의미범주 및 거리 가중치를 고려한 통계기반 동형이의어 분별 시스템

김준수[0] 김창환 이왕우 이수동 옥철영
울산대학교 전자계산학과
{jskim, edel, wwlee}@cic.ulsan.ac.kr {sdlee, okcy}@uou.ulsan.ac.kr

A Homonym Disambiguation System Based on Statistical Model Using Sense Category and Distance Weights

Jun-Su Kim[0] Chang-Hwan Kim Wang-Woo Lee Soo-Dong Lee Cheol-Young Ock
Dept. of Computer Science, University of Ulsan

요 약

본 논문에서는 Bayes 정리를 적용한 통계기반 동형이의어 분별 시스템에 대한 외부실험 결과를 분석하여, 정확률 향상을 위한 의미범주 가중치 및 인접 어절에 대한 거리 가중치 모델을 제시한다. 의미 분별된 사전 뜻풀이말 코퍼스(120만 어절)에서 구축된 의미정보를 이용한 통계기반 동형이의어 분별 시스템을 사전 뜻풀이말 문장에 출현하는 동형이의어 의미 분별에 적용한 결과 상위 고빈도 200개의 동형이의어에 대해 평균 98.32% 정확률을 보였다. 내부 실험에 사용된 200개의 동형이의어 중 49개(체언 31개, 용언 18개)를 선별하여 이들 동형이의어를 포함하고 있는 50,703개의 문장을 세종계획 품사 부착 코퍼스(350만 어절)에서 추출하여 외부 실험을 하였다. 분별하고자 하는 동형이의어의 앞/뒤 5어절에 대해 의미범주 및 거리 가중치를 부여한 실험 결과 기존 통계기반 분별 모델 보다 2.93% 정확률이 향상되었다.

1. 서론

"**의미**" 란 단어를 사전에서 찾아보면 "말이나 글의 뜻", "사물이나 행위의 드러나지 아니한 숨은 뜻" 으로 뜻풀이를 하고 있다. Yarowsky는 "의미란 명확하게 정의된 개념이 아니다" 라고 하였다. 이런 의미를 분별 한다는 것은 쉬운 일이 아니다.

일반적으로 문장을 분석 할 때 발생하는 의미 애매성을 해결하기 위하여 문장의 주제, 문맥, 사전 정보 등을 이용하며, 특히 인접 어휘들을 의미 해결의 실마리로 활용하고 있다.

본 논문에서는 의미적으로 관련이 있는 어휘들로 정의된 사전의 뜻풀이에서 의미분별을 위한 의미정보를 구축하고, 이를 통계 모형에 기반한 동형이의어 분별 모델에 활용하는 방안을 연구하고자 한다.

2. 통계기반 의미 분별 모델

2.1 Bayes 정리에 기반한 통계 모델

사전 뜻풀이말에서 추출된 동형이의어 의미범주[1]의 정보를 Bayes 정리의 사전 확률로 적용한 의미 분별 모델에서는, 임의의 문장 C 에서 나타나는 동형이의어 N 은 다음 수식에 의하여 의미 $Ns_1, Ns_2, \ldots, Ns_n$ 중 하나로 분별된다.

$$W(N,C) = \arg\max_{Ns_i} P(Ns_i, C) \qquad \text{수식(1)}$$

$$P(Ns_k, C) = \sum_{j=1}^{m} P(Ns_k \mid w_j) \qquad \text{수식(2)}$$

$$P(Ns_k \mid w_j) = \frac{P(w_j \cap Ns_k)}{\sum_{i=1}^{n} P(w_j \cap Ns_i)} \qquad \text{수식(3)}$$

수식(3)에서 Ns_k 는 동형이의어 N 의 k-번째 의미이며 문장 C 에서 출현하는 w_j는 Ns_k 의 의미범주에 속하는 어휘로 빈도 정보를 가지고 있다. 또한 w_j

[1] 본 논문에서 '의미범주'란 동형이의어 의미별 의미정보 집합에 들어 있는 어휘들을 말한다.

