한국어 어휘의미망을 이용한 비감독 어의 중의성 해소 방법의 성능 향상[1)]

권순호, 김민호, 권혁철
부산대학교 컴퓨터공학과
e-mail:soonhok7@pusan.ac.kr, karma@pusan.ac.kr, hckwon@pusan.ac.kr

An Enhanced Method for Unsupervised Word Sense Disambiguation using Korean WordNet

Soonho Kwon, Minho Kim, Hyuk-Chul Kwon
Dept of Computer Science & Engineering, Pusan National University

요 약

자연언어처리에서 어의 중의성 해소(word sense disambiguation)는 어휘의 의미를 정확하게 파악하는 기술로 기계번역, 정보검색과 같은 여러 응용 분야에서 중요한 역할을 한다.

본 논문에서는 한국어 어휘의미망(Korlex)을 이용한 비감독 어의 중의성 해소 방법을 제안한다. 의미 미부착 말뭉치에서 추출한 통계 정보와 한국어 어휘의미망의 관계어 정보를 이용함으로써 자료 부족 문제를 완화하였다. 또한, 중의성 어휘와 공기어휘 간의 거리 가중치, 의미별 사용 정보 가중치를 사용하여 언어적인 특징을 고려하여 본 논문의 기반이 되는 PNUWSD 시스템보다 성능을 향상하였다.

본 논문에서 제안하는 어의 중의성 해소 방법의 평가를 위해 SENSEVAL-2[2)] 한국어 데이터를 이용하였다. 중의성 어휘의 의미별 관계어와 지역 문맥 내 공기어휘 간의 카이제곱을 이용하였을 때 68.1%의 정확도를 보였고, 중의성 어휘와 공기어휘 간의 거리 가중치와 의미별 사용 정보 가중치를 사용하였을 때 76.9% 정확도를 보여 기존의 방법보다 정확도를 향상하였다.

1. 서론

자연언어에서는 하나의 단어가 여러 가지 의미로 해석될 수가 있다. 예를 들어, “a) 나는 요즘 밤마다 꿈을 꾼다.” “b) 이번 주말 온 가족이 밤을 따러 갔다.” a), b) 두 문장에서 ‘밤’은 다른 의미로 해석된다. a)는 저녁 어두운 뒤부터 새벽 밝기까지의 동안의 의미로 사용되고, b)는 밤나무 열매의 의미로 사용된다. 인간은 이러한 단어가 문장에 사용되었을 때, 직관적으로 어떤 의미인지 쉽게 알 수 있지만, 컴퓨터는 의미를 쉽게 구분하지 못한다. 따라서, 다수의 다른 뜻으로 쓰이는 어휘의 의미(이하 어의)를 정확하게 구분하는 방법이 필요하다. 이러한 방법을 어의 중의성 해소(word sense disambiguation)라고 하며, 기계번역, 정보검색과 같은 여러 응용 분야에서 중요한 역할을 한다.

어의 중의성 해소 방법은 기계 가독형 사전 등의 지식베이스를 이용한 지식 기반 어의 중의성 해소(knowledge based WSD), 대량의 말뭉치에서 추출한 통계적인 정보를 이용하는 말뭉치 기반 어의 중의성 해소(corpus based WSD), 말뭉치와 지식베이스를 함께 사용하는 혼합형 어의 중의성 해소(hybrid WSD) 등 크게 세 가지 유형으로 분류할 수 있다. 다시 말뭉치 기반 어의 중의성 해소는 의미 부착 말뭉치를 이용한 감독 중의성 해소(supervised disambiguation)와 의미 미부착 말뭉치를 이용한 비감독 중의성 해소(unsupervised disambiguation)로 분류할 수 있다.

