어휘의미분석 말뭉치 구축의 절차와 문제

신지현 최민우 강범모

고려대학교 언어과학과

{haoyou, minoo, bmkang}@korea.ac.kr

Procedures and Problems in Compiling a Disambiguated Tagged Corpus

Chihyon Shin, MinWoo Choi, Beom-mo Kang

Dept, of Linguistics, Korea University

요 약

동음이의어 간의 서로 다른 의미를 효율적으로 변별해 줄 수 있는 방법 중 하나로 어휘의미분석 말뭉치의 활용을 들 수 있다. 이는 품사 단위의 중의성을 해소해 줄 수 있는 형태소 분석 말뭉치를 기반으로, 이 단계에서 해결하지 못하는 어휘적인 중의성을 해결한 것으로, 보다 정밀한 언어학적 연구와 단어 의미의 중의성 해결(word sense disambiguation) 등 자연언어처리 기술 개발에 사용될 수 있는 중요한 언어 자원이다. 본 연구는 실제로 어휘의미분석 말뭉치를 구축하기 위한 기반 연구로서, 어휘의미분석 말뭉치의 설계와 구축 방법론 상의 제반 사항을 살펴보고, 중의적 단어들의 분포적 특징과 단어의 중의성 해결 단계에서 발생할 수 있는 문제점을 지적하고, 아울러 그 해결 방법을 모색해 보는 것을 목적으로 한다.

1. 서론

동음이의어 간의 서로 다른 의미를 효율적으로 변별해 줄 수 있는 방법 중 하나로 어휘의미분석 말뭉치의 활용을 들 수 있다. 이는 품사 단위의 중의성을 해소해 줄 수 있는 형태소 분석 말뭉치를 기반으로, 이 단계에서 해결하지 못하는 어휘적인 중의성을 해결한 것으로, 보다 정밀한 언어학적 연구와 단어 의미의 중의성 해결(word sense disambiguation) 등 자연언어처리 기술 개발에 사용될 수 있는 중요한 언어 자원이다. 본 연구는 어휘의미분석 말뭉치의 설계와 구축 방법론 상의 제반 사항을 살펴보고, 중의적 단어들의 분포적 특징과 단어의 중의성 해결 단계에서 발생할 수 있는 문제점, 아울러 그 해결 방법을 모색해 보는 것을 목적으로 한다. 구체적으로, 이 논문은 21세기 세종계획의 어휘의미분석 말뭉치 구축을 위한 기초 연구이다.

2. 관련연구

어휘의미분석 말뭉치는 구축에 있어 시간과 노력이 많이 든다는 결점은 있으나, 어휘 문맥과 문맥에서의 어휘의미를 함께 제공한다는 면에 있어서 장점을 지닌다. 이러한 말뭉치는 가장 기본이 되는 원시 말뭉치(raw corpus)에서 출발하여 그것을 가공한 형태소 분석 말뭉치의 단계를 거친 후 생성될 수 있다. 의미 분별을 위한 시도로 국내외 현황을 살펴보면 다음과 같다.

2.1 국외

영국에서는 Langcaster 대학에서 시장 조사 인터뷰 전사물 2만 어휘를 대상으로 말뭉치를 구축했으며, 독일에서는 'Hallig and von Warburg'의 의미 범주 체계(semantic category system)를 발전시키기 위한 수직 말뭉치의 형태로 의미분석 말뭉치를 구축하

였다. 또한, 체코에서는 교육부의 지원으로 Masaryk 대학에서 구축된 말뭉치가 있다. 미국 Princeton 대학에서 구축한 의미분석 말뭉치 SEMCOR는 Brown Corpus를 이용하여 구축되었다. 이것은 명사(고유명사), 동사, 형용사, 부사 등의 개방어(open class words)를 대상으로 의미분석에 착수하였으며, WordNet의 의미목록들을 참조로 하였다.

