(제 10회 한글 및 한국어 정보처리 학술대회)

가중치 정보를 이용한 한국어 동사의 의미 중의성 해소

임 수 종, 박 영 자, 송 만 석

연세대학교 컴퓨터과학과
서울시 서대문구 신촌동 134 우: 120-749,
{mefong, yjpark, mssong}@december.yonsei.ac.kr

Word Sense Disambiguation of Korean Verbs Using Weight Information from Context

Soojong Lim, Youngja Park, Mansuk Song
Department of Computer Science,
Yonsei University

요약

본 논문은 문맥에서 추출한 가중치 정보를 이용한 한국어 동사의 의미 중의성 해소 모델을 제안한다. 중의성이 있는 단어가 쓰인 문장에서 그 단어의 의미 결정에 영향을 주는 단어들로 의미 결정자 벡터를 구성하고, 사전에서 그 단어의 의미 항목에 쓰인 단어들로 의미 항목 벡터를 구성한다. 목적 단어의 의미는 두 벡터간의 유사도 계산에 의해 결정된다. 벡터간의 유사도 계산은 사전에서 추출된 공기 관계와 목적 단어가 속한 문장에서 추출한 거리와 품사정보에 기반한 가중치 정보를 이용하여 이루어진다. 4개의 한국어 동사에 대해 내부실험과 외부실험을 하였다. 내부 실험은 84%의 정확률과 baseline을 기준으로 50%의 성능향상, 외부 실험은 75 %의 정확률과 baseline을 기준으로 40 %의 성능향상을 보인다.

1 서론

의미 중의성 해소(word sense disambiguation)란 의미 중의성을 갖는 단어가 속해 있는 문맥을 이용하여 사전에서 정의한 의미를 판별해 주는 작업을 말한다[10]. 예를 들어 '타다'라는 동사가 사전에서 '액체에 다른 것을 넣어 섞다'와 '탈것. 짐승의 몸 높은 곳 따위에 몸을 올려놓다' 라는 두 가지 의미를 갖는다고 할 때, '나는 기차를 탔다' 라는 문장에 쓰인 '타다'에 대해서 시스템이 자동적으로 '탈 것' 에 관련된 의미임을 결정해 주는 것이다. 의미 중의성 해소는 의미 태깅, 정보 검색, 기계 번역, 대화 분석을 포함한 많은 자연어 처리 응용 프로그램에서 사용되고 있다[5].

최근의 의미 중의성 해소 연구는 크게 두 가지로 나눌 수 있다[9]. 첫째 지식기반 방법으로 기계 가독형 사전이나 시소러스를 이용하는 방법이 있다[1,10,11]. 둘째 말뭉치에 기반하는 방법으로 이 방법은 태깅된 말뭉치를 사용한 교사 학습[2,4]과 태깅되지 않는 말뭉치를 사용하는 비교사 학습[7]으로 나눌 수 있다.

한국어의 의미 중의성 해소 연구로는 지식 기반과 비교사 학습을 함께 사용한 방법[13,15]과 교사학습을 사용한 방법[12,14]이 있다. [12]는 의미 태깅된 말뭉치에서 추출한 지역 정보를 이용하였고 [14]에서는 의미 태깅된 말뭉치에서 추출한 분류 정보를 이용하였다. [12, 14]는 의미 태깅된 말뭉치를 사용하여 자료 부족 현상과 지식획득을 위한 병목 현상을 보이고 있다. [13, 15]는 사전의 의미 구분에 따라 의미 중의성을 해소하였다. [13]은 사전의 의미 항목과 문장의 유사도를 계산할 때 자료 부족 현상을 보완하기 위하여 n-차 문맥 유사도를 추출하였다. [15]는 명사, 동사 분포와 사전의 정의 부분에 나타나는 부류 개념과 유사어를 이용하여 자료 부족 현상의 해결을 시도하였다. [13]은 유사도 계산시에 문맥 윈도우에서 공기 관계 정보만을 사용하였고 [15]는 목적어-서술어 관계만을 고려하여 체언류나 자동사의 의미 중의성 해결 방법으로 확장이 어렵다.

