다의어 분별 정확률 개선을 위한 보조사의 통사격 결정

신준철○, 옥철영
울산대학교
ducksjc@nate.com, okcy@ulsan.ac.kr

Determining a Syntactic Case of Auxiliary Postposition for Improving Accuracy of Polysemy Word-Sense-Disambiguation

Joon-Choul Shin○, Cheol-Young Ock
Ulsan University

요 약

하위범주화는 술어와 보어간의 의존 관계를 정의하는 언어정보로서 다의어 태깅이나 이 외에 자연어처리의 다양한 곳에 이용될 수 있다. 그러나 하위범주화에서 다루는 필수논항은 격조사로 표현되어 실제로 한국어에서 자주 나타나는 보조사는 여기에 포함되지 않는다. 이런 문제 때문에 하위범주화네 나타난 격조사만을 그대로 이용하려고 하면 재현율에 큰 문제가 발생하게 된다. 본 논문에서는 문장에서 격조사 대신 보조사가 사용되었을 때 하위범주화의 필수논항으로 인정할 수 있는 방법을 제시하고, 특히 보조사에 적용할 경우에 생기는 이점을 실험으로 증명한다.

주제어: 다의어, Word-Sense-Disambiguation, 하위범주화, 보조사, 격조사, 필수논항

1. 서론

자연어처리 기술이 발전하면서 다양한 언어자원이 생겼고 그런 언어자원을 효율적으로 사용하려는 시도가 이뤄지고 있다. 대표적으로는 WordNet과 같은 것이 있으며, 한국어에는 비슷한 것으로 UWordMap이 존재한다. UWordMap에는 다양한 어휘 의미관계 정보가 포함되어 있는데 이중에서도 가장 중요한 것 중에 하나가 하위범주화 정보이다. 이 하위범주화는 술어와 필수논항의 의미제약을 정의한다. 따라서 하위범주화의 기본적인 정보단위는 트리플(Tripple)이 되며, 용언과 필수논항의 격 그리고 의미제약으로 이루어져 있다.

- 먹다 * 을/를 * 열매
- 먹다 * 을/를 * 먹이
- 가다 * 에 * 건물
- 가다 * 에 * 나라

하나의 트리플에는 하나의 필수논항이 포함되어 있는데, 예를 들어서 '먹다'와 '열매'의 관계에는 필수논항으로 격조사 '을/를'이 있다. 이 트리플의 문제점은 실제로 '먹다'가 활용될 때 논항으로 항상 격조사 '을/를'이 나타나지 않을 수 있다는 점이다. 예를 들어서 "열매도(까지) 먹었다." 처럼 보조사 '도/까지'가 사용될 수도 있다는 것이다. 이 예에서 '도/까지'는 목적격으로 사용된 것이기 때문에 위 트리플의 '을/를'과 격이 일치한다고 여겨져야 하지만 의미처리(의미역)를 통해 이런 판단을 내리기는 쉽지 않다.

본 논문에서는 위와 같은 경우에 트리플에 저장된 필수논항의 격조사가 실제 문장에 나타난 논항과 문자적으로는 다르더라도 격의 관점에서는 일부 또는 전체가 일치하는 것으로 처리하는 정량적인 방법을 제안하고자 한다.

2. 관련 연구

황화상(2015)은 의미역은 의미격 조사에 의해 부여된다고 하였고, 보조사의 의미 기능은 의미역이 부여된 전체 명사구와 관련된다고 하였다. 또한 보조사의 쓰임 여부가 선행 명사구의 의미역 자체에는 아무런 영향을 끼치지 못한다고 하였다[1]. 예를 들어서 '노래방에서만'에는 의미격 조사 '에서'와 보조사 '만'이 동시에 나오는데 의미격을 결정하는 것은 '에서'이며 '만'은 의미격에 영향을 끼치지 않는다는 것이다. 즉, 보조사만으로는 의미격을 한정할 수 없다는 것이다. 김수정(2013)은 소설을 중심으로 연구하였으며, 주어는 '이/가'나 '은/는' 혹은 다른 보조사가 결합한 형식으로 나타나기도 하고, 조사 없이 나타나거나 생략되기도 한다고 하였다[2]. 즉, 보조사 '은/는'은 주격으로 사용될 수도 있으며, 아무런 조사 없이 명사만으로 주격을 가지는 어절도 있다는 것이다.

