의미 중의성 해소를 위한 품사의 역할 : 영어와 한국어 비교

조 정 미
서강대학교 전자계산학과
서울시 마포구 신수동 1
우: 121-742
jmcho@nlprep.sogang.ac.kr

김 길 창
한국과학기술원 전산학과
대전시 유성구 구성동 373-1
우: 305-701
gckim@csking.kaist.ac.kr

서 정 연
서강대학교 전자계산학과
서울시 마포구 신수동 1
우:121-742
seojy@ccs.sogang.ac.kr

Role of POS Tags in Word Sense Disambiguation : A comparison of English and Korean

Jeong-Mi Cho
Dept. of Computer Science
Sogang Univ.

Gil Chang Kim
Dept. of Computer Science
KAIST

Jungyun Seo
Dept. of Computer Science
Sogang Univ.

요약

본 논문[1]은 의미 중의성 해소에 있어서 품사 태깅의 중요성을 언급한 Wilks의 논문 [6]을 근거로 하여 한국어 의미 중의성 해소에 있어서의 품사 태깅의 역할을 살펴보고, 영어의 경우와 비교, 분석한다. 한국어 사전과 코퍼스를 각각 대상으로 품사 태깅을 이용한 의미 중의성 실험 결과, 한국어의 경우는 영어의 경우보다 품사를 이용한 의미 중의성 해소율이 떨어지는 결과를 보이고 있다.

1 소개

의미 중의성 해소란 문맥 내에서 단어의 올바른 의미를 선택하는 작업으로, 이는 단어와 문맥에 대한 다양한 지식을 필요로 한다. 기존의 의미 중의성 해소와 관련한 연구들은 대량의 사전과 코퍼스를 이용한 추론과 학습을 통해 이러한 지식을 획득하였다[1,2,3,4,5,7]. 이 과정에서 지식 획득의 병목 문제(knowledge acquisition bottleneck)와 자료 부족 문제(data sparseness problem)가 발생한다.

[6]은 이러한 대량의 학습 없이 정확한 품사 태깅만으로 의미 중의성을 얼마나 해소할 수 있는가를 보여준다. 지금까지 단어에 대한 품사 태깅과 의미 태깅은 전혀 별개로 연구가 진행되어 온 반면, [6]에서는 품사의 중의성 해소는 의미 중의성 해소에 있어서 중요한 역할을 한다라는 가정을 세우고, 이와 관련된 몇 가지 흥미로운 실험 결과를 보여주었다. 실험 결과, 동형이의어(homograph) 수준의 중의성은 단어의 품사를 아는 것에 의해 충분히 해소 가능하다라는 결론을 내리고 있다.

본 논문에서는 [6]의 논문을 근거로 하여 한국어 의미 중의성 해소에 있어서의 품사 태깅의 역할을 살펴보고, 영어의 경우와 비교, 분석한다.

1) 본 연구는 과학 재단의 특정 연구 과제 "통계 모델을 이용한 한국어 처리"의 지원을 받은 것입니다.

2 단어의 유형 분류

homograph란 관련 있는 의미들(meanings)로 구성된 의미(sense) 집합으로, 표 1에서 단어 '쓰다'의 homograph 중 하나인 '쓰다[1]'은 'write'와 관련된 의미들로 구성되어 있다. homograph의 한국어 번역어인 동형이의어란 형태는 같으나 뜻이 다른 단어를 의미하며, homograph와는 의미에 있어서 약간의 차이가 있다. 표 1에서 '쓰다[1]', '쓰다[2]', '쓰다[3]'은 각각 동형이의어 관계에 있는 단어들이다.

표제어	품사	의미 정의
쓰다[1]	타동사	a.펜, 연필, 붓 따위로 획을 그어 글자를 이루다 b.글을 짓다
쓰다[2]	타동사	a.(모자를) 머리에 얹다. b.억울한 누명이나 죄를 입다. c.(먼지나 액체 따위를) 온통 몸에 받다.
쓰다[3]	형용사	a.맛이 소태와 같다. b.입맛이 없다 c.마음이 언짢거나 괴롭다.