2.2 국내

연세대학교의 의미주석 말뭉치는1998년도에 구성된 총 100만 어절을 대상으로 '한국어 교육용 말뭉치'를 수정·보완하여 구축 진행중이다. 김한샘(2000)에 의하면, 연세 의미주석 말뭉치는 '연세 한국어 사전'을 참조로 의미 주석을 시도하며 의미주석 반자동 프로그램을 통해 의미 주석을 붙이는 방식으로 진행된다.

또한, 국립국어연구원의 의미주석 말뭉치는 21세기 세종 계획의 1999년 결과 보고서에 포함된 '형태소 분석 말뭉치 구축 지침'에 따라 각 어휘들에 대한 품사태그를 붙인 후, 1990년~1999년에 간행된 문헌을 대상으로 진행한다. '표준국어대사전'을 참조로 각 항목의 번호(어깨번호)를 다는 방식으로 의미 주석 부착이 이루어지며, 반자동 의미 태깅 프로그램 'Word Anal 1.0'을 이용하여 작업이 이루어지고 있다. 형태소 분석 말뭉치를 의미 분석에 적합하게 어휘 간 분리 혹은 결합하는 방식을 취하고 있다.

마지막으로, 인천대학교의 의미분석 말뭉치는 하위범주화 사전과 어휘의 계층적 의미 관계를 나타낸 명사 시소러스를 데이터로 하여 구축된 말뭉치이다. 이수선 외(1999)에 의하면, 이 작업이 2만여 문장을 선택제약 알고리즘과 의미태거(Sense Tagger)를 이용하여 자동적으로 의미변별을 위한 표지를 붙인 후, 작업자가 오류를 수정하는 방식으로 진행되었다고 소개하였다.

3. 의미 분석 말뭉치 구축의 과정

3.1 기본 분석 원칙

어휘의미분석 말뭉치를 구축하는 데 있어서 기본적 분석 방법으로, 우선 말뭉치 내부에 의미 변별을 위한 표지를 부착해 주는 방식과 외부에 사전과 같은 참조 자료를 두고 말뭉치와 연결시켜 주는 방식을 생각해 볼 수 있다. 전자는 하나의 말뭉치에 모든 정보가 표시된다는 장점이 있으며, 그 의미 구분의 표지로서는 한자어 또는 영어 대역어를 이용하는 방법이 있을 수 있다. 그러나, 고유어의 의미 변별은 적당한 대역어를 선별하기 어려운 경우가 존재 할 소지가 있고, 동일한 한자어를 사용하면서 의미는 다른 경우[1] 등을 고려해 본다면 대용량의 의미분석 말뭉치에 사용하기에는 적절치 못한 방법으로 생각된다. 특수한 목적(기계번역 등의 자연언어 처리)을 위한 한정된 어휘의 말뭉치 구축에서는 고려해 볼 수도 있을 것이다.

따라서, 의미분석 말뭉치 구축의 일차 단계에서 외부의 사전을 참조하는 방식을 선택하였다. 기존의 전자화된 사전의 정보를 이용하면 말뭉치 자체에 부가되는 정보의 양을 줄이면서도 효율적인 의미 분석이 가능하다는 장점이 있다. 반면, 두 가지 이상의 자료에 정보가 담겨 있으므로, 활용의 측면에서 다소 복잡함을 감수해야 할 것이다.

3.2 분석 대상 : 한국어 텍스트의 중의성

한국어의 경우 교착어의 특성을 가지고 있기 때문에, 어절 단위의 분석을 시도하기에는 난점이 있다. 따라서 형태소 분석 표지를 먼저 부착 후, 그 결과를 바탕으로 의미 분별 표지를 부착하는 방법을 채택하였다. 본 논문은 21세기 세종계획의 형태소 분석 말뭉치(2000년도에 구축된 150만 어절 규모)를 이용하여 그 분석 결과를 외부사전[2]의 뜻풀이와 연결시켜 의미를 구분해 주는 작업을 바탕으로 하고 있다.

