한국어 어휘의미망에 기반을 둔 어의 중의성 해소 시스템의 구현[1)]

김민호[O] 황명진 신종훈 권혁철
부산대학교 컴퓨터공학과
{karma, holgabun, luna12s, hckwon}@pusan.ac.kr

Implementation of Word Sense Disambiguation System based on Korean WordNet

Minho Kim[O] Myeong-jin Hwang Jong-hun Shin Hyuk-Chul Kwon
Dept. Computer Science and Engineering Pusan National University

요 약

자연언어처리에서 어휘의 의미를 구분하는 것은 기계번역이나 정보검색과 같은 여러 응용 분야에서 매우 중요한 역할을 한다. 국내에서도 여러 어의 중의성 해소 시스템이 소개되었으나 대부분 시스템이 의미 부착 말뭉치를 이용한 감독 학습 방식을 기반으로 두고 있다.

본 논문은 한국어 어휘의미망을 이용한 비감독 어의 중의성 해소 시스템을 소개한다. 일반적으로 감독 어의 중의성 해소 시스템은 비감독 어의 중의성 해소 시스템보다 성능은 좋으나 대규모의 의미 부착 말뭉치가 있어야 한다. 그러나 본 시스템은 한국어 어휘의미망과 의미 미부착 말뭉치에서 추출한 어휘 통계 정보를 이용해, 의미 부착 말뭉치에서 추출한 의미별 통계 정보를 이용하는 감독 중의성 해소 방법과 같은 효과를 낸다.

본 시스템과 타 시스템의 성능 비교를 위해 'SENSEVAL-2' 평가 대회의 한국어 평가 데이터를 이용하였다. 실험 결과는 추출된 통계 정보를 바탕으로 우도비를 이용하였을 때 정확도 72.09%, 관계어 가중치를 추가로 이용하였을 때 정확도 77.02%로 감독 중의성 해소 시스템보다 높은 성능을 보였다.

키워드: 한국어 어휘의미망(KorLex), 어의 중의성 해소, 우도비

1. 서 론

자연언어처리에서 어휘의 의미(word sense, 이하 어의)를 구분하는 것은 기계번역이나 정보검색과 같은 여러 응용 분야에서 매우 중요한 역할을 한다[1]. 기계번역 시스템을 이용하여 '나는 맛있는 사과를 좋아한다.'라는 문장을 영어로 번역할 때, 사과의 의미를 'apology'와 'apple'로 구분하여야 시스템은 주어진 문장을 정확한 문장으로 번역할 것이다. 이처럼 다수의 다른 뜻으로 쓰이는 어휘의 의미를 정확하게 구분하는 것을 어의 중의성 해소(word sense disambiguation)라고 한다.

어의 중의성 해소를 위해서 다음과 같은 두 가지 기본적인 접근 방법이 있다. 하나는 선택 제약(selectional restriction)을 이용하는 것이고, 다른 하나는 통계적인 정보를 이용하는 것이다[2].

선택 제약이란 문장 안에서 두 어휘가 공기(共起)할 때 나타나는 제약이다. 예를 들어, '죽다'라는 단어는 그 쓰임이 '생물'에 한하고 '사망하다'라는 단어는 그 쓰임이 '사람'에 한하는 것처럼 두 어휘가 의존 관계를 맺을 때 의미상의 모순이 없도록 하는 제약이다. 선택 제약에 어긋나는 어휘의 의미를 제거함으로써 의미 분석 시 발생하는 어의 중의성을 줄일 수 있다.

또 다른 방법은 대량의 말뭉치에서 추출한 통계적인 정보를 이용하는 것이다. 이 방법에서 어의 중의성 문제는 기계학습에서의 통계적 분류 문제로 단순화되어 전통적인 여러 기계학습 기법(사례 기반 학습, 결정 트리, 베이지안 분류 등)을 적용하여 해결된다. 기계학습을 통한 어의 중의성 해소는 학습을 위하여 개별 의미를 부착한 어휘들로 이루어진 말뭉치(이하 의미 부착 말뭉치)를 이용하는지에 따라 감독 중의성 해소(supervised disambiguation)와 비감독 중의성 해소(unsupervised disambiguation)로 나누어진다. 감독 중의성 해소는 의미 부착 말뭉치를 이용하여 생성한 분류자(classifier)가 대상 어휘의 의미를 분류하는 것이고, 비감독 중의성 해소는 의미를 부착하지 않은 말뭉치의 데이터를 의미에 따라 몇 개의 클러스터로 집단화한 후에 대상 어휘와 가장 가까운 클러스터의 의미를 대상 어휘의 어의로 부여하는 것이다.

본 논문에서는 우도비(likelihood ratio)[3]와 한국어 어휘의미망(KorLex)[4, 5]을 이용한 비감독 중의성 해소 시스템을 제안하고자 한다. 한국어 어휘의미망의 의미체계를 기준으로 어의 중의성을 가지는 어휘(이하 중의성 어휘)를 골라내고, 대상 어휘가 가지는 각 의미와 지역문맥(local context)에 나타나는 어휘 간의 우도비를 측정하여

1) 이 논문은 2008년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2007-000-20517-0)

어의 중의성을 해소한다.

