論文
2-12-11

Bayes 정리에 기반한 개선된 동형이의어 분별 모델 (An Improved Homonym Disambiguation Model based on Bayes Theory)

김 창 환* 이 왕 우**
(Chang-Hwan Kim) (Wang-Woo Lee)

요 약

본 연구에서는 동형이의어 분별을 위하여 허정(2000)이 제시한 "사전 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템"이 가지는 문제점과 향후 연구과제로 제시한 문제들을 개선하기 위하여 Bayes 정리에 기반한 동형이의어 분별 모델을 제안한다. 의미 분별된 사전 뜻풀이말 코퍼스에서 동형이의어를 포함하고 있는 뜻풀이말을 구성하는 체언류(보통 명사), 용언류(형용사,동사) 및 부사류(부사)를 의미 정보로 추출한다.

동형이의어의 의미별 사전 출현 빈도수가 비교적 균등한 기존 9개의 동형이의어 명사를 대상으로 실험하여 비교하였고, 새로 7개의 동형이의어 용언(형용사, 동사)을 추가하여 실험하였다. 9개의 동형이의어 명사를 대상으로 한 내부 실험에서 평균 99.37% 정확률을 보였으며 7개의 동형이의어 용언을 대상으로 한 내부 실험에서 평균 99.53% 정확률을 보였다. 외부 실험은 국어 정보베이스와 ETRI 코퍼스를 이용하여 9개의 동형이의어 명사를 대상으로 평균 84.42% 정확률과 세종계획의 350만 어절 규모의 외부 코퍼스를 이용하여 7개의 동형이의어 용언을 대상으로 평균 70.81%의 정확률을 보였다

ABSTRACT

This paper asserted more developmental model of WSD(word sense disambiguation) than J. Hur(2000)'s WSD model. This model suggested an improved statistical homonym disambiguation Model based on Bayes Theory.

This paper using semantic information(co-occurrence data) obtained from definitions of part of speech(POS) tagged UMRD-S(Ulsan university Machine Readable Dictionary(Semantic Tagged)). we extracted semantic features in the context as nouns, predicates and adverbs from the definitions in the korean dictionary.

In this research, we make an experiment with the accuracy of WSD system about major nine homonym nouns and new seven homonym predicates supplementary. The inner experimentalresult showed average accuracy of 98.32% with regard to the most Nine homonym nouns and 99.53% for the Seven homonym predicates.

An Addition, we gave test on Korean Information Base and ETRI's POS tagged corpus. This external experimental result showed average accuracy of 84.42% with regard to the most Nine nouns over unsupervised learning sentences from Korean Information Base and ETRI Corpus, 70.81% accuracy rate for the Seven predicates from Sejong Project phrase part tagging corpus (3.5 million phrases) too.

* 정회원 : 울산대학교 컴퓨터정보통신공학부 대학원 석·박사 통합과정중
** 정회원 : 울산대학교 컴퓨터정보통신공학부 대학원 석사

논문접수 : 2001. 12. 15.
심사완료 : 2001. 12. 22.

1. 서론

자연어 처리 연구에 있어 가장 어려움을 겪는 중의성(ambiguity) 문제는 형태소 분석, 구문 분석 등 모든 분석 과정에서 필연적으로 발생된다. 일부 과정에 발생하는 중의성 문제는 어느 정도 성과를 거두고 있다. 의미 및 담화 분석에 대한 연구가 활발해지면서, 의미 중의성 해결(WSD: word sense disambiguation) 방안이 활발히 연구되고 있다. 의미 중의성 해결이란, 문장에 출현하는 단어가 둘 이상의 상이한 의미로 사용될 때 문맥상 옳은 하나의 의미로 분별하는 것을 말한다.

의미 중의성 해결을 위한 연구는 학습 데이터의 형태에 따라 사전을 이용하는 방법과 코퍼스를 이용하는 방법으로 나눌 수 있고, 방법론에 따라서 규칙을 이용하는 방법과 확률 통계를 이용하는 방법 및 의미 계층 구조를 이용하는 방법으로 크게 나눌 수 있다.[1][2][6]

사전을 이용하는 방법은 언어의 동적인 특성을 반영하기 어렵다는 단점이 있으나, 단어의 의미별 정보를 상세히 추출할 수 있다는 장점을 가지고 있다. 코퍼스를 이용한 의미 중의성 해결을 위해서는 다량의 의미 분별 코퍼스가 필요한데, 양질의 코퍼스를 구하기가 힘들고 코퍼스를 구축하기 위해서는 많은 비용과 시간이 소요 된다는 단점이 있다. 그러나, 언어의 동적인 특성을 잘 반영하는 장점을 가지고 있다.