본 논문에서는 한국어 어휘의미망(KorLex)[4]과 의미 미부착 말뭉치를 이용한 비감독 어의 중의성 해소 방법을 제안한다. 한국어 어휘의미망에서 중의성 어휘의 의미별 관계어와 중의성 어휘의 주위 문맥에 나타나는 어휘 간의 카이제곱검정(χ^2-test)에 의한 독립성 검정을 통해 어의 중의성을 해소한다. 또한, 언어적인 특징을 반영하기 위해, 중의성 어휘와 공기어휘 간의 거리 가중치, 의미별 사용 정보 가중치를 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 본 논문에서 제안하는 어의 중의성 해소 방법을 자세히 설명하며, 4장에서는 3장에서 제안한 방법의 성능을 평가한다. 마지막으로 5장에서는 결론 및 향후 연구 과제를 논하고자 한다.

2. 관련 연구

어의 중의성 해소 방법은 지식 기반 어의 중의성 해

1) 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 자원을 받아 수행된 연구임(No, 2010-0028784)

2) ACL SIGLEX와 EURALEX의 후원 하에 개최된 어휘 의미 중의성 해소 기술 평가 대회로 1998년 이래 3년마다 개최되고 있음.

소, 말뭉치 기반 어의 중의성 해소, 말뭉치와 지식베이스를 함께 사용하는 혼합형 어의 중의성 해소 등 크게 세 가지 유형으로 분류할 수 있다.

지식 기반 어의 중의성 해소 방법 중 가장 대표적인 방법은 Lesk[9]의 연구를 들 수 있다. Lesk는 기계 가독형 사전(MRD)에 나타난 중의성 어휘의 뜻풀이에 나타난 어휘와 중의성 어휘의 주위 문맥에 나타난 어휘 간의 중복치를 계산하여 중의성 어휘의 의미를 구분한다. Timothy[6]는 MRD에 나타난 중의성 어휘의 뜻풀이와 의미 간 연결 정보를 이용하여 중의성 어휘의 주위 문맥에 나타난 어휘와 간의 유사도를 다이스 계수(Dice coefficient)로 계산하여 중의성 어휘의 의미를 결정한다. Denis[7]는 Wikipedia의 어휘 간 연결 정보를 분석하고 그 연결 정보를 다이스 계수를 사용하여 중의성 어휘의 의미를 분류하였다. 지식 기반 어의 중의성 해소는 자료 부족 문제로 인해 말뭉치 기반 어의 중의성 해소와 비교하여 성능이 좋지 못하다.

말뭉치 기반 어의 중의성 해소 방법은 감독 중의성 해소와 비감독 중의성 해소로 분류할 수 있다. 감독 중의성 해소는 의미 부착 말뭉치를 이용하여 생성한 분류자(classifier)가 대상 어휘의 의미를 분류하는 것이다. Lucia[8]는 SENSEVAL-3 학습 데이터를 ILP(Inductive Logic Programming)로 학습한 모델을 이용하였다. Zhi Zhong[11]은 OntoNote[3] section 02-21의 데이터와 SemCor[4] 데이터를 SVM(Support Vector Machine)으로 학습한 모델을 이용하였다. 비감독 중의성 해소는 의미를 부착하지 않은 말뭉치에서 단어의 공기정보를 이용하여 중의성 어휘의 의미에 따라 클러스터를 생성하여 중의성 어휘의 의미를 선택하는 것이다. Schutze[10]는 의미 태깅이 되어 있지 않은 문맥을 벡터 공간으로 표현하고 이들 간의 유사도를 기반으로 EM(Expectation Maximization) 알고리즘을 사용하여 클러스터를 생성하는 모델을 제안하였다. 감독 중의성 해소가 비감독 중의성 해소보다 성능이 좋게 나타나지만, 대규모 의미 부착 말뭉치가 필요하다.

