의미주석말뭉치와 전자사전의 의미기술정보

서상규 김한샘
연세대학교 국어국문학과, 언어정보개발연구원
{misorado, star}@sentence.yonsei.ac.kr

Sense tagged Corpus and Definition Information in MRD

Sang-Kyu Seo Han-Saem Kim
Dept. of Korean Language & Literature
/ Center for Linguistic Informatics Development, Yonsei University

요 약

의미주석말뭉치는, 문맥에 출현하는 각 어휘의 의미를 특정 사전의 세부의미항목(sense)에 대응시켜 주석함으로써 구축한 말뭉치이다. 이 말뭉치 구축에 있어서의 태그셋은, '연세 한국어 전자사전'의 각 의미기술정보를 기호화하여 사용하였다. 사람에 의한 실제 주석 작업 단계에서, 전자사전 정보의 불완전함 때문에 발생한 문제를 해결함으로써 본래의 사전 정보가 대폭 수정되었다. 즉, 의미 주석 과정에서 문제가 되는 요소에 대한 검토를 통해서 품사 정보, 문법 정보 등을 수정하고 기존 sense를 통합, 추가, 재배열함으로써 기존의 사전 정보를 개선할 수 있었다. 이와 같은 말뭉치와 전자사전, 자연언어처리 시스템의 활발한 상호 작용을 통해서 언어정보처리 분야 연구의 질적 향상이 가능하다. 나아가, 인간이 직접 판단하여 주석한 대규모의 의미주석말뭉치를 분석하여 응용함으로써 텍스트내의 단어와 전자사전의 세부의미항목을 연결시키는 태거를 개발할 수 있을 것이다.

1. 서론

NLP 시스템을 개발하는 데에 있어, 다른 시스템과 변별되고 이전의 시스템보다 향상되었음을 보일 수 있는 결정적인 요소는 정보 처리의 핵심이 되는 알고리즘이라고 할 수 있을 것이다. 그러나 언어 정보 처리에 있어 이에 못지 않게 중요한 요소로 대두되는 것이 대량의 정밀한 언어 자료의 확보이다. 그러나 신뢰할 수 있는 언어 자료를 대량으로 확보하는 것은 쉽지 않다. 언어학의 분야에서는 언어학적 이론을 발전시키는 데에 초점이 맞추어져 있어 주로 예외적인 언어 사실에 관심을 가지며, 전산학의 분야에서는 다량의 자료를 확보하는 데에 중점을 둔 나머지 자료의 정교함이 결여되기 쉽기 때문이다. 최근 들어서야 이러한 문제를 해결할 수 있는 학문 분야가 국내에 싹트기 시작했다. 바로 말뭉치언어학(corpus linguistics)과 전자사전편찬학(computational lexicography)이다. 특정한 이론에 치우치지 않고 언어 전체의 양상을 드러내는 표본을 추출하여 전산화한 자료인 말뭉치와 언어학적인 연구 결과를 집대성한 자료를 구조적으로 저장하여 2차적으로 응용할 수 있게 한 전자사전은 자연언어처리 분야의 근간을 이루는 중요한 요소라 하겠다.

이 논문에서는 정교한 언어 자료의 구축과 재활용이라는 측면에 초점을 맞추어 의미주석말뭉치와 전자사전 정보의 활용에 대해 논의해 보고자 한다.

2. 말뭉치, 전자사전, NLP 시스템

말뭉치와 전자사전은 자연언어처리에 있어 가장 기본이 되는 언어 자료이다. 이 두 가지의 언어 자료와 실제 응용 시스템의 세 요소는 서로 상호보완적인 관계에 있다. 이 세 요소의 균형이 잘 이루어질 때 비로소 효율적인 언어 정보 처리가 가능한 것이다. Ooi 외(1999)에서 지적한 바와 같이 각 요소들의 재활용성(re-usability)은 언어정보처리 분야의 연구가 심화되면 될수록 중요한 개념이다. 각각의 정보가 서로 어떻게 재활용되는지를 살펴보면 왜 이 분야가 학제적 연구를 필요로 하는지를 짐작할 수 있다.

MRD는 NLP 시스템이 입력 자료를 분석하여 출력 결과물을 생성하는 과정에 있어 기준이 되는 데이터이다. MRD 데이터의 규모와 정확성이 NLP 시스템의 성능을 좌우할 수 있다. 따라서 NLP 시스템이 오류를 포함하는 결과를 생성하는 경우에는 입출력데이터를 분석하여 MRD의 결함을 발견하고 수정, 보완할 수 있다.

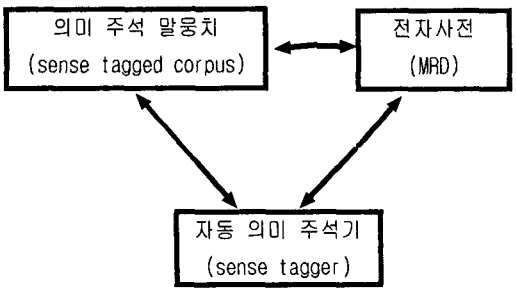
개발된 NLP 시스템의 성능을 실험하는 데에 있어서 말뭉치는 필수적이다. 기계 번역기의 예를 들자면, 여러 개의 말뭉치 표본을 대상으로 자동 번역을 수행해 보아야 어떤 오류가 발생하는지를 포착해 낼 수 있고, 반복되는 실험과 수정을 거쳐 상용 가능한 시스템을 만들어 낼 수 있다. 시스템의 평균적인 성공률도 말뭉치를 대상으로 한 테스트를 통해서 제시될 수 있다. NLP 시스템 중에는 말뭉치의 구축을 돕는 도구도 있다. 수많은 웹문서 중에서 필요한 목적에 따라 말뭉치를 자동적으로 수집하고 구조적으로 저장하는 시스템이나, 기존 말뭉치를 한 단계 높은 수준의 주석 말뭉치로 자동 변환해 주는 태거(tagger)도 있다.

현대적 개념의 사전 컨텐트를 구축하는 데에 있어서도 말뭉치는 필수적이다. 사전 구조를 형성하는 정보들을 기술하는 데에 말뭉치가 적극적으로 활용된다. 사전의 엔트리를 확정하고, 의미를 정의하고, 문법적인 정보를 부여하며, 어휘 간의 관계를 규명하고, 예문을 고르는 등 사전 정보 구축의 전반에 걸쳐 말뭉치가 활용되게 되는 것이다.

