사전 뜻풀이말에서 추출한 의미 정보에 기반한 의미 중의성 해결

허 정[0] 옥 철 영
orionhj@cic.ulsan.ac.kr okcy@uou.ulsan.ac.kr
울산대학교. 컴퓨터정보통신공학과

Word-Sense Disambiguation based on Semantic Informations extracted from Definitions in Dictionary

Jeong Hur, Cheol-Young Ock
Dept. of Computer Engineering and Information Technology.
University of Ulsan

요 약

본 연구에서는 사전의 뜻풀이말에서 의미 정보를 추출하고, 이 의미 정보를 확률 통계적 방법에 적용하여 의미 중의성을 해결하는 모델을 제안한다. 사전의 뜻풀이말에 동형이의어를 포함하고 있는 표제어와 뜻풀이말을 구성하는 보통 명사, 형용사와 동사를 의미 정보로 추출한다.

비교적 중의성이 자주 발생하는 9개의 동형이의어 명사를 대상으로 실험하였다. 학습에 이용된 데이터로 정확률을 실험하는 내부 실험의 결과, 체언류(보통 명사)와 용언류(동사, 형용사)의 가중치를 0.9/0.1로 주는 것이 가장 정확률이 높았다. 외부 실험은 국어 정보베이스와 ETRI 코퍼스를 이용하여 1,796문장을 실험하였는데, 평균 79.73%의 정확률을 보였다.

1. 서 론

의미 중의성 해결은 문맥 내에 출현하는 단어가 둘 이상의 의미를 지닐 때, 의미들 중 문맥상 옳은 하나의 의미를 분별하는 것으로, 자연 언어 처리의 가장 힘든 요인 중의 하나이다. 의미 중의성이 해결되면 기계 번역에서 올바른 대역어를 선정할 수 있으며, 정보 검색에서의 정확률을 크게 향상시킬 수 있다.

지금까지의 형태소 분석이나 구문 분석은 어느 정도의 성과를 거두고 있으나, 담화 분석에 대한 연구가 활발해지면서, 의미 중의성 해결의 중요성이 부각되고 있다.

의미 중의성 해결을 위한 연구는 학습 데이터의 형태에 따라서 사전을 이용하는 방법과

코퍼스를 이용하는 방법으로 분류할 수 있고, 방법론에 따라서 규칙을 이용한 방법, 확률 통계를 이용하는 방법과 의미 계층 구조를 이용하는 방법으로 분류할 수 있다.

사전을 이용하는 방법은 언어의 동적인 특성을 반영하지 못하는 단점이 있으나, 모든 단어를 의미에 따라 단어의 정보를 따로 기술하고 있기 때문에 의미 정보를 추출하기에 쉽다는 장점이 있다. 코퍼스를 이용한 의미 중의성 해결을 위해서는 대량의 의미 부착 코퍼스가 필요한데, 신뢰성이 보장된 이용 가능한 의미 부착 코퍼스를 구하기가 힘들고, 코퍼스를 구축하기 위해서는 비용이 많이 드는 단점이 있다. 그러나, 언어의 동적 특성을 잘 반영하는 장점 때문에 많이 사용되고 있다. 확률 통계를 이용한 연구는 자료 부족 문제가 발생하는 단점이 있으나, 어휘의 불규칙적인 특성을 잘 반영하는 방법이라는 점에서 많이 이용되고 있다. 의미 계층 구조를 이용한 연구들은 주로 Roget thesaurus나 Wordnet을 이용하는데, 의미별로 단어들을 잘 클러스터링하고 있어서 의미 중의성 해결을 위해 가장 활발히 이용되는 자원이다. 그러나, 영어권의 언어이므로 한국어에 적용하는 데에는 한계가 있다.