최근에는 지식 기반 어의 중의성 해소와 말뭉치 기반 어의 중의성 해소의 단점을 극복하기 위해 지식베이스와 대규모 말뭉치를 함께 사용하는 혼합형 어의 중의성 해소가 활발히 연구되고 있다. 김은진[2]은 기계 가독형 사전을 이용하여 의미부착 되어 있지 않은 웹 문서를 의미부착 데이터로 자동 변환하는 방법을 제안하였다. 김민호[1]는 한국어 어휘의미망과 의미 미부착 말뭉치에 추출한 통계 정보를 이용하여 중의성 어휘가 가지는 각 의미와 지역문맥에 나타나는 어휘 간의 연관성을 분석하여 어의 중의성을 해소하는 방법을 제안하였다. 본 논문과 비교 대상이 되는 어의 중의성 해소 방법은 2008년 부산대에서 개발한 어의 중의성 해소 시스템이다. 본 논문에서는 [1]의 알고리즘을 토대로 성능을 향상시킬 수 있는 방법을 제안한다.

3) Penn Treebank에 문장 분리, 토큰 분리, 품사, 의미 등의 정보가 부착된 말뭉치.

4) 프린스턴대 WordNet 프로젝트 연구팀에서 만든 의미 부착 말뭉치.

3. 한국어 어휘의미망을 이용한 어의 중의성 해소

일반적으로 감독 중의성 해소가 비감독 중의성 해소보다 성능이 좋게 나타나지만, 대규모 의미 부착 말뭉치가 필요하다. 본 논문에서는 구축 비용이 큰 의미 부착 말뭉치를 사용하지 않고, 500만 어절 규모의 세종 형태 분석 말뭉치를 사용하였다. 또한, 한국어 어휘의미망을 활용하여 의미별 관계어 확장을 통해 더 풍부한 통계 정보를 이용한다.

한국어 어휘의미망(KorLex)은 WordNet을 참조 모델로 하여 구축되었으며, 명사, 동사, 형용사, 부사, 분류사로 구성되며, 약 13만 개의 신셋과 약 15만 개의 어의를 포함하고 있다[4]. 신셋(synonym set)은 동일한 어의를 가지는 동의어 집합을 의미한다. 한국어 어휘의미망에서 2개 이상의 신셋을 가지는 어휘를 중의성 어휘로 본다. 예를 들어, '사과'는 'appology'의 뜻을 가진 '사과1'과 'apple'의 뜻을 가진 '사과2'와 같이 두 가지 신셋을 가지는 중의성 어휘다. 이러한 중의성 어휘의 의미를 구분하기 위해 의미별 관계어를 이용한다. 한국어 어휘의미망의 계층 구조에서 관계어들은 같은 성격을 지닌다. 예를 들어, '사과(apple)'와 '배(pear)'는 '과일(fruit)'의 하위어로 '먹다(eat)', '맛있다(delicious)' 등과 연관성이 있다. 따라서, 중의성 어휘의 의미별 관계어와 지역문맥에 나타난 공기어휘와의 연관성을 파악하여 어의 중의성을 없앨 수 있다[1]. 중의성 어휘 w_{amb}의 의미가 s_k이고, 지역문맥에 나타나는 공기어휘 v_j라 할 때, 두 어휘의 카이제곱 $\chi^2(w_{amb}=s_k, v_j)$은 s_k의 관계어 r_i에 의해 수식 1로 계산된다.

$$\chi^2(w_{amb}=s_k, v_j)_1 \equiv \frac{\sum_{i=1}^{l}\chi^2(r_i, v_j)}{l} \quad (1)$$

수식 1을 이용하여 중의성 어휘와 지역문맥에 나타나는 공기어휘 간 독립성 검정을 수행하여 의미별 연관어휘 집합을 생성한다. 만약 두 가지 이상의 의미와 관련이 있으면 카이제곱 값이 가장 큰 것의 의미로 분류한다. 표 1은 '사과'의 의미별 연관어휘 집합의 일부이다.