전통적인 사전 편찬에서는 말뭉치 없이 편찬자의 직관에 의존해 이런 정보들을 생성했지만, 말뭉치가 활용됨으로써 그 과정이 더욱 효율적이고 체계적으로 이루어질 수 있게 된 것이다.

반면 말뭉치가 없이는 얻을 수 없는 정보도 있다. 각 어휘 혹은 어휘의 세부 항목이 얼마나 자주 쓰이는가를 반영하는 빈도 정보를 제시하기 위해서는 일정 규모 이상의 말뭉치가 반드시 필요하다. 기존의 사전을 검정하고 보완하는 데에도 말뭉치가 활용되는 것은 물론이다. 역으로 사전 정보를 이용해 말뭉치를 구축하는 경우도 있다. 주로 이 때의 말뭉치는 주석 말뭉치(tagged corpus)를 의미한다. 사전의 문법 정보, 문형 정보, 의미 정보 등을 각각 이용하여 품사주석 말뭉치(POS tagged corpus), 구문 분석 말뭉치(parsed corpus /tree bank), 의미 주석 말뭉치(semantic annotated corpus) 등을 구축한다.

이 논문에서 다룰 의미주석 말뭉치와 전자사전, 그리고 이와 관련된 NLP 시스템의 관계를 도식화하면 다음과 같다.



< 그림 1. corpus, MRD, tagger의 관계도 >

의미주석말뭉치(sense tagged corpus)는 전자사전에 등록된 각 어휘의 세부 의미 항목을 태그셋으로 한 주석 작업을 통해서 구축된다. 이 과정에서 전자사전의 오류, 잉여적인 정보가 발견되어 기존 전자사전의 정보를 개선할 수 있다. 또한 의미주석말뭉치의 분석을 통해서 전자사전에 세부의미항목 및 어휘의 빈도 정보가 추가된다. 사람이 직접 주석하여 구축할 수 있는 말뭉치는 규모가 제한될 수밖에 없지만, 이미 구축된 의미주석말뭉치의 통계적 처리 결과를 기반으로 하여 자동 의미 주석기 개발을 모색할 수 있으며, 이를 통해 다시 대규모의 의미주석말뭉치를 구축할 수 있다. 자동의미주석기는 전자사전의 의미기술정보를 기준 데이터 및 태그로 사용한다. 3장, 4장에서는 의미주석말뭉치와 전자사전 정보의 상호 재활용에 대해 기술한다.

3. 전자사전 정보를 활용한 의미주석말뭉치 구축

3.1 의미주석말뭉치의 개념

의미주석말뭉치는 의미범주의 설정과 관련하여 유형을 분류할 수 있다. 의미 주석 말뭉치를 만드는 데에 있어 기준이 되는 의미 범주 설정은 결국 의미 태그 세트를 설정하는 것이다. 의미 주석 말뭉치는 어휘 형태에 대응하는 의미 태그의 내용이 어휘의 의미를 세분화하는 것인가, 아니면 어휘의 의미 전체를 포괄하여 다른 어휘들과의 관련성을 파악할 수 있는 것인가에 따라, 의미장에 따른 의미 주석(semantic tagging)을 통한 말뭉치와 어휘별 의미 범주를 기준으로 주석한 말뭉치(sense tagged corpus)로 구별할 수 있다.

전자는 어휘를 의미적 단일체를 구성하는 것끼리 분류하는 의미장 이론에 근거하여 의미 주석을 다는 말뭉치이다. 어휘마다 태그셋을 따로 설정할 필요가 없고 의미 범주의 영역이 넓기 때문에 태그셋만 확정이 되면 작업의 일관성을 유지하기 용이하다. 그러나 수많은 어휘를 포괄하는 의미장의 계열관계 모델을 설정하는 것이 매우 어려운 작업이다. 또한 실제 작업에 있어 혼란을 주지 않도록 각 의미장 즉 태그의 내용이 중복되지 않도록 조정하는 것도 쉽지 않다.

후자는 사전에서 동형어나 다의어로 구분하는 어휘 형태의 세분화된 의미 그 자체의 기호를 태그로 부착하여 구축한 말뭉치이다. 이 방식을 채택할 경우 태그셋으로 활용되는 사전의 어휘 기술 및 의미 분류의 완성도가 주석 말뭉치의 질과 직접적으로 연관된다. 이러한 방식으로 주석 말뭉치를 구축하는 경우에 자세하게 의미가 분석된 자료를 얻을 수 있는 반면, 일정 수준 이상의 언어학적 지식을 필요로 하기 때문에 초기 단계에서는 자동 주석이 거의 불가능하다. 그러나 주석 과정이 바로 어휘 의미 분류를 검토하는 과정이 되므로 사전의 의미 기술과 의미주석 작업 간의 피드백을 통해서 사전과 말뭉치의 질을 동시에 향상시키는 효과를 기대할 수 있다.

PPHS1	She	Z8
VVD	laughed	E4.1+
RR	disagreeably	04.2-
,	,	
VVG	squashing	A1.1.1
APPGE	her	Z8
NN1	cigarette	F3
#	in	Z5
AT	the	Z5
NN1	butter	F1
.	.	

< 그림 2. Semantic tagging에 의한 말뭉치의 예 >

짧은	짧	VJ	-②
시간이었지만	시간	NN	-Ⅰ①
그	그	AN	2-②
여자를	여자	NN	-x1
만난다는	만나	VV	-Ⅰ②
건	것	NX	-Ⅱ⑤
쉬운	쉽	VJ	-Ⅰ①
일이	일	NX	1-④
아니었다	아니	VJ	-①

< 그림 3. Sense tagging에 의한 말뭉치의 예 >

그림 2와 3을 보면 Semantic Tagging에 의한 말뭉치(그림 2)와 Sense Tagging에 의한 말뭉치(그림3)가 크게 달라 보이지 않는다. 대상이 되는 어휘에 그 품사 태그, 의미 태그를 단계적으로 부여한 구조가 동일하기 때문이다. 그러나 2와 3에 출현하는 의미 태그의 내용은 매우 다르다. 2의 태그들이 [HUMAN], [ANIMATE] 등과 같은 의미 범주를 나타내는 반면 3의 태그들은 '어떤 시간에서 다른 시각까지의 동안, 또는 길이', '(한때로부터 다음의 한 때까지의) 동안이 길지 않다.' 등과 같은 사전의 의미기술정보를 대신하는 것이다.