---

[1] 표준국어대사전(1999)에 따르면 다음과 같은 예가 있다. 문자1(文字): 한자로 된 숙어나 성구. 문자2(文字): 의사소통을 위한 시각적인 기호체계

[2] 국립국어연구원의 표준국어대사전(1999)을 이용하였다.

어휘의미분석 말뭉치의 분석 대상 품사는 우선적으로 기능어(function words)를 제외한 내용어(content words)를 중심으로 선정하였다. 기능어의 경우 동일 품사 내에서 동음이의어가 나타나는 경우가 거의 없기 때문이다. 그 결과, 형태소 분석 말뭉치의 분석 표지 중 작업 대상으로 다음과 같은 품사들을 선정할 수 있었다.

(1) 일반명사(NNG), 의존명사(NNB), 동사(VV), 형용사(VA), 어근(XR), 관형사(MM), 일반부사(MAG)[3]

150만 어절 형태분석 말뭉치의 품사별 통계 결과와 표준국어대사전의 표제어 목록을 비교한 결과 다음과 같은 동음이의어의 분포가 확인되었다. 이 결과 150만 어절의 형태소 분석 말뭉치에서 작업 대상 품사의 단어 종류는 45,802개(전체 단어 종류의 25.1%) 정도이고, 이 중 실제 동음이의어가 나타나는 표제어의 수(단어 종류)는 11,526개로 파악되었다. 또한 실제 중의적인 단어 출현(token)수는 전체의 약 45.4% 정도를 차지한다.

품사	단어 종류 (출현빈도)	동음이의어 단어 종류 (출현빈도)	비율 (%)
일반명사	37,387 (781,457)	10,595 (497,623)	28.3 (63.6)
동사	3,216 (235,903)	504 (74,232)	15.6 (31.4)
일반부사	2,766 (79,648)	367 (8,724)	13.2 (10.9)
의존명사	223 (106,496)	20 (23,084)	8.9 (21.6)
형용사	760 (57,548)	34 (2,859)	4.4 (4.9)
관형사	134 (51,136)	6 (1,226)	4.4 (2.4)
어근[4]	1,316 (23,657)	0 (0)	0 (0)
전체	45,802 (1,335,845)	11,526 (607,748)	25.1 (45.4)

표1. 형태소분석 말뭉치에 나타난 한국어 텍스트의 중의성

[표1]에 의하면, 말뭉치 전체를 대상으로 실질어 출현 빈도의 절반(45.4%)이 중의적인 어휘로 나타났으며, 실제로 어휘 종류의 약 4분의 1 가량이 중의성을 지닌 것으로 관찰되었다. 비율 면에서, 명사의 중의성 정도가 가장 높으며, 다음으로 동사>일반부사>의존명사>형용사>관형사 순으로 관찰되었고, 중의성을 갖는 어근은 없는 것으로 나타났다.

일반명사:
말(10) 사람(1) 때(4) 일(4) 사회(6) 생각(2) 속(6)
문제(3) 문화(4) 날(4) 집(3) 앞(1).....
동사:
하다(1) 있다(1) 되다(3) 대하다(2) 보다(1) 위하다(2)
가다(1) 받다(3) 알다(1) 따르다(1) 들다(5) 보이다(3) .....
일반부사:
더(1) 다시(1) 안(1) 잘(1) 가장(4) 없이(1) 함께(1)
바로(1) 모두(1) 못(1) 다(1) 아직(1) .....
의존명사:
것(1) 수(3) 등(1) 년(2) 때문(1) 일(1) 씨(1)
원(1) 중(1)데(1) 명(1) 번(1) .....
형용사:
없다(1) 같다(1) 그렇다(1) 크다(1) 많다(1) 좋다(1)
어떻다(1) 이러하다(1) 다르다(1) 이렇다(1) 새롭다(1)
높다(1).....
관형사:
그(1) 이(1) 한(1) 두(1) 그런(1) 지난(1) 이런(1) 모든(1)
다른(1) 어떤(1) 어느(1) 제(1).....