표 1. 표제어 '쓰다'에 대한 사전 내용

단어는 그 단어에 해당하는 의미 분류에 따라 동형이의어 수준의 분류와 다의어 수준의 분류로 나뉜다.

1) 동형이의어 수준에 따른 분류

• MH(MonoHomographic) : 한 단어 유형에 대해 하나의 동형이의어만을 갖는 경우로, 표 2에서 단어 '프로그램'과 같이 단어에 대한 표제어 수와 단어 유형의 수가 1 개로 일치한다.

• PH(PolyHomographic) : 한 단어 유형에 대해 둘 이상의 동형이의어를 갖는 경우로, 단어에 대한 단어 유형은 1 개, 표제어의 수는 2 개 이상이다.

2) 다의어 수준에 따른 분류

• MS(Mono Sense) : 하나의 표제어가 하나의 의미만 갖는다. 표 2에서 '기술^2'와 '기술^3'은 하나의 표제어에 하나의 의미만을 갖는 MS 유형에 속하는 예이다.

• PS(PolySemous) : 하나의 표제어가 둘 이상의 의미를 갖는다. 표 1과 표 2에서 하나의 표제어에 'a, b, c' 등으로 의미가 다시 세분화되는 단어가 이 유형에 해당한다.

또한 단어는 품사에 의한 의미 중의성 해소 정도에 따라 다음과 같이 분류할 수 있다.

3) 품사에 의한 중의성 해소에 따른 분류

• GD(Guaranteed Disambiguation): 각 동형이의어가 서로 다른 품사를 갖는 단어. GD 유형에 속하는 단어는 품사 정보에 의해 의미 중의성을 해소할 수 있다. 예를 들면, 표 2의 '익다'와 같이 2 개의 동형이의어를 갖는 단어가 각각 자동사, 형용사의 품사를 갖는다면, 정확한 품사 태깅에 의해 동형이의어 수준의 의미 중의성 해소가 된다.

• PD(Possible Disambiguation): 유일한 품사를 갖는 동형이의어가 적어도 하나 이상 있는 단어. PD 유형에 속하는 단어는 어느 품사로 태깅되는가에 따라 의미 중의성 해소가 되기도 하고 중의성 해소를 할 수 없기도 하다. 예를 들면, 표 1에서 '쓰다'와 같이 3 개의 동형이의어를 갖는 단어가 각각 타동사, 타동사, 형용사의 품사를 갖는 경우, 중의성 해소하고자 하는 '쓰다'가 문장에서 형용사이면, 중의성 해소가 가능하지만 타동사인 경우는 두 개의 동형이의어(쓰다^1, 쓰다^2) 중 어느 것인지 분별할 수 없다.

• ND(No Disambiguation): 유일한 품사를 갖는 동형이의어가 없는 단어. ND 유형에 속하는 단어는 품사만으로 의미 중의성 해소를 할 수 없다. 예를 들면, 표 2의 '기술'과 같이 3개의 동형이의어를 갖는 단어가 모두 품사 명사를 갖는 경우, 품사 태깅에 의해 중의성 해소를 할 수 없다.

GD 유형에 속하는 단어 수는 단어의 품사 정보만을 이용한 의미 중의성 해소의 이론적인 하한값이 되며, PD 유형에 속하는 단어 수는 품사에 의한 중의성 해소의 상한값이 된다. 이러한 3 가지 유형의 분포는 그 대상이 사전인가 코퍼스인가에 따라 다른 양상을 보인다. 사전의 표제어들을 대상으로 하는 경우, 모든 단어들이 고르게 한 번씩 고려된다. 즉, 일반 텍스트에서 자주 사용되는 단어들과 자주 사용되지 않는 단어들이 모두 같은 빈도수로 처리된다. 따라서 대상 언어(영어, 한국어)에 대한 3 가지 유형의 분포를 알 수 있고, 그 언어에 있어서 의미 중의성에서 품사의 역할을 예측할 수 있다.