본 논문에 이용된 사전의 뜻풀이말은 서로 의미적인 관계를 가지는 단어들로 구성되어 있으므로 서로의 공기 관계를 이용하여 단어의 의미를 분별하고자 한다.

2. 기존 모델

허정(2000)은 사전 뜻풀이말에서 의미정보를 추출하고, 이 의미 정보를 확률 통계적인 방법을 적용하여 동형이의어 중의성을 해결하는 모델을 제안하였다.

본 연구에서는 기존 모델이 가지는 문제점들을 제시하고 이를 개선한 Bayes 정리에 기반한 동형이의어 분별 모델을 제안한다. 더불어 기존 연구에서 향후 연구과제로 제시된 문제점들을 다루도록 한다.

기존 모델의 주요 문제점은 다음과 같다.

1) 확률 통계적 방법의 접근 문제.
2) 체언류와 용언류의 관련성에 관한 가중치 문제.
3) 의미 정보 집합간의 교집합에 대한 고려 문제.

2.1 기존 모델의 확률적 접근법에 대한 문제점

기존모델은 수식(1)에서 볼 수 있듯이 의미정보와 문장의 단어와 공기하는 단어 Wj 의 빈도를 동형이의어 Sj 가 가지는 모든 빈도로 나누어서 확률을 구하였다.

$$P(W_j|S_i)=\frac{Frequency\cdot of\cdot noun\cdot W_j\cdot in\cdot S_i}{total\cdot frequency\cdot of\cdot nouns\cdot in\cdot S_i}$$

- 수식(1)

하지만 위의 수식대로 확률을 계산하면 가능한 사상들의 확률의 합이 1이 되지 않는다. 즉, "상호배타적인 모든 가능한 사상들의 확률의 합은 항상 1이 되어야 한다"[10]는 확률의 기본공리를 만족하지 못하므로 기존 모델은 확률적 접근 방법에서 문제가 있었다. 그리고 기존 모델은 이런 문제가 있는 접근법을 해소하기 위해 아래의 수식(2,3)을 도입하게 되었다.

$$Noun(C,S_i)=Match(C_n,S_i)\times\sum_j P(W_{nj}|S_i)$$

-수식(2)

$$\Pr ed(C,S_i)=Match(C_v,S_i)\times\sum_j P(W_{vj}|S_i)$$

-수식(3)

수식(2,3)은 확률의 합에 공기 관계를 가지는 단어의 수를 곱해줌으로써 소수 단어의 확률이 높게 나오더라도 의미 분별하려는 문장과 의미정보의 공기 관계가 많은 의미에 비중을 주기 위해서 고려된 수식이었다[1]. 그리고 기존의 모델에서는 수식(2,3)이 동형이의어를 분별하는데 좋은 효과를 보였다. 그러나 Bayes정리를 이용한 모델에서는 특정한 단어의 확률이 항상 0에서 1사이의 값을 가지므로 이런 고려는 하지 않아도 된다. 본 논문에서는 Bayes모델을 이용하여 확률적 접근방법의 문제점을 해소하였다.

<표 2>는 기존 모델을 이용하여 확률을 구한 결과인데 각 의미정보가 가지는 확률의 합이 1이 되지 않음을 확인 할 수 있다. <표 3>은 Bayes모델을 이용하여 확률을 구한 결과인데 각 의미정보가 가지는 확률의 합이 1이 됨을 볼 수 있다.

< 표 1 > 동형이의어(다리)가 포함된 예문

<Table 1> Example Sentence contains homonym (다리) in definitions of dictionary.