는 다른 의미범주에서 다른 빈도로 출현하기도 한다[2]. 수식(2)는 수식(3)에서 구해진 각각의 출현 어휘들에 대한 확률값에서 의미 Ns_k 로 판단할 확률의 합을 나타낸다. 그리고 수식(1)은 수식(2)에서 구해진 의미별 확률의 합 중에서 가장 큰 값을 문장 C 에서의 출현하는 동형이의어 N 에 대한 의미로 분별하는 방법이다.

본 통계적 기본 모델(NB: Na\ve Bayes Model)을 사전 뜻풀이말에서 고빈도로 출현하는 동형이의어 중에서 선별된 명사 31개, 용언 18개의 동형이의어에 대하여 세종 계획 350만 어절 품사태깅 코퍼스를 대상으로 실험하였다. 선별된 49개의 동형이의어를 포함하는 50,703개의 문장에 대하여 문장 전체어절에 적용 결과 명사는 77.67%, 용언은 61.75%의 평균 정확률을 보였으며, 동형이의어의 인접한 앞/뒤 5어절에 적용한 결과 명사는 72.87%, 용언은 43.79%의 정확률을 보였다.

2.2 기본 통계모델에서의 오류 유형

통계적 기본 모델(NB)에서 오분석 하는 경우를 대상으로 문제점을 알아보고 개선 방법에 알아보고자 한다.

[예문1] 그 바람에 **배**보다 배꼽이 더 커버린 과자값이 **들**었지만 그뒤에도 카메라가 쥐어진 할애비만 보면 외손녀는 울듯 비죽거렸고, 그때마다 아내는 카메라를 동댕이치겠다고 위협했다.

[예문1]에는 두개의 동형이의어 '배[2]'와 '들다[3]'가 있다. 이 중 '배'는 신체 부위를 나타내는 '배_1(신체)'의 의미로 사용되었다. [예문1]에 나타난 '배'에 대한 의미 분별 정보 즉, 각 의미범주에 들어있는 어휘들과 이들의 빈도 정보는 [표1]과 같다.

[표 1] [예문1]에서 추출한 의미정보(앞,뒤 5어절)

배_1 (신체)	체언류	배꼽(2)
	용언류	크다(26), 들다(9)
배_3 (선박)	체언류	바람(9)
	용언류	크다(25), 들다(1)
배_4	체언류	바람(1), 과자(2)
	용언류	크다(1), 들다(1)
배_6	체언류	값(1)
	용언류	크다(5)

추출된 어휘와 빈도 정보를 수식(3), (2)에 적용한 결과는 [표2]와 같다. 결과적으로 기본 모델은 배_4(과일)' 로 결정하여 동형이의어 분별이 실패한다.

[표 2] 통계적 방법에 의해 구한 확률값

	배_1	배_3	배_4	배_6
배꼽	1.00	0.00	0.00	0.00
바람	0.00	0.48	0.52	0.00
과자	0.00	0.00	1.00	0.00
값	0.00	0.00	0.00	1.00
크다	0.31	0.26	0.11	0.32
들다	0.39	0.05	0.56	0.00
합계	1.70	0.79	2.19	1.32
의미 분별 결과 (실패) : 배_4(과일)				

분별을 실패하는 주요 원인은 첫째, 의미 분별에 사용된 의미정보의 빈도에 대한 확률 계산에서 동형이의어의 사용 빈도를 고려하지 못함으로써 발생되는 오류이다. 예를 들어, [표1]에서 사전뜻풀이말에서 추출한 의미정보 '들다' 의 경우 배_3과 배_4에서 1회씩 출현하였으나 실제 배_3과 배_4로 사용된 뜻풀이말의 개수는 24, 513로 다르다. 즉, 24 및 513 개수에서 1회씩 추출한 의미정보 '들다' 의 빈도를 정규화 하여야 한다.