표2. 품사별 상위빈도 단어 목록과 동음이의어 수

[3] 본 연구에 사용된 형태소 분석 표지는 21세기 세종계획(1998~)의 형태소분석 말뭉치의 분석지침(김흥규, 강범모 2000)을 따르기로 한다.

[4] 형태소 분석에 있어서 동사/형용사 파생접미사 '-하'앞에 붙는 어휘의 품사에 대하여 부사 혹은 어근의 여부에 대한 논란이 존재한다. 우선은 분석지침에 의거하여 작성된 어근의 목록에 근거한다.

[표2]는 각 품사별 최상위 빈도 단어들의 동음이의어 개수를 표시한 것이다. 이 표에서 알 수 있듯이, 둘 이상의 의미를 지닌 단어들의 수가 많고, 같은 품사 내에서 10가지가 넘는 다양한 의미를 지닌 어휘들이 종종 출현한다는 사실을 알 수 있다. 이렇게 중의적 어휘들의 출현 빈도가 높고, 한 단어가 지니는 의미 또한 다양함에도 불구하고, 텍스트를 접하는 독자들은 별 어려움 없이 글의 문맥을 이해해 나간다.

한편, 표준국어대사전의 동음이의어 목록과 현재 대상 말뭉치 내에서 추출한 단어들을 비교한 후, 목록에 등재되어 있지 않은 단어는 품사 단위 내에서 중의성이 없는 것이므로, 의미표지를 부착할 필요가 없다. 작업 대상은 동일 품사 내의 다른 의미를 지닌 단어들로 한정하며, 형태소 분석 말뭉치를 기반으로 할 경우에는 일차적으로 형태소 분석 단계에서 한 어휘 내 다른 품사들 간에 나타날 수 있는 중의성이 해소되기 때문에, 추가적인 의미 변별은 필요하지 않다.

3.3 선정 어휘에 대한 용례 추출 및 분석표지 부착

다음 과정으로, 작업 대상 어휘들을 확정 후, 각각의 표제어에 대한 용례를 추출한다. 동일한 표제어의 용례들 중 SemConc(조진현 2001)[5]를 이용해 대상 어휘를 기준으로 전후 다섯 어절씩을 추출하여,[6] 동일 환경에서 출현하는 단어들의 양상을 파악한다. 의미 표지를 부착하는 데 있어서는 컴퓨터 보조 도구를 이용하였다. 분석 대상 어휘 한 개 당 용례들을 한 파일로 만들어, 동음어 사전과 용례 파일들을 동시에 띄운 후 사전 목록의 적절한 의미를 검색해 해당 어깨번호를 선택하면 자동적으로 단어에 의미 표지가 입력되는 방식을 택하였다. 다음은 의미 태깅 도구의 작업 중 일례이다.

