의미역 태깅의 제문제

김윤정○, 옥철영
울산대학교 국어국문학과, 울산대학교 IT융합전공
jungi0006@ulsan.ac.kr, okcy@ulsan.ac.kr

Consideration of Semantic Role Tagging

Yun-Jeong Kim○, Cheol-Young Ock
Ulsan University, Ulsan University

요 약

본고는 기존 연구에서 상정한 의미역에 기반하여 의미역 태깅 작업 중 실제 문장에 의미역을 태깅하는 데 나타난 문제점들에 대해 재고해보았다. 의미역을 태깅하는 데에 격틀 사전을 이용한 반자동의미역태깅프로그램의 정상적인 구동을 위한 사전의 재정비와 실제 문장에서는 드러나지만 사전에서는 나타나지 않는 문형 정보를 상세히 검토해야 함을 알게 되었다. 이를 해결하기 위해 격틀사전의 기본 사전이 표준국어대사전의 통사정보 제시를 문제삼아 이를 해결하기 위한 방안을 모색하고, 실제 문장에서 격교체에 의해 나타나고 있는 논항정보교체에 대처하기 위한 방안을 마련하고자 한다.

주제어: 의미역, 격틀사전, 필수논항

1. 서론

본고는 한국어 문장에 의미역을 태깅하여 기계가 사람과 유사한 인지 능력을 갖도록 하는 데에 목적이 있다. 이를 위해 김윤정(2014)은 의미역을 상정하고 의미역을 태깅하기 위한 지침을 작성하였다[1]. 이를 기반으로 세종구구조말뭉치를 대상으로 의미역 태깅 작업을 하였다. 의미역이 주석된 말뭉치를 기반으로 기계처리를 하기 위해서는 좀더 다양한 문장에 정확한 의미역을 태깅할 수 있어야 한다. 기준으로 격조사와 격틀(문형) 정보를 조사하여 각각의 형태별 의미역을 연결하여 보았다. 한국어의 격조사가 문장에 나타날 때 어떤 격을 가지고 어떤 의미 기능을 주로 하는지에 따라 의미역을 태깅한 결과물을 의미역 태깅 프로그램에 탑재하였다. 또한 기준 사전으로 표준국어대사전을 정하고 사전 내 용언을 추출하여 용언의 문형 정보, 뜻, 용례 정보 등을 기준으로 의미역을 태깅하였다. 태깅된 결과물을 격틀사전이라고 칭하고 이를 의미역태깅 프로그램에 탑재하였다.
일차적으로 작업을 위한 기초 작업을 하였고, 이를 기반으로 실제 말뭉치 대상 의미역 태깅을 하였다. 의미역을 태깅하는 모든 기준들을 바탕으로 의미역을 태깅하도록 하였지만, 사람의 직관과 사전적 자료를 바탕으로 태깅할 때와 기계가 탑재된 격조사별 의미역, 격틀사전을 기준으로 태깅을 할 때 정확률에 차이를 보이고 있다[2]. 이렇게 차이를 보이는 이유는 우선 격조사에 의미역을 태깅한 것과 문장에서 드러나는 격조사의 상황이 완벽하게 일치하지 않거나, 격틀사전의 정보가 실제 문장에 대입되기에 많이 부족하였기 때문이다.
현재까지 의미역 태깅 결과물은 1차, 2차 합하여 100만 개 정도가 구축되었다. 2차에 걸친 태깅 작업을 통해서 발견된 문제점으로는 기계적으로 자동 처리하는 것에도 정확하게 의미역을 태깅하는 데에 부족함이 있고, 자료를 기반으로 사람의 직관으로 의미역을 태깅하는 데에도 부족함이 있었다는 점이다. 이를 정리해 보면 가장 큰 문제점은 기준으로 정한 격틀사전의 문형 정보가 표준국어대사전의 격틀(문형) 정보만을 따랐다는 점과 표준국어대사전의 자체의 문제점을 해결하지 못했다는 점이다. 다음으로는 실제 문장이 그렇게 규범적이지 않다는 점이다. 규범문법에서 정한 규칙을 제대로 지키는 문장보다 이를 지키지 못한 문장의 사례가 많아, 사전 정보와 문법서의 정보만으로는 정확한 의미역 태깅이 어렵다는 점이다. 그러므로 본고에서는 이 두 가지 문제점을 제시하여 해결 방안을 모색해보고자 한다. 이는 앞으로 의미역 태깅의 기계 처리가 가능하도록 하는 중요한 작업과정이 될 것이라고 생각된다.
본고의 구성은 다음과 같다. 2장에서는 관련연구를 제시하고, 3장에서는 의미역 태깅시 발생하는 문제점을 3.1. 의미역 태깅의 사전적 문제, 3.2. 문장 구조적으로 발생하는 문제에 대해 나누어 기술한 후 마지막으로 4장에서 결론을 기술하겠다.