둘째, 현재의 분별 모델은 구문구조를 분석하지 않고 단순히 동형이의어를 포함하는 문장이 어떤 의미정보를 가지고 있느냐에 따라서 동형이의어를 분별한다. 이는 동형이의어가 은유적이나 숙어적으로 사용되지 않는다면 동형이의어의 해당 의미와 의미적으로 관련이 있는 단어들과 함께 사용된다는 사실에 근거하며, 동형이의어가 단문에서 사용되었다면 구문구조를 분석하지 않더라도 대부분 분별이 가능하다. 그러나, 복문 혹은 중문에서는 추출된 의미정보가 해당 동형이의어의 의미 분별에 결정적인 역할을 하지 못할 수도 있다. 따라서, 본 논문에서는 동형이의어를 중심으로 인접 5어절에 대해서만 의미정보로 추출하여 사용하였다[1,3]. 그렇지만 이 경우도 동형이의어와 인접한 거리에 따라서 의미적으로 결합하는 정도가 다를 수 있다. 그러므로, 의미정보가 발견된 위치(동형이의어와의 어절간 거리)를 고려하는 방법을 고려해야 한다.

본 논문에서는 이 두 가지 문제점을 해결하기 위한 방법을 제시한다.

---

[2] 동형이의어 '**배**'의 의미

배_1[신체부위], 배_3[운송수단(선박)], 배_4[과일], 배_6[갑절, 곱절]

[3] 동형이의어 '**들다**'의 의미

들다_1[머물다, 빛깔이 배다], 들다_4[위로 올리다, (사실이나 예를)지적하다], 들다_5[앞의 명사가 나타내는 행동을 받아 하다]

3. 통계기반 확률 모델에 가중치 적용

3.1 의미범주 가중치를 고려한 빈도정보

사전 확률로 이용하는 사전 뜻풀이말 의미정보는 동형이의어 의미($Ns_1, Ns_2, ..., Ns_n$)들의 출현 빈도에 따라 그 어휘종류 및 빈도에서 매우 큰 차이를 가진다.

문장 C 에 출현하는 어휘 wj ($\in Ns_1 \cap Ns_2 \cap ... \cap Ns_k$)가 여러 의미정보 집합에 공통으로 나타날 경우 수식(3)에 의해 의미범주 내 어휘의 빈도 총합이 작은 의미가 높은 확률값을 가지게 되어 선택될 가능성이 높아진다. [표3]을 보면 '배_1(신체)' 과 '배_4(과일)' 에 동일한 어휘가 출현 한다면 '배_1(신체)' 에서의 어휘 출현 빈도가 '배_4(과일)' 의 약 15배 이상 되어야만 '배_1(신체)' 로 분별할 확률이 높게 된다. 그러나 의미범주 내에 출현 빈도 15회는 매우 높은 수치로 그 어휘 하나만으로도 충분히 동형이의어를 분별할 수 있는 중요한 어휘이다. 따라서, 어휘가 가지는 특수한 상황을 고려하여 기본 통계모델의 Bayes 정리에 의미정보의 어휘수와 빈도를 동시에 고려하는 방법이 필요하다.

본 논문에서는 의미범주에 들어 있는 어휘들이 중요한 해결방법으로 보고 의미범주별 어휘들을 이용하고자 한다.

[표 3] 사전 뜻풀이말에서 추출한 의미범주별 어휘수 및 총빈도합 예

어휘	의미	어휘수		총빈도합	
		체언류	용언류	체언류	용언류
배	배_1	639	307	2,323	1,313
	배_3	668	283	1,593	1,114
	배_4	130	67	164	102
	배_6	363	82	676	178
소계		1,800	739	4,756	2,707
들다	들다_1	662	466	1,608	1,304
	들다_4	515	226	1,013	516
	들다_5	84	15	146	19
소계		1,261	707	2,767	1,839

동형이의어 의미범주 $Ns_1, Ns_2, ..., Ns_n$ 각각에 포함된 체언과 용언의 어휘수를 이용하여 수식(4) 와 같은 의미범주 가중치 얻게 된다.

$$Cat(Ns_k) = \frac{\text{a number of word in } Ns_k}{\sum_{j=1}^{n} \text{a number of word in } Ns_j} \quad \text{수식(4)}$$

의미 Ns_k 의 어휘들은 $P(w_j \cap Ns_k)$ 의 사전 확률값을 가지고 있다. 의미범주 가중치 $Cat(Ns_k)$ 를 기존 확률에 곱하여 $P(w_j \cap Ns_k) \times Cat(Ns_k)$ 새로운 사전 확률을 구하게 된다.