본 논문은 다음과 같이 구성된다. 2장에서 어의 중의성 해소의 국내외 관련 연구에 대해서 정리하고, 3장에서는 본 시스템의 구조와 어의 중의성 해소 방법을 상세히 설명할 것이다. 4장은 실험 방법과 결과를 기술하고, 마지막으로 5장에서 결론 및 앞으로의 연구 방향에 대해 논할 것이다.

2. 관련 연구

서론에서 언급한 감독 중의성 해소와 비감독 중의성 해소는 중의성을 가진 각각의 어휘마다 분류자와 클러스터를 생성하여야 하기 때문에 많은 시간과 작업량을 요한다. 따라서, 대규모의 중의성 어휘를 처리하고자 사전 기반 어의 중의성 해소 기술을 중심으로 연구가 진행되었다.

Lesk[6]는 기계 가독형 사전에 나타난 대상 어휘의 뜻풀이에 나타난 어휘와 대상 어휘가 나온 특정 문맥 안의 어휘 간의 중복(overlap)을 이용하여 어의를 구분하였다. 그러나 어휘 간의 정확한 일치를 전제로 하기 때문에 자료 부족 문제를 가져왔다. Luk[7]는 이 문제를 최소화하고자 롱맨 현대 영어 사전에서 통제 어휘(common word)를 정의 개념으로 추출하고 이 정의 개념들에 대한 통계 정보를 브라운 말뭉치로부터 추출하여 중의성을 없애는 방법을 제안하였다. 이 밖에도 Lesk[6]의 알고리즘을 간략화하여 검색 공간을 줄인 모델과 이 알고리즘에 그래프에 적용된 page rank 기법을 결합한 모델도 제안되었다[8].

최근에는 WordNet[9]을 이용한 어의 중의성 해소 기술이 활발히 연구되고 있다. Resnik[10]는 WordNet의 IS-A 계층 관계에 있는 명사에 대해 의미 유사도(semantic similarity)를 측정하여 어의 중의성 해소에 이용하는 방법을 제안하였고, Mihalcea 외[11]는 두 어휘 간의 공기 통계 자료를 인터넷으로부터 얻고, WordNet을 통해 두 어휘 간의 의미 밀도(semantic density)를 측정하여 순위를 바탕으로 어의 중의성을 없애는 기술을 제안하였다. 이 밖에도 Ganesh[12] 외는 WordNet의 의미 풀이말(gloss)과 대상 어휘가 포함된 문맥의 어휘 간의 유사도를 코사인 유사도와 자카드 유사도를 이용하여 계산한 후 가장 유사도가 높은 풀이말의 신셋을 어휘의 의미로 결정하는 방법을 제안하였다.

국내에는 어의 중의성 해소 기술에 대한 연구가 아직 미흡한 실정이다. 이승우 외[13]는 한국어 명사의 중의성 해소를 위해 원시말뭉치로부터 국소 문맥을 추출하고 사전의 의미 풀이말로부터 공기 정보를 획득하여 국소 문맥 데이터베이스를 구축하고 의미 계층 구조로부터 의미 데이터베이스를 학습하는 모델을 제안하였다. Seo 외[14]는 불린 함수(=조건)와 부류(=의미, class)로 구성된 규칙으로 이루어진 결정 목록을 이용한 어의 중의성 해소 모델을 제안하였다. 허정 외[15]는 사전에 기반을 둔 Lesk 방법의 최대 단점인 자료 부족 문제를 완화하고자 한국어 명사 개념망(ETRINET), 복합명사 의미 분석 사전, 원시말뭉치로부터 추출한 상호정보량을 활용한 어의 중의성 해소 모델을 제안하였다.

그러나 상호정보량은 두 어휘의 독립성을 측정하는 데는 좋은 척도이나, 두 어휘의 의존성을 측정하는 데는 적당하지 않다. 또한, 이러한 통계적 접근법은 중의성 어휘의 의미별 출현 빈도가 필요하므로 대용량의 의미 부착 말뭉치가 있어야 한다.

본 시스템은 통계적 접근법을 이용한 어의 중의성 해소에서 발생하는 이 두 가지 문제를 해결하고자 우도비와 한국어 어휘의미망을 활용하였다.

3. 어의 중의성 해소

본 논문에서 제안하는 어의 중의성 해소 시스템은 우도비와 한국어 어휘의미망을 활용한다. 이번 장에서는 본 시스템의 전체 구조와 어의 중의성 해소 방법을 상세히 소개한다.