<표 1> '사과'의 의미별 연관어휘 집합

순위	사과(apple)		사과(apology)	
	연관어휘	χ^2	연관어휘	χ^2
1	곶감	155167.2	국민	154443.5
2	오렌지	136362.2	시	103707.1
3	복숭아	64651.49	대학	91711.53
4	주스	58532.84	표명	76186.53
5	수박	56702.43	정부	57129.32
6	참외	46292.52	꿇다	14776.39
7	대추	41792.23	인수	10682.51
8	바나나	38775.88	치명	10429.07
9	꿀	34190.19	다정	8123.73
10	보장	27673.78	내밀다	7499.758

본 논문에서 제안하는 어의 중의성 해소 방법은 두 단계로 이루어진다. 먼저, 중의성 어휘의 의미별 연관어휘 집합과 지역문맥 내 나타난 공기어휘 집합의 중복치(overlap)를 계산하여 가장 높은 중복치를 가진 의미가 중의성 어휘의 의미가 된다. 부가적으로 중의성 어휘와 공기어휘 간의 거리에 따라 반비례하는 가중치를 적용한다. 김준수[3]는 150만 어절 규모의 세종 의미 태그 부착 말뭉치를 분석한 결과, 중의성 어휘와 인접한 위치에 있는 어휘일수록 중의성 어휘의 의미 결정에 많은 영향을 미친다는 사실을 알아냈다. 수식 2는 의미별 연관어휘 집합을 이용한 의미 결정 수식으로 $RS(s_k)$는 의미 s_k의 연관어휘 집합으로써, v_j가 $RS(s_k)$에 속하면 $\sigma(RS(s_k), v_j)=1$이고, 아니면 0이다. 그리고 수식 3은 수식 2에서 거리 가중치를 추가한 것으로 $d(w_{amb}, v_j)$는 중의성 어휘와 공기어휘 간의 거리를 나타낸다.

$$WSD_1(w_{amb}, c) = \arg\max_{s_k} \sum_{v_j \in c} \sigma(RS(s_k), v_j) \quad (2)$$

$$WSD_2(w_{amb}, c) = \arg\max_{s_k} \sum_{v_j \in c} \frac{\sigma(RS(s_k), v_j)}{d(w_{amb}, v_j)} \quad (3)$$

만약 중복치가 같아 수식 2 또는 3으로 어의 중의성 해소가 안 되는 경우 중의성 어휘의 의미별로 지역문맥 내 나타난 공기어휘와 카이제곱을 계산하여 그 값의 비중의 곱이 가장 높은 것으로 의미를 선택한다. 부가적으로 카이제곱 값에 의미별 사용 정보를 가중치로 이용한다. 허정[5]은 의미 부착 말뭉치에서 추출한 의미별 사용 비율 정보를 가중치로 이용하여 높은 성능 향상을 보였다. 의미별 사용 정보는 중의성 어휘의 의미별 연관어휘 집합과 중의성 어휘가 공기하는 빈도의 합으로 나타낸다. 수식 4는 의미별 사용 정보 가중치 $W(w_{amb}=s_k)$이다.

$$W(w_{amb}=s_k) = \sum_{w_i \in RS(s_k)} freq(w_{amb}=s_k, w_i) \quad (4)$$

수식 5는 의미별 사용 정보 가중치가 적용된 중의성 어휘와 공기어휘의 카이제곱이다.

$$\chi^2(w_{amb}=s_k, v_j)_2 \equiv \frac{W(w_{amb}=s_k)\sum_{i=1}^{l}\chi^2(r_i, v_j)}{l} \quad (5)$$

수식 6, 7은 중의성 어휘의 의미별로 지역문맥 내 나타난 공기어휘와 카이제곱 값의 비중을 이용한 의미 결정 수식이다. 카이제곱이 0이 나와 수식 1 또는 5의 값이 0이 되거나 무한대가 되는 것을 방지하기 위해, Good-Turing frequency estimation을 이용하여 unknown data의 빈도를 추정하였다. 수식 8, 9는 중의성 어휘 w_{amb}의 의미 중 s_k의 카이제곱을 제외한 모든 의미의 카이제곱 합을 나타낸다. 수식 7, 9는 각각 수식 6, 8에서 의미별 사용 정보 가중치를 추가한 것이다.