이용되는 학습 데이터의 형태에 따라서 보면 서로의 단점을 보완하기 위해 사전과 코퍼스를 병행하는 연구가 활발히 진행되고 있다.[5, 6, 7, 8] 방법론적인 측면에서는 공기 정보를 이용한 확률 통계와 의미 계층 구조를 이용하는 연구가 활발히 진행되고 있다.[3, 6]

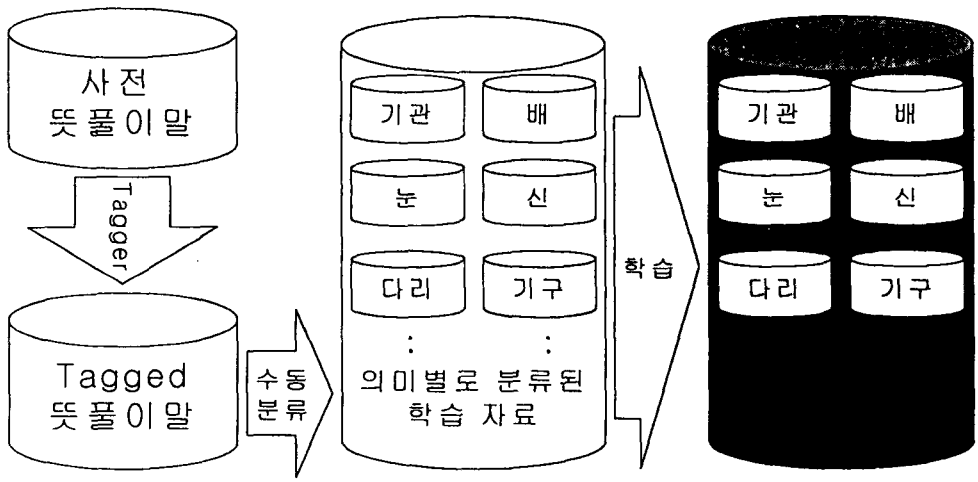
Yarowsky (1992)는 의미가 부착되지 않은 코퍼스와 시소러스(Roget thesaurus)를 이용한 통계 기반 의미 중의성 해결 방법을 제안하였다.[6] 이는 코퍼스로부터 시소러스의 의미 범주에 대한 단어 출현의 통계적 데이터를 구축하여, 시소러스의 의미 범주와 관련된 단어의 의미 중의성을 해결한다. 그러나, 상기의 방법론은 시소러스의 의미 범주를 일관성 있게 구축하기는 상당히 힘들다.

Alpha K, Luk (1995)는 코퍼스를 이용한 연구의 단점인 자료 부족 현상을 최소화하기 위해 LDOCE(Longman Dictionary of Contemporary English)의 사전에 정의된 1,792개의 통제 어휘(controlled vocabulary)를 기준으로 Brown 코퍼스에서 공기 정보에 대한 통계값을 추출하여 의미 중의성 해결 방법을 제안하였다.[7] 그러나, 상기의 방법론은 통제 어휘 자체가 의미 중의성을 지닌 동형이의어가 많음으로 인해서 한계점이 존재한다.

박영자 (1998)는 사전에서 추출한 의미 속성을 이용하여 명사 의미 클러스터링을 할 때, 의미 속성 추출에서 야기되는 의미 중의성 해결을 위해 사전의 의미 기술 문장을 대상으로 사용된 명사들의 공기 정보 통계값을 이용하여 의미 중의성을 해결하는 방법을 제안하였다.[4]

서희철 외 3 (1999)은 의미 계층 구조의 의미별 유사어들에 대한 유사어 벡터를 획득하여 다의어의 의미 중의성을 해결하는 방법을 제안하였다.[3]

조정미 (1998)는 지식 획득의 병목 문제(knowledge acquisition bottleneck problem)와 자료 부족 문제(data sparseness problem)를 효과적으로 처리하면서 의미 분별을 하는 방법을 제안하였다.[5] 지식 획득의 병목 문제를 해결하기 위해서 품사 부착 코퍼스로부터 선택 제한 지식을 추출하고, 자료 부족 문제를 해소하기 위해서 추출한 명사와 동사의 선택 제한 지식을 순환적으로 학습하는 방법을 사용하였다. 사전의 정보를 이용하여 의미 분별의 기준이 되는 의미 지시자를 선택하고 있다.