본 논문에서는 중의성이 있는 단어가 쓰인 문장에서 그 단어의 의미 결정에 영향을 주는 단어들로 의미 결정자 벡터를 구성하고, 사전에서 그 단어의 의미 항목에 쓰인 단어들로 의미 항목 벡터를 구성한다. 목적 단어의 의미는 두 벡터간의 유사도 계산에 의해 결정된다. 벡터간의 유사도

계산은 사전에서 추출된 공기 관계와 목적 단어가 속한 문장에서 추출한 거리와 품사정보에 기반한 가중치 정보를 이용하여 이루어진다.

한 단어의 의미는 의미를 나누는 사람의 주관에 따라서 의미가 세분되기도 하고 통합되기도 한다. 이러한 의미 구분의 특성 때문에 객관성을 확보하기 위해서 의미 중의성 해소를 위한 많은 연구가 기계 가독형 사전의 의미 구분에 따라서 진행되어왔고[5,6,12,13,14,15], 본 논문도 기계 가독형[1] 사전을 사용하여 의미 구분을 하였다. 의미 중의성 해소에 사용할만한 의미 태깅된 말뭉치가 부족한 현실을 고려하여 의미 태깅되지 않은 말뭉치를 사용하였다. 의미 분별에 필요한 문맥 정보에서 단순한 통계 빈도뿐 아니라 품사와 거리 정보를 포함하였다.

2 의미 결정자 벡터와 의미 항목 벡터의 정의

본 논문에서는 의미 중의성을 해소하고자 하는 단어가 속해 있는 한 개의 문장을 문맥으로 간주하였다. 그러나 문맥에 있는 모든 단어를 추출하는 것이 아니고 의미 결정자 벡터를 결정하는 과정에서 의미 결정에 영향이 없다고 판단되는 단어를 제거한다.

2.1 의미 결정자 벡터

의미 중의성을 해소하고자 하는 단어 w_a가 속해 있는 문장 i에 대한 의미 결정자 벡터, V_i,를 다음과 같이 구축한다. 문자 i에 나타난 명사, 동사, 형용사들 중에서 단어 w_a와 공기 관계 값, $Co(w_j, w_a)$,이 미리 주어진 임계값 , θ,을 넘는 단어들로 의미 결정자 벡터, V_i,를 식 (1)과 같이 구성한다. 이 실험에서는 임계값, θ,을 0.0001로 정하였다.

$$V_i = (w_1, w_2, \ldots, w_a, \ldots, w_{n-1}, w_n)$$

$$where \;\; Co(w_j, w_a) \geq \theta \;,\; 1 \leq j \leq n \quad \text{and}$$

$$Co(w_j, w_a) = \frac{f(w_j, w_a)}{f(w_j) + f(w_a) - f(w_j, w_a)} \quad (1)$$

$f(w_j)$: w_j의 빈도

$f(w_j, w_a)$: w_j, w_a 가 함께 나타난 빈도

예를 들어 '젊은 남자가 돈이 없어, 불도 피우지 못하고 밥도 먹지 못하였다.' 라는 문장이 있고 '피우다' 라는 단어의 의미 중의성을 해소하려 한다면 '젊다', '남자', '돈', '없다', '불', '피우다', '못하다', '밥', '먹다' 중에서 '피우다'와 공기 관계 값이 미리 설정한 임계값 0.0001을 넘지 못하는 '젊다', '없다', '못하다'를 제외하고 다음과 같이 의미 결정자 벡터 V_i가 결정된다.