의미	품사별 의미 빈도 합	다리 하나가 없는 사람
신체	체언류(1663)	하나/NNG(1), 사람/NNG(14)
	용언류(830)	없/W(17)
교각	체언류(467)	하나/NNG(1), 사람/NNG(5)
	용언류(185)	

< 표 2 > 기존 모델의 수식(1)을 이용한 확률

<Table 2> The Probability based on J. Hur(2000)'s Model using numerical formula(1)

단어 \ 의미	신체	교각	P(Si |Wj)
하나/NNG	1/1633(0.0006)	1/467(0.0021)	0.0027
사람/NNG	14/1633(0.0086)	5/467(0.0107)	0.0193
없/VV	17/830(0.0205)	0/185(0.0)	0.0205
기존 WSD	0.01861	0.02304	

< 표 3 > Bayes모델을 이용한 확률

<Tab l e 3> The Probability based on Bayes Model

단어 \ 의미	신체	교각	$\Sigma P(S_q \mid W_j)$
하나/NNG	0.2218	0.7782	1.0
사람/NNG	0.4456	0.5544	1.0
없/VV	1.0	0.0	1.0
Bayes WSD	1.6674	1.3326	

2.2 체언류와 용언류에 대한 가중치 적용의 문제점

기존의 연구에서는 체언류와 용언류의 관련성에 관한 가중치를 내부 실험을 통하여 체언류에 0.9와 용언류에 0.1을 결정하여 동형이의어 중의성을 해결하였다.

일반적으로 용언류의 의미 정보 집합보다 체언류의 의미 정보 집합이 상대적으로 크다. 그러므로 수식(1) 을 이용하여 확률값을 계산하면 의미 정보 집합이 적은 용언류의 확률값이 대부분 크게 되므로 용언류의 확률값이 의미 분별에 큰 영향을 끼친다.

이러한 용언류의 확률값을 조정하기 위해 용언류의 가중치를 적게 주면 체언류와 용언류의 영향을 균등히 반영할 수 있었다.

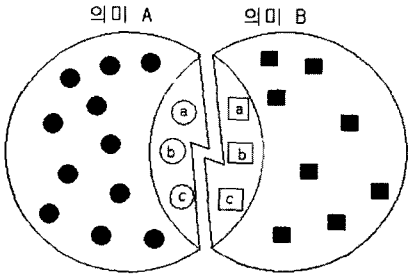
$$Rel(C,S_i)=w_n\times Noun(C,S_i)+w_v\times Pred(C,S_i) \quad \text{- 수식(4)}$$

본 연구에서 제시한 Bayes모델에서는 용언류의 의미 정보 집합이 상대적으로 적더라도 확률값의 합은 항상 1.00이므로 용언류의 확률값이 체언류의 확률값보다 커지지 않는다. 그러므로 Bayes모델에서는 수식(4)에서 사용한 용언류와 체언류의 관련성에 대한 가중치를 사용할 필요가 없어졌다.

기존 모델의 오류 유형을 분석해 보면 <표2>처럼 가중치 때문에 오분석되는 결과가 생겨났다. <표2>는 체언류와 용언류에 각각 가중치 0.9와 0.1을 곱하게 되어 다리의 의미를 교각으로 오분석하게 되는 예문이다. Bayes모델은 <표3>에서 볼 수 있듯이 기존의 오류 유형을 해결할 수 있다.

2.3 교집합 의미 정보에 대한 문제점

[그림 1] 을 보면 의미 정보들 사이에는 의미들 간에 중복되는 교집합 부류의 정보(A∩B)들이 있다.



[그림 1] 의미 A와 의미 B를 가지는 단어 W의 의미정보 집합 관계

[Fig. 1]The intersection between the set of Sense A and B

허정(2000)은 교집합 부류의 의미정보들이 의미별로 다른 빈도를 가지고 있어서 교집합 부류의 의미정보를 고려하지 않아도 된다고 제시했었다.

하지만 기존 모델의 오류 유형을 분석해 보면 교집합 부류의 의미정보와 각 의미의 총 빈도수의 크기가 달라 좀 더 세밀하게 다룰 필요가 있다.

기존 연구에서 제시한 모델은 이러한 교집합 부분에 대한 고려를 할 수가 없지만 Bayes 모델은 의미 정보들 사이에 있는 교집합 부분을 고려해 오분석되는 문제점을 해결 할 수 있다.

하지만 Bayes 정리를 이용하더라도 의미정보의 총빈도수 합이 서로 크게 차이가 날 경우에는 총 빈도수의 합이 적은쪽의 확률이 커지게 되어 오분석되는 결과가 있었다. 이를 해결하기 위해서는 총 빈도수의 합에 대한 가중치를 부여해 줄 필요가 있다.