[그림 1] 의미 정보 추출 과정

<표 1> 풀이말 분류의 기본 형태

뜻풀이말의 맨 끝에 핵심어가 있는 형태로 핵심어가 표제어의 상의어가 된다.		
표제어	뜻풀이말	하의어 < 상의어
나룻배 유조차 :	나룻터에서 사람이나 짐 등을 건내 주는 배. 유조 시설을 갖춘 차. :	나룻배 < 배 유조차 < 차 :

1) UMRD : Ulsan university Machine Readable Dictionary

<표 2> 학습 대상 자료와 의미 정보의 예(배:교통수단)

학습 대상 자료의 예(배:교통수단)	학습된 1차 의미 정보의 예(배:교통수단)
강/NNG+에서/JKB 쓰/VV+는/ETM 배/NNG+./SF 강철/NNG+로/JKB 만들/VV+ㄴ/ETM 배/NNG+./SF 손님/NNG+을/JKO 태우/VV+는/ETM 배/NNG+./SF 돛/NNG+없/VA+는/ETM 작/VA+은/ETM 배/NNG+./SF 아주/MAG 크/VA+ㄴ/ETM 배/NNG+./SF 나무/NNG+로/JKB 만들/VV+ㄴ/ETM 배/NNG+./SF 나포/NNG+하/XSV+ㄴ/ETM 배/NNG+./SF 낚시질/NNG+에/JKB 쓰이/VV+는/ETM 배/NNG+./SF : :	[체언류] <<NNG>>:가능(1):가죽(2):강(2):강력(1):강물(1):강배(1):강선(1):강철(1):객선(2):거루(1):거룻배(1):거리(1):거북(1):거북선(1) ..... [용언류] <<VV>>:가(1):가라앉(1):가라앉히(1):갈추(8):갈추어지(1):건네(1):건네주(1):걸(1):깨뜨리(1):끌(1):나르(13):나아가(1):나타나(1) ..... <<VA>>:가깝(1):가볍(3):같(3):깊(1):낡(1):낮(1):넓(1):다르(1):무겁(1):반듯하(1):비(2):비슷하(1):빠르(4):어떠하(1) .....

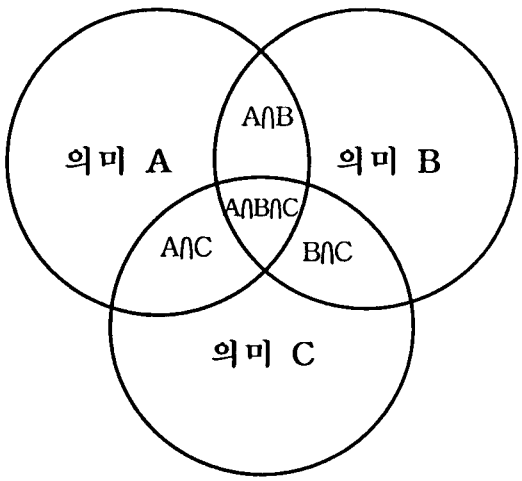
추출된 의미 정보의 형태는 품사 태그, 공기 관계에 있는 단어와 출현 횟수의 정보를 가진다.

<표 2>는 의미별로 분류된 학습 대상 자료와 의미 정보의 예를 보여 주고 있다.

<표 2>에서 알 수 있듯이 학습 대상 자료에서 의미 중의성 해결에 도움이 되지 않는 조사나 부사류 등의 정보는 의미 정보로 사용하지 않고, 보통 명사, 동사와 형용사만을 의미 정보로 사용한다. 그리고, 뜻풀이말의 표제어도 의미를 나타내는 하의어로써 의미정보의 자질을 갖추고 있으므로 의미 정보에 포함시킨다. <표 2>에 의해 학습된 의미 정보는 자료 부족 문제를 야기한다. 자료 부족 문제를 해결하기 위해 사전 뜻풀이말 중 의미 분별 대상 단어를 상의어로 갖지 않고 단지 뜻풀이말에 포함된 경우들을 [그림 1]의 방법을 반복하여 2차 의미 정보를 구축하는데, 뜻풀이말의 표제어는 의미 정보에 포함시키지 않는다. 1차 의미 정보와 2차 의미 정보의 합병을 통해 최종적인 의미 정보가 구축된다.

3. 의미 정보를 이용한 의미 중의성 해결

의미 정보는 [그림 2]와 같은 집합의 형태로 표현 할 수 있다.