코퍼스에 나타난 단어들을 대상으로 하는 경우는 사전을 대상으로 하는 경우와 상황이 매우 다르다. 첫째, 코퍼스에는 빈번하게 나타나는 단어도 있고, 한번도 나타나지 않는 단어가 있기 때문에 각 단어의 발생 빈도수를 고려하여야 한다. 둘째, 단어의 발생 빈도와 단어의 의미의 중의성간의 연관성을 고려하여야 한다. 자연 언어는 사용 빈도가 높은 단어들이 많은 의미 중의성을 갖는다[8]. 따라서 코퍼스를 대상으로 할 경우, PH 유형과 PS 유형이 차지하는 비율이 높아질 것이며, 의미 중의성 문제가 더욱 심각하게 된다.

표제어	품사	의미 정의
기술^1	명사	a.자연을 인간 생활에 적합하도록 이용하는 수단의 총체 b.어떤 일을 솜씨 있게 할 수 있는 방법이나 능력
기술^2	명사	교묘한 솜씨로 잠시 눈을 속여 재미있게 부리는 재주
기술^3	명사	사물의 내용을 그대로 기록하여 서술하는 것
프로그램	명사	a.진행 계획이나 순서. b.문제 해결을 위해 컴퓨터에 주어진 처리 방법과 순서를 기술한 일련의 명령문 집합
익다^1	자동사	a.다자라 여물다 b.뜨거운 열을 받아 연하게 되다.
익다^2	형용사	여러 번 경험하여 서투르거나 설지 않다.

표 2. 사전 내용의 일부

3 실험 및 평가

실험은 유형별 단어의 분포만을 조사하는 이론적 실험과 실제 의미 중의성 해소에 적용하는 실험으로 나누어 수행하였다[2].

3.1 이론적 실험

LDOCE와 그랜드 사전의 표제어, KAIST 코퍼스[9]에 나타난 단어들에 대해 GD, PD, ND, 3 가지 유형별의 분포를 분석하였다. 분석 결과는 표 3[3]과 같다.

LDOCE의 경우, PH 유형에 속하는 단어들 중

2) 실험 결과 중 영어에 대한 결과는 [6]에 근거한 것이다.

3) 사전의 표제어에서 기능어, 즉 영어에서는 전치사와 관사, 한국어에서는 조사, 어미 등을 제외하고, 코퍼스에 나타난 단어들 중 사전에 등록이 되어 있지 않은 단어들과 조사, 어미, 기호 등은 제외한다.

88%가 품사 정보만으로 의미 중의성 해소가 정확하게 되며, 95%는 중의성 해소가 가능하다. 사전 표제어 중 동형이의어가 단 하나뿐인 단어들(MH 유형)을 품사 정보로 의미 중의성 해소가 확실한 GD 유형으로 간주하면, LDOCE의 경우 GD 유형에 속하는 단어가 전체 표제어의 98.6%, PD 유형은 99.4%에 해당한다. 따라서 LDOCE는 품사 정보만으로 표제어의 99.4%에 대해 의미 중의성 해소를 할 수 있다.

	LDOCE	그랜드사전	KAIST 코퍼스
MS		76.6%	15.8%
PS		23.4%	84.2%
MH	88%	86.2%	29.0%
PH	12%	13.8%	71.0%
GD	88%	17.7%	32.3%
PD	95%	21.6%	82.4%
ND	5%	78.4%	17.6%
하한값	98.6%	88.7%	51.9%
상한값	99.4%	89.2%	87.5%
크기(단어)	36,000	117,730	203,312

표 3. 단어의 유형별 분석

영어와 한국어는 전체 단어에서 MH 유형과 PH 유형이 차지하는 비율은 크게 다르지 않으나, PH 유형 중 PD와 ND 유형의 분포에서는 상반되는 현상을 보인다. 영어는 PD와 ND의 비율이 95 : 5인데 비해, 한국어는 22 : 78이다. 한국어는 PH 유형 중 22%의 단어만이 품사 정보만으로 중의성 해소가 가능하기 때문에 의미 중의성의 문제가 더욱 심각하다라고 할 수 있다. 한국어 사전의 표제어 중 MH 유형에 속하는 단어들을 GD 유형으로 간주하면, GD 유형에 속하는 단어가 전체 표제어의 88.7%, PD 유형은 89.2%이다. 따라서 의미가 하나뿐인 단어까지 고려하더라도 영어와 비교하여 품사에 의한 중의성 해소의 가능성에 큰 차이가 있다.