3. 사전 뜻풀이말에서 추출한 의미 정보

의미 분별을 위해서 먼저 이루어지는 작업이 의미 정보 획득이다. 정보의 질과 양에 따라 의미 분별의 정확성이 크게 영향을 받는데, 사전 뜻풀이말 코퍼스는 각 의미에 따라 구성하는 단어들의 의미적 분포가 쉽게 구별됨을 알 수 있다. 기존의 연구에서는 사전 뜻풀이말의 이와 같은 특성을 이용한 공기정보를 의미 정보로 이용한다.

일반적인 통계 기반의 의미 중의성 해결은 통계 데이터의 자료 부족 문제가 야기되는데 이 문제은 의미 중의성 해결의 정확률에 큰 영향을 끼친다. 이러한 자료 부족 문제를 최소화하기 위하여 본 연구에서는 부사류에 대한 의미정보를 추가하여 의미정보를 확장하였다.

<표 4> 의미 정보 예(배 : 운송수단, 신체부위)

<Table 4> Example of semantic information (배: ship, abdomen)

배(운송수단)
[체언류](총 빈도수, 최고빈도 : 1990, 48) <NNG> : 갑판(7), 강(9), 고기잡이(5), 기관(5), 모양(4), 물(25), 승객(7), 운항(6), 정박(8), 짐(24), 항구(20), … [용언류](총 빈도수, 최고빈도 : 823, 44) <VV> : 가(16), 건너(5), 나르(20), 싣(44), 타(29), 내리(10), 태우(6), … <VA> : 가볍(5), 넓(3), 빠르(4), 크(17), 평평하(3), … [부사류](총 빈도수, 최고빈도: 80, 5) <MAG> : 가득(2), 두루(2), 빨리(1), 깊이(1), …
배(신체부위)
[체언류](총 빈도수, 최고빈도 : 1494, 95) <NNG> : 가슴(34), 동물(6), 등(68), 모양(14), 물(5), 병(10), 수술(6), 음식(6), … [용언류](총 빈도수, 최고빈도 : 544, 44) <VV> : 가르(7), 곯(3), 따(8), 앓(2), 차(4), … <VA> : 고프(8), 넓(4), 뚱뚱하(3), 부르(32), 아프(7), [부사류](총 빈도수, 최고빈도: 63, 6) <MAG> : 갑자기(1), 배불리(1), 조금(6), 마구(1), …

<표 4> 에서 기존의 연구는 체언류와 용언류만을 의미정보로 사용하였으나 본 연구에서는 의미정보를 확장하여 부사류에 대한 의미정보를 고려하였다. 분별하려는 동형이의어가 용언류일 경우, 부사류에 의해 의미 분별이 이루어지고 체언류와 더불어 용언류의 의미를 제한하게 된다.

명사에 대한 동형이의어 중의성 해결시에 부사류에 대한 의미정보를 고려하였을 때 평균 정확률이 99.24% 에서 99.37% 로 향상되었고, 용언의 경우는 평균 정확률이 99.22% 에서 99.53% 로 향상되었다.

체언류, 용언류 및 부사류, 각각은 품사 태그, 단어, 빈도로 구성이 된다. 빈도는 학습 코퍼스에서 출현한 횟수이다.

4. 개선된 동형이의어 분별 모델

사전에서 추출한 의미정보 빈도를 이용하여 문장상의 동형이의어를 분류하는데 이전의 확률을 사전확률로 이용한 Bayes 정리를 이용하여 사후확률을 구하는 모델을 제안한다.

4.1 Bayes Model

본 연구에서 제시한 Bayes 정리에 기반한 모델은 다음과 같다. Bayes 정리 :

$$P(B_x \mid A) = \frac{P(A \cap B_x)}{P(A)} = \frac{P(A \cap B_x)}{\sum_{i=1}^{n} P(A \cap B_i)}$$

위의 Bayes 정리를 이용한 중의성 해결 모델은 아래와 같이 정의한다.

$$WSD(C, H_{Si}) = \arg MAX_{H_{Si}} Sim(C, H_{Si}) \text{ - 수식(5)}$$

H_{si}는 동형이의어 H가 가지는 i번째 의미를 나타낸다. Sim (C, H_{si})는 문장 C와 의미 H_{si}의 관련성을 나타낸다. 수식(5)는 Sim(C, H_{si})의 의미 자질값들 중 최대인 값을 가지는 의미를 선택하여 의미 중의성을 해결한다. Sim(C, H_{si})는 아래 수식(6)에 의해서 구해진다.

