구문패턴과 순환 뜻풀이망을 이용한 동형이의어 분별

이왕우[0], 최호섭, 옥철영
울산대학교 컴퓨터정보통신공학부
{wwlee[0], hoseop, okcy}@mail.ulsan.ac.kr

Homonym disambiguation using syntactic pattern and recursive definition network

Wang-Woo Lee[0], Ho-Seop Choe, Cheol-Young Ock
Dept. of CIC, University of Ulsan

요 약

뜻풀이에서 추출한 의미 정보를 이용한 통계적인 방법의 기존 동형이의어 분별 시스템에는 불필요한 의미 정보들을 많이 가지고 있었다. 그리고 동형이의어간의 의미정보가 서로 교차하는 부분이 많아 확률적인 결정에 오류를 발생시켰다. 본 논문에서는 뜻풀이에서 구문패턴을 분석하여 보다 정제된 의미 정보를 추출하였고, 구문패턴에 속하는 어휘들의 하위어를 사전에서 자동 추출하여 부족한 의미 정보를 보완하였다. 또한, 구문패턴으로 분별할 수 없는 일부 동형이의어들은 순환 뜻풀이 망(RDN)을 이용하여 동형이의어를 분별하였다. 이러한 방법으로 동형이의어 분별을 통해 기존 연구보다 8%의 정확률 향상을 가져왔다.

1. 서 론

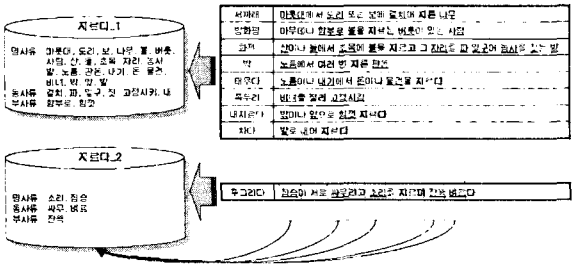
동형이의어를 분별하기 위해 여러 가지 방법이 시도되었는데, 초기에는 규칙을 이용한 방법이 주로 사용되었고 컴퓨터가 사전이나 대용량 코퍼스를 처리할 수준에 이르러서 코퍼스를 이용한 방법을 많이 연구하였다. 코퍼스를 이용한 방법은 학습 방법에 따라 의미 표지가 부착된 코퍼스를 이용하는 지도 학습과 의미 표지가 부착되지 않은 코퍼스를 이용하는 비지도 학습으로 나뉜다. 의미 표지가 부착된 코퍼스를 이용하는 방법은 통계적인 정보를 구축하여 의미 분별을 시도한다[1][2]. 또한, 시소러스나 WordNet, 뜻풀이에서 추출한 망을 이용한 방법이나 의미망(semantic network)이나 온톨로지(ontology)를 이용한 방법으로도 동형이의어를 분별하려는 시도도 있었다[5][8].

본 논문에서는 의미 표지가 부착된 코퍼스를 이용하여 구문패턴을 구축하고 Naï ve Bayes정리를 이용하여 통계정보를 마련하였다. 그리고 사전에서 추출한 하위어 집합과 순환 뜻풀이망을 의미 분별에 이용하려고 한다.

2. 기존 연구

기존 연구에서는 사전 뜻풀이에서 동형이의어와 공기하는 명사류, 동사류 그리고 부사류들을 의미정보로 구축하여 Naïve Bayes정리를 이용해서 통계정보를 마련하였다. 하지만 동형이의어와 공기하는 단어들을 의미정보로 구축하였기 때문에 동형이의어의 각 의미간에 공통되는 의미정보들이 많아 통계를 이용한 분별에 문제점이 있었다.

기존 연구에서는 동형이의어 분별을 위한 의미정보를 구축할 때 동형이의어와 공기하는 단어들을 수집하였다[2].



[그림 1] 기존의 의미 정보 구축 방법

그림 1처럼 의미정보를 수집하게 되면 [지르다]라는 동형이의어와 연관성이 높은 단어도 뽑히지만 [지르다]의 의미 분별과는 상관없는 표제어를 설명해주는 단어들도 뽑히게 된다. 그래서 동형이의어의 각각 의미마다 공통되는 의미정보가 많아지게 되고, 공통되는 의미정보가 많아지면 통계적으로 의미를 분별하는 데 어려움을 겪게 된다. 본 논문에서는 의미를 분별해주는 핵심적인 단어들만 뽑기 위해 구문패턴을 수집해서 의미 정보의 공통된 부분을 많이 줄일 수 있게 개선하였다.