이 논문에서 의미주석말뭉치는 말뭉치와 어휘별 의미 범주를 기준으로 주석한 말뭉치(sense tagged corpus)를 가리킨다.

3.2 전자사전의 의미기술정보를 이용한 의미주석

이 논문에서 소개하는 의미주석 말뭉치는 연세대 언어정보개발연구원에서 구축한 YSTC(Yonsei sense tagged Corpus)이며, 이에 기반이 된 전자사전은 같은 기관에서 편찬한 '연세 한국어사전'의 전자화된 데이터이다.

의미주석 말뭉치의 구축은 본격적으로 의미 주석에 들어가기 전에 몇 가지 전단계를 거치게 된다. 다음은 YSTC를 구축한 과정의 개요이다.

1) 말뭉치 구성

YSTC는 문화관광부 주관 '한국어 세계화 추진을 위한 기반 구축' 프로젝트(책임: 김하수)의 일환으로 구축되기 시작했다. '한국어 교육'이라는 특수한 목표가 있었기 때문에, '연세 의미 주석 말뭉치'는 일반적 균형 말뭉치의 특성을 유지하면서도, 외국인을 위한 한국어 교육에 필요한 기초 어휘들이 많이 포함될 수 있는 텍스트들을 중심으로, 총 100만 어절의 말뭉치를 구축하였다.

2) 품사 주석 말뭉치 구축

의미주석 말뭉치를 구축하기 위한 전단계로서 품사 정보를 부착한 말뭉치의 구축은 필수적이다. 의미 주석의 대상이 되는 어휘를 기준으로 텍스트의 어절을 분할하고, 분할된 단위에 대해 기본적인 품사 표지를 붙임으로써 의미 주석을 위한 기초 작업을 수행했다.

품사 주석 작업에서는, 먼저 자동 태거(HAM_KTS)를 이용해 자동적으로 태그를 부착했고, 자동 태거의 오류를 중심으로 사람이 자동 주석 결과를 수정, 보완하는 두 단계에 걸쳐 수행되었다.

3) 의미 주석 대상 어휘의 선정

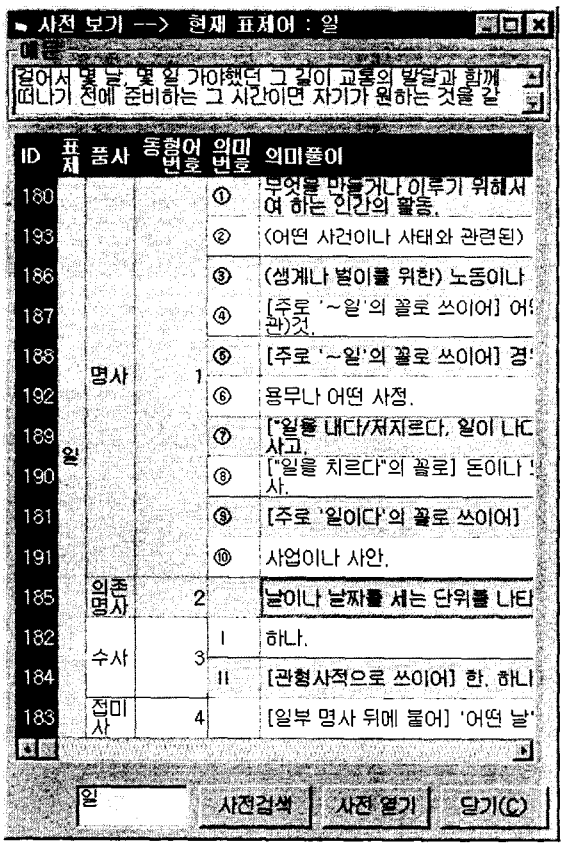
의미주석은 많은 노력과 시간을 작업이기 때문에 말뭉치에 출현하는 모든 어휘에 대해 한꺼번에 주석을 달기는 어렵다. 따라서 주석의 결과물이 효과적으로 이용될 수 있는 기본적인 어휘를 선정하여 1차적 주석 대상으로 삼았다. 대상 어휘 목록은, "현대 한국어 기본 어휘 후보 목록-일반어휘편-," 서상규(1998)'를 기반으로 하고, 1998년도의 "한국어 교육을 위한 기초 어휘 선정"(1998)의 연구 결과를 보완함으로써 작성되었다. 의미 태깅 작업이 진행됨에 따라, 처리 어휘 목록에 추가와 삭제가 계속되어 끊임없이 어휘 목록이 갱신되었고 최종적으로 1,000개의 주석 대상 어휘가 확정되었다. 이 목록은 향후 3,000개까지 확충할 예정이다.

4) 주석 보조 시스템의 개발

단순한 수동 주석 방식만으로는 100만 어절의 말뭉치에 의미 태그를 다는 것이 비효율적이라는 판단하에, 사용자의 편의를 위한 주석 보조 도구를 개발했다. '참뜻'이라는 보조 시스템은 의미 태그의 내용이 되는 사전의 효과적인 검색, 태그의 자동 부착, 어휘별 주석을 가능하게 함으로써 주석 작업을 반자동화하였다.

위에서 설명한 말뭉치 구성, 품사 주석 말뭉치 구축, 의미 주석 대상 어휘의 선정, 주석 보조 시스템의 개발의 과정을 거치고 나서야 본격적인 의미주석 작업이 가능하게 된다.