2. 관련 연구

본고에서 문제로 삼고 있는 표준국어대사전의 사전적 문제와 실제 문장이 기계처리에 문제가 되는 점에 관한 관련연구를 살펴보도록 하겠다.
2.1. 사전적 문제는 본고에서 바탕 사전으로 삼은 표준국어대사전의 문제이다. 표준국어대사전에 문제점은 표제항의 선정과 구조에서부터 통사제시문제, 용례의 미흡함으로 나눠볼 수 있다. 실제로 표준국어대사전은 표준어, 방언, 고어, 북한어, 비속어까지 포함시키는 표제항의 선정의 문제를 갖고 있다.

표준국어대사전의 용례의 미흡함을 제시한 논의는 다음과 같다. 정호성(2000)은 <표준>의 경우 주표제어와 부표제어를 합하면 50만 8천 여 항목 중 용례가 제시된 주

표제어는 74,092항목, 부표제어는 25,651항목으로 19% 수준에 머무르고 있다고 했다[3]. <표준>의 표제어 중 용례가 제시된 것은 20%밖에 되지 않고 인용례는 전체 표제어의 10%에 불과하다고 한다. 한영균(2006)은 현대국어 표준어 용언 표제항 중에서 용례가 있는 것은 48%에 불과하고, 51.4%에 달하는 항목에는 용례가 없다고 밝히고 있다[4]. 용례가 있더라도 평균 용례수는 평균 2개 안팎, 다의어의 경우에는 주표제항 다의어와 부표제항 다의어의 경우가 차이가 있고, 부표제항의 경우 소홀하게 다루어졌음을 지적하였다. 용언의 용례가 1개인 항목이 가장 많아서 전체의 37.3%를 차지했고 인용례는 22.85%, 작성례는 77.15%이어서 사전 용례의 충실성을 판단하기가 어렵다고 했다.

표제항 선정에 대해서는 조재수(2000:137)는 표준국어대사전에서 “다른 표제어에 딸리지 않고 배열되는 말”이라고 설정하여 일부 파생용언과 파생부사 따위의 경우 그 어근을 주표제어로 삼았다고 기술했다[5]. 이는 유현경(2010:223)에서도 표제어 단위 설정 문제 중 가장 논란이 되는 것으로 꼽고 있다[6]. 그중 일음절 불구 어근까지를 표제어로 삼는 것을 문제로 지적하였다. 이은령. 윤애선(2010)도 표제어 제시 방법과 연계된 문제를 지적하였다[7]. 표제어와 부표제어의 원칙에 기인한 문제로 어근 명사에만 의미 정보를 주고 부표제어인 동사에는 의미정보가 없이 문형과 용례만을 제시한 점이다.

통사제시정보에 대해서는 고석주(2003)에서 표준국어대사전과 연세한국어사전을 비교하여 문형 정보의 단순제기가 연세한국어사전에 비해 미흡한데 ‘교차정보보어구문’을 이루는 서술어가 가지는 문법적 특성을 명확히 밝히지 못하고 있기 때문이라고 지적하였다[8]. 또한 격틀(문형) 정보를 단순히 조사의 형태로만 제시하는 방식은 실제 문장에서 조사를 대입했을 때 문장의 정문 여부를 확인하기 어렵다고 했다. 이은령.윤애선(2010)은 표준국어대사전의 통사정보 제시는 “주어를 제외한 용언의 필수 성분만을 격조사나 어미로 표시한다.”라는 원칙에 따르기 때문에 문형 정보에 쓰인 격조사의 통사, 의미적 주해는 일러두기에 명시될 뿐 이를 통해 동사의 가능한 모든 통사 구조는 알 수 없다고 하였다. 게다가 문형정보가 용례의 문형을 반영하는 데에 지나치게 충실하여 일반적인 문형과는 거리가 있다고 보았다.

이은령.윤애선(2010:181,190)에서는 표준국어대사전의 다의구분과 의미정보에서 의미정보의 구성에 문제가 있음을 지적하였다. 표준국어대사전은 공통 문형 정보를 앞세워 동일 부류의 의미를 함께 묶고, 이후 하위 단계에서 각 어휘의미를 번호로 구분하며 각각의 의미에 해당하는 표제어 용례를 제시하고 있으며, <명사+접사>동사의 파생현상이 국어에서 매우 빈번하기 때문에 의미기술의 경제성에 근거하여 <명사+접사>파생동사의 의미를 호도하는 결과를 낳았다고 했다.

2.2. 실제 문장의 비규범적인 문제점으로는 주로 띄어쓰기의 문제점을 살펴볼 수 있다.