수식(5)에 가중치가 고려된 확률값을 통계적 기본 모델의 수식(1)과 수식(2)에 적용하여 의미범주 가중치를 고려한 통계적 모델(CA : statistical model with CAtegory weight)을 완성한다.

$$P_{cat}(Ns_k \mid w_j) = \frac{P(w_j \cap Ns_k) \times Cat(Ns_k)}{\sum_{i=1}^{n} P(w_j \cap Ns_i) \times Cat(Ns_i)} \quad \text{수식(5)}$$

통계적 기본 모델에서 분별 실패한 [예문1]에 대하여 [표1]에서 추출된 어휘에 대해 의미범주 가중치를 고려한 통계적 모델(CA)을 적용한 결과 올바르게 분별함을 [표4]에서 볼 수 있다.

[표 4] 카테고리 가중치를 고려한 확률값

	배_1	배_3	배_4	배_6
배꼽	1.00	0.00	0.00	0.00
바람	**0.00**	**0.83**	**0.17**	**0.00**
과자	0.00	0.00	1.00	0.00
값	0.00	0.00	0.00	1.00
크다	0.47	0.36	0.04	0.13
들다	**0.70**	**0.08**	**0.22**	**0.00**
합계	2.17	1.43	1.27	1.13
의미 분별 결과 (성공) : 배_1(신체)				

[표3]과 [표4]를 비교하면 '배_1' 의 의미범주에 속하는 '들다' 는 빈도 9로 '배_4' 에서의 빈도 1보다 높다. 기존 통계 모델에서는 '배_1' 로 0.39 확률을 가져 '배_4' 의 확률 0.56 보다 낮았다. 의미범주 가중치를 고려한 [표4]의 결과를 보면 0.70 으로 '배_1' 의 확률값이 올라감을 볼 수 있다. 따라서 의미범주 가중치를 고려한 통계적 모델이 빈도 정보를 적절히 고려할 수 있는 장점을 가지게 된다.

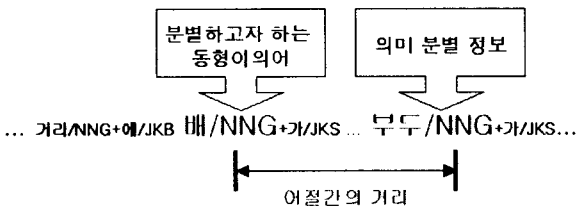
3.2 어절간 거리에 대한 고려

문장에 출현하는 동형이의어 의미 분별에서 구문구조를 이용할 수 있다면 분별시에 양질의 의미정보를 선별하여 불필요한 요소를 줄일 수 있을 것이다.

본 의미분별 모델은 사전의 의미정보를 기반으로 간결한 통계적 모형을 기반으로 한 모델이다. 따라

서 인접 어절에 대한 정보를 효율적으로 이용하여 해결 하고자 한다.

앞/뒤 5어절에서의 분별 정확률이 전체어절을 고려한 경우와 크게 차이가 없다는 것은 본 논문에 사용되는 의미정보가 인접 어절에서 많이 발견된다는 것이다. 특히, 앞/뒤 5어절 내에서도 동형이의어와 더욱 인접한 어휘가 의미 분별에 큰 영향을 준다는 점은 자명하다. 따라서 인접 거리에 대한 가중치를 적절히 적용해 보고자 한다.



[그림 1] 동형이의어와 의미정보 어휘간의 거리

동형이의어와 의미정보로 사용되는 어휘간의 절대거리 $|d(N)-d(w_j)|$ 를 고려하여 수식(6)의 가중치를 만들게 된다. 거리 가중치 Dis(N, w_j)를 카테고리 가중치를 고려한 새로운 확률값인 수식(5)에 적용하여 어절간 거리에 대한 가중치를 고려하여 어절 거리가 멀어 질수록 의미 분별에 미치는 영향이 감소 하게 된다. 따라서 인접한 어절에서 어휘가 발견 되면 높은 확률을 유지하게 되며 인접도가 떨어 질수록 감소하는 방법을 적용하게 된다.