본 논문에서는 UWordMap의 하위범주정보를 사용하기 위해서 격조사가 없는 경우(보조사만 나타나거나, 조사

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (R0101-16-0176, Symbolic Approach 기반 인간모사형 자가 학습 지능 원천 기술 개발)

가 없는 경우)에도 문장에서 그 어절의 통사격(논항)을 추측한다[3]. UWordMap은 표준국어대사전을 기반으로 다의어 수준에서 구축한 어휘의미망으로 명사의 상위어, 하위어 관계와, 용언의 하위범주화 정보가 포함되어 있다[4].

UWordMap을 이용한 연구로는 동형이의어와 다의어 중의성 해소에 관한 연구가 존재한다[5, 6]. 신준철(2016)은 기존에 존재하던 동형이의어 분별 시스템의 정확률을 개선하기 위해서 UWordMap을 이용하는 연구를 하였다. 또한 UWordMap을 이용하여 다의어의 중의성을 해소하는 연구도 하였다.

3. 다의어 분별 알고리즘

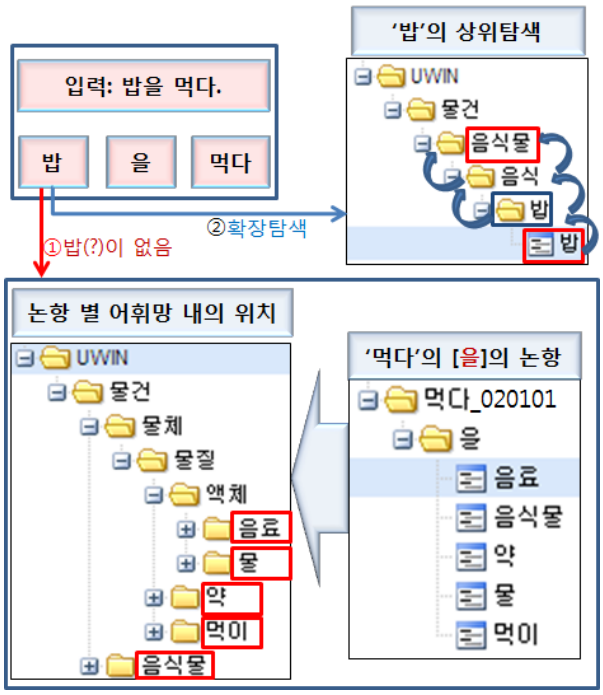

그림 1. 용언의 다의어 분석 예(밥을 먹다)

본 논문에서는 보조사의 격을 결정하는 알고리즘의 효과를 증빙하기 위해 이 알고리즘을 다의어 분별에 응용한다. 구체적으로는 신준철(2015)의 다의어 중의성 해소 알고리즘을 수정하여 개선여부를 확인한다[6]. 기존의 알고리즘은 다음의 3단계로 구성된다.

(1) 입력된 문장에서 용언를 찾고, 그 용언과 인접한 (또는 의존관계에 있는) 명사를 찾는다.
(2) 용언과 명사의 관계를 UWordMap에서 찾으며, 이 과정에서 명사의 상위어를 사용할 수도 있다.
(3) 문장에서 사용된 논항과 UWordMap에서 찾은 논항이 일치하는지 확인한다.

단계 (1)은 문장에서 용언과 연관된 명사를 찾는 과정으로, 주로 이 둘은 인접한 경우가 많다. (2)에서는 UWordMap의 하위범주화 사전에서 (1)의 용언-명사를 가진 트리플을 찾는다. 이 과정에서 상위어를 사용할 수 있는데, 이것이 [그림 1]에 나타나있다. '밥'의 상위어 '음식물'이 '먹다'와 관련되어 있음을 알 수 있다.

그리고 단계 (3)에서 트리플에 있는 논항과 문장에서 수집한 논항이 일치하는지 확인한다. 만약 일치한다면 해당 트리플에 등록된 용언과 명사의 다의어 번호에 (예: 그림 1의 먹다_020101) 점수가 높게 계산된다. 점수 계산에는 자질 함수방식이 사용되며, 총 4가지 자질 함수가 있다. 이것은 <표 1>에 모두 나타나있다. 그리고 수식 1은 다의어의 점수를 계산하기 위해 자질함수의 값을 사용하는 것이다. poly는 다의어이며 i는 자질함수의 번호이고(1~4), w는 가중치 f는 자질함수이다. 자질함수가 4개이기 때문에 가중치도 4가지이고, 각각의 가중치는 실험을 반복하여 실험자가 직접 수작업으로 정할 수 있다. 모든 다의어 중에서 가장 점수가 높은 다의어로 태깅한다.