한영균(2003)은 말뭉치를 통한 어휘 계량적 분석, 특히 어휘 빈도 조사에 있어 가장 큰 골칫거리 중 하나가 띄어쓰기 문제이지만, 이를 해결하는 것은 참으로 어렵다고 주장했다[9]. 띄어쓰기의 문제는 합성어의 판별기준 모호하기 때문이라고 했다. 한영균(2003:73)에서 또한 어휘 빈도 조사에서는 어떤 사전의 표제항을 준거로 하던 표제어로 등재되지 않은 형태와 등재된 형태가 공존한다. 특히 복합 구성 문법 단위는 사전 표제어로의 등재 여부와 관계없이 항상 띄어 쓴 예와 붙여 쓴 예가 출현한다. 따라서 왜곡되지 않은 어휘빈도 조사를 위해서는 조사 대상 단위를 한정할 필요가 있고, 아울러 하나의 단위를 띄어 쓴 예들을 처리하기 위한 별도의 방법이 개발되어야 한다고 주장하였다.

3. 의미역 태깅 문제점

3.1. 사전적 문제점

격틀사전의 문제점은 표준국어대사전이 안고 있는 통사정보제시의 문제와 표제항 구성을 주표제항과 부표제항으로 구성하는 데에서 발생하는 문제점을 그대로 안고 있다는 데에 기인한다. 2.1.절 관련연구에서 제시한 것과 같이 표준국어대사전은 통사정보정보제시 문제와 표제항의 구성의 문제, 용례 제시의 문제가 있다.

표준국어대사전의 용언을 구축 사전을 기반으로 삼고 북한어, 고어, 방언을 배제한 나머지 표제항을 모두 대상으로 하였다. 비표준어인 ‘~의 잘못’과 같은 표제항은 배제하지 않고 그대로 두었다. 선별된 모든 용언의 기존의 동형이의어 수준에서 다의어 수준으로 세분화한 뒤 개별어에 각각의 의미역을 태깅하는 작업을 시행하였다. 이때 먼저 부딪치게 된 것이 표제항의 격틀정보 심각할 정도로 부족하다는 점이었다. 현재 용언 중 격틀정보가 없는 것이 37,114개로 54%를 차지하고 있다. 게다가 용례가 없는 것은 97%를 차지하고 있다[1]. 사전의 정보를 기반으로 의미역을 태깅하였지만, 실제로 46%의 격틀정보만으로 의미역을 태깅했다고 볼 수 있다. 그 외에 격틀정보가 없는 표제항에 대해서는 해당 표제항이 부표제항으로 존재하는 경우가 많았고 대체로 ‘-하다’, ‘-되다’의 형태로 어근형의 의미와 용례에 기반해서 의미역을 태깅하기도 하였다. 표준국어대사전 내의 정보로는 의미역을 찾거나 태깅하기 어려운 경우에는 관련 사전을 찾아 참고하거나[2], 한국어 문장의 가장 기본이 되는 1항 술어 형식으로 대상역을 태깅하였다.

이렇게 구축된 격틀사전의 태깅된 의미역을 실제 작업 말뭉치에 적용한 결과 중 출현 여부에 따른 조사 결과는 다음과 같다. 격틀사전의 용언 수는 동형이의어 단위로

1) 여기에서 용례가 없다는 97% 비율은 어근형의 부표제항으로 제시된 사례의 경우도 포함된 결과이다. 용례정보에 대해서는 한영균(2006:296)에서도 언급한 바 있다. 표준어만을 대상으로 삼은 경우에도 48%정도가 용례가 없다고 제시하였다.

2) 문형정보와 용례정보가 없는 표제항에는 인터넷 검색, 연세한국어사전, 고려대 민족한국어사전이 탑재된 다음사전, 백과사전 등을 검색하여 최대한 관련어가 들어 있는 예문을 찾아 의미역을 태깅할 수 있도록 하였다.

계산하면 67,941개이다. 이중 실제 말뭉치에서 출현빈도 0인 용언은 58,547개로 격틀사전 전체 용언의 86%를 차지하고 있다. 실제 작업 말뭉치에서 사용된 용언의 수는 9,394개로 격틀사전의 전체 용언 중 14%만이 사용되었다. 사용된 용언도 출현 빈도가 10 이상인 용언 3,007개로 32%의 비중을, 출현빈도가 1인 용언은 2,385개로 25%의 비중을 차지하고 있다. 격틀사전에서 실제 말뭉치에 사용된 용언 중 상위빈도 10위까지 정리한 표는 다음과 같다.