$$\mathrm{Dis}(N, w_j) = \frac{1}{\sqrt{|d(N)-d(w_j)|}} \qquad \text{수식 (6)}$$

$$P(Ns_k, C) = \sum_{j=1}^{m} P_{cat}(Ns_k \mid w_j) \times \mathrm{Dis}(N, w_j) \qquad \text{수식 (7)}$$

[표 5] CA 가중치 적용 후 편차가 20%이내 거리 가중치를 적용한 결과

	배_1	배_3	배_4	배_6
배꼽	**1.00**	**0.00**	**0.00**	**0.00**
바람	0.00	0.83	0.17	0.00
과자	**0.00**	**0.00**	**0.50**	**0.00**
값	**0.00**	**0.00**	**0.00**	**0.50**
크다	0.27	0.20	0.02	0.08
들다	0.32	0.04	0.10	0.00
합계	1.59	1.07	0.79	0.58
의미 분별 결과 (성공) : 배_1(신체)				

본 방법의 효율성을 높이기 위해 의미범주 가중치를 고려한 통계적 모델(CA)을 적용한 결과에서 분별력의 차이가 근소한 경우(편차가 20% 이내)에 한해서 적용하는 방법을 택하도록 한다.

[표4]의 경우 최고 값을 받은 경우가 36.17% 다음 값이 23.83%로 20%내이다. 따라서 본 거리 가중치를 고려한 결과[표5] 올바르게 분별됨을 볼 수 있다.

4. 실험 및 분석

4.1 분별 동형이의어 선정 및 실험 문장 추출

사전 뜻풀이말에 나타나는 동형이의어 200개를 1차로 선정하고 의미별로 균형 있게 사용되는 체언 31개, 용언 18개, 총 49개 어휘를 선별하여 의미 분별 모델에 적용한다[표 6].

[표 6] 의미 분별 실험에 사용된 동형이의어

체 언 (31개)	거리, 결정, 경기, 국, 기구, 기원, 날, 눈, 대, 독, 등, 못, 배, 부정, 비, 상, 성, 의사, 의지, 이상, 장기, 장수, 절, 주장, 중, 지도, 차, 창, 철, 판, 표
용 언 (18개)	갈다, 고르다, 괴다, 끼다, 달다, 들다, 말다, 맞다, 묻다, 붓다, 쉬다, 싸다, 쓰다, 이르다, 지다, 차다, 켜다, 타다

세종계획 품사 부착 코퍼스(약 350만 어절)를 대상으로 동형이의어를 포함하는 약 50,700개의 문장을 추출하여 통계적 기본 모델(NB)을 이용하여 자동 의미 분별을 수행하였다. 자동 분별된 문장을 후처리 작업을 통하여 올바른 의미로 분별 후 정확률을 비교 하였다.

4.2 기본 모델과 의미범주 가중치 모델 비교

기본 통계적 모델의 평균 정확률은 아래 [표7] 같다. 동형이의어별 분석 결과는 [부록1]와 같다.

본 실험에서는 기본적으로 전체어절과 앞/뒤 5어절[4]을 이용한 의미 분별을 시도 하였다.

체언의 경우 기존 통계적 모델(NB)에서 전체어절과 앞/뒤 5어절 분별시 4.8%의 정확률 차이를 보이고 있다. 이는 다량의 의미정보가 인접한 어절에

[4] 본 논문에서 제안하는 의미 분별 모델은 구문 분석 과정을 거치지 않고 있다. 따라서 인접한 앞/뒤 5어절 정도를 이용한 분석을 시도한다.

분포한다고 볼 수 있으며, 장문에 출현하는 동형이의어 의미 분별시 인접한 어절에서 충분히 의미정보를 추출할 수 있다. 또한 전체어절에서 정보를 추출할 경우 발생할 수 있는 불필요한 정보를 차단하는 효과도 있다.