주석 작업에 있어 가장 중요한 요소 중 하나가 태그셋의 설정이다. YSTC는 어휘별 의미 범주를 기준으로 주석한 말뭉치이므로, '연세한국어사전'의 의미기술 정보를 태그셋으로 사용했다. 의미 주석 작업에 있어 전자사전에 구조적으로 저장된 모든 데이터가 필요한 것은 아니기 때문에 표제어, 품사, 동형어번호, 의미항목번호, 의미 기술 정보, 예문의 여섯 가지 정보만이 표시되도록 데이터베이스를 구성한 다음 의미주석에 이용했다. 이 중, 동형어번호와 의미항목번호를 주석의 태그 표지로 정했고, 의미풀이(의미 기술 정보)는 태그 표지가 가리키는 실제 내용으로, 주석자가 말뭉치상의 주석 대상 어휘를 분석하는 기준이 되었다. 예문 역시 주석자의 판단에 중요한 역할을 한다. 의미풀이보다는 훨씬 직관적인 이해가 쉽기 때문이다. 다만 주석 보조 프로그램 상에서는, 작업의 편의를 위해서 예문은 의미 항목을 클릭하면 윗부분의 창에 표시되도록 조정하였다.



<그림 4. YSTC 구축에 사용된 '참뜻'의 전자사전 창 >

의미주석 말뭉치 구축에 활용된 전자사전 DB는 그림 4의 샘플과 같으며, 의미 주석 작업을 통해 구축된 말뭉치의 결과물은 다음 <그림5>와 같다.

No	어절 번호	본문 어절	순서	분석	품사	의미주석
11174	6875	@	1			
11175	6876	의상은	1	의상	NNX	
11176	6876	의상은	2	은	PA	
11177	6877	아무	1	아무	AN	2-
11178	6878	일도	1	일	NN	1-②
11179	6878	일도	2	도	PA	
11180	6879	없이	1	없이	AV	-표⑧
11182	6880	잠을	1	잠	NN	-①
11183	6880	잠을	2	을	PA	
11184	6881	잤는데	1	자	VV	- I ①
11185	6881	잤는데	2	ㅏ쓰는데	EF	
11186	6882	자기만	1	자기	NP	3-①
11187	6882	자기만	2	만	PA	
11188	6883	목이	1	목	NN	1-①x16
11189	6883	목이	2	이	PA	
11190	6884	타는	1	타	VV	1- I ④
11191	6884	타는	2	는	EF	
11192	6885	듯하여	1	듯하	VX	-
11193	6885	듯하여	2	여	EF	
11194	6886	깬	1	깨	VV	1- I ①
11195	6886	깬	2	ㄴ	EF	
11196	6887	것도	1	것	NX	
11197	6887	것도	2	도	PA	
11198	6888	이상하였다	1	이상하	VJ	-②
11199	6888	이상하였다	2	였다	EF	
11200	6889	.	1	.	SP	
11201	6890	@	1			

<그림 5. YSTC의 일부 >

실제로 YSTC를 구축하는 과정에서 어휘별 의미 범주를 기준으로 주석하는 말뭉치의 문제점이 드러났다. 어휘의미체계에 대한 연구를 통해 의미 범주 태그를 새로 개발해야 하는 부담을 줄이기 위해, 이미 존재하는 전자 사전의 정보를 태그셋으로 재활용했지만 기존 사전의 의미 분류가 하나의 체계적인 태그셋으로 활용하기에는 미흡하다는 것이다. 또, 전자사전의 항목별 의미가 어휘별로 자세하게 기술되었고 용례까지 보조 기제로 사용할 수 있음에도 불구하고 역시 주석자들 간의 개인차를 극복하기는 어려웠다는 것이다.

품사 태깅 단계에 적용된 품사 분류 체계와 전자사전의 품사 분류 체계가 일치하지 않았던 것도 작업자들의 혼란을 초래했다. 이는 처음에 각기 다른 목적으로 구축, 개발되었던 말뭉치, 전자사전, 태거가 함께 활용됨으로써 생긴 문제이다. 2장에서 지적한 바대로 이 세 요소의 상호 작용에 대한 이해와 효율을 최대로 높이기 위한 최적화 작업이 필요했던 것이다. 다음 장에서는 기준으로 적용되던 전자사전의 의미기술 정보의 오류가 의미주석말뭉치 구축 작업을 통해서 어떻게 발견되고 수정되었는지에 대해 살펴보기로 한다.

4. 의미주석말뭉치 구축을 통한 전자사전 개선

의미 주석 작업은 말뭉치에 출현한 용례의 문맥을 바탕으로 해당 어휘의 의미가 세분화되어 있는 의미 항목 중 어느 것과 관련이 있는지를 결정하는 것이다.

그러나 연세 의미주석말뭉치에서 의미 분류의 기준으로 삼고 있는 '연세 한국어사전'은 집필 당시 의미 주석 말뭉치의 의미 분류 기준으로 쓰일 것을 염두에 두지 않았고, 풍부한 용례를 중심으로 한 세밀한 의미의 기술에 신경을 썼기 때문에 실제 작업에 있어 많은 어려움이 발생했다.

한 어휘가 가지는 의미의 영역을 전체로 보았을 때, 세밀하게 의미를 분류한 것이 넓은 의미의 몇 가지의 항목으로 기술한 것보다 오히려 표현하는 의미 영역이 좁을 수 있기 때문이다. 특히 '가다, 나다' 등 원형 의미에서 파생된 확장 의미가 다양한 기본 어휘들의 경우에는 이런 정교한 의미 분류가 의미 주석에 있어 방해 요소로 작용할 수 있다.

의미주석 작업 상의 어려움을 해결하기 위해서 관리자가 주석자들로부터 사전의 오류를 보고 받아 전자사전의 정보를 수정, 보완하는 절차를 거쳤다.

이 과정을 통해서 전자사전의 업그레이드가 이루어졌다. 국내의 자연언어처리 분야에서 사용되고 있는 전자사전의 대부분이 일반 언어사전을 전자화한 것임을 고려할 때 이런 피드백 과정을 필수적이라 하겠다. 기존의 일반 언어사전이 편찬자의 직관에 많은 부분 기대고 있어 정보의 체계성이 떨어지기 때문이다. 각 소절에서 전자사전 정보를 업그레이드한 실례를 유형별로 논의하기로 한다.

4.1 품사 정보의 수정

의미의 기술은 제대로 되어 있는데 품사 정보만 틀린 것을 발견하는 것은 실제로는 극히 드문 일이다. 말뭉치의 어휘와 전자사전의 의미 항목을 대응시키는 데에 집중하다 보면 간과하기 쉽기 때문이다. 그러나 문맥이나 활용의 제약 등을 분석해 보면 오류를 발견할 수 있다.