문장에서 동형이의어의 의미를 결정해주는 데 큰 도움을 주는 요소는 격이 가지는 단어들일 것이다. 소성미(1998)는 타동사의 목적어 관계의 선택 제한 지식을 의미 분별 지식으로 이용하여 동형이의어를 분별하였다[4].

박영자(1997)는 사전의 의미 기술관계로부터 구축된 네트워크인 의미 참조망(sense reference network)을 이용하여 의미속성을 추출하고 클러스터링을 하였다[6]. 본 논문에서는 순환 뜻풀이망을 동형이의어 의미 분별에 이용한다.

3. 사전에서 추출한 의미 정보

사전에서 구문패턴과 하위어집합 그리고 순환 뜻풀이망을 구축하였다.

3.1 동형이의어 분별을 위해 뜻풀이에서 구문패턴 추출

다음은 사전 뜻풀이에서 뽑은 구문패턴의 종류들이다.

3.1.1 주격, 목적격, 보격, 부사격 패턴

문장에서 동사의 의미를 분별하는 데 큰 도움을 주는 단어는 격이 가지는 체언류라고 가정하고 구문패턴을 사전 뜻풀이에서 뽑았다.

뜻풀이에서 동사와 연관되는 격의 체언류들을 뽑기 위해 먼저 격의 체언류를 뽑을 범위를 설정하였다. 범위는 동사와 동사(동형이의어) 사이로 정하였다. 그리고 범위 사이에 들어있는 격의 체언류들을 구문패턴으로 구축하였다.

3.1.2 ' 명사 - 동사(동형이의어)' 패턴

' 명사 동사' 패턴을 고려한 이유는 일부 문장에서 격의 생략 현상이 나타나기 때문에 격이 나타나지 않아도 의미 분별할 ' 동사' 앞에 있는 ' 명사' 는 의미 분별에 큰 도움이 된다.

예문) 권총을 쏘고, 그림을 그리고, 불 지르던, 그 영화는 간 데 없다.

3.1.3 ' 동사(동형이의어)+관형형 전성어미 - 명사' 패턴

수식어구는 명사류와 수식관계에 있는 용언류 중 관형형 전성어미가 부착된 것들을 패턴 정보로 구축하였다. 수식받는 명사가 수식하는 동사의 의미 분별에 영향을 주기 때문에 의미정보로 활용하였다.

예문) 앞사람들이 맞고 지르는 비명 같았다.

3.1.4 ' 동사(동형이의어) - 동사(동형이의어)' 패턴

동사의 의미를 분별하기 위해 코퍼스를 분석해보면 격이 없어 의미 정보를 얻기가 어려운 경우가 있다. 이런 부분을 해결하는 데 이어진 동사 패턴이 의미 분별에 유용하게 사용될 수 있다.

예문) 사람의 몸에 비치기 때문에 늘 바르게 살아야 해.

3.2 사전에서 추출한 하위어 집합

동형이의어를 분별하는데 구문패턴만 이용하다 보면 의미정보의 부족현상이 나타난다. 의미 정보의 부족현상을 해소하기 위해 사전을 이용하여 구문패턴에 나오는 단어의 하위어들을 뽑아 의미 정보로 활용하였다.

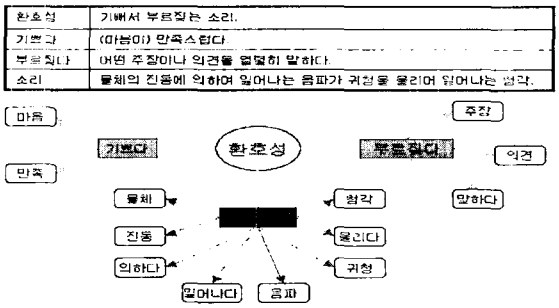
조평옥(1997)은 뜻풀이의 형태를 11가지로 분류하였는 데 그 중 첫번째 형태(뜻풀이의 맨 끝에 핵심어가 있는 형태)를 이용하여 하위어를 구축하였다.[3]

[표 1]. 사전에서 하위어 추출

표제어	뜻풀이
환호성	기뻐서 부르짖는 소리.
고함	크게 외치는 소리.
횃불	홰에 켠 불.