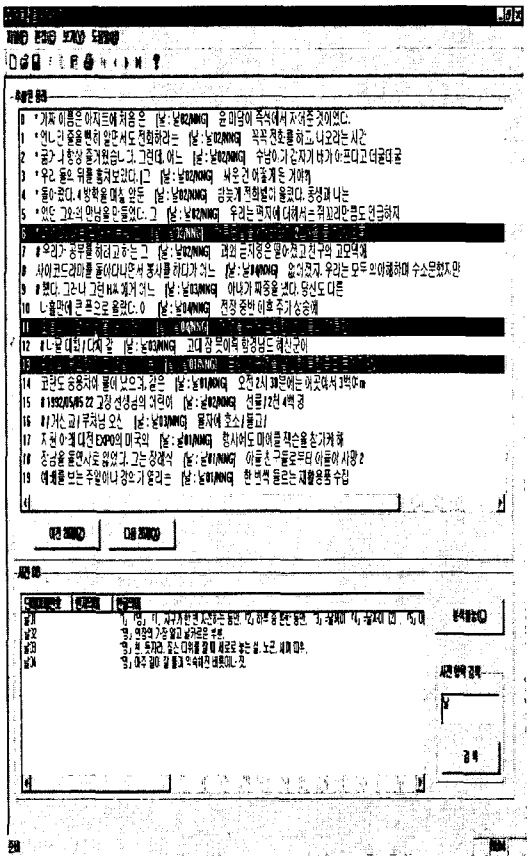

그림1. 의미분석 말뭉치의 작업 도구

이 때, 일반적으로 체언과 수식언 등은 뒤에 오는 서술어, 혹은 피수식어와의 연어 관계를 보일 확률이 높고, 반대로 서술어로 쓰이는 용언은 앞에 오는 형태소와 의미적으로 연관성이 있게 된다. 따라서 품사별로 그에 적합한 정렬을 통해 작업을 진행할 수 있다.

3.4 통계 분석 및 오류 수정

위의 과정을 통해 구축된 어휘의미분석 말뭉치로부터 각종 통계 정보를 추출하고, 태깅 과정에서의 오류를 수정한 후 최종적인 의미분석 말뭉치를 완성한

[5] 중심어 인접 어휘들의 어절 수를 지정해서 용례를 추출하는 프로그램

[6] Kaplan(1955)은 실험결과로, 일반적으로 사람들이 문맥에서 중심어의 전후 두 단어를 통해 의미 식별이 가능하다고 하였다.

다.

4. 결과 및 검토

4.1 통계적 분포

지금까지 작업결과 나타난 사실들에 대하여 살펴 본다

(2) 앉아서 곰곰이 생각해 봐도 정호의 [말 : 말01/NNG]한마디가 마음에 자꾸 걸리고 불쾌했습니다.

(3) 못하는 새로운 먹이에 덤벼든다. 그리고 [말 :말05/NNG] 안장같이 생긴 여치의 가슴을 물고,

(4) 구주 여행사라고 했다. 배낭 여행에 [드는 : 들01/VV + 는/ETM] 비용은 이백오륙십만 원 정도라 했다.

(5) 게 고작이다. 왜 그렇게 안 [드시나고 : 들04/VV + 시/EP + 나고/EC] 여쭈면 입맛이 없다고 하신다. 술이

말뭉치 내에서 가장 높은 빈도수를 보이는 일반명사 말'은, 표준국어대사전에는 일반명사로서 총 10개의 항목으로 등재되어 있다. 총 7535번 중 7371번(총 빈도수의 97.8%)이 연어'(말01)의 의미로, 155번(2%)이 똥물'(말05), 나머지 9번(0.1%)이 분석불능[7]으로 나타났다. 단어 들다'의 경우, 동사로써 등재되어 있는 표제어의 가짓수는 다섯 가지이나, (4)에서와 같은 의미(들다01)의 경우 1503번(58.9%), (5)에서 쓰인 의미(들다04)의 경우 1029번(40.3%) 기타 형태소 분석오류 혹은 분석 불능은 27번(1.0%) 출현하는 것으로 나타났다. 작업대상 단어들이 가지는 의미가 다양함에도, 이처럼 말뭉치 내에서 출현하는 표제어들의 의미는 한정되어 있다. 동일한 품사 내에서 동음이의어로 분류되는 단어들 중에는 방언, 오표기, 고어, 전문용어 등으로 분류되는 표제어들 또한 존재하며, 이들 항목을 가지는 어휘들 역시 중의성이 있는 단어라는 기준을 정하였다. 150만 어절의 형태분석 말뭉치는 현대국어 기초자료 말뭉치 구축을 목적으로 완성된 것이며, 1900년대부터 현재까지 발간된 신문, 잡지, 책 등 주변에서 비교적 접하기 쉬운 텍스트를 토대로 구축된 것이다. 따라서, 고어 혹은 방언, 오표기, 전문용어 등이 나올 확률이 적으므로 그만큼 말뭉치에서 출현하는 빈도가 낮아지게 된다.