3.1 본 시스템의 구조와 처리 흐름

어의 중의성 해소 시스템은 그림 1과 같이 구성된다. 한국어 품사 태거를 이용하여 입력한 문장에 대해 품사 부착을 한 후 명사에 대해서만 한국어 어휘의미망 검색기[2]를 이용하여 의미를 검색한다. 한국어 어휘의미망 검색 결과에서 여러 개의 의미가 있는 명사는 동음이의어로 판단하여 어의 중의성 해소를 실행한다.

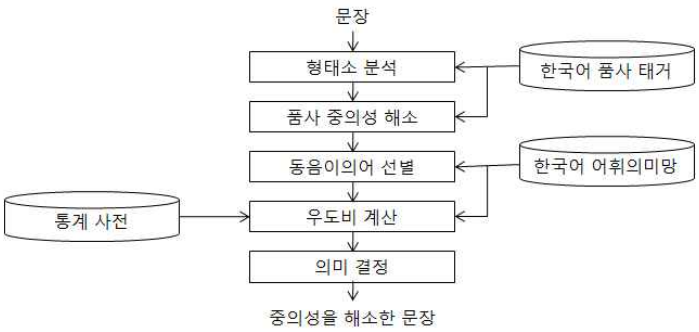

그림 133 어의 중의성 해소 시스템 구조와 처리 흐름

본 시스템에서 사용한 통계 사전은 세종 형태 분석 말뭉치[3]에서 명사, 동사, 형용사만을 대상으로 추출한 출현 빈도로 구성되어 있다. 통계 사전 구성에 사용된 세종 말뭉치는 품사 정보를 본 시스템에서 사용한 한국어 품사 태거의 품사 정보에 맞게 변환한 것으로 말뭉치 크기는 약 400만 어절이다.

3.2 우도비(Likelihood Ratio)

어의는 특정한 문맥 자질(contextual features)과 강하게 연관성을 가진다[16]. 본 시스템에서는 중의성 어휘의 각 의미와 주위 문맥에 나타난 단어들과의 의미적 연관성을 객관적으로 판단하고자 우도비를 이용하였다.

$$C = w_1, w_2, \cdots, w_{amb}, \cdots, w_{n-1}, w_n$$

---

2) 본 연구실에서 개발한 MS Jet DB(MS Access / ODBC)로 구성된 기존 한국어 어휘의미망을 개선한 검색 시스템으로 Pentium 4 2.9GHz에서 초당 5만 8천 단어를 검색할 수 있음.

3) 21세기 세종계획 1·2단계 사업성과물의 일부인 '21세기 세종계획 천만어절 색인 말뭉치'를 사용하였다.

C는 중의성 어휘 w_{amb}를 포함한 문맥이고, w는 문맥 내 공기어휘들이다. w_{amb}가 m개의 의미를 지닐 때, w_{amb}의 의미 $w_{amb_j}(1<j<m)$와 공기어휘 $w_i(1<i<n)$의 의존 관계에 대해 다음과 같은 두 가지 가설을 세웠다.

- 가설 1(H_1): $P(w_i|w_{amb_j}) = p = P(w_i|\neg w_{amb_j})$
- 가설 2(H_2): $P(w_i|w_{amb_j}) = p_1 \neq p_2 = P(w_i|\neg w_{amb_j})$

가설 1은 w_i의 출현과 w_{amb}가 의미 w_{amb_j}로 사용되는 것이 상호독립적이라는 것을 뜻하고, 가설 2는 w_i의 출현이 w_{amb}가 의미 w_{amb_j}로 사용되는 것에 종속적이라는 것을 뜻한다. c_1을 w_i의 출현 빈도, c_2를 w_{amb}가 w_{amb_j}로 사용된 빈도, c_{12}를 w_{amb}가 w_{amb_j}로 사용될 때 w_i의 출현 빈도라고 할 때, 확률 p는 최대우도추정에 의해 다음과 같이 구한다.

$$p = \frac{c_2}{N} \quad p_1 = \frac{c_{12}}{c_1} \quad p_2 = \frac{c_2 - c_{12}}{N - c_1} \tag{1}$$

이때, N은 원시말뭉치의 전체 어절 수이다.

우도비 λ는 전체 모수 공간상의 최대 우도 함수값에 대한 가설에 따라 표현되는 일부 공간상의 최대 우도 함수값의 비율로써 다음과 같이 표현된다.

$$\lambda = \frac{\max_{\omega \in \Omega_0} H(\omega; k)}{\max_{\omega \in \Omega} H(\omega; k)} \tag{2}$$

이때, ω는 모수 공간 Ω의 한 점이고 Ω_0는 가설에 따라 표현되는 일부 공간이다. 우도비의 중요한 특징은 $-2\log\lambda$가 점근적으로 x^2-분포를 따른다는 것이다. 특히, 이항분포에서 이 점근선은 매우 빠르게 접근하게 된다[3].

