상호정보량을 이용한 동형이의어 분별용 의미정보의 정제

김준수[0*] 이왕우* 김창환** 옥철영*
울산대학교 컴퓨터정보통신공학부*, 춘해대학 컴퓨터정보과**
{endstart[0], wwlee, okcy}@mail.ulsan.ac.kr* chkim@choonhae.ac.kr**

Refinement of Semantic-Information for WSD

Using Mutual Information

Jun-Su Kim[0*] Wang-Woo Lee* Chang-Hwan Kim** Cheol-Young Ock*
Dept. of Computer Engineering & Information Technology, University of Ulsan*
Dept of Computer Techology, Choonhae College**

요 약

사전 뜻풀이에서 추출된 기존의 의미정보는 동형이의어가 포함된 뜻풀이에서 명사, 용언을 모두 추출하는 방법을 이용하여 단어 중의성 해소에 부적절한 정보를 상당수 포함하게 되었다. 이러한 부적절한 정보 때문에 오분석이나 과분석이 발생하게 된다. 그러므로 기존의 의미정보에서 동형이의어 분별에 유용한 정보만을 선택하는 기준이 필요하게 되었다. 본 논문에서는 사전 뜻풀이에서 동형이의어와 의미정보 사이의 상호정보량을 계산하고 임계치를 설정하여 의미정보를 선택제약하는 방법을 이용하였다. 임계치에 의해 제한된 의미정보의 효율성을 실험하기 위한 다양한 동형이의어 분별 실험들을 수행하였다.

1. 서 론

일반적으로 문장을 분석 할 때 발생하는 의미 애매성을 해소하기 위하여 문장의 주제, 문맥, 사전 정보 등을 이용하며, 특히 단어 의미 중의성 해소(word sense disambiguatuon : WSD)를 위하여 인접 어휘들을 의미 해결의 실마리로 활용하고 있다. 사전 뜻풀이(dictionary definitions)에 출현하는 동형이의어와 함께 나타나는 명사, 용언의 정보를 의미 분별 정보로 이용하는 연구가 진행되었으며[1], 이 의미정보(semantic-information)를 통계적으로 활용하는 방법이 연구되고 있다[1][2][3]. 하지만 기존의 연구들에 이용되는 의미정보는 두 가지 큰 문제를 가지고 있다. 첫째, 동형이의어가 포함된 사전 뜻풀이 내의 명사, 용언을 여과 없이 추출하게 되어 의미 분별의 실마리 역할을 하지 못하는 다수의 어휘들을 포함하게 된다는 문제이다. 따라서 의미 분별에 불필요한 어휘들을 적절히 선별하는 방안이 모색되어야 한다. 다음으로, 사전의 뜻풀이는 일상 생활에서 사용하는 다양한 어휘를 포함하지 못한다는 문제를 가지고 있다. 이는 의미정보의 어휘 부족 현상을 야기한다.

본 논문에서는 동형이의어 분별에 유용한 의미정보를 정제해 내는 방법을 제시하고자 한다. 먼저 기존 연구에 사용되는 의미정보를 살펴보고, 다음으로 용어나 문서의 연관성 측정에 자주 이용되는 상호정보량(mutual information)을 이용하여 의미정보를 단순 공기빈도가 아닌 정량화된 계수값으로 나타낸다. 마지막으로 동형이의어 의미 분별 실험을 통해 의미정보의 정제 효율성을 검증한다.

2. 의미정보 추출과 정제의 필요성

기존 동형이의어 분별을 위한 목적으로 만들어진 의미정보의 추출 방법을 알아보고 추출된 의미정보를 통해 정제의 필요성을 살펴본다.

사전 뜻풀이의 형태는 매우 다양하며 조평옥(1999)은 11가지 유형으로 분류하고 있다[4]. 그 중 대표적인 2가지 유형을 동형이의어 분별용 의미정보 추출 기준으로 이용하였다. 우선, 동형이의어가 뜻풀이의 마지막에 나타나서 표제어와 상위-하위의 관계를 가지는 경우, 그리고 동형이의어가 뜻풀이 문장의 처음이나 가운데에 사용되는 경우를 이용하였다.