V_i =(남자, 돈, 불, 피우다, 밥, 먹다)

2.2 의미 항목 벡터

사전에서 w_a가 k개의 의미를 갖는다고 할 때 w_a의 의미를 설명해 놓은 문장에서 명사, 동사, 형용사를 추출하여 다음과 같이 의미 항목 벡터를 구성한다.

$$v_1 = (w_{11}, \ldots, w_{1l_1})$$

$$\vdots$$

$$v_k = (w_{kl}, \ldots, w_{kl_k})$$

3 의미 중의성 해소에 사용한 정보

본 논문에서 의미 중의성을 해소하기 위하여 사용한 문맥 정보는 다음과 같다.

3.1 공기 관계 정보

본 논문에서는 [13]에서 추출한 공기 관계 정보를 이용하였다. 약 300만 어절의 기계 가독형 사전에서 공기 관계 정보를 획득하였으며 공기 관계는 같은 의미 항목에서 쓰였음을 뜻한다. 사전에서 사용된 단어는 의미를 명시적으로 설명하는데 사용한 단어들이기 때문에, 같은 의미 항목에서 사용된 단어들은 서로 의미적으로 연관도가 매우 높다. 공기 관계는 동사의 의미를 결정하는데 중요한 역할을 하는 명사, 동사, 형용사에 대해서만 추출을 하였다.

3.2 품사 정보

품사 정보만을 사용하면 완전한 의미 중의성 해소를 할 수는 없으나 의미 태깅을 위한 의미있는 단계로 진입할 수 있다[9].

동사의 의미를 구분하는데 있어서 타동사의 경우는 목적어로 사용된 명사가 의미 중의성을 해소하는데 중요한 정보가 되고 자동사의 경우도 주어에 사용된 명사를 포함하여 명사가 중요한 역할을 한다. 유사도 계산에 단순히 공기 관계만을 참조하는 것보다는 품사까지 고려하는 것이 의미 중의성을 해소하는데 도움이 된다. 동사의 의미 중의성을 해소할 경우 명사에 가중치를 둔다.

3.3 거리정보

1) 한글 학회의 우리말 큰 사전(1991)

의미 중의성을 해소하려는 단어와 문장에서 얼마만큼의 거리를 갖는가도 의미 중의성 해소에 중요한 정보가 된다. 동사의 경우는 특히 주위의 단어가 의미를 결정하는데 중요한 역할을 한다.

본 논문에서 사용된 거리 정보의 예는 다음과 같다.

예 2) 추위를 몹시 타는 여우가 불을 피웠다

예 2)에서 동사 '피우다' 의 의미 중의성을 해소할 때 '추위' '타다' '여우' '불' 등의 단어가 유사도 계산시 의미 결정자 벡터의 후보가 되나 '피우다'로부터 거리적으로 가깝게 쓰인 '불'이라는 단어가 사전에서 정의한 '피우다'의 네가지 의미[2] 중에서 예2) 문장에 쓰인 의미를 결정하는데 중요한 역할을 함을 알 수 있다. 일반적으로 한국어의 명사구는 가장 가까운 동사구에 부착됨을 감안하여 의미 중의성을 해소하고자 하는 단어와의 거리 정보를 유사도 계산시에 반영을 한다.

4 가중치 정보와 중의성 해소 알고리즘

공기 관계 정보를 바탕으로 유사도를 계산하는 방법[6,13]과 본 논문의 차이점은 가중치 정보에 있다. 의미 결정자 벡터와 의미 항목 벡터의 유사도를 계산할 때 문맥에서 얻을 수 있는 가중치 정보를 공기 관계 유사도 값에 반영하여 시스템의 성능을 향상시키고자 한다. 여기에서 사용한 가중치 정보는 목적 단어와의 거리 정보, 의미 결정자 벡터와 의미 항목 벡터의 품사 정보이다.