3.3 순환 뜻풀이 망

사전의 뜻풀이는 동형이의어의 의미를 분별하는 데 실마리를 제공해 준다[7]. 동형이의어의 뜻풀이와 문장에 출현하는 단어의 뜻풀이간에 관련성을 정의할 수 있다면 그 관련성을 의미 분별에 사용할 수 있다.



[그림 2] 순환 뜻풀이 망(RDN)

그림 2는 뜻풀이를 이용하여 순환망을 구성한 것이다. 두 단어를 순환망으로 범위를 넓혀 가면 매칭되는 부분이 생기는데 매칭되는 개수를 의미정보의 빈도로 사용하였다. 매칭되는 빈도가 높은 의미를 의미 분별 결과로 결정한다.

4. 동형이의어 분별

동형이의어 의미 분별 과정은 다음과 같다.

1. 패턴에 적용되는 빈도값을 Naïve Bayes 확률값으로 바꿔 점수를 매긴다.
2. 패턴을 적용한 결과값이 모두 0일 경우 하위어 집합을 이용하여 점수를 매긴다.
3. 2단계의 결과값도 0일 경우 순환 뜻풀이 망을 이용하여 패턴정보와 동형이의어와의 관련성을 구한다.

4.1 구문패턴 적용

뜻풀이 말에서 추출한 구문패턴을 의미 분별할 동형이의어가 포함된 문장에서 추출한 구문패턴과 비교하여 일치하는 의미정보에 대한 점수를 계산한다.

$$WSD(C, H_{S_i}) = \arg MAX_{H_{S_i}} Sim(C, H_{S_i}) \quad \text{- 수식(1)}$$

Hs_i 는 동형이의어 H가 가지는 i번째 의미를 나타낸다. $Sim\ (C,\ Hs_i)$는 문장 C와 의미 Hs_i 의 관련성을 나타낸다. 수식(1)은 $Sim\ (C,\ Hs_i)$의 의미 자질값들 중 최대인 값을 가지는 의미를 선택하여 동형이의어를 분별한다. $Sim\ (C,\ Hs_i)$는 아래 수식(2)에 의해서 구해진다.

$$Sim(C,H_{S_i}) = Case(C,H_{S_i}) + NV(C,H_{S_i}) + VN(C,H_{S_i}) + VV(C,H_{S_i})$$

- 수식(2)

*Case(C, Hsi)*는 문장 *C* 에서 출현하는 격이 가지는 어휘들과 의미 Hs_i의 관련성이고, NV(*C, Hsi*)는 문장 *C*에서 ' 명사-동사' 패턴이 출현할 때 명사와 의미 Hs_i의 관련성이며, VN(*C, Hsi*)는 문장 *C*에서 ' 동사(동형이의어)+관형형 전성어미 - 명사' 패턴이 출현할 때 명사와 의미 Hs_i의 관련성이다. 그리고 VV(*C, Hsi*)는 문장 *C*에서 ' 동사(동형이의어)-동사(동형이의어)' 패턴이 출현할 때 동사와 의미 Hs_i의 관련성이다.

$$Case(C,H_{S_i}) = \sum_{j=1}^{n} P(H_{S_i} \mid W_{case_j})$$

- 수식(3)

$$NV(C,H_{S_i}) = \sum_{j=1}^{n} P(H_{S_i} \mid W_{nv_j})$$

- 수식(4)

$$VN(C,H_{S_i}) = \sum_{i=1}^{n} P(H_{S_i} \mid W_{vn_j})$$

- 수식(5)

$$VV(C,H_{S_i}) = \sum_{i=1}^{n} P(H_{S_i} \mid W_{vv_j})$$

- 수식(6)

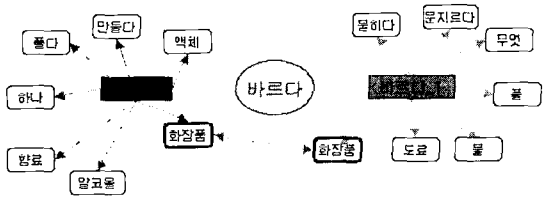
$P(Hs_i \mid W_j)$는 동형이의어(H)가 포함된 문장에서 의미정보에 속하는 단어(W_j)가 나타났을 때, Hs_i의 의미로 해석될 확률을 나타낸다. 의미정보에 단어(W_j)가 나타나지 않으면 하위어 정보를 이용하여 확률을 계산한다.