의미정보는 위의 두 가지 경우에 해당하는 뜻풀이와 그 표제어를 UMRD-S[1]에서 선택하고 동형이의어와 함께 사용된 명사, 용언을 추출하고 공기빈도(co-occurrence)를 계산하여 의미정보로 이용하였다. 아래 [표 1]은 동형이의어 '배_3(ship)'이 들어있는 뜻풀이들이다. 그 표제어 및 밑줄 친 단어(체언, 용언)들을 의미정보로 추출한다.

[표 1] '배_3(ship)'이 들어있는 뜻풀이 문장

동형이의어	표제어	뜻풀이
배_3 (ship)	객선	손님을 태우는 배
	유조선	유종 시설을 갖춘 배
	난파	배가 항해 중에 폭풍우를 만나 깨어짐
	기항	항해 중인 배가 목적지가 아닌 항구에 잠시 들름
	운하	육지를 파서 강을 내고 배가 다니게 한 수로
	...	...

사전 특성상 뜻풀이에 고빈도로 출현하는 명사(사람, 일, 말, 때, …), 동사(하다, 되다, 있다, 이르다,…), 형용사(없다, 있다, 크다, 작다, 같다, 다르다, …)등과 같이 동형이의어 분별에 직접적인 영향이 적은 단어들 또한 높은 공기빈도로 의미정보에 들어 있게 된다. 그러나 단순 공기빈도와 공기빈도의 총합을 이용하는 기존의 통계적 모델[1][2][3]에서는 이렇게 불필요한 어휘들이 통계적 확률계산에 큰 영향을 주게 되었다. [표 2]를 보면 '하다, 일, 사람, 말' 등이 높은 공기빈도로 들어 있음을 알 수 있다. 그러나 '정박'의 경우 전체 뜻풀이에서의 빈도가 11로 낮지만, '배_3'과의 공기빈도는 10으로 높은 공기빈도를 가지며 동형이의어 '배'의 분별에 중요한 정보로 인식된다. 따라서 단순 공기빈도를 동형이의어 분별에 사용하기에는 어려움이 많다. 의미정보에서 이들 단어를 선별하는 기준 마련이 필

1) UMRD-S : 울산대학교 기계 가독형 사전(UMRD)에 있는 뜻풀이(1,172,00어절)에 출현하는 동형이의어(14,500여개)에 대하여 사전에 등재된 번호를 기준으로한 의미 주석(sense tag)을 414,800개의 어절에 부착

요하다.[4]

[표 2] 의미정보 '배 3(ship)'의 공기빈도값

의미정보	빈도	공기빈도	의미정보	빈도	공기빈도
하다	9,557	47	젓다	33	8
일	8,842	27	띄우다	62	6
사람	8,261	22	운항	20	5
말	7,169	15	원양	7	4
정박	11	10	입항	3	1
…	…	…	…	…	…

3. 의미정보의 상호정보량 측정

일반적으로 단순 공기빈도를 이용하는 방식 대신 개별 단어의 빈도와 전체 빈도를 함께 이용하여 단어 사이의 통계적인 연관성을 객관적으로 평가하는 상대 공기빈도 방식이 주로 이용된다[5][6]. 본 논문에서는 단어 관계에 대한 통계적 언어 모델링에 자주 이용되는 정보이론에 기반한 상호정보량을 이용해 보고자 한다.

3.1 상호정보량

상호정보량이란 두 독립사건의 확률변수 X와 Y 사이의 의존 관계를 정량적으로 나타내는 것으로 공식은 다음과 같다.

$$MI(x,y) = \log \frac{P(x,y)}{P(x)\times P(y)} \qquad \text{수식(1)}$$

$$\approx \log \frac{Nf(x,y)}{f(x)\times f(y)} \qquad \text{수식(1)'}$$

의미정보의 상호정보량을 계산하기 위해 위의 수식(1)′ 을 이용한다. 수식(1)′ 에서 $f(x)$, $f(y)$, $f(x, y)$는 각각 x의 빈도, y의 빈도, x와 y의 공기빈도를 나타내며, 마지막으로 N은 전체 뜻풀이 수이다. 상호정보량은 두 확률이 완전히 독립적일 경우(공기빈도가 0인 경우) 0이 되고 의존 관계가 깊을수록 높은 값을 가진다.