<표1> 격틀사전의 용언 중 실제 말뭉치 출현 빈도 상위 10위

순위	용언	출현 빈도
1	하_01	13,314
2	있_01	11,429
3	되_01	7,461
4	없_01	6,437
5	아니	4,315
6	보_01	3,891
7	대하_02	3,637
8	같	3,609
9	위하_01	2,669
10	받_01	2,327

<표1>에 제시된 바와 같이 격틀사전에서 사용된 동사 중 상위 10위 안에 드는 것으로는 '하다, 있다, 되다, 없다, 아니다, 보다, 대하다, 같다, 위하다, 받다'는 한국어 문장에서 가장 많이 사용하는 기능동사이다. 이들 동사들 중 '하_01'을 살펴보도록 하겠다. '하_01'를 다의관계로 나누어 보면 본동사만 10개의 격틀 정보가 있고, 10개의 격틀 정보 하위에 각각의 문형별 다의정보를 따로 구분하고 있다. 본용언과 보조용언이 같이 구성되어 있고, 보조용언은 보조동사와 보조형용사 각각을 갖추고 있다. 이렇게 '하_01'의 전체 격틀은 12개이고 다의정보까지 세분하면 41개의 개별 의미를 가지고 있다. 이렇게 세분되어 있지만, 실제 작업말뭉치에서는 하나의 동형이의 수준으로만 제시가 되니 사전의 41개 의미역을 일일이 대입해 보아야 하는 어려움이 있다. 이는 기계처리에도 큰 걸림돌이 되고, 사람의 직관으로 작업을 하는 데에도 어려움이 따른다.

<표2> 표준국어대사전의 '하_01' 다의정보별 의미역 태깅

하다 010101	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010102	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010103	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010104	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010105	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010106	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010107	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010108	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010109	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010110	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010111	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010112	[동사]	◁…을▷	{X:행동주 Y:대상-을/를}
하다 010200	[동사]	◁…을 -게▷	{X:행동주 Y:대상-을/를 Z:방법-게}
하다 010301	[동사]	◁…을…으로▷	{X:행동주 Y:대상-을/를 Z:착점-으로/로}
하다 010302	[동사]	◁…을…으로▷	{X:행동주 Y:대상-을/를 Z:방향-으로/로}
하다 010303	[동사]	◁…을…으로▷ ◁-기로▷	{X:행동주 Y:대상-을/를 Z:내용-으로/로} {X:행동주 Y:대상-을/를 Z:내용-기로}
하다 010400	[동사]	◁…을 -고▷	{X:행동주 Y:대상-을/를 Z:내용-고}
하다 010501	[동사]	◁…으로▷	{X:행동주 Z:원인-으로/로}
하다 010502	[동사]	◁…으로▷	{X:행동주 Z:경로-으로/로}
하다 010601	[동사]		{X:대상}
하다 010602	[동사]		{X:대상}
하다 010701	[동사]	◁-고▷	{X:행동주 Z:내용-고}
하다 010702	[동사]	◁-고▷	{X:행동주 Z:내용-고}
하다 010800	[동사]	◁…에/에게 -게▷	{X:행동주 Z0:착점-에/에게 Z1:방법-게}
하다 010900	[동사]	◁-게▷	{X:행동주 Z:내용-게}
하다 011001	[동사]		{X:대상}
하다 011002	[동사]		{X:대상}
하다 011003	[동사]		{X:대상}
하다 011004	[동사]		{X:대상}
하다 011005	[동사]		{X:대상}
하다 011006	[동사]		{X:대상}
하다 011007	[동사]		{X:대상}
하다 011101	[동사][보조]		
하다 011102	[동사][보조]		
하다 011103	[동사][보조]		
하다 011104	[동사][보조]		
하다 011105	[동사][보조]		
하다 011106	[동사][보조]		
하다 011107	[동사][보조]		
하다 011201	[형용사][보조]		
하다 011202	[형용사][보조]		

<표2>에서 빈칸이 있는 것은 격틀정보가 없다는 것을 의미한다. [11]과 [12]에 해당하는 보조용언은 의미역 태깅 대상이 아니다. 그러나 실제 문장에서는 하다의 형태 중 사전의 격틀정보의 적용이 용례와 맞지 않는 것도 많고, 제시된 용례가 전부가 아니기 때문에 제시한 표대로 의미역 태깅이 되는 것은 아니다. 이 부분이 격틀사전을 수정보완해야 하는 가장 큰 이유이다.

문형 정보 중에서 [4], [7], [11], [12]번 동일한 보문 정보를 갖는다.

[4]번 문형 [◁…을 -고▷]
뜻풀이 '이름 지어 부르다.'
용례 '꿀을 얻기 위해 벌을 치는 것을 양봉이라고 한다.'
[7]번 문형[◁-고▷]

뜻풀이 ① '간접 인용의 경우에는 '-고'가, 직접 인용의 경우에는 '-라고'가 쓰인다. 이르거나 말하다.' ② '주로 '하는' 꼴로 쓰이는데 '-고 하는'은 '-는'으로 줄기도 한다. 다른 사람의 말이나 생각 따위를 나타내는 문장의 내용을 받아 뒤에 오는 체언을 꾸미는 기능을 나타내는 말.'
용례: [010701] 그 책에서는 세계는 이제 정보화의 전쟁에 돌입했다고 했다.
[010702] 그가 거짓말을 했다고 하는 증거는 있다.