[그림 2] 의미 A,B,C를 가지는 단어 W의 의미 정보 집합 관계

[그림 2]와 같이 의미 정보들 사이에는 의미들 간에 중복되는 교집합 부류의 정보(A ∩ B, B ∩ C, A ∩ C, A ∩ B ∩ C) 들이 있다. 교집합 부류의 정보들에 대한 자질값을 부여하는 방법에 따라 의미 중의성 해결에 중요한 영향을 미칠 수 있다. 그러나, 본 연구에서는 각각의 의미에 대해 개별적으로 의미 정보를 가지고 있고, 교집합 부류의 의미 정보들도 의미별로 다른 빈도를 가지고 있으므로, 교집합 부류의 의미 정보에 대해 고려할 필요가 없다.

의미 중의성 해결은 <표 3>의 수식 (1)을 이용하여 이루어진다.

$Rel(C, S_i)$는 문장 C 과 의미 S_i의 관련성을 나타낸다. $Rel(C, S_i)$의 의미 자질값 중 최대인 값을 가지는 의미를 선택하여 의미 중의성을 해결한다. $Rel(C, S_i)$는 수식 (2)에 의해서 구해진다. $Noun(C, S_i)$는 문장 C 에서 출현하는 체언류와 의미 S_i의 관련성이고, $Verb(C, S_i)$는 문장 C 에서 출현하는 용언류와 의미 S_i의 관련성이다. 그리고 w_n 과 w_v는 체언류와의 관련성과 용언류와의 관련성을 수식에 적용할 가중치이다. 가중치는 내부 실험에 의해서 구해진다. w_n 과 w_v의 합은 1 이다.

$Noun(C, S_i)$와 $Verb(C, S_i)$는 수식 (3)과 수식 (4)에 의해서 구해진다. $P(W_j \mid S_i)$는 S_i의 의미를 가진 단어가 발생했을 때, W_j의 단어가 발생할 확률을 나타낸다.

$C = C_n + C_v$, $C_n = \sum_j W_{nj}$ 이고, $C_v = \sum_j W_{vj}$ 이다. $Match_{(C_n, S_i)}$는 문장 C에서 의미 S_i와 공기 관계를 가지는 체언류의 개수이고, $Match_{(C_v, S_i)}$는 용언류의 개수이다.

4. 실험 결과

본 연구에서는 의미 중의성을 가지는 동형이의어 명사 9개를 선정하여 실험을 하였다. 의미 중의성 실험을 하기 전에 앞 단락에서 언급된 w_n 과 w_v의 가중치를 결정하기 위해 내부 실험을 하였다. <표 4>은 실험 대상 명사와 각각의 의미이고, <표 5>와 [그림 3]은 내부 실험 결과이다.

<표 3> 연구에 사용된 수식

$$WSD(C, S_i) = \arg MAX_{S_i} Rel(C, S_i) \quad \text{수식(1)}$$

$$Rel(C, S_i) = w_n \times Noun(C, S_i) + w_v \times Verb(C, S_i) \quad \text{수식(2)}$$

$$Noun(C, S_i) = Match_{(C_n, S_i)} \times \sum_j P(W_{nj} \mid S_i) \quad \text{수식(3)}$$

$$Verb(C, S_i) = Match_{(C_v, S_i)} \times \sum_j P(W_{vj} \mid S_i) \quad \text{수식(4)}$$

<표 4> 실험 대상 단어, 의미, 학습 데이터 수

단어	의미	학습 데이터 수 (단위:sentence)	단어	의미	학습 데이터 수 (단위:sentence)
기관*1	몸(器官)	212	기구	장치(機具)	472
	조직(機關)	453		조직(機構)	78
	장치(機關)	112	다리	교각(橋脚)	81
병	그릇(甁)	48		발(下肢)	326
	병사(兵)	3	배	과일(梨)	24
	질병(病)	880		운송수단(船)	513
눈	신체부위(目)	553		신체부위(腹)	322
	식물(木)	31	비	청소도구	29
	기상현상(雪)	181		기상현상(雨)	288
차	운송수단(車)	107		비석(碑)	21
	음료(茶)	56		비율(比)	70
	차이(差)	61	신	신발	89
				종교(神)	236