코퍼스의 경우는 사전과 비교하여 PS 유형과 PH 유형의 비율이 월등하게 크다. 의미 중의성이 많은 단어들이 많이 사용된다라는 것을 입증한 것이다. 또한 PH 유형에 속하는 단어들 중 품사 정보만으로 중의성 해소가 가능한 비율 역시 사전과 비교하여 월등하게 크다. 코퍼스에 나타난 단어들 중 MH 유형에 속하는 단어들을 GD 유형으로 간주하면, GD 유형에 속하는 단어가 전체 단어의 51.9%, PD 유형에 속하는 단어가 87.5%이다. 따라서 코퍼스는 의미 중의성이 있는 단어가 어떤 품사를 갖는가에 따라 중의성 해소에 많은 영향을 받는다.

그림 1은 LDOCE와 그랜드 사전, KAIST 코퍼스를 대상으로 전체 단어에 대한 MH, GD, PD, ND 유형에 속하는 단어의 분포를 비교한 것이다. 그림에서 보듯이, 영어보다는 한국어가, 사전보다는 코퍼스에서의 의미 중의성 문제가 심각함을 알 수 있다.

표 3에서 크기에 있어서 LDOCE와 그랜드 사전의 차이가 큼을 알 수 있다. 이것은 [6]에서 대상으로 한 LDOCE가 학생들을 대상으로 한 36,000 단어의 축소판 사전이기 때문이다. 따라서 영어와 한국어간의 정확한 비교를 위해서는 사전의 크기와 종류 면에서 좀더 다양한 실험을 필요로 한다.

3.2 의미 중의성 실험

3.1에서 살펴본 이론적 실험 결과와 실제 품사만을 이용한 의미 중의성 해소 결과를 비교하기 위하여 KAIST 코퍼스 일부에 대해 품사 태깅만에 의한 의미 중의성 해소 실험을 수행하였다.

KAIST 코퍼스는 품사 부착 코퍼스로, 사전에 등록되어 있지 않은 미등록어와 조사나 어미와 같은 기능어는 제외하고 실질어만을 대상으로 하였다. KAIST 코퍼스의 품사 집합은 간단한 대응 테이블에 의해 그랜드 국어 사전에서 사용된 품사 집합으로 대응시킬 수 있다.

단어의 여러 동형이의어 중 같은 품사를 갖는 동형이의어가 하나 이상인 경우 품사만으로 단어에 대한 유일한 의미를 부여할 수 없다. 이런 경우, 사전에 나타난 첫번째 동형이의어 의미를 그 단어의 의미로 선택한다. 예를 들면, '배'라는 단어는 8 가지의 동형이의어가 사전에 등록되어 있고, 그 중 6 가지가 품사, '명사'를 갖는다. 코퍼스에서 '배'가 '명사'로 품사 태깅이 된 경우, 이들 6 가지의 동형이의어 중 사전에 첫번째로 등록된 동형이의어의 의미를 '배'에 대한 중의성 해소의