[표 7] 기본 통계적 모델(NB, 정확률)

	문장수	전체	5어절	정확률 (전체어절)	정확률 (5어절)
체언	30,451	23,652	22,189	77.67%	72.87%
용언	20,252	12,506	8,868	61.75%	43.79%
합계/평균	**50,703**	**36,158**	**31,057**	**71.31%**	**61.25%**

용언은 사전 뜻풀이말에서 추출한 의미정보만으로는 동형이의어 분별에 부족함이 많다. 예를 들어 '붓다'의 경우 잘 어울려 사용되는 '쏟다' (쏟아 붓다)가 의미정보에 포함되어 있지 않았다. 이를 의미정보에 추가하여 분석한 결과 6%나 정확률이 증가함을 볼 수 있다. 이는 사전 뜻풀이말의 기술 방법에 의해 사용되는 어휘가 제한적인데 원인이 있다. 따라서 효율적으로 필요한 의미정보를 추가하는 방법이 차후에 연구되어야 할 것이다.

[표 8] 의미범주 가중치(CA)를 고려한 NB 모델

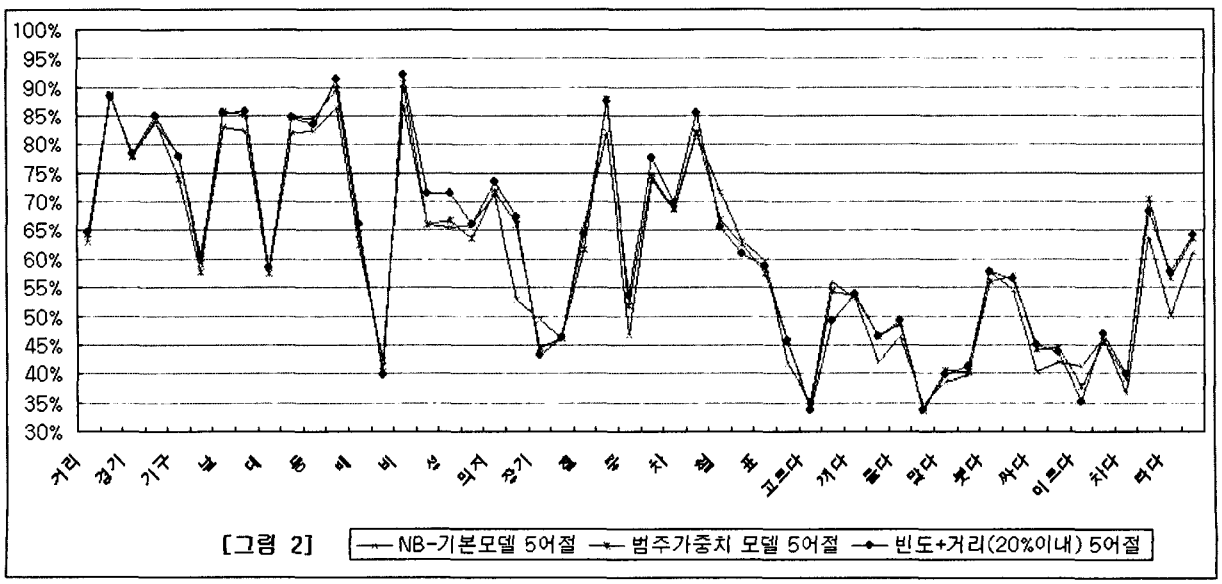
	문장수	전체	5어절	정확률 (전체어절)	정확률 (5어절)
체언	30,451	24,412	23,034	80.17%	75.64%
용언	20,252	12,557	9,238	62.00%	45.62%
합계/평균	**50,703**	**36,969**	**32,272**	**72.91%**	**63.65%**

[표8]는 의미범주 가중치(CA)를 고려하여 새로운 확률을 얻는 모델의 분석 결과이다. 가중치를 고려한 결과 전체어절 분석에서는 1.7%, 앞/뒤 5어절 분석에서는 2.4% 정확률이 증가 하였다. 전체 어절을 분석한 경우 29개의 동형이의어 정확률이 증가했으며, 앞/뒤 5어절에서는 31개의 동형이의어 정확률이 증가하였다. 기본 통계적 모델에 의미범주 가중치를 적용하는 것이 의미 분별에 효율적임을 알 수 있다. 또한, 앞/뒤 5어절을 고려한 경우에서 효과적으로 가중치가 적용됨을 볼 수 있다.