보조동사인 문형 [11]번은 7개의 다의정보를 담고 있다. 7개의 정보 중 ⑥ '동사 뒤에서 '-고 하다' 구성으로 쓰여 앞말의 사실이 뒷말의 이유가 됨을 나타내는 말.', ⑦ ' 동사 뒤에서 '-고는 하다', '-곤 하다' 구성으로 쓰여 앞말이 뜻하는 행동을 습관처럼 하거나 앞말이 뜻하는 상황이 반복되어 일어남을 나타내는 말.' 이다.
보조형용사인 문형 [12]는 2개의 다의 정보가 있는데 그 중 ② '형용사 뒤에서 '-고 하다' 구성으로 쓰여 앞말의 사실이 뒷말의 이유가 됨을 나타내는 말.' 이다.

이상의 비교 결과를 통해 알 수 있는 것은 보조용언과 본용언이 유사한 보문 구성을 가지고 있어 둘의 구분이 어렵다는 점이다. 이렇게 세분화된 사전의 다의 정보가 실제 작업 말뭉치에서는 동형이의정보 수준의 '하-01'로만 제시되어 나타나니 처리하기가 성가시다고 아니할 수 없다. 또한 실제 작업 말뭉치와 완전히 일치되지 않아서 41개의 변수에 또 다른 정보를 요구하게 된다[3].

용례 부분에서 용례가 없는 것이 97%라는 점이다. 이는 실로 사전으로서의 기능을 무시한 처리라고 볼 수 있다. 한영균(2006:296)은 현대국어 표준어 용언 59,299개 항목 중에서 용례가 제시되어 있는 것은 48.6%에 불과한 28,791개 항목뿐이다[10]. 게다가 용언 항목 수의 분포 중 용례가 1개인 항목이 가장 많고, 전체의 37.3%를 차지하며, 인용례와 작성례의 비율을 확인하였는데 인용례의 비율이 22.85%, 작성례의 비율이 77.15%였다. 인용례의 많고 적음에 따라서 사전 용례의 충실성을 판단하기는 어렵지만, <표준>이 역사사전을 지향한 것이 아님을 인용례의 비율을 통해서도 확인하여 지적했다.

3.2. 실제 문장에서의 구조적 문제점

실제 문장의 비규범적인 문제점으로는 좁게는 실현된 문장의 단어, 어휘 구성의 문제이고, 넓게는 띄어쓰기, 문장의 확대라고 볼 수 있다. 어휘 구성이나 문장 구성의 띄어쓰기 문제가 우선 작업의 어려움이었다. 그 다음으로는 문장의 확대로 인한 의미역 태깅의 어려움이나 격틀정보에 없는 격교체에 의한 문장의 생성이 있다.
이러한 실제 작업 말뭉치의 문장 구조적 문제점은 격틀 사전에서는 해결해 줄 수 없는 부분이다.

용언 중 합성어의 문제는 띄어쓰기의 문제와 실제 말뭉치에선 하나의 단어로 처리하고 있는 것들이 사전에서는 미등재어인 문제가 있다. 특히, 사전에서 하나의 단어로 인정해서 의미역을 태깅해 놓은 것이 실제 작업 말뭉치에서는 띄어쓰기가 되거나 단어 사이에 조사가 붙어 실현된 경우가 많아서 전체 문장에서의 태깅 기준에 어려움을 겪는다.

'몰아넣다는 표준국어대사전에 등재된 합성어이다. 그러나 실제 작업 말뭉치에서는 <그림1>처럼 '몰아 넣을'의 형태로 띄어쓰기를 하는 경우가 많다. 이런 경우 프로그램에서는 각각을 형태소 단위로 보고 '몰다'와 '넣다'를 개별 단어로 인지한다. 그런 경우 지침서대로 '몰다'와 '넣다' 각각에 대상역 '화기를', 착점역에 '단전에'로 태깅하고 있다. 그러나 이런 경우 이렇게 의미상 둘을 동일하게 태깅하는 것은 큰 무리가 없으나, 실제로 '몰아넣다'를 위한 의미역 태깅이 아니므로 결과물에 대한 새로운 처리가 필요하다.

<그림1> 합성어의 실현 사례 - 몰아넣다

의미역부착 | 형태소·의존관계

순서	의존	어절	3 마치_02	7 몰_01	8 넣	10 하_01	12 통일_02시키
1	2	이_05/MM					
2	3	동작_03/NNG+을/JKO	THM				
3	4	마치_02/VV+고/EC					
4	7	나_01/VX+아서/EC					
5	7	화기_02/NNG+를/JKO		THM	THM		
6	7	단전_01/NNG+에/JKB		GOL			
7	8	몰_01/VV+아/EC					
8	9	넣/VV+는/ETM					
9	10	호흡/NNG+을/JKO				THM	
10	12	하_01/VV+아/EC					
11	12	정신_12/NNG+을/JKO					THM
12	12	통일_02/NNG+시키/XSV+ㄴ다/EF+./SF					

이와 반대의 문제도 존재한다. 사전에는 미등재어인데, 사전에서 접사로 인정한 '-하다, -되다, -받다, -당하다' 류는 하나의 단어로 실제 작업 말뭉치에 실현된다. 이러한 대상에 대해서는 사전에서 의미역 태깅 기준을 제시하지 못하고 있다.