지나치다 형
2 ① (어떠한 것이) 일정한 한도를 넘어서 있다.
¶ 당신은 술이 너무 지나쳐요./부모가 설령 도에 지나친 요구를 한다 하더라도 최대한 그것을 들어주는 것이 자식된 도리다.
② ('지나치게'의 꼴로만 쓰이어) 알맞지 않을 만큼 심하게. 너무나.
¶ 생나무라 단단하기는 할 테지만 지나치게 가늘어 보였다./영식이는 지나치게 뛰어다닌다고 주의를 들었다.

위의 용례를 보면 '지나치다'는 분명히 형용사의 기능을 하고 있다. 기존의 사전 정보에서는 동사의 의미 항목 중 일부로 포함시킨 오류를 수정하여 분리된 표제어로 등재했다

4.2 의미항목(sense)의 추가

사전정보를 경신하는 데에 있어 가장 많은 비중을 차지하는 것이 바로 의미 항목의 추가이다. '연세 한국어사전'이 말뭉치를 기반으로 개발되기는 했지만 말뭉치가 모든 언어 현상을 반영할 수 없을 뿐더러, 편찬자의 직관이 개입되었기 때문에 텍스트에서 출현한 수많은 어휘의 쓰임을 모두 기술하지 못했다. 따라서 의미주석 작업을 통해서 발견되는 새로운 의미들을 추가하는 것은 매우 중요한 작업이며 또한 어려운 작업이라 하겠다. 기존의 태그, 즉 사전에 기술된 어휘 항목에 해당하지 않는 어휘의 쓰임이 출현했을 때 이를 어떻게 처리할 것인가는 작업자들에게 큰 고민이 아닐 수 없다.

가끔 초보 작업자들이 해 오는 질문 중의 하나가 '가. 필름, 나. 사진, 다. 사진기, …' 등과 같은 항목의 나열에서 볼 수 있는 '가', '나', '다'와 같은 것을 어떻게 처리할 것인가 하는 문제이다. 물론 이런 경우에는 따로 의미 항목을 추가하면 안 된다. 이 때의 '가', '나', '다' 등은 의미나 내용물을 갖추지 못한 기호일 뿐이기 때문이다.

의미 주석이 불가능한 또 하나의 케이스는 해당 문맥에서만 일회적으로 쓰인 용법인 경우이다. 개인이 발화상에서 언어 유희를 구사하거나, 어휘의 의미를 정확히 모르고 사용하거나, 혹은 비유적으로 어휘를 사용한 경우에 사전의 의미 기술 정보에서 적당한 항목을 찾을 수 없다. 이 경우에도 역시 이런 일회적인 쓰임을 위해서 따로 의미 항목을 설정하는 것은 비합리적이다.

위에 언급한 두 가지 경우와 주석자의 분석 오류를 제외하고 나면 실지로 의미 항목을 추가하는 경우는 그리 많지 않다. 그러나 의미 항목을 추가할 때에는 기존의 어휘 분류 체계와의 균형을 고려하여 의미를 기술해야 하므로 아예 새로 의미를 분류하는 것보다 더 힘든 경우가 많다. 다음은 형용사 '깊다'의 누락된 의미를 보충한 예이다.

깊다 형
① 수면에서 밑바닥까지의 거리가 멀거나 길다. ¶ 물이 깊어야 고기가 모인다./장쇠는 장화를 산중의 깊은 연못에 빠뜨렸다.
② (겉이나 밖에서) 거리가 멀다. ¶ 산이 깊고 길이 좁아 걸음이 빠르지 못하였다./오진암은 계곡도 깊고, 본사에 못지 않게 경내도 넓었다.
③ (주위보다 바닥이) 상대적으로 낮거나 패여 있다. ¶ 눈밑의 잔주름도 깊어 보였다.
④ (생각이나 마음이) 듬직하고 신중하다. ¶ 그는 이 방문이 선생님의 깊은 생각으로 이루어진 것임을 느끼게 되었다./우리는 그의 신실하고 깊은 마음에 감동을 느꼈다.
⑤ (정도나 수준이) 매우 높고 대단하다. ¶ 부친의 사랑에 출입하는 이들 중에는 그런 방면에 조예가 깊은 사람들이 많았다./자세히 들어 보면 그렇게 깊고 높은 지식 얘기를 하는 것도 아니었다.

⑥ [동사적으로 쓰이어] 시간의 흐름을 따라 상태가 더욱 심하다. ¶ 일행은 저녁을 먹고 밤이 깊어서야 떠나는 기차에 다시 올랐다./늦가을이 깊어 과수원도 다른 때보다 훨씬 더 썰렁하게 느껴졌다.

⑦ (시간이 오래되어 병의 상태가) 더욱 심하다. ¶ 여윈 몸에 병이 깊으니 신음소리가 절로 난다./일중독증이 깊은 나에게는 자유란 그저 무료함에 지나지 않았다.

⑧ 자욱하거나 짙다. ¶ 산에는 깊은 그늘이 드리워 보랏빛 음영을 띠고 있었다.

⑨ (관계나 관련 등이) 가깝거나 밀접하다. ¶ 난 이미 그와 남남이 될 수 없는 깊은 인연을 느끼고 있어./충치의 발생 빈도와 부모들의 치아 건강에 대한 교육은 관련이 깊다.

⑩ [주로 '깊은' 의 꼴로 쓰이어] (담겨진 내용이나 생각이) 진지하고 풍부하다. ¶ 그 깊은 내용을 제자들은 알 수 없었다./물론 스승으로서는 무슨 깊은 뜻이 있었을 것이다.

⑪ [주로 '깊은, 깊게' 의 꼴로 쓰이어] (쉽게 깨지 않을 정도로, 잠이) 든 상태에 있다. ¶ 요즈음은 열대야 현상이 계속되어 깊은 잠을 자기가 어렵다.

⑪x1 (어떤 행동의) 정도가 심하다. ¶ 깊은 포옹/깊은 입맞춤/사려 깊은 말씀/주의 깊게 살폈다./

⑪x2 (뿌리, 근원 따위가) 깊숙하다. ¶ 뿌리 깊은 전통

무려 11개의 의미 항목으로 기술했음에도 불구하고 굵게 표시한 부분인 ⑪x1, ⑪x2의 의미가 기존 사전 정보에는 빠져 있었다. ⑪x1과 ⑪x2의 분포가 전체 '깊다' 쓰임의 17%에 달하는 것을 고려해 볼 때, 이러한 피드백 과정이 얼마나 중요한 지를 알 수 있다.