[표 9] CA 가중치 적용 후 편차가 20%이내 거리 가중치를 적용한 모델

	문장수	전체	5어절	정확률 (전체어절)	정확률 (5어절)
체언	30,451	24,653	23,322	80.96%	76.59%
용언	20,252	11,814	9,219	58.33%	45.52%
합계/평균	**50,703**	**36,467**	**32,541**	**71.92%**	**64.18%**

[표9]의 결과는 의미범주 가중치를 고려하여 1차로 분별을 시도하고 분별력의 편차가 20% 이내의 경우에 앞/뒤 5어절에 대해 거리 가중치를 적용한 결과이다. 전체 어절 정확률은 기본 통계적 모델(NB)에 비해 0.61%의 증가에서 머물며 의미범주 가중치적용 보다 감소함을 볼 수 있다. 앞/뒤 5어절 분석 결과에서는 가장 높은 정확률을 보이고 있다. 따라서 본 논문에서 제시하는 두 가지 가중치를 복합적으로 결합한 모델이 효율적임을 알 수 있다.



[그림 2]

4.3 오류 유형

의미범주 및 어절간 거리 가중치를 고려한 모델에서 전체어절 및 앞/뒤 5어절에 모두에서 정확률이 떨어지는 경우를 분석한 결과 의미정보 부족이 가장 큰 원인으로 나타났다. 이는 사전에서 광범위한 어휘의 사용을 자제하고 제한된 어휘로 뜻풀이를 구성하는 이유 때문으로 생각된다.

5. 결론 및 향후 과제

본 논문에서 제안한 의미범주 및 어절간 거리 가중치 정보를 고려한 통계기반 모델을 실험한 결과를 분석해 보면 적절한 가중치가 의미 분별에 도움이 된다는 결론을 얻었으며, 더욱 견고한 가중치를 찾아야 한다.

둘째, 사전 뜻풀이에서 추출한 의미정보 정제 및 확장을 위한 연구를 진행하여야 한다. 정제의 경우 뜻풀이말의 고유의 특성상 고빈도로 발생하는 명사(사람, 일, 말, 때, …), 동사(하다, 되다, 있다, 이르다, …), 형용사(없다, 있다, 크다, 작다, 같다, 다르다, …)등이 의미 분별에 미치는 영향을 조사하여 불필요한 의미정보의 경우 적절한 제거 방안을 마련하여야 할 것이다. 의미정보 확장을 위해 의미 태깅 툴을 만들어 대량의 의미 분별 코퍼스 구축과 함께 필요한 의미정보를 추출하는 방법과 의미 계층망 등을 이용한 정보의 확장을 연구하여야 할 것이다.

6. 참고 문헌

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

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

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

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

[5] Alpha k. Luk(1995), "Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions", 33rd Annual Meeting of the ACL

[6] R. Bruce(1994)," Word-Sense Disambiguation Using Decomposable Models" 32rd Annual Meeting of the ACL, pp. 139-145

[7] P. Brown, V. Della Pietra, S. Della Pietra and R. Mercer(1991) Word sense disambiguation using statistical methods. In Proceedings of the 29th Annual Meeting of the ACL, pp.264-270

[8] G. Cottrell(1989) A Connectionist Approach to Word sense Disambiguation. Pitman, London

[9] E. Brill(1993) A Corpus-Based Approach to Language Learning. Ph.D. thesis Computer and Information Science, University of Pennsylvania

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

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

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

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

[부록] 기본통계모델(NB), 의미범주 가중치(CA)