<그림2> 표준국어대사전 미등재어 : 작품하다

의미역부착 | 형태소·의존관계

순서	의존	어절	4 보이_01	7 보_01	8 작품_01하	10 일하	11 보_01	14 보이
1	2	개인적/NNG+이/VCP+ㄴ/ETM						
2	8	차원_01/NNG+에서/JKB						
3	8	눈_01/NNG+에/JKB	THM					
4	5	보이_01/VV+지/EC						
5	6	않/VX+는/ETM						
6	7	관객/NNG+들_09/XSN+과/JKB		COM	COM			
7	8	마주_01/MAG+보_01/VV+며/EC						
8	21	작품_01/NNG+하/XSV+다가/EC+./SP						

이런 류의 단어들은 띄어쓰기에 일관성이 없고 실제 말뭉치에서의 처리 또한 어려움을 보이는 것들이다.
실제 작업 말뭉치 내에서 사전미등재어인 -하다, -되다, -받다, -당하다' 류를 추출한 결과는 다음과 같다.

3) 여기에는 두 가지 이유가 있다. 형태소분석 단계에서부터 본용언인지 보조용언인지 판별이 어려운 것에서 오류가 생겨서 제시된 경우가 발생한다. 이 경우 만약 본용언인데 보조용언으로 태깅이 되어 있다면 태깅 정보를 수정하고 의미역을 그에 맞춰 태깅해야 한다. 또 다른 경우는 본용언으로 태깅이 되어 있는 경우는 보조용언으로 태깅 수정을 하면 된다. 이 과정에서 의미역 태깅의 오류가 발생하게 된다.

<표3> 사전미등재어 -하다, -되다, -받다, -당하다' 개수

개수	-하-	-되-	-당하-	-받-
미등재어	213	316	137	357
중복어 삭제	165	207	94	170

미등재어임에도 실제 작업 말뭉치에 출현한 개수는 상당히 많은 수이다. 이외에도 더 많은 단어들이 존재하지만, 대표로 파생어 4개만을 다루어보았다. 중복어를 삭제한 각각의 단어를 합하면 636개로 적지 않은 수이다. 격틀사전에서 실제 작업 말뭉치에 출현하지 않은 단어보다, 실제 작업 말뭉치에 출현한 이와 같은 단어들에 대한 의미역 태깅 작업이 필요하다고 본다.

또한 하나의 단어가 띄어쓰기의 일관성이 깨짐으로 생기는 문제는 이외에도 상당히 많다.

격틀사전에서 '-하다', '-되, -받다, -당하다'의 어휘쌍으로 존재한다고 하더라도 이들이 능동과 피동의 의미 관계로 연결되는 경우는 많지 않다. 대체로 한자어 어근에 '-하다', '-되다'를 붙여 단어쌍을 이루는 경우였다. 용례가 대체로 어근형의 의미를 그대로 사용하게 되어 있어 개별적인 '-하다', '-되다'의 고유한 용례가 없는 경우가 많아서 의미역을 제대로 태깅하기가 어렵다.

이러한 격틀사전의 문제점은 실제 작업 말뭉치에 격틀사전을 이용한 반자동 태깅이 좋은 결과를 만드는 데에 걸림돌이 되고 있다[4].

동일한 격틀정보에 의한 문장이지만, 격틀 앞에 오는 체언이 무엇이냐에 따라 의미역 태깅이 달라지므로 이 부분도 기계처리에 어려움이 될 수 있다.

(1) 김치를[대상] 안주로[착점] 하다.
(2) 유럽 통합의 꿈은[기점] 신중한 발걸음을[대상] 필요로[방법] 하다.
(3) {투표를 바탕으로}[방법] 하다.

(1)~(3)의 예문은 술어가 '하다'이고, 격틀이 [~를 ~로 ~하다]로 동일한 문장 구조를 가지고 있다고 볼 수 있다. 그러나 동일한 구조가 아닌 다른 구조로 처리를 해야 하고 의미역도 다르게 태깅해야 하는 예시이다.

(2)의 경우는 '유럽 통합의 꿈을 이루기 위해서는 신중한 발걸음이 필요하다'라는 문장으로 재구해볼 수 있다. 이렇게 재구된 문장에서는 '꿈을 이루기 위해서는'은 목적역이고, '신중한 발걸음이' 대상역이 된다. 그러나 술어 '필요하다'가 '필요로 하다'로 나뉘면서 의미역 태깅에 어려움이 생긴다. '필요로 하다'가 하나로 묶이지 않으면 의미역이 생성되지 못하게 된다. 또는 이를 묶기 위해서 (2)의 결과와 같이 억지스럽게 의미역을 태깅해야 한다. 이 문제는 한국어의 단어에 대한 정의가 어렵다는 점과 하나의 단어 내부에 조사가 개입되고 띄어쓸 수 있다는 특이한 부분이 있기 때문이다. 이를 해결하지 않는다면 자연스러운 의미역 태깅은 쉽게 이루어지기 어려울 것이다.