추가해야 할 의미 항목들 가운데 눈이 띄는 것은 바로 관용 표현에 대한 정보이다. 관용 표현은 둘 이상의 어휘가 모여 한 단위를 이루되 의미의 확장 또는 변이를 동반하므로, 관용 표현의 구성 요소인 어휘가 주석의 대상이 되었을 때에는 일반적인 어휘의 의미에 대응시키기가 곤란하다. 물론 태그셋으로 삼은 전자사전의 정보에 관용 표현에 대한 항목이 일부 기술되어 있다. 그러나 실제 텍스트에 나타난 어휘의 의미를 분석하는 의미주석 작업의 특성상 기존에 발견해 내지 못했던 관용 표현들이 수없이 목록에 추가되었다.

가슴 명 x1 ['가슴을 쓸어내리다'"의 꼴로] (근심스럽거나 불안하던 느낌이 없어지고) 마음이 놓이다. 안심이 되다. ¶ 나는 '후우'하고 가슴을 쓸어내렸다./전화를 통해 들려오는 '합격'이라는 말에 나는 가슴을 쓸어내렸다.

x2 ['가슴을 울리다'"의 꼴로] 감동시키다. ¶ 여인의 애처로운 사연은 모든 이들의 가슴을 울렸다.

x3 ['가슴을 치다'"의 꼴로] 일이 마음대로 되지 않아 답답해 하거나 억울해 하다. ¶ 가슴을 치고 땅을 치고 싶은 심정이었다.

x4 ['가슴을 태우다'"의 꼴로] 몹시 애태우다. ¶ 기철은 떠나간 그녀가 오기만을 가슴 태우며 기다렸다.

x5 ['가슴이 내려앉다'"의 꼴로] 깜짝 놀라다. 몹시 놀라다. ¶ 가슴이 철렁 내려앉고 등 위로 식은땀이 흐르는 얘기들이다.

x6 ['가슴이 미어터지다'"의 꼴로] 마음이 매우 아프게 느껴지다. ¶ 모든 동지들로부터 진실을 캐는 듯한 눈빛을 받을 때마다 가슴이 미어터질 것 같았다.

x7 ['~ 가슴을 안고'"의 꼴로] 어떠한 감정이나 느낌을 가지고. ¶ 나는 터져 버릴 것 같은 가슴을 안고 인간 답게 살기를 외쳤다./나는 설레는 가슴을 안고 집 안 청소를 하고 있었다.

위의 예는 '가슴'의 의미 항목 중 기존 사전정보에 포함되어 있던 관용 표현과 관련된 항목이다. 비교적 여러 개의 관용 표현이 기술되어 있지만 여기에 '가슴이 뛰다', '가슴이 막히다', '가슴이 철렁하다', '가슴을 펴다' 등 많은 관용 표현이 추가되었다. 신체의 일부를 나타내는 명사들이 주로 기초 어휘이기 때문에 주석 대상 어휘에 많이 포함되어 있는데 이들이 또한 여러 논문에서 지적한 대로 관용 표현을 빈번하게 구성하므로 추가된 관용 표현의 거의 대부분을 이들이 차지했다.

꿩 명 닭과 비슷하나 꼬리가 길고, 몸빛은 불그스름하고 몸에 알락달락한 검은 점이 있는 큰 새. 단 한 마리의 꿩도 잡아 본 적이 없었지만, 얼마나 꿩들을 그들은 뒤쫓았던가?

x1 ["꿩 구워 먹은 소식"] 전혀 소식을 모름. ¶그가 간 지 두 시간이나 됐는데도 꿩 구워 먹은 소식이기 때문에 더욱 마음이 산란했다.

x2 ["꿩 먹고 알 먹다"] 한꺼번에 두 가지 이익을 보다. ¶그는 부잣집 여편네들이 판 벌이고 앉아 있는 곳이라면 꿩 먹고 알 먹기 십상이라고 하였다./잘만 하면 꿩 먹고 알 먹는 결과도 절로 굴러올 수 있어.

x3 ["꿩 대신 닭"] 적당한 것이 없을 때 그와 비슷한 것으로 대신하는 경우를 비유적으로 이르는 말. ¶꿩 대신 닭이라고, 너라도 같이 가자

관용 표현 중에는 위의 x1~x3과 같은 속담도 있다. 속담을 구성하는 어휘들에 대한 정보도 따로 기술해 줘야 함은 물론이다.

괜찮다 형 I① 꽤 좋다. ¶고등어 고기가 맛이 괜찮았다./시험을 떠칠 앞둔 어느 날 형우가 반에서 성적이 괜찮은 몇몇 아이를 모았다.

I② (몸과 마음의 건강 상태가) 별탈이 없다. 정상이다. ¶재귀열은 고열이 오르는 일차 일 주일 정도의 고비를 넘기면 한 일 주일 정도 괜찮다가 다시 열이 오르는 병이었다.

II 지장이 없다. 거리낄 것이 없다. 허락할 수 있다. ¶ 어려운 사람을 돕는 데에 드는 돈이라면 기꺼이 써도 괜찮다./젊은 시절에는, 낭비해도 좋다든지 우습게 보내도 괜찮을 정도로 무가치한 시간이란 단 1분도 없다.

IIx1 부드럽게 거절하는 뜻. ¶제 차 타고 가세요. - 아뇨, 괜찮아요. 걸어 가겠어요.

IIx2 다른 사람의 사과 따위를 받아들이는 뜻. 천만에요. ¶정말 미안합니다. - 괜찮아요.

위의 '괜찮다'에서 Ⅱx1, Ⅱx2 등은 빈도가 높은 구어적인 표현을 사전정보에 반영한 것이다. '글쎄', '마', '아' 등의 감탄사나 '그냥', '그만', '그저' 등의 부사의 사전 정보에도 구어적 용법을 반영하는 의미 항목이 추가되었다.