연관성 분석에 상호정보량을 이용할 때 문제점으로는 저빈도 단어 사이의 상호정보량이 고빈도 단어 사이의 정보량보다 상대적으로 과대 평가되는 경향이 있다.

3.2 상호정보량 측정

사전 뜻풀이에 고빈도로 출현하는 200개의 동형이의어 중 의미별로 고른 분포를 가지는 48개를 대상으로 상호정보량을 측정하였다[2].

[표 3] 의미정보 '배_3(ship)'의 상호정보량 측정

의미정보	공기빈도	상호정보량	의미정보	공기빈도	상호정보량
하다	47	0.1954	젓다	8	1.8308
일	27	-0.0321	띄우다	6	1.5339
사람	22	-0.0419	운항	5	1.8442
말	15	-0.2235	원양	4	2.2032
정박	10	2.4144	입항	1	1.9691
…	…	…	…	…	…

[표 3]은 [표 2]에 나타난 의미정보들에 대한 상호정보량들로 '하다, 일, 사람, 말'과 같이 뜻풀이 내에 폭넓게 사용하는 단어들은 상호정보량이 낮음(음수 값)을 볼 수 있다.

유용한(의미 분별에 중요한) 의미정보를 상호정보량이 0이상인 경우로 단정하기는 어렵다. 그래서 [표 4]와 같이 상호정보량 임계치를 기준으로 선별(정제)된 의미정보를 동형이의어 분별 실험에 적용해 보아야 할 것이다. 48개 동형이의어가 들어있는 뜻풀이 문장은 총 21,163개이며, 의미정보는 총 48,172개이다. 임계치를 통해 의미정보의 수를 감소시키면 분별 가능한 뜻풀이 문장역시 줄어들게 된다. 임계치가 1.1인 경우 25,076개(52.06%)의 의미정보로 18,243개(86.20%)의 문장을 분석할 수 있게 된다. [표 4]의 비율을 비교하면 의미정보의 감소 비율이 적용 가능한 문장의 감소 비율보다 상대적으로 높음을 알 수 있다.

[표 4] 상호정보량(MI)에 의한 의미정보 개수 및 적용 가능 문장수(동형이의어 48개에 대해)

MI(x)	적용 가능 문장		의미정보	
MI(x)≥y	문장수	비율	정보수	비율
All	21,163	100.00%	48,172	100.00%
0.0	20,928	98.89%	45,462	94.37%
0.1	20,854	98.54%	44,427	92.23%
0.2	20,674	97.69%	43,154	89.58%
0.3	20,212	95.51%	41,593	86.34%
0.4	20,141	95.17%	39,892	82.81%
0.5	19,815	93.63%	38,068	79.03%
0.6	19,656	92.88%	36,042	74.82%
0.7	19,251	90.97%	33,894	70.36%
0.8	18,903	89.32%	31,724	65.86%
0.9	18,674	88.24%	29,508	61.26%
1.0	18,488	87.36%	27,289	56.65%
1.1	18,243	86.20%	25,076	52.06%
1.2	17,747	83.86%	22,961	47.66%
1.3	16,583	78.36%	20,774	43.12%
1.4	15,779	74.56%	18,711	38.84%
1.5	14,338	67.75%	16,891	35.06%

4. 정제된 의미정보를 이용한 WSD 실험 및 분석

상호정보량 임계치를 기준으로 선별(정제)된 의미정보가 동형이의어 분별에 어떤 영향을 미치는지 알아보기 위해 본 연구에서는 세 가지 WSD 방법을 이용하였다.

4.1 WSD 실험 방법

실험 ① : 상호정보량의 합을 이용한 WSD

뜻풀이 문장 C 에 나타나는 동형이의어 H는 아래의 수식(2)에 의하여 $H_{S_1}, H_{S_2}, \cdots, H_{S_n}$ 중 하나로 분별된다.(임계치 이상의 제한된 의미정보를 이용)

$$W(H, C) = \arg\max_{H_S}\left(\sum MI(H_{S_i}, x_j)\right) \qquad \text{수식(2)}$$

위의 수식(2)에서 $\sum MI(H_{S_i}, x_j)$는 동형이의어 의미별로 계산되어진 상호정보량의 합을 의미한다.