$$WSD_3(w_{amb}, c) = \arg\max_{s_k} \prod_{v_j \in c} \frac{\chi^2(w_{amb}=s_k, v_j)_1}{\chi^2(w_{amb} \neq s_k, v_j)_1} \quad (6)$$

$$WSD_4(w_{amb}, c) = \arg\max_{s_k} \prod_{v_j \in c} \frac{\chi^2(w_{amb}=s_k, v_j)_2}{\chi^2(w_{amb} \neq s_k, v_j)_2} \quad (7)$$

$$\chi^2(w_{amb} \neq s_k, v_j)_2 = \sum_{i=1}^{n}\chi^2(w_{amb}=s_i, v_j)_1 - \chi^2(w_{amb}=s_k, v_j)_1 \quad (8)$$

$$\chi^2(w_{amb} \neq s_k, v_j)_2 = \sum_{i=1}^{n}\chi^2(w_{amb}=s_i, v_j)_2 - \chi^2(w_{amb}=s_k, v_j)_2 \quad (9)$$

4. 실험 및 평가

본 논문에서 제안한 어의 중의성 해소 방법을 기존의 논문들과 비교 평가하기 위해 SENSEVAL-2의 한국어 학습 데이터를 대상으로 실험하였다. 어의 중의성 해소를 위해 중의성 어휘 주위 문맥에 나타난 공기어휘를 이용할 때, 문맥의 윈도우 사이즈를 고려하여야 한다. 윈도우 사이즈는 중의성 어휘를 기준으로 좌우의 어휘의 수를 의미하는 것으로, 윈도우 사이즈가 커짐에 따라 가파르게 정확도가 상승하다가 어느 이상 커지면 정확도의 변화가 거의 없어진다[1][3]. 본 논문에서는 통계 사전의 크기를 고려하여 윈도우 사이즈 5를 기본값으로 선택하였다.

표 2는 기본 알고리즘에 거리 가중치와 의미별 사용 정보 가중치를 적용한 알고리즘의 정확도를 비교한 것이다. 1) 거리 가중치를 적용하였을 때, 2) 의미별 사용 정보 가중치를 적용하였을 때, 그리고 3) 두 가중치를 모두 사용하였을 때 각각 3.1%, 7.4%, 8.8%의 정확도 향상을 보였다.

<표 2> 가중치 적용에 따른 정확도

어휘	정확도			
	기본 알고리즘 (수식 2+6)	가중치 1 (수식 3+6)	가중치 2 (수식 2+7)	가중치 1+2 (수식 3+7)
눈	87.2	88.0	86.5	88.0
손	93.9	93.9	97.0	97.7
말	39.0	39.0	51.0	47.0
바람	62.2	62.2	71.4	70.4
거리	51.1	56.5	58.0	63.4
자리	87.1	82.2	95.0	95.0
의사	40.6	46.1	59.4	61.8
목	92.0	93.0	96.0	96.0
점	65.7	71.7	70.7	73.7
밤	70.3	87.1	75.2	79.2
전체 정확도	68.1	71.2	75.5	76.9

표 3은 SENSEVAL-2 한국어 데이터로 평가한 시스템별 정확도이다. WAMID[5]는 '상호 정보량과 복합명사 의미사전에 기반한 동음이의어 중의성 해소' 모델로 최소한의 의미 부착 말뭉치만을 이용하는 감독 중의성 해소와 비감독 중의성 해소의 중간 모델이다. PNUWSD[1]는 '한

국어 어휘의미망에 기반을 둔 비감독 어의 중의성 해소' 모델로 본 논문의 기반이 되는 시스템이다.

<표 3> 시스템 간 정확도 비교

시스템	정확도
제안 알고리즘 (NEWPNUWSD)	76.9
PNUWSD	75.6
WAMID	68.0

본 논문에서 제안한 'NEWPNUWSD'는 비감독 어의 중의성 해소 방법임에도 감독 중의성 해소와 비감독 중의성 해소의 중간 모델인 'WAMID'에 비해 높은 성능을 보였다. 또한, 본 논문에서 사용한 말뭉치와 지식베이스가 같은 'PNUWSD'에 비해서도 1.3%의 성능을 향상하였다.