$$Sim(C,H_{Si}) = Noun(C, H_{Si}) + Pred(C,H_{Si}) + Adv(C, H_{Si})$$
- 수식(6)

$Noun(C, H_{si})$는 문장 C에서 출현하는 체언류와 의미 H_{si}의 관련성이고, $Pred(C, H_{si})$는 문장 C에서 출현하는 용언류와 의미 H_{si}의 관련성이며, $Adv(C, H_{si})$는 문장 C에서 출현하는 부사류와 의미 H_{si} 의 관련성이다.

$$Noun(C, H_{Si}) = \sum_{j=1}^{n} P(H_{Si} \mid W_{nj}) \quad \text{- 수식(7)}$$

$$\Pr ed(C, H_{Si}) = \sum_{j=1}^{n} P(H_{Si} \mid W_{vj}) \quad \text{- 수식(8)}$$

$$Adv(C, H_{Si}) = \sum_{j=1}^{n} P(H_{Si} \mid W_{aj}) \quad \text{- 수식(9)}$$

$P(H_{si} \mid W_j)$는 동형이의어(H)가 포함된 문장에서 의미정보에 속하는 단어(W_j)가 나타났을 때, H_{si}의 의미로 해석될 확률을 나타낸다.

$$P(H_{Si} \mid W_j) = \frac{P(W_j \cap H_{Si})}{\sum_{i=1}^{n} P(W_j \cap H_{Si})} \quad \text{- 수식(10)}$$

<표 5> Bayes 정리에 기반한 용언의 형이의어 (바르다) 분별의 예

<Table 5> The process Example of WSD model based on Bayes Theory for homonym predicates "바르다"

의미	품사별 의미 빈도 합	물건을 기름에 담그거나 발라 흠씬 배게 하다
바르다_1 (붙이다)	체언류(1,333)	물건/NNG(13), 기름/NNG(29)
	용언류(83)	담그/W(1), 배/W(3), 하/W(21)
	부사류(57)	흠씬/MAG(2)
바르다_2 (참되다)	체언류(1,494)	물건?NNG(1)
	용언류(823)	하/W(26)
	부사류(39)	

	체언류(수식7)		용언류(수식8)		
	물건	기름	담그	배	하
바르다_1	0.85	1.00	1.00	1.00	0.30
바르다_2	0.15	0.00	0.00	0.00	0.70
$\sum P(H_{Si} \mid W_j)$	1.00	1.00	1.00	1.00	1.00

	부사류(수식9)	수식(6)
	흠씬	-
바르다_1	1.00	5.15
바르다_2	0.00	0.85
$\sum P(H_{Si} \mid W_j)$	1.00	6.00

<표 5>는 용언의 의미 중의성 해결에 부사류에

대한 의미정보(수식 9)가 사용되어 의미 분별되는 과정을 보여주고 있다.

4.2 내부실험에 사용된 실험 데이터

내부 실험에 사용된 데이터는 총 5,246문장이고, 어절 수는 38,296 개로써, 평균 한 문장이 7.3개의 어절로 구성되어 있다. 이 중 1차 의미 정보의 구축에 사용된 문장은 2,065 문장이다.

실험에 사용된 단어는 명사가 9개로써 "기관","기구","눈","다리","병","배","비","신","차"이다. 각 단어당 의미수는 평균2.7개이다.

용언은 7개로써 "끼다","말다","붓다","지다","지르다","바르다","세다"이다. 각 단어당 의미수는 평균2.4개이다.

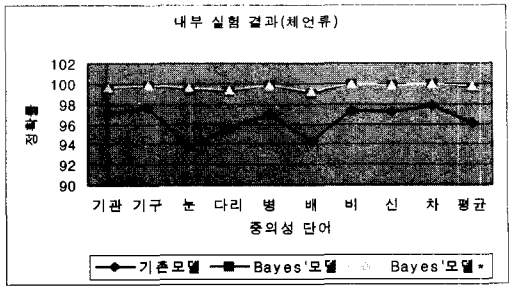
4.3 내부 실험의 결과 및 평가

<표 6> 내부 실험 결과(체언류)

<Table 6> The result of Inner Experiment(Nouns)

모 델	기관	기구	눈	다리	병
기존모델	97.04	97.64	93.73	95.58	96.99
Bayes모델	99.47	99.90	99.39	99.23	99.85
Bayes모델*	99.55	99.90	99.58	99.39	99.89