시간을 나타내는 부사 중에는 오늘, 어제 등과 같이 명사이지만 조사없이 부사처럼 쓰이는 용법을 가지는 어휘들이 있다. 대부분 이런 부사적 용법을 기술하였지만 혹시 빠진 경우에는 이를 보충했다. 아래 '그저께'의 'x1'이 추가된 항목이다.

그저께 명 어제의 전날. 이틀 전. ¶그분이 그저께부터 우리 사슴목장 일을 도와 주시기로 했지요.
x1 [부사적으로 쓰이어] 어제의 전날에. 이틀 전에. ¶졸업식은 그저께 있었다지?

태그셋에서 누락된 어휘의 용법이 기존에 기술된 어휘의 의미기술 내용과 문법적 특성을 달리하여 아예 전자사전의 표제어가 추가되는 예도 있다. 이런 경우에는 동형어를 따로 세면 주석의 대상이 되는 어휘의 수까지 변하게 된다.

가지다 2 보동
① 〔 '-아/어 가지고'의 꼴로 쓰이어〕 어떤 행위를 끝내거나 상태를 유지하여. ¶ 우물에서 물을 퍼 가지고 지붕 위로 끼얹는 사람도 있었다./그는 너무 기뻐 가지고 말이 안 나왔다.
② 〔 '-아/어 가지고'의 꼴로 쓰이어〕 어떤 일의 결과로 인해서, 또는 그 결과를 이용하여. ¶ 대학교를 나와 가지고도 직장을 잡기가 힘든 세상이다./널 대학에 보내 가지고 출세시키는 게 이 엄마의 소원이다.
③ 〔 '-아/어 가지고'의 꼴로 쓰이어〕 [부정적인 의미로] 어떤 행위의 결과나 특정한 상태에 처해 있어서는. ¶ 세상이 이래 가지고 어디 사람 살겠나./저래 가지고는 사람이 될까 싶었다.
④ ['그래 가지고, 이래 가지고'의 꼴로 쓰이어] 말을 단순히 이어가는 뜻. '그래서. 이래서'로 바꿀 수 있음. ¶ 아 근데, 그래 가지고 내가 있으려니까…/옆에서는 빵빵거리고 차는 안 나가고, 이래 가지고, 어쩔 줄 몰랐다.

기존의 사전의 '가지다' 정보에는 위에 기술한 의미 항목이 포함되어 있지 않았다. 그런데 의미 항목을 추가하려고 보니 추가해야 할 '가지다'의 의미는 용언과 결합할 때에만 나타나므로 이를 따로 보조동사로 기술해야 할 필요가 있었던 것이다. 결과적으로 '가지다'는 동사와 보조동사, 두 개의 동형어를 통해 기술되었다.

4.3 의미항목의 분리

드물기는 하지만 기존의 의미 항목을 두 개의 항목으로 분리해야 하는 경우도 있다.

데리다 동 (아랫사람을) 자기와 함께 있게 하다. 자기를 따라오게 하다.
¶ 장쇠는 장화를 외가로 데리고 가는 척 하다가 산중의 깊은 연못에 빠뜨려 죽인다./내 아들이 데몬지 뭔지 하다가 여기로 붙잡혀 온 모양인데 그 놈을 데리러 왔어요./궁지에 몰린 명수와 철수는 결국 동생들을 데리고 고아원에 들어가기로 결심을 했다.

기존 사전에서 '데리다'는 위와 같이 기술되었다. 그러나 위의 사전정보는 몇 가지 문제점을 안고 있다. 우선 '(아랫사람을)'이라는 제약적인 정보는 '자기와 함께 있게 하다.'의 의미일 때만 유효하다. 또, 예문으로 제시된 세 개의 문장은 모두 '자기를 따라오게 하다'라는 의미에만 해당한다.

이런 문제점을 해결하기 위해 의미 항목을 분리하고 의미기술을 일부 수정한 결과는 다음과 같다.

데리다 동 x① (아랫사람을) 자기와 함께 있게 하다. ¶ 판수가 일행을 데리고 떠들고 있다./날더러 너희들을 데리고 살라는구나.
x② 자기를 따라오게 하다. ¶ 장쇠는 장화를 외가로 데리고 가는 척 하다가 산중의 깊은 연못에 빠뜨려 죽인다./내 아들이 데몬지 뭔지 하다가 여기로 붙잡혀 온 모양인데 그 놈을 데리러 왔어요./궁지에 몰린 명수와 철수는 결국 동생들을 데리고 고아원에 들어가기로 결심을 했다.

4.4 의미항목의 수정

사전의 의미 기술이 틀렸거나 불완전할 때에는, 주석 과정에서 의미 항목을 수정하게 된다.

아 감 2
① 가벼운 놀라움이나 당황한 느낌 등을 나타낼 때 쓰는 말. ¶아, 영이는 결혼까지 생각하고 있구나!/아, 어떻게 이런 일이 일어날 수 있단 말인가./아, 벌써 그렇게 되었군요.
② 감동적인 느낌을 나타낼 때 쓰는 말. ¶ 아, 예쁘다!/아 이거였구나!/아, 정말 멋있다!
③ 한탄, 걱정, 근심, 유감 등의 느낌을 나타내는 말. ¶ 아, 인제는 고국에 다시 못 올 것 같은 생각이 듭니다그려./아, 우울하구나!/아 모두 쓸데없는 생각들이다.
④ 모르던 것을 깨닫게 되었을 때 감탄하듯이 소리내는 말. ¶ 아, 그래요 딸이 하나 있지요. 대대장이 한 번 상처했거든요./아, 이제야 생각나요.
⑤ 남에게 말을 걸거나 할 때 남의 주의를 끌려고 말에 앞서 내는 말. ¶ 아, 강 선생 내가 만약 죽지 않고 산다면 말이야./아, 나 여기 있어.
⑥ 말을 처음 시작하거나 이어갈 때 약간 말을 끌면서 내는 말. ¶ 아, 아까운 시간이 다 흘러갔어요./아, 이거?/아, 저 말이에요.

위에 제시한 감탄사 '아'의 사전정보를 보면 ④의 의미와 예문이 일치하지 않는 것을 알 수 있다. 문제가 발견된 후에 사전의 의미 기술 정보는 '문득 생각이 나거나, 모르던 것을 깨닫게 되었을 때 감탄하듯이 소리내는 말.'로 수정되었다.