4. 결론 및 향후 연구

본 논문에서는 [1]에서 제안한 한국어 어휘의미망을 이용한 비감독 어의 중의성 해소 방법을 토대로 성능을 향상할 수 있는 방법을 제안하였다. 한국어 어휘의미망에서 관계어는 서로 같은 성격을 지니기 때문에 중의성 어휘의 의미별 관계어와 지역문맥 내 공기어휘와의 연관성을 파악하여 의미별 연관단어 집합을 생성하고, 생성된 의미별 연관어휘 집합을 이용하여 중의성 어휘의 의미를 분류하였다. 그리고 중의성 어휘와 공기어휘의 거리에 따라 반비례하는 가중치를 적용하고, 의미별 연관어휘 집합을 이용하여 의미별 사용 정보를 가중치로 둠으로써 기존 방법보다 성능을 향상하였다.

본 연구에서 중의성 어휘의 관계어가 많이 쓰이지 않는 어휘로 구성될 때, 자료 부족 문제가 발생하여 성능이 나빠진다. 이처럼 통계 정보로 해결되지 않는 분석을 위해 선택 제약과 같은 전처리 작업을 수행하는 방안에 대한 추가 연구가 필요하다.

참고문헌

[1] 김민호, "한국어 어휘의미망에 기반을 둔 비감독 어의 중의성 해소", 부산대학교 컴퓨터공학과 공학석사 학위 논문, 2009.

[2] 김은진, 이수원, "의미그룹을 이용한 단어 중의성 해소", 정보과학회논문지: 컴퓨팅의 실제 및 레터, 제16권 제6호, pp. 747-751, 2010.

[3] 김준수, 최호섭, 옥철영, "가중치를 이용한 통계 기반 한국어 동형이의어 분별 모델", 정보과학회논문지: 소프트웨어 및 응용, 제30권 제7호, pp. 659-671, 2003.

[4] 윤애선, 황순희, 이은령, 권혁철, "한국어 어휘의미망 「KorLex 1.5」의 구축", 정보과학회논문지: 소프트웨어 및 응용, 제36권 제1호, pp. 92-108, 2009.

[5] 허정, 서희철, 장명길, "상호정보량과 복합명사 의미사전에 기반한 동음이의어 중의성 해소", 정보과학회논문지: 소프트웨어 및 응용, 제33권 제12호, pp. 1073-1089, 2006.

[6] Baldwin, T., Kim, S., Bond, F., Fujita, S., Martinez, D., and Tanaka, T., "A Reexamination of MRD-Based Word Sense Disambiguation.", ACM Transactions on Asian Language Information Processing (TALIP), vol. 9, no. 1, pp. 1-21, 2010.

[7] Denis Turdakov and Pavel Velikhov, "Semantic Relatedness Metric for Wikipedia Concepts Based on Link Analysis and its Application to Word Sense Disambiguation", Proceedings of Colloquium on Databases and Information Systems, 2008.

[8] Lucia Specia, Ashwin Srinivasan, Ganesh Ramakrishnan and Maria das Graças Volpe Nunes, "Word Sense Disambiguation Using Inductive Logic Programming", Inductive Logic Programming, pp. 409-423, 2007.

[9] Michael Lesk, "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream cone", Proceedings of the 5th annual international conference on Systems documentation, pp. 24-26, 1986.

[10] Schütze, H., "Automatic word sense discrimination", Computational Linguistics Archive, Vol.24, No.1, pp. 97-123, 1998.

[11] Zhong, Z., Ng, H. T., and Chan, Y. S., "Word sense disambiguation using OntoNotes: an empirical study." In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Annual Meeting of the ACL. Association for Computational Linguistics, pp. 1002-1010, 2008.