4.1 가중치 유사도 계산

식 (2)에 의해 의미 결정자 벡터 V_i와 v_1부터 v_k까지 유사도를 계산한다.

$$S_m = \sum_{i=1, i\neq a}^{n} \sum_{j=l_1}^{l_m} sim(w_i, w_{mj}) \quad (1 \le m \le k) \quad (2)$$

$$sim(w_i, w_{mj}) = weight_p(w_i) * weight_d(w_i) * Co(w_i, w_{mj})$$

$$weight_p(w_i) = \begin{cases} 1.5 \text{ if } w_i \text{ is noun} \\ 1.0 \text{ otherwise} \end{cases}$$

$$weight_d(w_i) = \begin{cases} 1.5 \text{ if } |distance(w_i, w_a)| \le 3 \\ 1.0 \text{ if } 4 \le |distance(w_i, w_a)| \le 5 \\ 0.1 \text{ otherwise} \end{cases}$$

유사도를 계산할 때 공기 관계 정보에 기반하여 계산을 하고 거리 정보와 품사 정보에 가중치를 둔다. 여기서는 ±3의 거리는 실험 가중치 값 1.5를 이용하였고, ±4~±5는 실험 가중치 값 1.0을 이용하였으며, 그 이상의 거리는 실험 가중치 값 0.1을 사용하여 가까운 거리일수록 목적 단어의 의미 결정에 영향을 주도록 하였다. 이 수치를 사용한 이유는 실험 문장의 한 문장당 평균 어절수가 내부 실험인 경우 6.5개이고 외부 실험의 경우 12.8개로 비교적 짧은 실험 문장을 사용하였기 때문이다. 품사 정보의 실험 가중치 값은 동사와 명사의 유사도 계산시에 1.5를 사용을 하였다.

4.2 알고리즘

1. 의미 중의성을 해소하고자 하는 단어 w_a가 속해 있는 문장 i에서 의미 결정자 벡터 V_i를 구성한다.
2. 사전에서 w_a의 의미 항목에서 의미 항목 벡터 v_m를 구성한다.(m= 1, ..., k)
3. 식 (2)를 사용하여 V_i와 v_m의 가중치 유사도를 계산한다.
4. S_m이 가장 높은 값을 갖는 의미 항목을 w_a의 의미로 선택한다.

5 실험 및 결과

동사 '감다', '피우다', '빠지다', '타다' 4개의 단어에 대해서 사전의 의미 구분을 다의적 중의성 수준으로 실험하였다. 실험 대상 말뭉치는 내부 실험을 위해서 공기 관계를 추출한 사전에서 문장을 추출하였고, 외부 실험을 위해서 연세 말뭉치[3] 중에서 초등학교 교과서 말뭉치에서 해당 단어가 발생하는 문장을 추출하여 수작업을 통하여 해당 단어가 두 번 나오는 문장과 형태소 분석이 잘못된 문장, 앞 뒤 문장을 보지 않고는 사람도 의미를 구분하기 어려운 문장을 제거하였다. 내, 외부 실험과 더불어 기존의 가중치와 의미 결정 벡터를 사용하지 않은 실험[4]과 본 논문의 실험[5]을 병행하였다. 각 단어에 대한 사전의 의미 항목 개수와 실험 문장에서 실제로 사용된 의미의 개수는 표 1과 같다.

		감다	피우다	빠지다	타다	평균
의미항목		3	4	15	18	10.00
실제 사용	실험1	3	4	11	14	8.00
	실험2	3	4	11	13	7.75

[표 1] 의미 항목 수와 실제 사용 수

본 논문에서 baseline은 실험 대상 문장에서 가장 빈번하게 발생하는 의미를 단어의 의미로 결

2) 1. 꽃이나 불을 피게 하다.
2. 어떤 상태나 태도 따위를 나타내다.
3. 담배 연기를 들이마시었다 내보내다.
4. 먼지나 냄새를 일으키거나 풍기다.