수식 (2)는 이항분포 $b(k;n,x)$에서 수식 (3)으로 변환될 수 있다.

$$\begin{aligned} \log\lambda &= \log\frac{L(H_1)}{L(H_2)} \\ &= \log\frac{b(c_{12},c_1,p)b(c_2-c_{12},N-c_1,p)}{b(c_{12},c_1,p_1)b(c_2-c_{12},N-c_1,p_2)} \end{aligned} \tag{3}$$

실제로 가설 1과 가설 2에서

$$L(H_1) = b(c_{12};\, c_1,p)b(c_2 - c_{12};\, N - c_1, p) \tag{4}$$
$$L(H_2) = b(c_{12};\, c_1,p_1)b(c_2 - c_{12};\, N - c_1, p_2) \tag{5}$$

이므로 수식 (3)은 수식 (4), (5)에 의해 다음과 같이 변환된다.

$$\begin{aligned} \log\lambda = {} & \log L(c_{12},c_1,p) + \log L(c_2-c_{12},N-c_1,p) \\ & - \log L(c_{12},c_1,p_1) - \log L(c_2-c_{12},N-c_1,p) \end{aligned} \tag{6}$$
$$L(k,n,x) = x^k(1-x)^{n-k} \tag{7}$$

위에서 언급했듯이 $-2\log\lambda$는 자유도가 1인 x^2-분포를 따르고, 유의수준이 0.005일 때의 임계값(critical value)은 7.88이다. 만약, w_i와 w_{amb_j}의 $-2\log\lambda$가 7.88보다 크다면 기각영역에 속하므로 가설 1은 기각되고, 반대로 $-2\log\lambda$가 7.88보다 작다면 채택영역에 속하므로 가설 1은 채택된다.

지역문맥 내의 어떤 w_i가 오직 하나의 의미 w_{amb_j}에 대해서만 기각영역에 속하는 $-2\log\lambda$을 가진다면, 이 w_i는 w_{amb}의 의미를 w_{amb_j}로 구분해주는 어휘가 된다. 반면에, 여러 개의 의미에 대해 기각영역에 속하는 $-2\log\lambda$을 가지거나, 모든 의미에 대해 채택영역에 속하는 $-2\log\lambda$을 가진다면 이 w_i는 w_{amb}의 의미를 구분하는 어휘가 아니다.

3.3 한국어 어휘의미망의 활용

일반적으로 감독 중의성 해소는 비감독 중의성 해소에 비해 높은 성능을 보이지만, 대규모의 의미 부착 말뭉치가 필요하다. 본 논문은, 구축에 큰 비용이 소요되는 의미 부착 말뭉치를 사용하지 않고, 어휘 의미 중의성 해소에 필요한 통계 정보를 의미 미부착 말뭉치와 어휘의미망을 이용해 구하였다. 또한, 어휘의미망을 활용하면 의미별 관계어 확장을 통해, 기존처럼 말뭉치에서 추출한 단순한 통계 정보보다 더 풍부한 통계 정보를 얻을 수 있는 이점이 있다.

본 논문에서 시스템은 2007년 11월에 공개된 한국어 어휘의미망 KorLex 1.5를 활용였다. 현재 KorLex 1.5는 명사, 동사, 형용사, 부사와 분류사로 구성되며, 약 13만 개의 신셋과 약 15만 개의 어의를 포함하고 있다[4].

3.3.1 형제어를 이용한 의미별 통계정보 획득

중의성 어휘의 각 의미와 공기어휘 간의 $-2\log\lambda$를 계산하려면 의미별 사용 빈도가 필요하다. 하지만, 한국어 어휘의미망에서 추출한 형제어 정보를 통해 의미별 통계정보가 없더라도 우도비를 이용한 어의 중의성 해소가 가능하다. 예를 들어 "나는 맛있는 사과를 좋아한다."라는 문장에서 '사과'는 한국어 어휘의미망에서 다음과 같은 두 가지 신셋을 가진다[그림 2].

- 사과1: 지은 죄나 잘못에 대하여 용서를 빔.
- 사과2: 사과나무의 열매

이 예문에서 '사과'와 공기하는 어휘는 '나', '맛있다', '좋아한다'이므로 '사과1'과 각 공기어휘와의 우도비를 계산하여, 공기어휘 중에 '사과1'의 의미를 구분해주는 어휘가 존재하는지 판단하여야 한다.