문장의 확대로 인한 의미역 태깅의 어려움이 있다. 단문을 2개 연결해서 하나의 복문을 만든 경우에 발생하는 문제이다. 대체로 실제 작업 말뭉치의 경우 글의 장르별로 이러한 문제가 많이 나타나는 것과 적게 나타나는 것이 있다. 작업자의 직관에 혼동을 줄 수 있어 문제가 되기도 했다. 게다가 이런 경우 앞 문장과 뒤 문장의 접속 후 비문이 된 경우가 많이 발생하였다. 시제가 맞지 않다든지, 아니면 문장의 구조가 맞지 않다든지, 아니면 앞뒤 문장의 생략된 성분으로 인해 의미 파악이 어려운 등의 다양한 문제가 발생했다.

작업 말뭉치의 문장 중에 복문 중 의미역 태깅의 중의성이 발생하는 다음 문장을 살펴보자.

'이러한 주장은 초기의 데카르트와 그의 친구이자 신부였던 메르센트가 잘 보여주는데 여기서는 메르센느를 중심으로 살펴보기로 한다.' <BSH00127.txt 623번 문장>

위 예문은 복문으로 앞 문장의 서술어는 '보여주는데'이고 뒤 문장의 서술어는 '살펴보기로 한다.'이다. 여기에서 문제가 되는 부분은 뒤 문장이므로 뒤 문장의 서술어 '살펴보기로 한다.'를 중심으로 기술하겠다. '살펴보기로 한다.'는 '살펴보다'와 '하다' 두 개의 서술어가 추출된다.

(4-1) 여기서는[처소] 메르센느를[대상] 중심으로[방법] 살펴보다.
(4-2) 여기서는[처소] 메르센느를 중심으로[방법] 살펴보다.
(5) 여기서는[처소] 살펴보기로[내용] 한다.

위의 '살펴보다'의 논항을 두 가지 관점에서 구분해보았다. 즉, 문장 자체만으로 의미역을 부여했을 때와 전체 문장의 의미를 기준으로 의미역을 부여했을 때로 구분해보았다. 선행문이 없었다면 (4-1)의 의미역이 맞으니 선행문에서의 의미를 반영한다면 (4-2)의 의미역이 맞다. 동일한 문형을 가지고 있지만, 단문일 때와 선행문일 때의 의미역이 달라지는 예이다.

3.3. 해결방안

격틀사전에 탑재한 의미역의 재검토와 수정보완이 필요하다. 격틀사전을 구축하기 위해 대상으로 삼은 표준국어대사전의 용언 중 전체를 대상으로 격틀사전이 마련되어 있으니 2차로 실제 말뭉치에 출현한 용언 대상으로 새로운 격틀사전이 마련되어야 한다.

사용되고 있는 문장을 용례에 추가하고 이를 대상으로 의미역의 수정 보완을 시행해야 한다. 또한 앞에서 언급한 바와 같이 실제 작업말뭉치에 있는데 격틀사전에 없었던 용언을 격틀사전에 추가하는 것이 필요하다. 이렇게 실제 작업 말뭉치에 맞춰진 격틀사전을 마련하고, 여기에 계속해서 보완해 나가야 한다. 3.1.에서 언급했듯이 실제 단어의 다의정보 수준까지의 세분화는 상당히 좋은 접근이었다. 그러나 의미역태깅 프로그램에서 형태소분석단위로 제시하고 있는 것은 동형이의 수준까지이

4) 김완수(2015)는 의미역반자동태깅프로그램에서 격틀사전을 활용한 의미역 태깅 정확률은 72.3% 정도임을 제시하였다.

다. 이들의 차이는 기계처리에 오히려 어려움이 생긴다는 것이다. 형태소분석단위의 다의수준까지의 작업이 필요하다. 만약 다의수준으로의 형태소 분석이 어렵다면, 격틀사전의 의미역 정보를 실 사용례를 기준에 맞춰 의미역 태깅 정보를 제공해야 한다.

표준국어대사전의 격틀 정보의 미흡함을 해결한다면 이는 본고에서 구축한 격틀사전의 의미역정보도 같이 해결할 수 있을 것이다. 그 다음으로 교체정보보어구문의 문제점을 해결하는 것이다. 이 부분에 대해서는 격에 대해 논의한 격교체에 대한 논의[5]에서 하나의 용언이 여러 통사 구조를 가지면서 논항의 교체 현상을 보여주었다[11]. 이러한 논의를 바탕으로 대상 용언의 격의 다양성을 격틀에 부여하는 것을 검토할 것이다.