3) 연세대학교 언어정보개발연구원 한국어 사전 편찬실에서 1991년부터 수집한 약 4,500만 어절의 말뭉치

4) 실험 1 이라 명칭

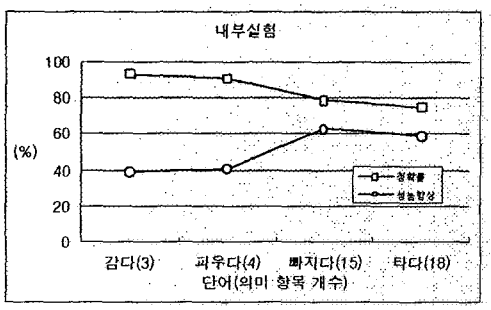
5) 실험 2 라 명칭

다의어		내부실험		
		baseline	정확률	정확률향상
감다	실험1	54	77.6	23.6
	실험2		92.7	38.7
피우다	실험1	50.0	75.0	25.0
	실험2		90.6	40.6
빠지다	실험1	15.4	30.8	15.4
	실험2		78.4	63.0
타다	실험1	16.1	41.1	25.0
	실험2		75.0	58.9
평균	실험1	33.9	56.1	22.3
	실험2		84.2	50.3

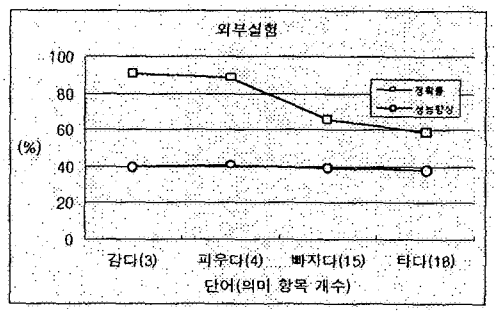
[표 2] 내부 실험 결과

다의어		외부실험		
		baseline	정확률	정확률향상
감다	실험1	51.0	79.4	28.4
	실험2		90.6	39.6
피우다	실험1	47.0	82.2	35.2
	실험2		88.0	41.0
빠지다	실험1	26.8	14.4	-12.4
	실험2		66.0	39.2
타다	실험1	20.2	32.3	12.1
	실험2		58.6	38.4
평균	실험1	36.3	52.1	15.8
	실험2		75.8	39.6

[표 3] 외부 실험 결과



[그림 1] 단어별 정확률과 정확률 향상 관계



[그림 2] 단어별 정확률과 정확률 향상 관계

정했을 때 정확률을 말한다[6]. 실험 결과 내부 실험은 84.2%의 정확률에 50.3%의 성능 향상 효과를 보였고, 의미 결정자 벡터를 구성하지 않고 가중치를 고려하지 않은 경우보다 정확률 면에서 27%의 정확률이 향상되었다. 외부 실험에서는 75.8%의 정확률에 39.6%의 정확률 향상 효과를 보였다. 의미 결정자 벡터를 구성하지 않고 가중치를 고려하지 않은 경우보다 22%의 정확률이 향상되었다. 가중치를 사용하지 않은 실험과 본 연구를 반영한 실험의 내부, 외부 실험 비교는 표 2, 3과 같다.

실험에서 사용한 단어와 단어에 대한 의미 구분이 다른 실험의 결과는 다음과 같다. 같은 baseline을 방식을 채택한 [14]는 내부 실험 중에서 동사만을 고려하면 99%의 정확률을 보였으며 35%의 정확률 향상을 보였고 외부실험 중에서 동사만을 고려하면 82%의 정확률에 13%의 정확률 향상을 보였다. 내부 실험만을 행한 [13]은 72%의 정확률, 42%의 정확률 향상을 보였다. baseline을 제시하지 않고 외부실험만을 한 [15]는 86.3%의 정확률을 보였다.

본 논문은 baseline을 기준으로 높은 정확률 향상을 보였다. 그림 1, 2는 각각 내부, 외부 실험에서 단어별 의미 항목의 개수, 정확률, 정확률 향상의 연관관계이다. 그림 1, 2에서 의미 항목의 개수가 많아질수록 정확률은 떨어지나 정확율 향상은 의미 항목 개수가 18개까지는 일정하거나 오히려 높아짐을 알 수 있다. 단어의 의미를 이용하는 응용 프로그램마다 단어의 의미 구분 정도는 달라질 수 있고 의미가 세분될수록 정확률이 떨어질 수 밖에 없다는 사실을 고려할 때 시스템의 성능은 정확률과 성능 향상률을 함께 고려하는 것이 바람직하다.