모 델	배	비	신	차	평균
기존모델	94.29	97.31	97.23	97.77	96.11
Bayes모델	99.00	100	99.80	100	99.24
Bayes모델*	99.17	100	99.80	100	99.37



[그림 2] 내부 실험 결과 차트(체언류)

[Fig. 2] The result chart of Inner Experiment(Nouns)

내부 실험은 기존 모델과 Bayes 정리에 기반한 모델의 정확률을 비교하였다. 부사류에 대한 의미정보의 영향을비교하기 위해 체언류와 용언류의의미정보만을 이용하는 모델을 Bayes 모델 이라 하였고 부사류에 대한 의미 정보까지 확장한 모델을 Bayes 모델* 이라 하였다.

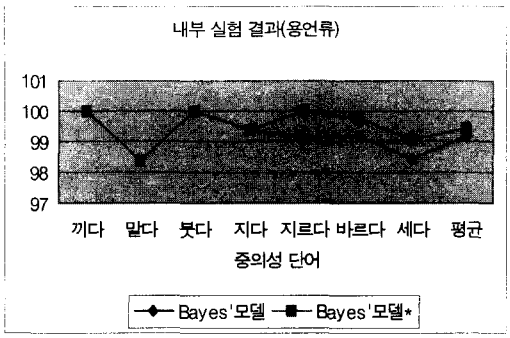
<표 6>은 명사 9개의 내부 실험 결과이며 중의성 단어 9개 모두에 대해서 기존의 모델보다 정확률의 향상을 보였다. 부사류에 대한 의미정보를 고려하였을 때 명사 9개의 경우도 정확률의 향상을 보였는데 용언류에 비해 낮은 수치를 보였다. 위의 결과는 부사류가 용언류와 어울려 쓰이며 용언류의 의미를 제한하는 경우가 많은 것으로 분석된다.

<표 7> 내부 실험 결과(용언류)

<Table 7> The Result of Inner Experiment(predicates)

모 델	끼다	말다	붓다	지다
Bayes모델	100	98.41	100	99.37
Bayes모델	100	98.41	100	99.37

모 델	지르다	바르다	세다	평균
Bayes모델	99.13	99.19	98.47	99.22
Bayes모델*	100	99.79	99.08	99.53



[그림 3] 내부 실험 결과 차트(용언류)

[Fig. 3] The result chart of Inner Experiment (predicates)

<표 7>은 용언류(동사,형용사) 7개의 내부 실험 결과이며 평균 정확률은 99.53% 으로 높은 정확률을 보였다. 이것은 Bayes 모델이 용언류에 대해서도 상당한 의미 분별력이 있는 것으로 분석 할 수 있다. 부사류에 대한 의미 정보를 고려 하였을 때 0.31%의 정확률 향상을 보였다.

4.4 내부 실험에서 실패한 예와 원인 분석

내부 실험에 사용된 문장은 평균 7.3어절로 구성되어 있어, 비교적 문장 길이가 짧고 대부분 단문이다. 따라서, 구문구조에 의한 오류보다는 빈도수가 높은 소수의 단어에 의해 발생하는 오분석이 오류의 가장 큰 원인이다.

<표 8> 오분석의 예

<Table 8> Example of Errors

오류 발생 문장	의미분별 대상단어	오분석 결과	올바른 결과
말이나 소의 등에 실은 짐을 배와 얼러 매는 줄.	배	교통	몸
동물이 공간적인 이동을 위해 사용하는 기관의 총칭	기관	조직	몸
빛의 세기를 나타내는 단위	세다	수효	힘

<표 8>은 빈도수가 높은 소수의 단어에 의해 발생한 오분석의 유형을 나타낸 것이다. "빛의 세기를 나타내는 단위" 에서 "단위"라는 단어가 "수효"의 의미 정보 집합에서 높은 빈도로 나타나다. 이로 인해 "힘"으로 분별 되어야 할 것이 "수효"로 오분석 되었다. 이와 같은 오류는 고빈도로 출현하는 단어의 빈도가 비율로 적용되므로 빈도 비율에 가중치를 부여하는 방법의 연구가 진행되면 해결이 가능 할 것이다. 각 예에서 오분석을 야기하는 고빈도 단어들을 밑줄로 표시하였다.