4.5 의미항목의 통합

의미가 매우 다양한 경우에 각 항목간에 의미가 중첩되는 경우가 있다. 이런 경우에 겹치는 부분에 해당하는 어휘를 주석하기가 곤란해진다. '하다'의 예를 보면 이런 어려움을 이해할 수 있다.

하다 동
I ① (어떤 동작이나 행위를) 행하다. ¶ 너 여기 앉 아서 이것 좀 해라./학교 갔다 오면 숙제부터 하고 놀아야 한다./너는 말을 너무 많이 하는 것 같애.
I ② (앞의 명사가 나타내는 움직임이나 활동 등을) 동작으로 나타내다. ¶ 옷차림이 남루한 사내아이 몇이 구슬치기를 하고 있었다./그는 쓴웃음을 지으면서, 옥상 바닥에 엎드려 팔굽혀펴기를 했다.
I ③ (앞의 명사가 나타내는 일, 행위, 현상, 상태 따위를) 행하다. ¶ 우리 회사는 세계 시장으로 진출을 했다./이렇게 오셔서 축하를 해 주시니 감사합니다.
(중략)

I ①과 I ③을 비교해 보면 공통된 부분이 있다는 것을 알 수 있다. '(행위를) 행하다'의 의미를 가지는 경우에는 I ①과 I ③ 중 어떤 것으로 주석해야 할지 망설이게 될 것이다. '하다'의 의미 항목은 '1이 2를 하다'라는 문형에 한해서만도 21가지에 달할 정도로 많다. 또 의미 주석 과정에서 주석자들의 질문이 가장 많았던 어휘가 바로 '하다'였다. 이런 점을 고려할 때 의미 주석을 위해서는 너무 자세한 분류보다는 통합적인 의미기술 정보가 효과적임을 미루어 짐작할 수 있다.

5. 결론

지금까지 의미주석말뭉치와 전자사전 정보가 서로 어떻게 상호보완적으로 활용될 수 있는 지를 살펴보았다. 전자사전의 의미기술 정보는 의미주석말뭉치의 분석 기준인 태그셋의 역할을 수행하고, 의미주석말뭉치 구축 작업을 통해서 전자사전 정보의 문제점이 발견되고 수정되었다. 이렇게 말뭉치와 전자사전의 정보가 상호보완적으로 꾸준히 재활용된다면 자연어처리, 언어교육 등의 응용 분야에 이들을 효율적으로 활용할 수 있을 것이다. 피드백을 통해서 정제된 의미주석말뭉치와 전자사전 정보를 기반으로 하여 자동으로 의미를 분석하는 의미주석도구(sense tagger)를 개발하는 것이 남겨진 과제라 하겠다.

6. 참고 문헌

[1] A, Harley & G, Glennon, 1997, Sense tagging in Action, Proceeding of Special Interest Group on the Lexicon of ALC, Washington D. C., USA.

[2] Bo Svensen, 1993, Practical Lexicography, Oxford Univ.

[3] CLID ed, 1998, Yonsei Korean Dictionary, Dusan Donga.

[4] Cristopher D, Manning & Hinrich Schutze, 1999, Foundations of Statistical Natural Language Proceessing, The MIT Press.

[5] Geart Van Der Meer, 2000, Core, subsense and the New Oxford Dictionary of English, On how meanings hang together, and not separately, Proceeding of Euralex, Stuttgart, Germany.

[6] Hwee Tou Ng, 1997, Getting serious about word sense disambiguation, Proceeding of Special Interest Group on the Lexicon of ALC, Washington D. C., USA.

[7] J, Sinclair, ed, 1987, Looking Up, Collins.

[8] J, Thomas & M, Scott, eds, 1996, Using Corpora for Language Research, Longman,
Michael West, ed, 1967, A General Service List of English Words, Longmans.

[9] Vincent B. Y. Ooi, 1998, Computer, Corpus, Lexicography, Edinburgh Univ. Press

[10] Y, Wilks & M, Stevenson, 1997, Sense tagging : semantic tagging with a lexicon, Proceeding of Special Interest Group on the Lexicon of ALC, Washington D. C., USA.

[11] 고석주 외 (1999), "한국어 교육을 위한 기초 어휘 의미 빈도 사전의 개발", 『언어정보의 탐구 1』, 연세대학교 언어정보개발 연구원.

[12] 서상규(1998a), 『현대 한국어의 어휘 빈도(상 · 하)』, 연세대학교 언어정보개발연구원 내부 보고서(CLID-WP-98-02-28).

[13] 서상규 · 남윤진 · 진기호(1998a), 『한국어 교육을 위한 기초 어휘 선정 [1] - 기초 어휘 빈도 조사 결과 -』(한국어 세계화 추진을 위한 기반 구축 사업 1차년도 결과 보고서), 문화관광부/한국어세계화추진위원회.

[14] 서상규 · 남윤진 · 진기호(1998b), 『한국어 교육을 위한 기초 어휘 선정 [2] - 교재 8종의 어휘 사용 실태 조사 -』(한국어 세계화 추진을 위한 기반 구축 사업 1차년도 결과 보고서), 문화관광부/한국어세계화추진위원회.

[15] 서상규 편(1999), 『언어 정보의 탐구』 1, 연세대학교 언어정보개발연구원.

[16] 서상규 · 최호철 · 강현화(1999), 『한국어 교육 기초 어휘 의미 빈도 사전의 개발』(1999년도 한국어 세계화 추진을 위한 기반 구축 사업 결과 보고서), 문화관광부/한국어세계화추진위원회.

[17] 서상규 · 한영균(1999), 『국어 정보학 입문』, 태학사.

[18] 서상규(2000), "한국어교육 말뭉치와 학습사전의 개발", 제1차 한국어교육 국제학술대회(한국어세계화추진위원회/이중언어학회 공동주최, 2000년 11월 18일~19일) 발표논문.

[19] 서상규 · 강현화 · 유현경(2000), "한국어 교육 기초 어휘 의미 빈도 사전의 개발"(한국어 세계화 추진을 위한 기반 구축 사업 결과 보고서), 문화관광부.

[20] 임칠성 · 水野俊平 · 北山一雄(1997), 『한국어 계량 연구』, 전남대 출판부.