표 1 자질 함수들

번호	자질 함수
1	해당 동형이의어에서 이 다의어가 첫 번째 의미일 경우에 1을 반환. 그 외에는 0을 반환. (첫 번째라 함은 의미번호가 가장 빠른 것을 뜻한다.)
2	하위범주화에서 용언-명사 관계가 있음을 확인하면 1을 반환. 그 외에는 0을 반환. (즉, 트리플에서 논항을 제외하고 일치 여부를 확인한다.)
3	하위범주화에서 용언-논항-명사의 관계가 있음을 확인하면 1을 반환. 그 외에는 0을 반환.
4	트리플에 등록된 명사의 깊이를 반환 예) '음식물'은 최상위 명사까지 거리가 2이므로 2를 반환

$$Score(poly) = \sum_i w_i f_i(poly) \qquad \text{수식 1}$$

본 논문에서는 자질 함수 3을 수정하였다. 본래는 1 또는 0을 반환하지만 수정된 자질 함수 3은 논항이 일부 유사한 것으로 판단되면 0대신에 0.5와 같이 0과 1 사이의 값을 반환할 수 한다. 본 논문에서는 일단 보조사에 한정하여 단순한 방식으로 적용하였다. 보조사는 세종태그셋이 JX인 것을 의미한다.

은, 는, 이라도, 도, 마저, 처럼, 같이, 만 ...

보조사는 필수논항의 격조사와는 문자적으로 다르기 때문에 자질 함수 3에서는 항상 0이 반환되지만 수정된 자질 함수 3에서는 0.5가 반환된다. 예를 들어서 "밥도 먹었다."는 단계 (2)에서 '밥'의 상위어 '음식물'을 찾게 되고, '음식물'이 '먹다'와 관계가 있으며 이 때 등록된 논항은 '을'임을 알게 된다. 그리고 단계 (3)에서 자질 함수 3을 계산하게 되는데, 이 때 논항이 일치하지 않지만 문장에서 나타난 논항 '도'는 보조사이고 목적격 '을'과 일부 유사한 것으로 판단하여

수정된 자질 함수 3은 0.5를 반환한다.

종종 신문 제목과 같이 짧게 서술하고자 할 때에는 논항의 격조사가 생략되는 경우도 있다. 예를 들어서 "밥 먹었니?"에서 '밥'에는 조사가 없다. 이런 경우에도 수정된 자질 함수 3은 보조사가 있었던 것처럼 0.5를 반환하도록 한다.

예외적으로 격조사 '에서'의 경우에는 일반적으로 이를 대체할만한 보조사가 없는 것으로 보인다. 예를 들어서 "공장에서 가공된다."를 "공장도 가공된다."로 바꿀 수는 없다. 강제로 바꾼다면 그 문장의 의미가 완전히 바뀌게 된다. 따라서 트리플에 등록된 논항의 격조사가 '에서'라면 문장에서 사용된 논항이 보조사가 쓰여도 수정된 자질 함수 3은 0을 반환한다. '에서'를 제외하고는 모든 보조사는 다른 논항으로 결정할 수 있으며 모두 자질함수 3에서 모두 일괄적으로 0.5를 반환한다.

4. 실험결과 및 분석

실험환경은 신준철(2015)의 것과 동일하게 표준국어대사전의 용례를 말뭉치로 사용하였다. UWordMap 브라우저에서 단어를 검색하고 다의어 번호를 클릭하면 나타나는 용례 항목을 통해서 확인할 수 있으며[3], 다만 여기에는 원형만 표시하고 있다. 이 말뭉치에는 총 314만개의 형태소가 있고 정확률 측정에는 품사가 NNG, VV, VA인 것만 대상으로 하였다. 그래서 실제로 정확률 측정의 대상이 되는 형태소의 수는 총 89만개이며 정확률 측정은 형태소 단위로 하였다. 형태소 분석과 품사 그리고 동형이의어 분석을 위해서는 UTagger 를 사용하였고, baseline으로 삼기 위해 신준철(2015)의 기존 알고리즘을 그대로 사용하여 다의어 분별 정확률을 측정하였다. 이 정확률은 본래 65.58%였으나 UWordMap이 1년간 계속 추가 구축되었고, 소스코드의 일부 오류를 수정하여 정확률이 조금 향상되어 66.17%로 측정되었다.