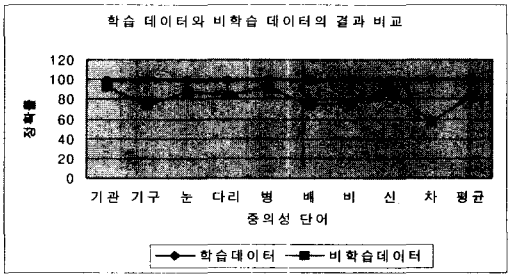
5. 실험 및 평가

본 연구에서는 내부 실험에서 언급한 의미 중의성을 가지는 동형이의어 명사 9개와 용언 7개를 실험하였다. 외부 실험에 사용된 실험 데이터는 명사의 경우는 국어정보 베이스(ver 1.0)와 ETRI품사 부착 코퍼스를 사용하였고 용언은 세종계획의 품사 부착 코퍼스를 사용하였다.

5.1 외부실험에 사용된 실험 데이터

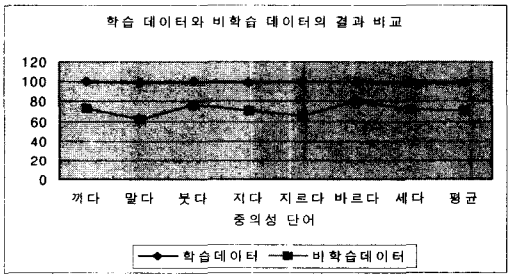
실험에 사용된 데이터는 국어 정보 베이스(ver 1.0)와 ETRI의 품사 부착 코퍼스 그리고 세종계획의 품사부착된 코퍼스를 사용하였다. ETRI의 코퍼스는 총 1,796 문장에 38,266 어절로 한 문장 당 평균 어절 수는 21.3개이다. 세종계획의 품사부착된 코퍼스(350만 어절)중 테스트에 사용된 문장은 3,114 문장이다. 외부 실험에서는 내부 실험(학습 데이터)에서 사전 뜻풀이 말에서 추출한 의미정보가 외부실험(비학습 데이터)에서도 적절한지를 알기 위해 내부 실험에서와 동일하게 적용하였다.

5.2 외부 실험의 결과 및 평가



[그림 4] 학습 데이터와 비학습 데이터의 정확률 비교(체언류)

[Fig. 4] The accuracy rate comparison between supervised learning data and unsupervised learning data(Nouns)



[그림 5] 학습 데이터와 비학습 데이터의 정확률 비교(용언류)

[Fig. 5] The accuracy rate comparison between supervised learning data and unsupervised learning data (predicates)

[그림 4]과 [그림5]에서 보면 알 수 있듯이, 외부 실험의 경우는 학습 데이터에 대한 정확률보다 떨어지나 명사에 대한 평균 정확률 84.42% 와 용언에 대한 평균 정확률 70.81% 로 비교적 높은 수치를 보였다. 위의 결과로 사전 뜻풀이말에서 추출한 의미 분별 정보가 외부 실험에서도 의미를 가질 수 있는 것을 알 수 있다.

5.3 외부실험에서 실패한 예와 원인 분석

외부실험의 경우는 코퍼스가 한 문장 당 평균 어절수가 21.3개로 문장 길이가 긴 장문이 대부분이다. 그래서 구문 구조를 고려하지 않아 발생하는 오분석이 가장 큰 원인이다.

<표 9> 오분석의 예

<Table 9> Example of Errors

오류 발생 문장	의미 분별 대상 단어	오분석 결과	올바른 결과
남극에서는 꽃 피는 식물 주종류인 남극잔디와 남극개미자리 눈이 녹기 시작하면 파릇파릇 돋아나 남극의 여름을 상징하는 꽃을 피운다.	눈	식물	현상
몸을 가지고 숨쉬는 모든 것이 한 쌍씩 노아와 함께 배에 올랐다	배	몸	교통

<표 9>에서 보면 알 수 있듯이 위의 문장에서 "눈"의 의미는 "현상"이 되어야 하나 "꽃 피는 식물","파릇파릇 돋아나","꽃을 피운다"라는 구문들이 "식물"에 대한 의미 정보 집합에 해당하는 공기 정보를 많이 가지고 있으므로 "식물"에 대한 의미 정보 자질 값이 크게 되어 오분석 결과가 발생한다. 이와 같은 오분석은 구문 구조에 따른 거리 가중치 부여에 대한 연구가 진행되면 해결이 가능하다.

오분석을 발생시키는 구문은 밑줄로 표시하였다.