실험 ② : 통계적 계산법을 이용한 WSD

$$W(H, C) = \arg\max_{H_S} \sum_{j=1}^{m} P(H_{S_k} \mid w_j) \qquad \text{수식(3)}$$

$$P(H_{S_k} \mid w_j) = \frac{P(w_j \cap H_{S_k})}{\sum_{i=1}^{n} P(w_j \cap H_{S_i})} \qquad \text{수식(3)'}$$

수식(3)′ 에서 H_{S_k}는 동형이의어 H의 k-번째 의미이며 w_j는 문장 C 에 출현하는 H_{S_k}의 의미정보에 속하는 어휘로 공기빈

도를 가진다. 수식(3)은 수식(3)' 에서 계산되어진 의미별 확률의 합 중에서 가장 큰 값을 동형이의어 H의 의미로 분별하는 방법이다.(임계치 이상의 제한된 의미정보를 이용)

실험 ③ : 체언과 용언을 고려한 통계적 WSD[3]

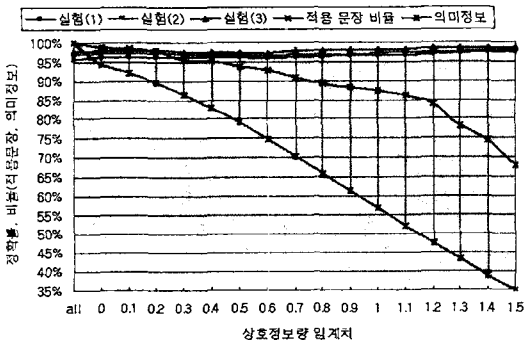
수식(4),와 수식(5)를 이용하여 문장 C 에 출현하는 체언에 해당하는 확률을 계산하고, 수식(3)에 적용하여 동형이의어를 분별하는 방법이다.(용언도 동일한 방법이며 지면상 수식 생략)

$$P(H_{S_k} \mid w_{N_i}) = \frac{P(w_{N_i} \cap H_{S_k})}{\sum_{i=1}^{n} P(w_{N_i} \cap H_{S_k})} \quad \text{수식(4)}$$

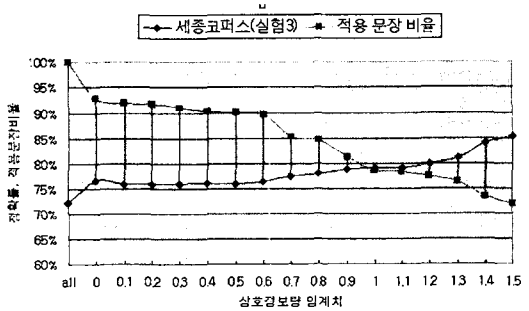
$$P(w_{N_i} \cap H_{S_k}) = \frac{\text{체언 } w_{N_i}\text{의 공기빈도}}{H_{S_k}\text{의 체언 공기빈도의 합}} \quad \text{수식(5)}$$

4.2 분석

상호정보량 임계치에 의해 선별(정제)된 의미정보를 동형이의어 분별에 적용한 결과이다. [그림 1]은 학습 문장(사전 뜻풀이)에 대한 세 가지 WSD 결과를 보여주고 있으며 [그림2]는 비학습 문장(세종 코퍼스)에 대해 실험(3)을 적용한 결과이다.



[그림 1] 상호정보량 임계치에 따른 사전 뜻풀이 문장 분석 결과(의미정보 Not found 문장을 제외한 정확률)



[그림 2] 상호정보량 임계치에 따른 세종 코퍼스 문장 분석 결과(의미정보 Not found 문장을 제외한 정확률)

[그림 1]을 보면 임계치 1.2 이하에서는 분석 가능 문장과 분별 정확률이 완만하게 변하지만 1.2 이상에서는 분석 가능 문장이 급하게 줄어듦을 볼 수 있다. 이는 상호정보량 1.2 이상의 의미정보가 학습 문장(사전 뜻풀이)의 동형이의어 분별에 중요한 정보임을 알 수 있다.