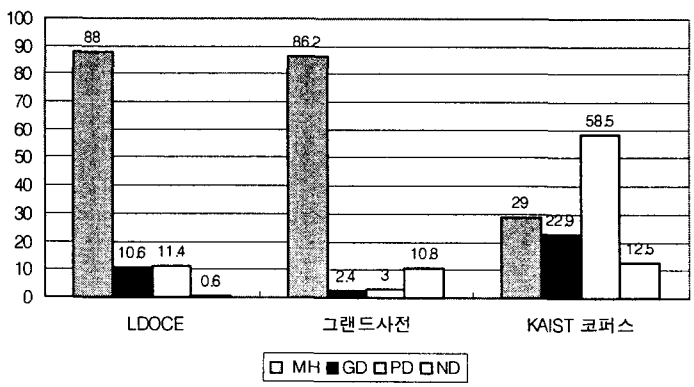

그림 1. LDOCE, 그랜드 사전, KAIST 코퍼스간의 분포 비교

결과로 선택한다. 이것은 한 단어가 여러 가지 뜻을 가질 때는 사용 빈도가 높은 것, 또는 역사적으로 근원이 된 것부터 사전에 배열한다는 원칙에 근거한 것이다.

KAIST 코퍼스에서 270 문장을 추출하고, 이 중 미등록어와 기능어를 제외한 1784 개의 단어를 대상으로 동형이의어 수준의 의미 중의성 해소를 수행하였다. 실험 대상에 대한 유형 분석은 표 4와 같다.

KAIST 코퍼스 중 1784 단어	
MS	19.9%
PS	80.1%
MH	39.9%
PH	60.1%
GD	34.0%
PD	69.1%
ND	30.9%

표 4. 1784 단어의 유형별 분석

1784개의 단어 중 PH 유형에 속하는 단어는 60.1%에 해당하고, 의미가 하나뿐인 단어를 GD 유형에 속한다고 간주하면, GD 유형에 속하는 단어가 60%, PD 유형에 속하는 단어가 81%이다. 즉, 품사 정보만을 이용할 경우, 최고 81%의 중의성 해소 결과를 얻을 수 있다는 것이다. 실제 품사 태깅에 의한 의미 중의성 해소 실험 결과는 표 5와 같다. 표 5는 영어의 Wall Street Journal을 대상으로 한 의미 중의성 해소 실험 결과와 비교한 것이다[4]. 표에서 보듯이, 한국어의 경우 품사 정보만을 이용한 것은 72%, 품사 정보와 첫 번째 의미 선택 휴리스틱을 함께 이용한 것은 84.3%의 정확률을 보였다. 이것은 같은 실험 대상에 대해 사람이 한 것과 비교한 결과이다. PH 유형에 속하는 단어들 중에서는 73.9%가 정확한 의미 중의성 해소 결과를 내고 있다.

	WSJ	KAIST코퍼스
전체 실험 단어 수	1,700	1,784
품사만 이용		72.0%
품사+첫번째의미선택	92.0%	84.3%
PH에 속하는 단어	57.0%	60.1%
품사+첫번째의미선택	87.4%	73.9%

표 5. 품사를 이용한 의미 중의성 해소 결과

영어의 경우는 전체 단어 중 57%가 PH 유형에 속하고, 이 중 품사와 첫번째 의미 선택 휴리스틱에 의해 87%가 의미 중의성이 해소된다. 전체 단어로 환산하면 92%가 품사와 첫번째 의미 선택 휴리스틱에 의해 의미 중의성 해소가 가능한 것이다. 전체 단어를 대상으로 하면 8%, PH 유형에 속하는 단어만을 대상으로 하면 13% 가량 영어에 비해 한국어의 중의성 정확률이 떨어지는 결과를 보인다. 따라서 한국어가 영어에 비해 품사에 의한 의미 중의성 해소가 더욱 어렵다라고 말할 수 있다.

4 토의 및 결론

[6]에서는 품사만을 적용하여 문장의 모든 단어들의 의미 중의성을 해소하는 실용적인 중의성 해소 모델을 제안하였다. 실제로 Wall Street Journal을 대상으로 하여 품사만을 이용한 중의성 해소 실험을 해 본 결과, 동형이의어 수준으로 실질어에 대해서는 92%, 실질어와 기능어를 모두 대상으로 한 경우에는 94%의 중의성을 해소하였다. [6]에서는 동형이의어 수준의 중의성 해소라면 품사만을 이용하여 효과적인 중의성 해소를 할 수 있다라고 결론을 내리고 있다.

