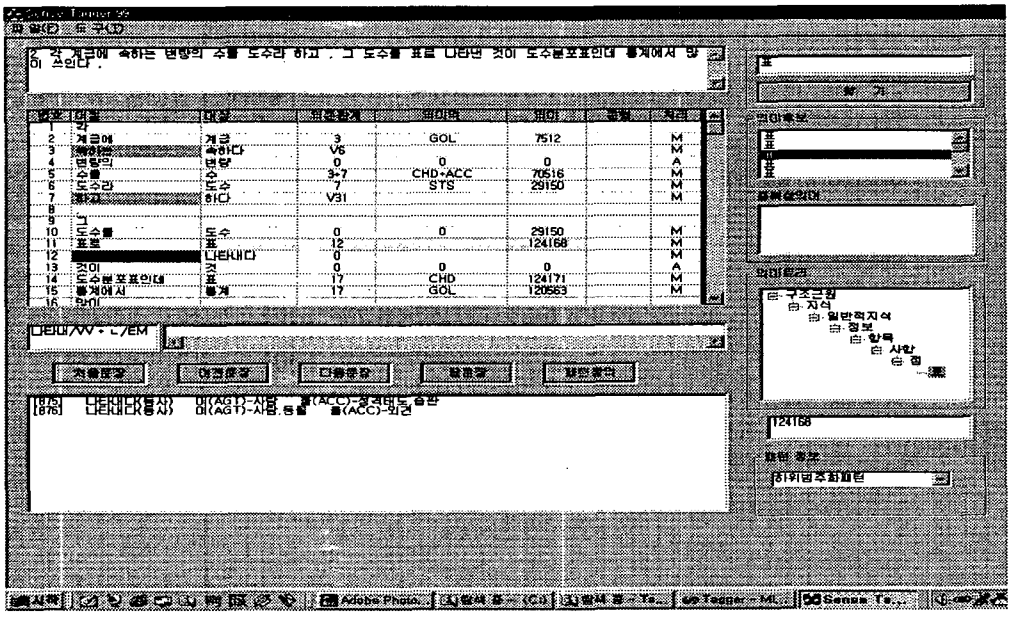

그림 6. 말뭉치의 의미 태깅 작업 일례

역할과 개념 정보가 부여되며, 술어에는 하위범주 패턴 정보를 통해 술어의 의미정보가 기술된다.

이러한 과정을 거쳐 2만여 문장에 대해 의존 관계와 명사의 의미를 태깅한 말뭉치를 구축하였다.

4. 실험 및 결과 분석

4.1 실험

하위범주화 사전이 구문과 어휘의 중의성을 어느 정도 해결하는지를 보기 위해 반자동적으로 태깅된 2만여 문장의 말뭉치를 통해 검증 작업을 수행하였다.

태깅된 말뭉치는 자동 의미 태깅 과정에서 이미 하위범주화사전과 명사 시소러스를 이용하였다. 그러나 앞에서 기술한 바와 같이 자동 매칭 알고리즘상의 오류로 인하여 하위범주화사전과 정합이 이루어졌음에도 불구하고 찾아주지 못하는 경우가 발생할 수 있고 잘못된 하위범주 패턴을 올바른 패턴이라고 제시할 수 있는 경우가 생긴다. 이것은 자동 매칭 알고리즘의 오류이지 하위범주화 사전 자체의 문제라고는 볼 수 없다.

따라서 의미태깅된 말뭉치에서 자동 매칭 정보와 수동 태깅 정보를 단순 비교하면 하위범주화 사전의 올바른 정합률을 찾아낼 수 없다. 따라서 수작업을 통해 정정해 준 의미 태깅된 말뭉치와 하위범주화 사전이 어느 정도 정합되는지를 다시 살펴보아야 하위범주화 사전의 성능을 검증할 수 있다.

[단계 1] 하위범주화 사전에 있는 용언 포함 문장 추출

먼저 의미 태깅된 문장은 여러 개의 용언을 포함하는 장문이므로 각 용언별로 의존관계가 있는 명사 어절을 포함하는 단문으로 분리한다.

2만 문장 중 용언이 의존 관계를 포함하고 있어 단문으로 추출된 문장의 수는 43,067개이다.

이 중 하위범주화 사전의 용언과 일치하는 수는 42,899개로 이를 [단계2]의 대상으로 삼는다.

[단계 2] 하위범주화 사전과 정합

용언 42,899개중 문장의 보어 성분이 하위범주화 사전의 패턴과 하나라도 일치하는 경우는 42,471개이다. 이 중 계층적 명사 의미 시소러스를 이용한 의미 정합을 통해 의미 범주까지도 고려된 용언의 수는 34,401개이다.