6 결론 및 향후 연구

본 논문은 사전의 의미 분류에 기반하여 사전에서 추출한 공기 관계에 문맥에서 얻은 품사정보와 거리 정보를 사용하여 한국어 동사의 의미 중의성 해소 방법을 제안하였다. 제안한 방법은 baseline을 기준으로 한 성능 향상율이 높다. 향후 연구 계획은 제안한 방법이 자료 부족 문제를 좀더 극복할 수 있도록 하고 한국어 체언류의 의미 중의성 해소에 이용할 수 있도록 가중치 정보를 좀더 다양하게 연구하는 것이다.

감사의 글

본 연구를 수행하는데 필요한 자료 조사와 많은 조언을 해 주신 울산대학교 국어국문과 한영균

6) 여기서 사용한 baseline 이외에 random하게 답을 결정해서 맞는 확률로 정하는 방법이 있다[9].

교수님의 도움에 감사를 드립니다.

참고문헌

[1] Bruce R. and L. Guthrie, "Genus Disambiguation: A Study in Weighted Preference ," Proceedings of COLING-92, pp. 1187- 1191, 1992

[2] Bruce R. and J. Wiebe, "Word-sense Disambiguation Using Decemposable Models," Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 139-145, 1994

[3] Gale B, K. Church and D. Yarowsky, "One Sense per Discourse," Proceedings of the DARPA Speech and Natural Lan- guage Workshop, pp. 233-237, 1992

[4] Hwee T. N. and Hian B. L, "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exmplar-Based Approach," Proceedings of the 34th annual Meetings of the Association for computational Linguistic, pp. 40-47, 1996

[5] Jen Nan Chen and Jason S. Chang, "Topical Clustering of MRD Senses Based on Information Retrieval Techniques," *Computational Linguistics*, Vol. 24 No.1, pp. 61-95, 1998

[6] Luk A, "Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions," Proceedings of ACL 95, pp. 181-188, 1995

[7] Ted Pedersen and Rebecca Bruce, "Distinguishing Word Senses in Untagged Text," Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, 1997

[8] Philip Resnik, "Selectional Preference and Sense Disambiguation," Proceedings ACLS-IGLEX Workshop, 1997

[9] Yorick Wilks and Mark Stevenson, "Combining Independent Knowledge Sources for Word Sense Disambiguation." Proceedings RANLP 97, 1997

[10] Yael Karov and Shimon Edelman, "Similarity-based Word Sense Disambiguation," *Computational Linguistics*, Vol. 24 No.1, pp. 41-59, 1998

[11] Yarowsky D, "Unsupervised Word Sense Disambiguation Rivaling Supervised methods," Proceedings of ACL95, pp. 189-196, 1995

[12] 김봉섭, 이종혁, 이근배. 말뭉치를 기반으로 한 한국어 명사의 의미 중의성 해소. 1997년 한국정보과학회 가을 학술 발표논문집 24권. No2 pp. 227-230. 1997

[13] 박영자. 사전을 이용한 단어 의미 자동 클러스터링: 유전자 알고리즘 접근법. 연세대학교 컴퓨터 과학과 박사 학위 논문. 1998

[14] 이호, 백대호, 임해창. 분류 정보를 이용한 단어 의미 중의성 해결. 정보과학회 논문지 (B) 제 24권 제 7호, pp. 779-789, 1997

[15] 조정미, 조영환, 김길창. 코퍼스와 사전을 이용한 한국어 동사 의미 분별. 1997년 한국정보과학회 가을 학술 발표 논문집 Vol. 24. No2, pp. 235-238. 1997