한국어에 대해서도 같은 실험을 수행해 본 결과, 영어와는 다른 결과를 보이고 있다. 의미 중의성 해소에 있어서 품사의 유용성이 이론적 분석과 실제 실험 모두에서 영어에 비해 한국어가 떨어지는 결과를 보이고 있다. 이것은 실험에 이용된 사전에 의존적인 문제일 수도 있다. 그러나 코퍼스를 대상으로 한 실험에서도 영어에 비해 한국어가 전체 단어를 대상으로 한 경우에는 8%, PH 유형에 속하는 단어만을 대상으로 한 경우에는 13% 정도 중의성 해소율이 떨어진다라는 것은 영어에 비해 한국어의 의미 중의성 정도가 더욱 심하다라는 것을 의미한다.

또한 [6]에서는 기능어는 모두 MH 유형으로 분류하여, 기능어를 포함시킬 경우 의미 중의성 해소율이 더욱 올라가는 결과를 보이고 있으나, 한국어의 기능어는 이렇듯 단순하지 않다. 예를 들어 표 6에서와 같이 조사 '으로'는 하나의 품사를 갖는 MH 유형에 속하기는 하지만, 실제 문장에서의 역할은 7 가지이다. '으로'에 대해 이러한 역할을 구분하지 않은 의미 중의성 해소란 매우 실용적이지 못하다. 따라서 의미 중의성의 첫 단계로 품사를 이용하는 것은 의미가 있으나, 정확하고 실질적인 의미 중의성 해소를 위해서는 사전과 코퍼스, 시소러스 등 다양한 지식원을 함께 이용한 학습이 필요할 것이다.

표제어	품사	의미 정의
으로	조사	a.수단, 방법을 나타냄
		b.재료를 나타냄
		c.방향을 나타냄
		d.변화를 나타냄
		e.자격을 나타냄
		f.원인을 나타냄
		g.시간을 나타냄

표 6. 조사 '으로'에 대한 사전 내용

감사의 글

본 연구에서 중요한 부분을 차지하는 전자화된

4) [6]에서는 Wall Street Journal 중 5개의 기사를 선택하여 실질어만을 대상으로 Brill의 태거를 이용하여 품사 태깅한 후 Brill의 태거의 품사 태그와 LDOCE의 품사 태그는 수동으로 매핑하였다.

그랜드 사전의 이용에 도움을 주신 한국과학기술원 전산학과 최기선 교수님께 감사드립니다.

참고문헌

[1] I. Dagan and A. Itai. Word Sense Disambiguation using a second language monolingual corpus. *Computational Linguistics*, Vol. 20, No. 4, pp. 563-596, 1995.

[2] W. Gale, K. Church and D. Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and humanities*, 26, pp. 415-439, 1993.

[3] Y. Karov and S. Edelman. Learning similarity-based word sense disambiguation from sparse data. In *Proceedings of the Workshop on Very Large Corpora*, pp. 42-55, 1996.

[4] A. K. Luk. Statistical sense disambiguation with relatively small corpora using dictionary definitions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 181-188, 1995.

[5] P. S. Resnik. Selectional preference and sense disambiguation. In *ANLP Workshop, "Tagging Text with Lexical Semantics: Why, What, and How?"*, 1997.

[6] Y. Wilks and M Stevensen. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Journal of Natural Language Engineering*, Vol. 4, No. 3, 1997.

[7] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 189-196, 1995.

[8] G. Zipf. The meaning-frequency relationship of words. *Journal of General Psychology*, Vol. 3, pp. 251-256, 1945.

[9] 김재훈, 김길창, 한국어에서의 품사 부착 말뭉치 작성 요령 : Kaist 말뭉치. 기술보고서 CS/Tr-95-99, 한국과학기술원 전산학과, 1995.

[10] 김상형, 그랜드 국어 사전, 금성출판사, 1995.