임계치를 증가하면 [표 4]에서와 같이 의미정보가 감소하게 된다. 특정 동형이의어와 관계없이 여러 문장에 사용되는 고빈도 어휘들은 상호정보량이 낮아 일반적으로 쉽게 제거 될 수 있다. 그러나 상호정보량은 낮지만 동형이의어 분별에 유용한 의미정보 역시 제거되는 문제점을 가지고 있다.

[표 5] 상호정보량 임계치를 1.2로 지정했을 때 '배 3(ship)'에서 제거되는 단어중 의미분별에 유용한 의미정보

상호정보량	의미정보
0.0≤X<0.4	향하다(0.18), 돌다(0.21), 세우다(0.38), 지나가다(0.39), ...
0.4≤X<0.8	조정(0.42), 허안(0.58), 움직이다(0.58), 뜨다(0.65) 물고기(0.67), 바람(0.71), 그물(0.77), ...
0.8≤X<1.2	파도(0.87), 내리다(0.85), 선박(0.97), 군함(0.90) 가라앉다(0.93), 보트(1.03), 연안(1.15), 바다(1.18)
1.2≤X	다니다(1.25), 타우다(1.21), 운하(1.23), 운행(1.32) 흘러가다(1.37), 해상(1.43), 함선(1.44), 실리다(1.45) 승무원(1.46), 잠기다(1.45), 뛰우다(1.53), 젓다(1.83) 입항(1.97), 원양(2.20), 정박(2.41)

[표 5]를 살펴보면 '향하다, 세우다, 지나가다, 움직이다, 내리다' 등은 시소러스나 개념망상에서 '배'의 상위어에 해당하는 운송수단의 의미자질에 해당하며 '뜨다, 가라앉다'등은 '배'의 의미자질에 하게 된다. 그리고 '해안, 파도, 연안'등은 '배'가 움직이는 공간인 '바다'의 하위어에 해당하는 단어들이다. 그러므로 시소러스나 개념망을 활용한다면 낮은 상호정보량에 의해 제거된 의미정보를 추가 할 수 있으며 부족한 의미정보도 확장할 수 있을 것이다.

5. 결론

본 논문에서는 의미분별에 불필요하거나 영향이 적은 의미정보를 정제하기 위해 단순 공기빈도가 아닌 상호정보량에 의해 정량화된 계수값을 구하고, 임계치를 통해 제한(정제)된 의미정보를 WSD 실험에 적용하였다. 실험 결과 1.1~1.2 사이의 임계치에 의해 정제된 약 50%의 의미정보 만으로 동형이의어를 효율적으로 분별할 수 있음을 보여준다.

본 연구는 앞으로 의미 분별에서 제외된 문장(의미정보를 찾을 수 없는 경우) 및 의미 분별에 실패한 문장을 분별하기 위해 유의어, 개념망 등을 이용한 의미정보 확장 방안을 지속적으로 연구해야 할 것이다.

6. 참고문헌

[1] 허정, 사전 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템, 울산대 석사학위논문, 2000

[2] J.S. Kim 외, A Korean Homonym Disambiguation System Based on Statistical model Using weights, Proceeding of The 16th Pacific Asia Conference, p166~176, 2002

[3] 이용우 외, Bayes 정리에 기반한 개선된 동형이의어 분별 모델, 제13회 한글 및 한국어 정보처리 학술대회, p465~471, 2001

[4] P.O. Cho, A Korean Noun Semantic Hirarchy based on Semantic Features, Proceeding of the 18th ICCPOL Vol1, 1999

[5] 옥은주 외, 의미속성 기반의 개념망을 위한 어휘 연관도 측정, 제13회 한글 및 한국어 정보처리 학술대회, p146~154, 2001

[6] 정영미 외, 한국어 텍스트 내 용어연관성 분석을 위한 기초연구, 제5회 한국정보관리학회 학술대회논문집, 1998