말01「명」「1」사람의 생각이나 느낌 따위를 표현하고 전달하는 데 쓰는 음성 기호.「2」음성 기호로 생각이나 느낌을 표현하고 전달하는 행위. ……
말02「명」툽질을 하거나 먹줄을 그을 때 밑에 받치는 나무.
말03「I」「명」곡식, 액체, 가루 따위의 분량을 되는 데 쓰는 그릇.「II」「명」「의」부피의 단위.
말04「명」『민』십이지에서 오05(午)를 상징적으로 나타내는 말.
말05「명」『동』말과의 포유동물.
말06「명」「식」「1」물속에 나는 은화식물을 통틀어 이르는 말.「2」가랫과의 여러해살이 수초(水艸). 「3」=해조05(海藻).
말07「명」『운』「1」고누, 윷 따위의 판에서 놀이의 도구로 쓰는 물건. 「2」=마06(馬).
말08「명」『방』망울01[5] 의 방언(경북).
말09「명」『방』마루폭 의 방언(경상).
말10「명」버선 의 잘못.

표3. 표준국어대사전에서 일반명사 말'의 표제어

4.2 자동 분석을 위한 규칙 정립

형태소 분석 말뭉치는 규칙이나 통계정보에 의한 자동 분석과 표지 부착을 통하여 수작업의 비중을 줄일 수 있다.

(6) 임금의 은덕을 걸부시켜 각 한 [수씩 : 수17/NNB + 씩/XSN]4 수로 옳었다. 매수 첫머리는

(7) 행복한 나라다 하고 그 이유로서 [먹을수 : 먹/VV + ㄹ/ETM + 수02/NNB] 있는 것이면 개구리건 달팽이건 놓치질

[7] 전후 다섯 어절만을 보고도 의미분별을 할 수 없는 경우는 분석불능의 표지를 붙인 후, 최종적으로 수합하여 전체 말뭉치의 문맥을 참고로 2차 분석을 하기로 한다.

의존명사는 단위성 의존명사와 비단위성 의존명사로 나누어질 수 있는데, (6)에서 알 수 있듯이, 관형사 또는 숫자, 수사 뒤에 나타나는 의존명사는 단위성 어휘이며, (7)처럼 용언에 첨가된 관형사형 전성어미 -'ㄹ' 다음에 나오는 의존명사는 비단위성 어휘들이다. 따라서, 의존명사의 경우는 선행하는 어휘의 품사 태그를 고려하여, 규칙을 통해 처리할 수도 있다.

그럼에도 불구하고, 의미분석 말뭉치는 그러한 규칙 파악이 쉽지 않다. 가령, '말01'(언어)의 경우 경동사인 -'하다' 등의 용언과, '말05'(동물)의 경우 타다' 등의 용언과 결합할 수 있기 때문에 규칙을 설정하는데 도움이 될 수는 있으나, 규칙적용이 일정한 단위를 대상으로 하는 것이 아니라 개별 표제어 별로 규칙을 파악해야 하기 때문에 규칙을 정립하기에 어려움이 따른다.

위의 예들에서 보이는 바와 같이 의미분석의 경우, 그 규칙성을 결합하는 다른 형태와의 관계에서 파악하는데 있어서, 개별 어휘의 의미에 따라 다른 모든 경우를 파악하여 규칙화 하는 것은 작업상에 있어서 큰 이익을 보기 어렵다. 따라서 의미분석 말뭉치 작업은 사람에 의한 수동 부착 방식이 주가 될 것이다. 그러나, 출현 빈도가 높은 특정한 어휘가 일정한 규칙성을 지닐 경우에는 부분적으로 문맥규칙을 이용한 처리를 하여 작업을 효과적으로 진행할 수 있는 가능성은 있을 것이다.

4.3 문제점 및 해결방안

4.3.1 형태소 분석상의 태깅 오류

(8) 바람 창문이 덜컹이는 소리. 소리. [바람에 : 바람01/NNG + 에/JKB] 춤을 추는 나무. 나무 그림자.

(9) 그녀는 남자 친구의 손목을 놓치는 [바람에 : 바람01/NNG + 에/JKB] 그를 더욱 곤란하게 만들기 일쑤였다.

