

어휘 공기 집합과 시소리스를 활용한 한국어 동형이의어 분별

울산대학교 컴퓨터정보통신공학부
이왕우[†] · 최호섭 · 김준수 · 옥철영

Disambiguation of Korean Homonym Using Lexical Co-occurrence Set and Thesaurus

Wang-Woo Lee, Ho-Seop Choe, Jun-Su Kim, Cheol-Young Ock

Department of Computer Engineering and Information Technology, University of Ulsan, Ulsan, Korea

요 약

본 논문은 한국어 정보처리에서 발생하는 어휘 중의성 문제 중에 동형이의어 분별의 중의성을 해결하기 위하여, 유용한 구문 패턴을 바탕으로 사전 뜻풀이와 150만 어절의 말뭉치에서 어휘 공기 집합을 추출하여 동형이의어의 분별에 이용하였다. 특히, 용언류 동형이의어를 분별할 때에는 어휘 공기 집합의 자료 부족문제를 해결하기 위하여 시소리스를 이용한 어휘 공기 집합의 확장 방법을 제시한다. 시소리스 확장을 통한 분석에서 동형이의어의 분별이 실패할 경우 제한된 어절을 대상으로 통계적인 분석을 시도하여 동형이의어를 분별한다. 중의성이 높은 469개 동형이의어에 대하여 2가지 실험을 통해 각각 90.05%와 92.23%의 정확률을 얻을 수 있었다.

서 론

언어처리에는 여러 가지 중의적인 문제들이 발생하고, 이를 해결하기 위해 많은 연구가 진행되고 있다. 해결되어야 할 중의성은 크게 형태, 품사, 구문, 의미 중의성으로 나눌 수 있다. 중의성의 해결 방안으로 쉽게 사용할 수 있는 방법은 공기관계를 이용하는 것이다.^{4,5,8,10)} 지금까지 많은 자연언어처리에서 공기 관계를 이용하여 중의성을 해결하였으며 사용된 사례를 살펴 보면, ⁵⁾에서는 명사간의 유사성을 코퍼스에서 공기관계를 이용하여 학습하고 병렬 구문과 무표지 격의 중의성 해결에 사용하였다. ¹⁰⁾에서는 격관계와 수식관계의 구문 공기 정보를 이용하여 영-한 번역의 역어 선택에 활용하였다. ⁴⁾에서는 의미별로 수집한 문맥 정보를 기저 벡터화하는 방법을 이용하여 한국어 명사의 의미 분별을 시도하였다. 이처럼 공기 정보는 한국어 분석의 여러 가지 중의성 해결에 사용이 가능하다. 본 논문에서도 구문 패턴에서 추출한 어휘간의 공기정보를 이용하여 동형이의

어 분별을 시도한다.

¹³⁾에서는 ⁵⁾에서 사용되었던 공기 유사도의 문제점으로 다의어와 정보의 부족을 제시했는데, 정보의 부족 현상은 시소리스를 활용하여 일부 해소하였다. 본 논문에서도 어휘 공기 집합의 정보 부족 현상을 시소리스를 활용하여 해결을 시도한다.^{15,16)}

2장에서는 어휘 공기 집합의 추출을 위한 품사 패턴에 대해 설명하고, 3장에서는 어휘의 부족 현상을 해결하기 위한 시소리스와 개념망의 사용방법에 대해 설명한다. 4장에서는 동형이의어 분별을 위한 전체적인 과정을 설명한다. 5장에서는 실험 및 결과에 대해서 설명하고 6장에서는 결론 및 향후 연구에 대해 논한다.