보조사의 격 결정 효과를 실험하기 위해서 3장에서 상술한 알고리즘을 사용하였으며, 그 결과로 정확률이 66.35%가 나왔다. <표 2>가 이것을 나타낸다. 이것은 기존 정확률에서 약 0.18%가 향상된 것으로 매우 작은 변화이다. 이 변화가 의미가 있는 것인지 아니면 단순한 우연인지 확인하기 위해서 알고리즘이 수정되면서 발생한 태깅 결과의 변화를 분석하였다. 알고리즘이 수정되면서 올바르게 태깅된 형태소의 개수는 총 4천개이고, 수정 전에는 올바르게 태깅되다가 알고리즘 수정으로 인해 오답으로 태깅된 형태소는 2,452개이다. 이것들을 분석해본 결과 모두 보조사가 사용되거나 조사가 생략된 문장들이였다. 올바르게 변화한 것이 반대의 것에 비해서 1.6배로 많기 때문에 단순한 우연으로 정확률이 향상된 것이 아닌 것으로 판단된다.

표 2 정확률 비교

	수정 전	수정 후
정확률	66.17%	66.35%

부작용으로 2,452개가 틀리게 태깅된 것은 단순하게 모든 보조사에 대해서 동일하게 0.5로 취급하였기 때문으로 판단된다.

5. 결론 및 향후 연구방향

본 논문에서는 실제 문장에서 필수논항의 격이 나타나지 않고 보조사가 나타난 경우 하위범주정보의 의미제약정보를 이용하여 보조사의 격을 결정하는 방법을 제안하였다. 본 논문에서는 간단하게 보조사 또는 조사가 생략된 경우에 한정해서 다른 필수논항과 50%의 유사도가 있다고 간주하도록 하였고, 이 방법이 효과가 있음을 확인하기 위해서 다의어 중의성 해소에 실험하였다. 그 결과 약간의 효과가 있음을 확인하였다.

본 논문에서 제안하는 알고리즘은 비록 간단하지만 하위범주화를 조금 더 효과적으로 사용할 수 있음을 보였다. 특히 현재까지 구축된 한국어 하위범주화 정보는 실제로 응용하기에는 다소 부족한 면이 있으며 보조사에 대해서는 전혀 다루고 있지 않기 때문에 이러한 연구가 앞으로 반드시 필요하다. 따라서 향후에는 보조사의 여러 현상들을 통계적으로 접근할 필요가 있으며, 기존의 하위범주화를 더욱 효과적으로 응용할 수 있는 방향으로 연구해야할 것이다.

참고문헌

[1] 황화상, "보조사의 주변 범주", 국어학회논문지, 제37호, pp.309-334, 2015.

[2] 김수정, 최동주, "소설 텍스트에서의 주어의 실현 양상", 한민족어문학논문지, 제64호, pp.37-69, 2013.

[3] UWordMap, http://klplab.ulsan.ac.kr/doku.php?id=uwordmap, 울산대학교 자연어처리연구실.

[4] 배영준, 옥철영, "한국어 어휘지도(UWordMap)와 API 소개", 한국정보과학회언어공학연구회 제 26회 한글 및 한국어 정보처리 학술대회, pp.27-31, 2014.

[5] 신준철, 옥철영, "한국어 어휘의미망(UWordMap)을 이용한 동형이의어 분별 개선", 정보과학회논문지, 제43권, 제1호, pp.71-79, 2016.

[6] 신준철, 옥철영, "어휘지도(UWordMap)를 활용한 명사와 용언의 다의어 중의성 해소, 제27회 한글 및 한국어 정보처리 학술대회 논문집, pp.216-219, 2015.

UTagger는 울산대학교 한국어처리연구실에서 개발한 품사 및 동형이의어 태깅시스템이다.