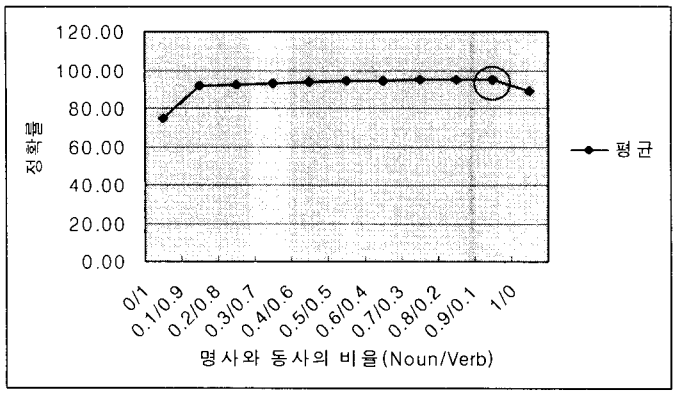
*1 : 조직과 장치의 의미가 다의어이므로 한자가 동일함.

<표 5> 내부 실험 결과

(단위 : %)

비율*1	기관	병	눈	차	기구	다리	배	비	신	평균
0/1	55.47	85.61	77.65	79.91	72.73	85.51	73.11	83.09	73.23	75.36
0.1/0.9	88.42	96.56	89.54	97.32	92.73	93.61	91.62	93.63	96.61	92.69
0.2/0.8	91.38	96.77	90.2	97.32	93.27	93.61	91.50	93.87	96.61	93.31
0.3/0.7	93.44	96.89	90.85	97.77	93.64	93.61	92.08	94.36	96.61	93.93
0.4/0.6	94.08	96.89	91.11	97.77	94	93.61	92.78	95.1	96.61	94.27
0.5/0.5	94.73	**96.99**	91.89	97.77	94.37	94.84	93.36	95.35	**96.93**	94.77
0.6/0.4	95.88	96.89	92.94	98.21	94.73	94.84	93.71	95.83	**96.93**	95.22
0.7/0.3	96.26	96.78	93.46	98.21	95.27	94.84	**94.06**	**97.55**	96.61	95.57
0.8/0.2	96.65	96.78	**93.72**	98.21	96.18	**95.34**	93.55	**97.55**	96	95.67
0.9/0.1	97.17	96.78	93.33	**98.22**	**97.46**	**95.34**	93.80	**97.55**	96.92	**95.92**
1/0	**97.81**	89.69	78.69	91.97	95.64	87.47	87.78	86.77	92.62	89.47

*1 : 비율은 체언류(w_n)/용언류(w_v)임.



[그림 3] 내부 실험 결과 그래프

[그림 3]을 보면 용언류의 공기 정보만으로 실험을 했을 때는 정확률이 평균 75.36%인 반면, 체언류만을 이용한 실험에서는 정확률이 평균 89.47%였다. 그리고, 체언류와 용언류의 가중치를 0.9/0.1로 하였을 때, 평균 92.69%로 용언류만을 이용한 실험보다 17.33%의 정확률 향상을 보였다. 위의 결과는 다음 세가지의 이유로 분석될 수 있다.

첫째, 용언류의 의미 정보 집합이 체언류의 의미 정보 집합보다 작음으로 해서 자료 부족 현상이 체언류보다 심하다고 할 수 있다.

둘째, 용언류의 의미 정보 집합이 체언류의 의미 정보 집합에 비해 교집합에 해당하는 부분이 크다.

셋째, 동사와 형용사는 대부분 인접한 단어들과 공기 관계를 밀접하게 이루고 있다. 그러므로, 의미 분별하고자 하는 단어가 포함된 문장에서 의미 분별을 할 때, 의미 분별의 대상이 되는 단어와의 거리에 따라 weight을 주는 것이 고려가 되어야 한다. 그러나, 본 연구에서 이것을 고려하지 않았다.

<표 6> 실험 결과(국어 정보 베이스, ETRI 코퍼스)

단어	의미	학 습 데이터 수	정확률	단어	의미	학 습 데이터 수	정확률
기관	몸(器官)	17	88.24	기구	장치(機具)	24	75
	조직(機關)	185	90.81		조직(機構)	98	89.8
	장치(機關)	2	100	다리	교각(橋脚)	21	71.43
병	그릇(甁)	12	16.67		발(下肢)	58	84.48
	병사(兵)	0	0	배	과일(梨)	6	33.33
	질병(病)	151	85.43		운송수단(船)	92	75
눈	신체부위(目)	431	77.26		신체부위(腹)	50	62
	식물(木)	1	100	비	청소도구	1	0
	기상현상(雪)	79	78.48		기상현상(雨)	86	79.07
차	운송수단(車)	46	52.17		비석(碑)	10	30
	음료(茶)	39	46.15		비율(比)	1	0
	차이(差)	12	83.33	신	신발	2	100
					종교(神)	372	86.83

내부 실험 결과 본 연구에서 사용할 체언류(w_n)와 용언류(w_v)의 가중치는 0.9/0.1로 선택하였다.