"지르다"의 경우 평균 정확률 66.00% 를 보였는데 대부분의 오분석이 "지르다"가 가지는 의미 정보 집합에 없는 "비명"을 포함한 문장에서 발생하였다. 이러한 오분석을 해결하기 위해 의미 정보를 확장하는 연구가 진행되어야 한다.

6. 결론 및 향후 연구

본 연구에서는 허정(2000)이 제시한 "사전 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템"이 가지는 문제점과 향후 연구과제로 제시한 문제들을 개선하여 통계학의 Bayes 정리에 기반하여 의미 분별 정보와 동형이의어와의 발생빈도(사전확률)를 이용하여 사후확률을 구하는 모델을 제안한다.

기존 모델이 의미 정보 집합간의 교집합에 대해

고려를 하지 못한 점과 체언류와 용언류의 관련성에 관한 가중치를 부여한 문제점들을 개선하였다.

의미 분별된 사전 뜻풀이 말 코퍼스에서 동형이의어를 포함하고 있는 뜻풀이 말을 구성하는 체언류, 용언류 및 부사류를 의미 정보로 사용하였으며, 실험 데이터로 용언류 7개를 추가하여 연구하였다.

본 연구에서 제시한 Bayes 모델은 기존의 모델보다 높은 정확률을 보이고 있으며, 내부 실험에서 부사류에 대한 의미 정보를 고려했을 때 정확률의 향상을 통해, 부사류의 의미 정보도 중의성을 가지는 문장에서 공기하는 체언류, 용언류와 더불어 중요한 의미 정보가 됨을 알 수 있다.

끝으로 향후 연구에서는 내부 실험에서 언급한 빈도 비율에 대한 가중치 문제와 외부 실험에서 구문 구조에 따른 거리 가중치 문제, 의미 정보를 확장하는 문제 등의 연구가 진행되어야 한다.

※ 참고 문헌

[1] 허 정, 2000. "사전 뜻풀이말에서 추출한 의미 정보에 기반한 동형이의어 중의성 해결 시스템", 울산대학교 석사 학위 논문.

[2] 박성배, 장병탁, 김영택, "의미 부착이 없는 데이터로부터의 학습을 통한 의미 중의성 해소", 한국 정보과학회 '2000 봄 학술 발표 논문집 B', 제 27 권 1호, pp330 - 332, 2000.

[3] 송영빈, 최기선, "동사의 애매성 해소를 위한 시소러스의 이용과 한계", 제 12 회 한글 및 한국어 정보처리 학술대회 발표논문, pp.255 - 261, 2000.

[4] 이창기, 이근배, "의미 애매성 해소를 이용한 WordNet 자동 매핑", 제 12 회 한글 및 한국어 정보처리 학술대회 발표논문, pp.262 - 268, 2000.

[5] 조정미, "코퍼스와 사전을 이용한 동사 의미 분별", Ph.D. these, 한국과학기술원, 1998.

[6] D. Yarowsky(1992), "Word-Sense Disambiguat-ion Using Statical Model of Roget's Corpora", COLING-92

[7] J. Hur(2001), "A Homonym Disambiguation System based on Semantic Information extracted from Definitions in dictionary", ICCPOL-2001

[8] G. Rigau(2000), "Nave Bayes and Examplar-based Approaches to Word Sense Disambiguation Revisited", ECAL

[9] L. Marquez(2000), "Machine Learning and Natural Language Processing"

[10] 유지성, 오창수 , 현대 통계학 , 전영사

김 창 환

1997년 울산대학교
전자계산학과(공학사)
1997년~1999년 8월
삼보정보시스템
정보기술연구소(연구원)
1999년 울산대학교
컴퓨터정보통신공학부
대학원 석-박사통합과정
2000년 ~ 현재 ㈜시리울산
선임연구원
1999년 ~ 현재 울산대학교
컴퓨터 정보통신공학부
대학원 석·박사통합과정중
관심분야 : 자연언어처리, 문서
분류, 정보검색, 전자도서관,
인터넷 비즈니스,
디지털경영 등

이 왕 우

2001년 울산대학교
전자계산학과(공학사)
2001년 울산대학교
컴퓨터정보통신공학부
대학원 석사과정(입학)
2000년 6월 ~ 현재
㈜시리울산 연구원
2001년 ~ 현재 울산대학교
컴퓨터 정보통신공학부
대학원 석사(재학중)
관심분야 : 자연언어처리,
인공지능, 정보검색 등