3.2에서 <그림1>을 통해 드러나 합성어의 일관성 없는 출현에 대해서는 한영균(2003)에서는 작업 전 말뭉치에 가공이 필요하다고 하였다. 그러나 그 또한 어려움이 있다고 언급한 것처럼 이를 말뭉치에 가공처리하지 않은 상태에서 모든 작업을 실시하였으므로, 반대로 작업한 결과물을 대상으로 띄어쓴 합성어를 하나로 묶어 찾은 결과물을 합성어와 동일하게 처리할 수 있도록 해야 할 것이다.

사전에는 미등재어인데, 사전에서 접사로 인정한 '-하다, -되다, -받다, -당하다' 류는 하나의 단어로 실제 작업 말뭉치에 실현된다. 이러한 류의 단어들은 말뭉치 대상으로 추출된 결과물을 모아 격틀사전에 추가하는 것을 고려해야 한다.

의미역의 중의성이 발생하는 경우에는 격틀사전에 해당 동사의 의미역이 단문일 경우와 복문일 경우를 모두 주석하여 작업의 기반 사전으로써 역할을 할 수 있도록 해야 한다.

그렇게 함으로써 의미역 태깅을 하는 작업자에게도 반자동 태깅에도 좋은 길잡이라 될 수 있도록 수정보완을 해야 한다.

4. 결 론

본 논문에서는 상정된 의미역을 실제 문장에 태깅할 때 발생하는 문제점을 논하였다. 태깅에 가장 큰 어려움은 표준국어대사전을 기반으로 만든 격틀사전의 기능이 그리 원활하지 못했고, 실제 문장 또한 규범 문법에 맞게만 나타나지 않아서임을 알 수 있었다. 이에 격틀사전을 수정보완하기 위한 방안과 띄어쓰기에 의한 문제점과 구조적으로 확장된 문장을 처리하기 위한 방안을 제시해 보았다.

5) 남승호(2008:48)는 의미 구조의 핵심인 사건구조와 논항구조 사이의 유기적 관계가 표면의 격틀을 포함한 통사 구조를 실현시키는 원리를 제공하고 있다고 했다. 이러한 의미-통사 구조의 상관성을 찾기 위해, 한국어 술어의 논항 교체 현상을 검토하고, 통사 구조의 차이에 따른 의미 차이를 기술하였다.

감사의 글

"이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (R0101-15-0176)"

참고문헌

[1] 김윤정, 김완수, 옥철영, "전산언어학에서의 한국어 필수논항의 의미역 상정과 재고", 언어와 정보 제18권 제2호, pp. 169-199, 2014.

[2] 김완수, 옥철영, "한국어 격틀 사전과 의미역 빈도 정보를 사용한 한국어 의미역 결정", 한국정보과학회 학술발표논문집, 한국정보과학회, pp. 651-653, 2015.

[3] 정호성, "『표준국어대사전』 수록 정보의 통계적 분석", 새국어생활 제10권 제1호, 국립국어연구원, pp. 55~72, 2000.

[4] 한영균, "《표준국어대사전》의 용례에 대한 사전학적 검토", 국어학 제48권, pp. 289~312, 2006.

[5] 조재수, "문제점이 많은 표준국어대사전", 새국어생활 제10권 제1호, 국립국어연구원, pp. 133~149, 2000.

[6] 유현경, "한국어대사전 편찬에 대한 새로운 제안 - 표준국어대사전에 대한 평가를 기반으로", 한국사전학 제15권, pp. 220~246, 2000.

[7] 이은령, 윤애선, "표준국어대사전의 동사 정보 개선을 위한 연구-한국어 어휘의미망의 구축에서 나타난 문제점을 중심으로-", 한민족어문학 제31권, pp. 157-194, 2010.

[8] 고석주, "사전의 문법 정보에 대하여", 언어사실과 관점 제12권 제13호, pp. 181~216, 2000.

[9] 한영균, "어휘 계량적 분석과 띄어쓰기 문제", 한국문화 제31권, pp.49~71, 2003.

[10] 한영균, "한국어 어휘 교육·학습 자료 개발을 위한 계량적 분석의 한 방향", 어문학 제94권, pp.119~146, 2006.

[11] 남승호, "한국어 술어의 사건 구조와 논항 구조", 서울:서울대학교출판부, 2008.

[12] 윤준태, 정의석, 송만석, "명사간 어휘 정보를 이용한 한국어 복합 명사 분석",정보과학회논문지(B) 제25권 제11호, pp. 1716~1725, 1998.

[13] 고려대학교 민족문화연구원, 『고려대한국어대사전』, 고려대 출판부, 2009.

참고 사이트

[1] 국립국어원, "표준국어대사전"
<http://stdweb2.korean.go.kr/search/View.jsp>

[2] 연세대학교 언어정보연구원, "연세 현대 한국어사전"
<https://ilis.yonsei.ac.kr/dic>