용언 42899개 중		
[구문정합]이 이루어진 용언	42471개	정합도 99%
[의미정합]이 이루어진 용언	34401개	정합도 80%

그림 7. 하위범주화 사전과의 정합 결과

[단계 3] 하위범주화 사전과의 정합도 및 빈도정보, 연어정보 추출

의미 정합이 이루어진 전체 34,401개 용언 중 이 용언에 의존되는 명사가 하위범주화 사전의 명사와 의미가 모두 일치하는 경우는 22,360개이고 명사의 의미가 일부분만 일치하는 수는 12,041개이다.

용언에 의존하는 명사별로 하위범주화 사전의 의미와 매칭된 수를 찾아보면 전체 대상명사는 68,012개이고 이 중 하위범주화사전의 의미와 정합된 수는 55,089개이다.

이 과정에서 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 d 일치하는 경우 패턴 일치 빈도 정보의 값을 증가시켰다. 또한 실험 문장의 명사들이 하위범주화 사전의 명사들의 의미 마커 중 어느 것과 매칭됐는지 그 빈도정보를 포함시켰다.

용언 34401개 중		
명사의 의미 모두 일치	22360개	정합도 65%
명사의 의미 부분 일치	12041개	정합도 35%

그림 8. 용언 기준 의미 정합 결과

전체명사 68012개 중	
의미정합된 명사 55089개	정합도81%

그림 9. 전체 명사의 의미 정합도 결과

4.2 결과 분석

실험 2에서 하위범주화사전과의 구문정합도가 99%인데 비해 의미정합도가 80%인 원인은 두 가지로 볼 수 있다. 첫째는 용언에 의존되는 명사가 대명사나 의존명사일 경우 그 의미가 시소러스에 들어있지 않아 의미 매칭에 실패한 경우이다. 이는 실험 3의 정합률을 떨어뜨리는 이유이기도 하다. 둘째는 말뭉치 자체의 형태소 태깅 오류로 올바른 명사의 의미를 줄 수 없는 경우와 고유명사의 경우 인명, 지명 등 10만 어휘를 포함하는 고유명사 사전을 이용하여 의미를 찾았으나 그 사전에서도 의미 매칭에 실패한 경우 의미를 줄 수 없는 경우이었다.

따라서 그러한 명사들의 의미를 별도로 처리하면 정합률을 향상시킬 수 있다. 보다 자동적인 방법에 의해 통계 정보를 부여하는 연구가 필요하다고 생각되며, 따라서 이러한 연구를 수행중인 상태이다.

5. 결론

본 연구에서는 한국어 용언 중심의 하위범주화 사전을 설계하고 구축하였다. 구축된 하위범주화 사전과 계층적 명사 시소러스를 이용한 선택제약 알고리즘을 이용해서 문장의 의미를 자동으로 처리하는 모듈을 설계하고 이를 이용해 의미 태깅된 말뭉치를 반자동으로 구축했다.

하위범주화 사전이 한국어 분석의 중의성 해소에 얼마나 유용한지 알아보기 위해 의미 태깅된 말뭉치와 정합실험을 했다. 실험 결과 태깅된 문장과 하위범주화 사전과의 구문 정합도는 99%를 보였고 의미 정합도는 81%를 보였다. 앞에서 지적했듯이 대명사나 의존명사, 고유명사 등에 대해 별도의 처리를 하면 의미 정합도는 보다 향상될 것으로 보인다. 또한 정합 실험 과정에서 얻은 명사의 연어 정보와 하위범주 패턴의 공기정보 등을 하위범주화 사전 DB에 포함시킴으로써 하위범주화 사전의 성능을 향상시켰다.

이렇게 구축된 하위범주화 사전은 의미 사전과의 연동으로 구문 분석 후보의 수를 줄이며, 동시에 어휘의 의미를 결정할 수 있어, 구문과 어휘 의미의 중의성을 해결할 수 있다. 또한 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] 신효필, "HPSG를 기초로 한 한국어 동사의 하위범주화", 언어학연구 제 7호

[4] 김봉모, "한국어 문장 분석을 위한 용언의 하위범주화", 부산대학교, 국어공학센터, 시스템공학연구소 보고서, 1996년 8월

[5] 추교남 외, "어휘와 구문의 중의성 해소를 위한 한국어 하위범주화사전 구축",한국어정보처리학회 추계 학술발표논문집 제 5권 제2호, 1998

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

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

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

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

[10] 윤평헌, "국어 명사의 의미 관계에 대한 연구", 한국과학재단 보고서, 1995

[11] 이현아, 이종혁, 이근배, "구문분석과 공기정보를 이용한 개념 기반 명사구 색인 방법" , 제 7회 한글 및 한국어 정보처리 학술발표논문집, 1995

[12] 김나리, "패턴 정보를 이용한 한국어 구문 분석", 서울대학교 컴퓨터공학과 석사학위논문, 1996