어휘 공기 집합의 추출을 위한 품사 패턴

⁹⁾에서 사용되었던 어휘 공기 집합(Lexical Co-occurrence Set, 이하 LCS)은 몇몇 동형이의어를 분별하기 위한 품사 패턴을 참고하여 LCS를 구축하였다. 본 논문에서는 몇몇 동형이의어가 아닌 전체 동형이의어 약 14,000개에 대해서 의미 분별에 유용한 품사 패턴을 추출한다. 의미를 분별하는 데 도움을 주는 어휘를 어절 단위로 사람이

[†]E-mail : wwlee@mail.ulsan.ac.kr

E-mail : hoseop@mail.ulsan.ac.kr

E-mail : kimjunsu@mail.ulsan.ac.kr

E-mail : okcy@mail.ulsan.ac.kr

직접 수집하였고, 이 정보를 기반으로 품사 패턴을 새롭게 추출하였다. 그리고 추출된 품사 패턴으로 뜻풀이에서 어휘 공기 집합을 추출하게 된다. 의미 분별 정보를 수집하기 위한 작업은 도구를 이용하였으며, 작업한 대상은 세종계 획에서 제공하는 의미 주석된 코퍼스 150만 어절을 사용하였다. Fig. 1은 의미 분별에 필요한 정보를 수집하는 데 사용된 도구로써, 동형이의어를 기준으로 앞뒤 10어절씩을 보여주고 동형이의어를 분별하는 데 유용한 어휘들을 어절 단위로 선택할 수 있게 만들어졌다.

1. 의미 분별 어휘의 어절 위치

의미 분별에 필요한 어휘를 수집한 결과 150만 어절중 631,779개의 의미 분별 어휘가 수집되었다. 동형이의어를 기준으로 63만 개의 의미 분별 어휘가 수집된 어절상의 위치는 Fig. 2와 같다.

Fig. 2를 보면 동형이의어를 기준으로 앞뒤 5어절 이내에 의미 분별 어휘가 위치하는 빈도는 621,872개이다. 이는 전체 의미 분별 어휘의 98%가 동형이의어의 앞뒤 5어절 이내에서 나타난다고 할 수 있다.²⁾에서는 어절의 제한을 이용하여 동형이의어 분별 모델을 제시하였는데 위 빈도 정보를 볼 때 타당한 방법론이라고 볼 수 있다.¹⁷⁾ 본 논

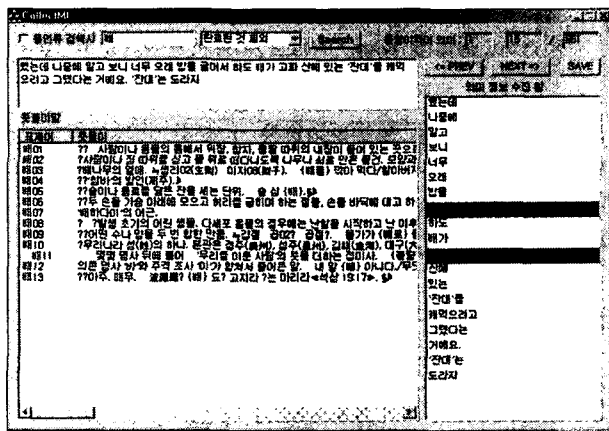


Fig. 1. 의미분별정보 수집도구.

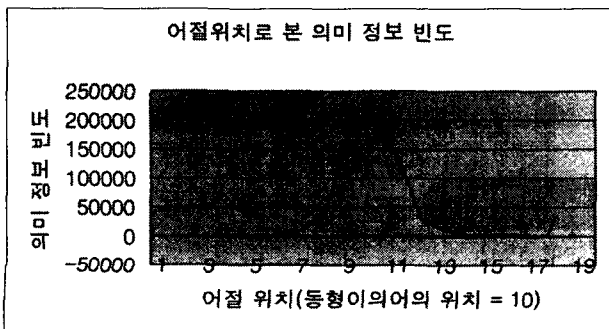


Fig. 2. 어절 거리와 의미 분별 정보의 빈도.

문에서도 어휘 공기 집합과 시소러스를 활용하여 동형이의어를 분별하지 못할 경우 5어절 이내의 단어들을 정보로 활용하여 동형이의어 분별을 시도한다.