공기의 움직임' 혹은 간절한 마음 상태' 등의 단독으로 사용될 수 있는 일반명사 바람'을 제외하고는, 뒷말의 근거나 원인'을 나타낼 시에는 의존명사로 형태소 분석이 이루어져야 한다. 그러나, 대용량의 자료를 자동 처리하는 과정에서 이러한 종류의 오류가 간혹 발생할 수 있다. 따라서, 형태소 분석 단계에서 일어나는 이러한 종류의 오류는 예측이 불가하다. 의미 분석의 정확성을 위해서는 분석의 오류를 수정하면서, 동시에 의미 분석을 병행해야 할 것이다.

4.3.2 직관의 불명확성

(10) 전전대통령의 측근들은 항간의 시비에 대해 [일일이 : 일일이01/MAG] 맞대응을 하지 않고 감사원의 감사를

(11) 협력 사업에 따른 손실을 정부가 [일일이 : 일일이02/MAG] 보전해 주게 되면 국내 업체들이

(12) 김효은 경찰청장은 앞으로의 수사와 관련, ["부정 : "/SS + 부정02/NNG] 입학생의 학부모가 누구인지 나에게는 일절

(13) 독재 정권 군사 정권의 불의와 [부정 : 부정06/NNG] 부패에 저항-, 더러는 엉들고 더러는

(14) 대한 그의 비판을 '예술 자체의 [부정'이라는 : 부정09/NNG + '/SS + 이/VCP + 라는/ETM] 식으로 단순화해서는 안될 것이다. 그것은

사전에 등재되어 있는 의미들을 이용하여 변별 작업을 진행하는데 있어서 의미의 구분이 명확함에도 불구하고, 작업자의 혼동을 일으키는 항목들이 출현한다. 사전에 기술된 일반부사 '일일이'의 경우 '일마다 모두' 라는 전체를 강조한 의미와, '하나하나'와 같이 각각의 부분을 강조한 의미 두 가지로 나누어질 수 있다. 그러나, 의미가 두 가지로 나누어져 있음에도 불구하고, 말뭉치 내에서는 제시된 어휘의 의미 경계가 모호하다. (10), (11)에서 보는 바와 같이, 의미의 비중을 어디에 두는가에 따라 분석표지를 구별해 주어야 하므로 작업자 개인의 직관에 의존하기에는 분석 오류의 가능성이 높다. (12), (13), (14)에서 각각 '不正', '不正', '否定' 의 의미로 쓰이는데, 작업자들이 문맥과 사전의 의미를 참조하면서도 이 세 가지 다른 의미의 쓰임을 명확히 파악하기에 어려웠다. 이러한 경우에, 작업자들 간의 사전

협의를 거쳐 가장 적합한 의미를 선택하였다.

4.3.3 사전과 형태소 분석이 다른 경우

(15) 현지의 값싼 노동력(임금이 우리 나라의 [10분의 : 10/SN + 분x15/NNB + 의/JKG] 1 이하라고 한다)으로 물건을 만들어

형태소 분석 말뭉치의 표지와 표준국어대사전의 표지가 일치하지 않는 경우가 있었다. 이는 형태소 분석 말뭉치는 엄밀한 사전적 정의를 목표로 하는 것이 아니고, 정해진 기준에 따른 형태소의 결합을 보여주는 것이 목적이기 때문이다. 예를 들면, '분'이라는 형태소의 경우 '10분의 1' 같은 표현에서의 '분'은 사전의 정의에 의하면 접미사이다. 하지만 형태소 분석 말뭉치에서는 이 단어를 접미사가 아닌 의존명사로 분류하고 있다. 이러한 경우의 처리 방법으로 별도의 표지를 부착하여 추후 처리하는 것으로 해결하였다.