동형이의어	의미수	NB-기본모델		CA 가중치 모델		CA+거리(20%이내)	
		전체어절	5어절	전체어절	5어절	전체어절	5어절
거리	3	64.20%	62.96%	63.58%	64.30%	65.53%	64.71%
결정	2	94.63%	88.35%	94.83%	88.68%	94.76%	88.48%
경기	3	79.33%	78.25%	77.95%	77.95%	80.09%	78.41%
국	2	88.03%	83.76%	88.89%	84.19%	91.45%	85.04%
기구	3	81.58%	78.22%	78.42%	73.86%	80.00%	77.82%
기원	3	58.16%	59.22%	58.51%	57.80%	59.93%	60.64%
날	2	93.52%	83.11%	95.73%	85.74%	95.44%	85.60%
눈	2	84.58%	82.28%	87.68%	85.20%	88.62%	85.75%
대	4	64.89%	57.45%	56.38%	57.45%	59.57%	58.51%
독	2	89.93%	82.01%	91.37%	84.89%	91.37%	84.89%
등	3	83.66%	82.35%	87.15%	84.53%	86.71%	83.66%
못	3	88.59%	86.47%	91.50%	89.57%	92.44%	91.46%
배	4	63.38%	62.54%	64.17%	64.27%	66.88%	66.09%
부정	3	39.66%	42.37%	37.05%	40.26%	38.86%	39.76%
비	4	87.09%	87.09%	90.08%	90.55%	91.97%	92.13%
상	4	62.37%	66.13%	61.83%	66.13%	71.51%	71.51%
성	4	70.71%	65.67%	71.97%	66.77%	74.49%	71.50%
의사	2	69.44%	66.03%	65.48%	63.71%	67.80%	66.17%
의지	2	75.80%	71.28%	76.20%	71.81%	76.86%	73.54%
이상	3	60.39%	52.90%	74.36%	66.09%	74.22%	67.36%
장기	3	54.64%	49.34%	49.34%	44.37%	46.36%	43.38%
장수	3	49.66%	46.21%	48.97%	46.21%	48.28%	46.21%
절	3	68.81%	66.10%	60.34%	61.69%	63.73%	64.41%
주장	3	94.65%	81.99%	98.09%	87.93%	97.60%	87.59%
중	3	55.49%	46.71%	64.16%	51.93%	61.86%	53.50%
지도	2	82.90%	74.04%	83.10%	74.45%	82.09%	77.67%
차	3	70.41%	68.65%	70.53%	68.77%	72.51%	69.71%
창	2	78.25%	81.75%	78.60%	82.11%	85.61%	85.61%
철	3	65.64%	71.78%	62.58%	66.87%	65.03%	65.64%
판	3	67.24%	63.22%	66.09%	62.64%	63.22%	60.92%
표	3	63.09%	59.78%	60.88%	57.58%	62.26%	58.68%
갈다	3	64.25%	41.90%	60.89%	45.81%	62.01%	45.81%
고르다	3	57.86%	35.22%	52.83%	34.28%	48.11%	33.65%
괴다	2	77.97%	55.93%	79.66%	54.24%	72.88%	49.15%
끼다	2	73.20%	53.10%	74.19%	53.60%	72.21%	53.85%
달다	4	64.75%	42.04%	67.36%	46.48%	68.67%	46.48%
들다	3	61.92%	46.36%	63.00%	48.54%	59.77%	49.05%
말다	3	48.51%	34.70%	49.25%	33.58%	45.90%	33.58%
맞다	3	55.45%	38.63%	55.38%	40.44%	52.49%	39.70%
묻다	3	64.86%	39.71%	66.85%	40.43%	60.71%	41.28%
붓다	2	76.92%	57.69%	76.92%	56.04%	75.27%	57.69%
쉬다	3	79.19%	54.53%	80.92%	57.03%	76.69%	56.65%
싸다	3	67.77%	40.40%	63.13%	44.15%	60.93%	45.03%
쓰다	3	53.78%	41.93%	56.68%	44.45%	52.18%	43.86%
이르다	3	60.50%	41.31%	49.24%	37.97%	44.07%	35.21%
지다	4	70.22%	46.22%	64.00%	45.78%	63.11%	46.96%
차다	4	60.74%	36.60%	62.20%	39.39%	58.36%	39.79%
켜다	2	83.82%	63.97%	87.50%	70.59%	83.82%	68.38%
타다	5	72.52%	50.23%	77.93%	56.83%	75.08%	57.81%
평균	**2.94**	**71.31%**	**61.25%**	**72.91%**	**63.65%**	**71.92%**	**64.18%**