2. 동형이의어(명사)를 분별하기 위한 의미 분별 어휘의 품사 패턴

동형이의어(명사)에 대한 품사 패턴의 추출 과정은 동형이의어를 기준으로 앞에 위치한 의미 분별 정보의 품사 패턴 추출과 뒤에 위치한 의미 분별 정보의 품사 패턴 추출로 나뉜다. 동형이의어를 기준으로 앞쪽에 위치한 의미 분별 정보의 품사 패턴을 추출하기 위하여 의미 분별 정보의 어절과 동형이의어가 포함된 어절의 품사 쌍을 추출한다. 추출된 품사 쌍의 종류는 총 21,456개이고, 각 종류의 빈도에 대한 합은 202,229개이다.

추출된 패턴들의 종류가 2만 개를 넘어 품사 패턴 정보로 활용하기가 어렵기 때문에 품사 패턴을 단순한 형태로 줄이는 작업을 한다. Table 1의 품사 쌍을 보면 'VA+ETM' - 'NNG+JKB'와 'VA+ETM' - 'NNG+JKO'가 각각 953개, 927개의 빈도를 차지하고 있는데, 실제 패턴을 활용하는 측면에서는 'NNG+JKB'나 'NNG+JKO'를 하나의 품사 패턴으로 봐도 상관이 없으므로 'NNG+...' 형식으로 변경해서 품사 패턴을 단순화시킨다. Table 2는 단순화 과정을 거친 품사 패턴에서 고빈도 패턴들만 정리한 것이다.

Table 2를 살펴보면 동형이의어와 복합명사를 이루는 'NNG-NNG' 패턴이 의미 분별에 가장 큰 비중을 차지

Table 1. 동형이의어(명사)의 앞에 위치한 의미 분별 정보의 품사 패턴의 일부

의미분별어절품사	동형이의어품사	빈도
	...	
VV+ETM	NNG+JKS	1,043
NNG+JKB	NNG+JKO	1,033
VV+ETM	NNG+JX	1,028
VV+ETM	NNG+JKO	1,003
NNG+JKO	NNG+JKB	991
VV+ETM	NNG	989
VA+ETM	NNG+JKB	953
VA+ETM	NNG+JKO	927
NNG	NNG+JC	916
MM	NNG+JKO	809
NNG	NNG+JKB+JX	799
NNG	NNG+XSV+ETM	766
NNG	NNG+SP	762
MM	NNG+JX	749
NNG+JKS	NNG+XSV+EC	738
NNG	NNG+XSV+EC	737
	...	

하고 있다. 동형의의어가 'NNG+XSV...' 형태일 경우, 동형의의어가 서술성 명사이기 때문에 의미분별에 영향을 주는 품사가 격조사를 취하는 명사, 부사 그리고 동사가 된다.

Table 3은 동형의의어(명사)의 뒤에 위치한 의미분별정보의 품사 패턴을 단순화시켜서 고빈도 패턴들만 정리한 것이다.

Table 3을 보면 동형의의어의 의미를 분별해 주는 중요한 품사들은 주로 용언류와 'NNG+XSV' 형태임을 알 수 있다. 그리고 동형의의어가 'NNG+XSV+ETM' 형태일 경우 수식을 받는 명사가 의미 분별에 중요한 역할을 하게 된다. 동형의의어 의미 분별에 중요한 역할을 하는 고빈도의 품사 패턴들은 의미 분별 정보의 수집과 동형의의어의 분별에 활용된다.