통계사전을 이용해 공기어휘의 사용 빈도는 쉽게 획득할 수 있지만, '사과1'과 '사과2'의 사용 빈도는 획득할 수 없다. 하지만, 한국어 어휘의미망에서 형제어는 서로

4) http://corpus.fr.pusan.ac.kr/korlex/

같은 성격을 지니기 때문에 '사과1'이 특정 어휘와 연관 관계가 있다면, '사과1'의 형제어 역시 그 어휘와 연관 관계가 있다. 따라서, '사과1'을 직접적으로 이용해 지역문맥 내의 공기 어휘와 연관성 여부를 판단하는 대신, '사과1'의 형제가 지역문맥 내의 공기 어휘와 연관성이 있

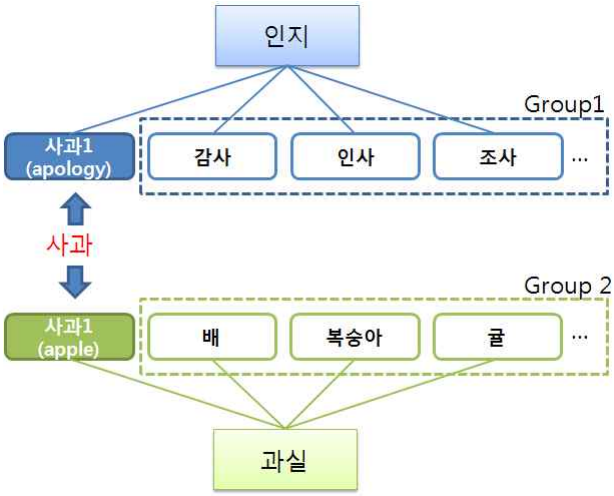

그림 134 '사과'의 형제어

는지를 판단하면 되므로, sw_k가 의미 w_{amb_j}의 형제어라고 할 때, 3.2의 수식 (1)에서 출현 빈도 c_2, c_{12}는 다음과 같이 구할 수 있다.

$$c_2 = \sum_{k=1}^{l} freq(sw_k),\ c_{12} = \sum_{k=1}^{l} freq(sw_k, w_i)$$

이때, 중의성이 없는 형제어만 이용하여 c_{12}를 계산함으로써 c_{12}의 신뢰도를 높인다. 만약, 중의성이 없는 형제어가 존재하지 않는다면 모든 형제어 빈도를 사용하되, 출현 빈도를 의미 분화 수로 나누어 계산에 이용한다.

.표 1은 위의 c_2, c_{12}를 이용하여 '사과'의 의미와 주위 문맥 내 공기한 어휘인 '나', '맛있다', '좋아한다'와의 우도비를 계산한 것이다.

표 12 '사과'의 의미와 문맥 내 공기 어휘들과의 우도비에 따른 연관성

의미	문맥 내 공기 어휘	$-2\log\lambda$
사과1 (apology)	나	78.7
	맛있다	0
	좋아한다	1.04
사과2 (apple)	나	64.8
	맛있다	37.05
	좋아한다	16.84

표 1에서 '나'는 모든 의미와 연관성을 지니므로 '사과'의 의미를 구분하는 어휘가 될 수 없다. 하지만, '맛있다'와 '좋아한다'는 모두 기각영역에서 속하는 값을 가지므로 '사과'의 의미를 '사과2'로 구분하는 어휘가 된다.

이처럼, 의미 미부착 말뭉치에서 추출한 통계 정보와 한국어 어휘의미망의 형제어 정보를 이용하여 공기 여부를 결정함으로써 대용량의 의미 부착 말뭉치 없이 어의 중의성을 해결할 수 있다.

3.4 관계어 빈도 가중치 적용

허정[15]은 의미 부착 말뭉치에서 추출한 의미별 사용 비율 정보를 가중치로 이용하여 높은 성능 향상을 보였다. 본 시스템에서는 의미별 사용 비율 정보를 이용하는 것과 같은 효과를 내고자 중의성 어휘 w_{amb}에 대해 다음과 같은 가정을 세웠다.

- 중의성 어휘의 관계어 빈도의 합에 대한 해당 의미별 관계어 빈도의 합의 비율은 의미별 사용 비율에 비례한다.

이 가정을 바탕으로 의미 w_{amb_j}에 대해 다음과 같이 가중치 W를 계산하였다.

$$W(w_{amb_j}) = \frac{\sum_{k=1}^{l} freq(w_{amb_j}, sw_k)}{\sum_{j=1}^{n}\sum_{k=1}^{l} freq(w_{amb_j}, sw_k)} \quad (7)$$

$$W(w_{amb_j}) = \frac{\sum_{k=1}^{l} freq(w_{amb_j}, cw_k)}{\sum_{j=1}^{n}\sum_{k=1}^{l} freq(w_{amb_j}, cw_k)} \quad (8)$$

수식 (7)은 형제어(sw) 빈도를 이용한 가중치이고, 수식 (8)은 자식어(cw) 빈도를 이용한 가중치이다. 이 가중치를 $-2\log\lambda$에 곱하여, 값을 보정한다.

4. 실험

본 논문에서는 평가를 위해 'SENSEVAL-2[5]'의 한국어 평가데이타를 이용하였다. SENSEVAL-2 한국어 데이타의 의미체계와 KorLex의 의미체계가 서로 일치하지 않기 때문에 부록 A를 바탕으로 의미체계를 KorLex에 매핑 하여 실험을 진행하였다.

4.1 윈도우 사이즈

어의 중의성 해소를 위해 주변에 함께 나타난 어휘를 이용할 때, 중의성 해소에 이용되는 앞/뒤 어절의 수(윈도우 사이즈)를 고려하여야 한다. 그림 3은 윈도우 사이즈에 따른 정확도 변화이다.

5) SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems

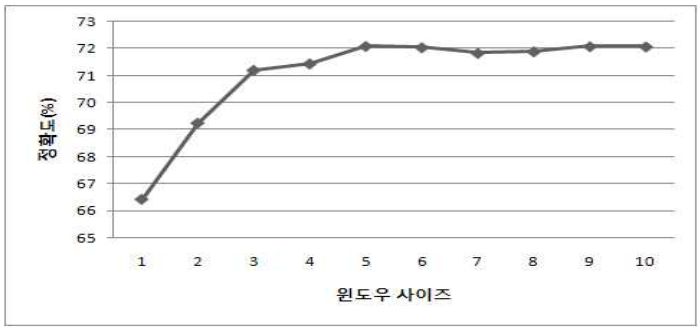

그림 135 윈도우 사이즈에 따른 정확도 변화

윈도우 사이즈 내에 의미를 구분하는 공기 어휘가 존재하지 않을 때에는 $-2\log\lambda$의 총합이 가장 큰 의미를 선택하였다.

그림 3에서 알 수 있듯이 중의성 해소에 이용되는 앞/뒤 어절의 수(윈도우 사이즈)가 1에서 5로 늘어날수록 정확도가 점점 높아지다가 6부터 거의 변화가 없는 것을 알 수 있다. 본 시스템에서는 통계 사전의 크기를 고려하여 윈도우 사이즈 5를 기본값으로 선택하였다. 그림 4는 윈도우 사이즈가 5일 때 SENSEVA-2 한국어 데이터에 나타난 어휘별 정확도이다.

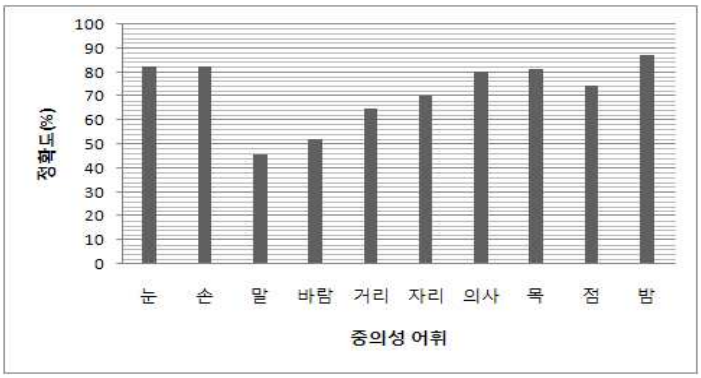

그림 136 중의성 어휘별 정확도 (윈도우 사이즈 = 5)

4.2 관계어 빈도 가중치 적용

표 2는 통계사전에서 추출한 의미별 관계어 빈도를 이용한 가중치를 적용하였을 때의 정확도이다. 형제어 가중치를 적용하였을 때와 자식어 가중치를 적용하였을 때 각각 2.96%, 4.93%의 정확도 향상을 보였다.

표 13 가중치 적용에 따른 정확도

어휘	정확도		
	가중치 미적용	형제어 가중치 적용	자식어 가중치 적용
눈	82.4	89.3	81.8
손	82.7	80.1	86.6
말	45.6	44.2	43.8
바람	52.1	48.4	46.7
거리	64.8	64.2	63.9
자리	70.3	91.1	92.8
의사	80.1	74.3	89.9
목	81.2	76.4	81.5
점	74.6	89.4	90.1
밤	87.1	93.1	93.1
전체 정확도	72.09	75.05	77.02

4.3 타 시스템과의 비교

표 3은 SENSEVAL-2 한국어 데이터로 평가한 시스템별 정확도이다. KUNLP[14]는 불린 함수(=조건)와 부류(=의미)로 구성된 규칙으로 이루어진 결정 목록을 이용한 감독 중의성 해소 시스템이다. WAMID[15]는 최소한의 의미 부착 말뭉치(의미별 사용 비율 정보, 복합명사 의미사전)만을 이용하는 감독 중의성 해소와 비감독 중의성 해소의 중간상에 있는 시스템이다. 반면에 본 논문에서 제안한 PNUNLP는 의미 미부착 말뭉치에서 추출한 통계정보만을 이용하는 비감독 중의성 해소 시스템이다.