(15)에서 밑줄 부분은 사전의 15번 어깨번호가 붙은 부분의 의미가 이 분석에 해당하나 실제로 사전에는 NNB(의존명사)로 기술되어 있지 않음을 나타낸다. 이처럼 사전과 형태소분석 말뭉치간의 차이가 나타날 수 있는 경우는 일반명사와 의존명사, 의존명사와 접미사, 관형사와 접두사등에서도 찾아 볼 수 있다.

4.3.4 사전에 미등록된 어휘

(16) 12일자 '길'난에 보도된 '교도소 가는 [부정'의 : 부정/NNG + '/SS + 의/JKG] 인태군 씨에게 전해 달라며 박찬수-

(16') 1992/01/15 22 '길' 보도 인태군 씨에 성금 조선일보 1월 12일자 '길'난에 보도된 '교도소 가는 부정'의 인태군 씨에게 전해 달라며 박찬수 씨(45· 서울· 강남구 개포동 현대아파트)가 6만원을 14일 조선일보사에 맡겼다.

사전에 등재되지 않은 어휘는 작업에 있어서 가장 큰 문제가 되고 있다. 반의어인 '모정'(母情)이라는 단어는 존재하나, 실재로 태깅 과정에서 (16)과 같이 '부정'(父情)이라는 단어가 출현할 경우, 사전에서 지원하는 어휘가 아니기 때문에, 적절한 분석표지가 모색되어야 한다.

5. 결론

본 논문에서는 어휘의미분석 말뭉치 구축을 위한 방법론과 진행, 그리고 문제점과 해결방안에 대한 제안을 하였다. 실제 어휘의 중의성을 해결해 줄 수 있는 방대한 양의 의미분석 말뭉치가 부재한 실정이며, 구축에 있어 방법론이나 중의성 해소에 대해 논란이 존재하고 있다. 어휘의미분석 말뭉치의 쓰임은 실로 유용하다 할 수 있겠다. 무엇보다도, 용례들을 보여주는 말뭉치를 기반(example-based)으로 하여 기계번역 등 각종 자연언어처리 분야에 응용할 수 있으며, 말뭉치에 나타난 단어들의 의미를 각종 통계 분석을 통해 계량적으로 한국어 현상에 대한 연구에도 사용될 수 있다.

현재 말뭉치 구축에 있어 수작업의 비중이 절대적으로 큰 실정이다. 의미분별에 있어서, 수작업의 비중을 줄이고, 최대한 규칙 적용을 통한 자동 처리에 적합하도록 기계가독형 사전(Machine Readable Dictionary)의 연어정보를 추출하여 적용하려는 시도가 있다.

앞으로 작업도구를 이용해 말뭉치에서 품사별로 나타나는 모든 단어들의 중의성을 해결하고, 말뭉치의 규모를 확장하는 것이 향후과제로 남아있다.

6. 참고문헌

[1] 김도완, 이경순, 김길창 (1999), 의미관계와 문형정보를 이용한 복합명사 해석", 제11회 한글 및 한국어 정보처리 학술대회 논문집, 310-315.

[2] 김홍규, 강범모 (2000), 「한국어 형태소 및 어휘 사용 빈도의 분석 1」, 서울: 고려대학교 민족문화연구원.

[3] 김홍규 외 (1998-2000), 「21세기 세종계획 연구보고서」, 문화관광부.

[4] 김한샘 (2000), 말뭉치의 주석과 활용: 의미 주석 말뭉치의 활용을 위한 기초 연구", 언어정보

연찬회 발표 논문집 14, 연세대학교 언어정보개발원, 17-27.

[5] 이수선, 박현재, 우요섭 (1999), "한국어 분석의 중의성 해소를 위한 하위범주화 사전 구축", 제11회 한글 및 한국어 정보처리 학술대회 논문집, 257-264.

[6] 조진현 (2001), SemConc Program Ver.1.0, 고려대학교 언어과학과.

[7] 국립국어연구원 (1999). 「 표준국어대사전 」, 서울:두산동아.

[8] Leacock, Claudia and Yael Ravin (eds.) (2000). *Polysemy*, New York: Oxford University Press.