실험 데이터는 국어 정보 베이스(ver1.0)과 ETRI 의 품사 부착 코퍼스를 대상으로 선정하였다. <표 6>은 실험의 결과를 나타낸다.

비교적 낮은 정확률을 보이는 데이터들은 크게 두 종류의 오류에 의한 것으로 분석된다. 첫째, 학습데이터의 오류에 의한 부적당한 의미 정보, 둘째, 실험 데이터에서 의미 중의성 해결을 하고자하는 단어와의 거리에 관계없이 동일한 공기 관계를 가진다고 보고 실험한 것이 많은 오류를 야기했다.

6. 결 론

본 연구에서는 사전의 뜻풀이말만을 학습 데이터로 이용하여, 의미 중의성을 해결하는 방법을 제안하였다. 1,796문장을 실험하여 평균 79.73%의 정확률을 보였다. 실험 데이터의 부족으로 정확률에 절대적인 의미를 줄 수는 없지만, 사전 뜻풀이말을 이용한 단어의 의미 공기 정보가 의미 중의성 해결에 중요한 지표가 될 수 있다는 것을 확인할 수 있었다. <표 6>의 결과에서 의미수가 많은 단어와 적은 단어의 정확률을 비교해 보면 의미의 수와 정확률은 반비례의 관계에 있다는 것을 확인 할 수 있었다. 이것은 의미 정보의 집합에서 교집합의 부류가 어느 정도의 비중을 차지하는가와 직접 관련되는 것으로 의미수가 많으면 대체로 교집합의 부류가 상대적으로 많은 비중을 차지하는 것과 관련되는 것이다.

의미 중의성 해결의 향상을 위해 의미 계층망을 이용한 의미 정보의 확장에 대한 연구가 진행되어야 할 것이고, 사전에서 의미 정보를 추출할 때, 구문 구조의 활용을 통해 어떻게 정확한 의미 정보를 추출할 것인가에 대한 연구도 진행되어야 할 것이다.

본 연구에서는 문장내의 모든 단어를 동일한 공기 관계를 가진다는 가정 하에 동일하게 취급하였으나, 이로 인해 많은 오류가 발생하였다. 이러한 문제를 해결하기 위해서는 의미 분별하고자 하는 문맥 구조 관계에 따라 weight를 어떻게 줄 것인가에 대한 연구도 진행되어야 할 것이다.

참 고 문 헌

[1] 조평옥, 옥철영. "한국어 명사 의미 계층 구조 구축", 제 9 회 한글 및 한국어 정보 처리 학술대회 발표 논문 pp.129~135. 1997.

[2] 이수광, 조평옥, 안미정, 옥철영, 박재득, 박동인. "의미속성에 기반한 한국어 명사 의미 TAG에 관한 연구", 제 10 회 한글 및 한국어 정보 처리 학술대회 발표 논문 pp.412~418. 1998.

[3] 서희철, 이호, 백대호, 임해창. "유사어를 이용한 단어 의미 중의성 해결", 제 11 회 한글 및 한국어 정보 처리 학술대회 발표 논문 pp.304~309. 1999.

[4] 박영자. "사전을 이용한 단어 의미 자동 클러스터링.: 유전자 알고리즘 접근법." 연세대학교 컴퓨터과학과. Ph.D. thesis. 1998.

[5] 조정미. "코퍼스와 사전을 이용하 동사 의미 분별" 한국과학기술원. Ph.D. thesis. 1998.

[6] Yarowsky, D. "Word-sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora." In Proceedings of COLING92, pp.454~460. 1992

[7] Alpha K, Luk "Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions" 33rd Annual Meeting of the ACL. pp.181~188. 1995.

[8] William B. Dolan "Word Sense Ambiguation : Clustering Related Senses" In Proceedings of COLING94, 1994