표 14 시스템 간 정확도 비교

시스템	방법론	자원	평가데이타	정확도
KUNLP (2001)	의미 부착 말뭉치를 기반으로 학습한 결정 목록을 이용한 감독 중의성 해소 시스템.	SENSEVAL-2 학습데이타	SENSEVAL-2 평가데이타	69.8%
WAMID (2006)	복합명사 의미사전과 최소한의 의미 분석 말뭉치를 이용한 감독과 비감독 중의성 해소 중간상의 시스템.	ETRI 한국어 명사 개념망, 복합명사 의미분석 사전, 세종 의미 분석 말뭉치	SENSEVAL-2 학습데이타	68.04%
PNUNLP (2008)	한국어 어휘의미망과 의미 미부착 말뭉치를 이용한 비감독 중의성 해소 시스템.	한국어 어휘의미망, 세종 형태 분석 말뭉치	SENSEVAL-2 학습데이타	77.02%

표 3에서 KUNLP의 정확도는 SENSEVAL-2 한국어 데이터의 학습데이타를 기반으로 학습을 하고, 평가데이타에 대해 평가한 결과이다. 반면에 PNUNLP와 WAMID의 정확도는 학습데이타를 대상으로 한 결과라 절대적인 비교가 불가하지만, 비감독 중의성 해소 시스템인 PNUNLP가 감독 중의성 해소를 이용하는 다른 두 시스템과 비교하여 월등히 높은 성능을 보인다는 것을 알 수 있다.

SENSEVAL-2 한국어 데이터의 학습데이타는 중의성 어휘를 포함하는 1,729개의 문맥으로 구성된다. 따라서, 이를 사용하여 학습한 KUNLP와 WAMID의 결과는 자료부족문제 때문에 성능이 높지 않았을 것으로 판단된다. 그러나 PNUNLP는 한국어 어휘의미망과 의미 미부착 말뭉치에서 추출한 어휘 통계 정보를 이용해, 동일한 크기의 의미 부착 말뭉치에서 추출한 의미별 통계 정보를 이용하는 감독 중의성 해소 방법과 같은 효과를 낸다. 본 실험에서는 400만 어절의 의미 미부착 말뭉치에서 통계정보를 추출하였기 때문에 비교적 자료부족문제가 심각하지 않았다.

또한, WAMID는 어휘 간의 연관계수로써 상호 정보량을 이용하는데, 상호 정보량은 두 어휘의 독립성을 측정하는 데는 좋은 척도이지만, 두 어휘의 의존성을 측정하는 데는 적당하지 않은 것으로 알려졌다[17]. 반면에 PNUNLP는 두 어휘의 의존성을 측정하는 데 많이 이용되는 우도비를 이용하여 신뢰성을 더 높였다.

5. 결론 및 향후 연구

본 논문에서는 한국어 어휘의미망을 이용한 중의성 해소 기술을 제안하였다.

본 논문에서는 어휘 간의 연관 관계를 통계적으로 판단하고자 한국어 어휘의미망을 이용하여 의미별 관계어의 빈도를 통계 정보로 활용하였다. 이는 의미 부착 말뭉치의 의미별 통계 정보를 직접적으로 활용하는 것과 같은 효과를 보였을 뿐만 아니라, 오히려 한국어 어휘의미망을 이용한 어휘 확장을 통해 통계 정보를 더 풍부하게 함으로써 감독 중의성 해소 시스템보다 더 높은 성능을 보였다. 또한, 한국어 어휘의미망 상의 관계어의 빈도를 가중치로 활용하였다.

본 연구의 의의는 다음과 같다. 첫째, 큰 비용을 요구하는 의미 부착 말뭉치를 구축하지 않아도 의미 부착 말뭉치를 이용한 것과 같은 효과를 볼 수 있다는 것이다. 둘째, 중의성 어휘의 관계어 빈도의 합에 대한 해당 의미별 관계어 빈도의 합의 비율은 의미별 사용 비율에 비례한다는 가정을 실험을 통해 증명하였다는 것이다.

앞으로 연구되어야 할 것은 다음과 같다. 첫째, 다른 평가데이타를 이용하여 다양한 중의성 어휘에 대해 평가함으로써 시스템의 신뢰성을 더 높여야 할 것이다. 둘째, 통계정보와 어휘 간의 연관계수를 이용하여 어의를 통계적으로 구별해 줄 수 있는 유의미 영역의 어휘를 미리 구하여 사전으로 저장함으로써 어의 중의성 해소 처리 속도를 높여야 할 것이다.

참고문헌

[1] Nancy Ide, Jean Véronis, "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art," Computational Linguistics, Vol.24, No.1, pp.1-40, 1998.

[2] 김영택 외 공저, 자연언어처리, p.296, 생능출판사, 2001.

[3] Ted Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," Computational Linguistics, Vol.19, No.1, pp.61-74, 1993.

[4] 이은령, 윤애선, "피동 정보를 통한 한국어 동사 어휘의미망 정제," 한국어학, 28권, pp.139-166, 2005.

[5] 황순희, 윤애선, "의미자질을 고려한 명사 어휘의미망 구축(1): 남/여성 자질을 중심으로," 한국어학, 제29권, pp.309-338, 2005.

[6] 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.

[7] Alpha K. Luk, "Statistical Sense Disambiguation with Relatively Small Corpora using Dictionary Definitions," In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pp.181-188, 1995.