$$P(H_{S_i} \mid W_j) = \frac{P(W_j \cap H_{S_i})}{\sum_{i=1}^{n} P(W_j \cap H_{S_i})}$$

- 수식(7)

4.2 순환 뜻풀이 망 적용

수식(1)에 의해서 동형이의어를 분별하지 못할 경우, 즉 확률값이 0일 경우 동형이의어를 분별하기 위해 순환 뜻풀이 망을 이용한다.



[그림 3] 순환 뜻풀이 망을 이용한 동형이의어 분별

그림 3은 "향수를 바르다"라는 문장을 의미 분별하는 데 '바르다'의 의미 정보에 '향수'라는 단어가 없어 RDN을 이용하여 동형이의어의 의미 분별 방법을 보여준다. '향수'의 뜻풀이와 '바르다' 각각 의미의 뜻풀이와 일치하는 단어의 개수가 많은 의미를 선택한다.

5. 실험 및 평가

기존 연구[2]와의 비교를 위하여 의미 분별할 동사는 기존 연구에서 사용하였던 동사 중 일부를 테스트하였다. '지르다', '지다', '바르다'를 테스트하여 기존 연구와 비교하였다.

5.1 실험에 사용된 데이터

실험에 사용된 데이터는 세종계획의 의미 표지가 부착된 코퍼스를 사용하였다. 세종계획의 의미 표지 부착된 150만 어절에서 702문장을 뽑아 테스트 문장으로 사용하였다.

5.2 실험 결과

표 2를 보면 '지르다', '지다', '바르다'중 '지르다'가 가장 높은 정확률을 보였다. '지르다'는 코퍼스에 나타난 구문패턴들이 다양하지 않아 의미 정보가 부족하지 않았는데, '지다'나 '바르다'는 코퍼스에서 나타나는 구문패턴들이 사전에서 나타나는 구문패턴보다 다양해서 정확률이 높지 않았다. 하지만 기존 연구와 비교해 볼 때 전체적으로 평균 8%정도 정확률이 높아졌다.

[표 2] 실험 결과

동형이의어 (의미 개수)	맞는 개수	총 테스트 문장	정확률(평균)
지르다(2)	132	149	88.6%
지다(5)	276	368	75%
바르다(2)	143	185	77%

6. 결론 및 향후 연구

의미 정보의 추출방법을 개선하여 통계적인 방법을 사용할 때 교집합 의미정보 때문에 일어나는 문제점들이 개선되었다.

구문패턴을 이용할 때 동일한 격에서 사용되는 어휘는 아니지만 테스트 문장에 나타난 어휘가 의미 정보에 들어 있을 때 계산하는 방법을 연구해야 한다. 그리고 순환 뜻풀이망을 사용할 때 명사와 동사(동형이의어)의 관련성 외에 명사와 명사(구문패턴)와의 관련성도 고려하는 방안을 모색해야 한다.

7.참고문헌

[1] 허 정, 2000, "사전 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템", 울산대학교 석사학위 논문.
[2] 이왕우, " Bayes 정리에 기반한 개선된 동형이의어 분별 모델", 제 13회 한글 및 한국어 정보처리 학술대회 발표논문, 2001
[3] 조명옥, 1996," 한국어 명사의 의미 계층 구조 구축", 울산대학교 석사학위 논문.
[4] 조정미, 1998, " 코퍼스와 사전을 이용한 동사 의미 분별", 한국과학기술원 박사학위 논문.
[5] 강신재, 2002, " 실용적인 온톨러지의 반자동 구축 및 어휘 의미 중의성 해소를 위한 응용", 포항공과대학교 박사학위 논문
[6] 박영자, 1997, " 사전을 이용한 단어 의미 자동 클러스터링: 유전자 알고리즘 접근법", 연세대학교 박사학위 논문
[7] Jean Veronis, Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries, COLING, August 1990
[8] Xiaobin Li, A WordNet-based Algorithm for Word Sense Disambiguation, 1995