[8] Namhee Kwon, Michael Fleischman, Eduard Hovy, "Senseval Automatic Labeling of Semantic Roles using Maximum Entropy Models," In Proceedings of SENSEVAL-3, 2004.

[9] Christiane Fellbaum, WordNet: An Electronic Lexical Database, The MIT Press, 1998.

[10] Philip Resnik, "Disambiguation Noun Groupings with Respect to WordNet Senses," In Proceedings of the Third Workshop on Very Large Corpora, 1995.

[11] Rada Mihalcea, Dan I. Moldovan, "A Method for Word Sense Disambiguation of Unrestricted Text," Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999.

[12] Ganesh Ramakrishnan, B.Prithviraj, Pushpak Bhattacharyya, "A Gloss-centered Algorithm for Disambiguation," In Proceedings of SENSEVAL-3, 2004.

[13] 이승우, 이근배, "국소 문맥과 공기 정보를 이용한 비교사 학습 방식의 명사 의미 중의성 해소," 정보과학회논문지 : 소프트웨어 및 응용, 제27권, 제7호, pp.769-783, 2000.

[14] Hee-Cheol Seo, Hae-Chang Rim, Soo-Hong Kim, "KUNLP system in Senseval-3," Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pp.222-225, 2004.

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

[16] David Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 1995.

[17] Christopher D. Manning, Hinrich Schùtze, Foundations of Statistical Natural Language Processing, p.182, The MIT Press, 2001.

부록 A. SENSEVAL-2와 KorLex의 의미체계 매핑

어휘	SENSEVAL-2 의미번호	의미	KorLex 의미번호	자료 수	
				학습	평가
눈	k00011	보는 감각을 가진 사람, 동물의 기관	눈1,눈6	125	66
	k00012	그물 같은 물건의 코와 코를 연결한 부분	-	1	
	k00013	하늘에서 내리는 눈	눈3	7	
손	k00021	사람의 팔목 끝에 달린 부분	손1	129	66
	k00022	손아랫사람	-	0	
	k00023	손해	-	0	
	k00024	성	-	1	
	k00025	후손	-	0	
	k00026	딴 데서 임시로 찾아오거나 가는 사람	손3	1	
	k00027	자신의 힘이나 역량을 이르는 말	-	1	
말	k00001	사람의 생각과 느낌을 표현하는 수단	말3	0	50
	k00002	끝	말6	0	
	k00003	집짐승	말1	11	
	k00004	곡식, 액체의 분량	말9	57	
	k00005	고누, 윷 따위의 판의 군사	말2	33	
바람	k00031	기압 변화로 일어나는 공기의 흐름	바람3	1	49
	k00032	어떤 일이 이루어지기를 기다리는 마음	바람1	0	
	k00033	어떤 일에 더불어 일어나는 기세	바람2	97	
	k00034	몸에 차려야 할 것을 차리지 않고 나서는 차림 또는 그 행실	-	0	
거리	k00041	사람과 차가 다니는 길	거리1	70	65
	k00042	일거리, 읽을거리	-	16	
	k00043	만드는데 주가 되는 물건	-	0	
	k00044	큰 이익	-	0	
	k00045	음악, 연극 등에서 단락, 과장, 마당을 뜻함	-	0	
	k00046	두 곳 사이의 먼 정도	거리2, 거리4	45	
자리	k00051	물체가 있거나 그것을 둘 수 있는 공간	자리3	97	49
	k00052	어떤 일에 종사하여 활동하는 직위나 지위	자리4	2	
	k00053	어떤 사람들이 모이도록 한 경우나 기회	-	2	
	k00054	십진법에 따른 숫자의 자리	자리10	0	
의사	k00061	의료 활동에 종사하는 사람	의사2	111	81
	k00062	동물이 움직이지 않고 죽은 체하는 일	-	0	
	k00063	의를 위해 죽음	-	0	
	k00064	마음먹은 생각	의사1	12	
	k00065	의심스러운 말	-	0	
	k00066	의논할 사항	-	0	
	k00067	신라시대의 으뜸 벼슬	-	0	
	k00068	의학이나 의료에 관한 일	-	0	
	k00069	실제와 비슷함	-	42	
목	k00071	머리와 몸의 잘록한 부분	목1,목3	99	49
	k00072	목과 비슷한 부분	목1,목3	1	
	k00073	통로 같은 곳으로 빠져나갈 수 없는 중요하고 좁은 곳	목2	0	
	k00074	나무 (木)	목7	0	
점	k00082	작고 둥글게 찍은 표	점5	5	49
	k00085	어느 속성이나 측면의 개별적인 부분, 요소	점3	89	
	k00086	물품의 가짓수를 셀 때 쓰는 말	-	1	
	k00087	살코기 따위의 작은 조각들을 셀 때 쓰는 말	-	4	
밤	k00091	밤나무의 열매	밤1	29	50
	k00092	저녁 어두운 뒤부터 새벽 밝기까지의 동안	밤2,밤3	72	