Table 2. 동형의의어(명사)의 앞에 위치한 의미 분별 정보의 고빈도 품사 패턴

의미분별정보의 품사	동형의의어의 품사	빈도
NNG		31,480
NNG+XSV+ETM	NNG+...	5,013
MM		6,908
VV+ETM VA+ETM	NNG+J...	8,824
VX+ETM	NNG+J...	1,231
NNP	NNG	1,178
MAG		1,473
NNG NNG+(JKO, JKB, JKS, JX)	NNG+XSV+...	17,699
VV+EC		1,006
NNG+JKG	NNG+(JKO, JKB, JX, JKS) NNG	8,266

Table 3. 동형의의어(명사)의 앞에 위치한 의미 분별 정보의 고빈도 품사 패턴

동형의의어의 품사	의미분별정보의 품사	빈도
	NNG+...	27,198
NNG	VV+...	2,746
	VA+...	641
	VV+...	5,468
NNG+JKS	VA+...	1,873
	NNG+XSV+...	1,460
NNG+JKO	VV+...	16,397
	NNG+XSV+...	4,525
NNG+JKG	NNG+...	8,832
NNG+JKB	VV+...	6,990
	NNG+XSV+...	2,594
NNG+JX	VV+...	2,338
	VA+...	799
NNG+XSV+ETM	NNG+...	2,025
	VX+...	1,354
NNG+XSV+EC	VV+...	348

3. 동형의의어(용언)의 품사 패턴

동형의의어(용언)의 품사 패턴 추출 과정도 품사가 명사인 동형의의어의 품사 패턴 추출 과정과 동일하게 작업하였다.

Table 4의 품사 패턴들은 Table 3의 품사 패턴과 유사한 부분이 많이 있다. 이것은 격어휘와 동사와의 관계가 의미적으로 관계가 많음을 보여준다.

Table 5에서 'V류+ETM' - 'NNG+...' 형태는 Table 2에도 나왔던 패턴이고, 'VV+EC' 형태는 바로 뒤 1어절 이내에서만 품사 패턴을 추출해야 좋은 공기 관계를 뽑을 수 있다.

이전 연구에서는 용언의 공기 집합만 추출한 결과로 291,412개가 추출되었는데 Table 4, 5에 나온 품사 패턴으로 새롭게 용언의 공기 집합을 추출한 결과 414,915개의 공기 관계가 추출되었다.⁹⁾

공기 어휘 집합은 동형의의어뿐만 아니라 모든 어휘에 대해서 공기 관계를 추출하여 구축하였다. 동형의의어 분별뿐만 아니라 의존 관계를 이용한 구문 분석 등에 사용될 수 있기 때문이다.

시소러스

¹²⁾에서는 사전의 뜻풀이에서 시소러스를 구축하기 위하여 뜻풀이의 패턴 11가지를 정의하고 Bottom-up 방식으로 시소러스를 구축하였다. 사전의 뜻풀이에서 시소러스를 구

Table 4. 동형의의어(용언)의 앞에 위치한 의미 분별 정보의 고빈도 품사 패턴

의미분별정보품사	동형의의어품사	빈도
NNG+JKO	VV+...	11,187
NNG+JKB	VV+...	7,339
NNG	VV+...	2,152
NNG+JKS	VV+... / VA+...	2,092
MAG	VV+...	1,547
VV+EC	VV+...	1,248
NNG+JX	VV+...	946
VV+ETN	VV+...	918
NNG+XSV+ETN	VV+...	841

Table 5. 동형의의어(용언)의 앞에 위치한 의미 분별 정보의 고빈도 품사 패턴

동형의의어품사	의미분별정보품사	빈도
VV+ETM	NNG+...	1,679
VV+EC	VX+...	1,420
VA+ETM	NNG+...	1,139
VV+EC	VV+...	686

축하게 되면 사람의 주관이 개입되지 않아 객관적인 시소러스가 구축될 수 있으나 뜻풀이의 비일관적인 기술방식 때문에 시소러스의 구조 역시 일관성을 지니지 못하게 되는 문제점이 있다. 이런 문제를 개선하기 위하여 사전의 뜻풀이에서 상위개념(중심어)를 수식하는 어절과 용언의 의미호응관계에서 의미 속성을 추출하여 이들 의미 속성에 의한 명사의미체계를 구축한 시소러스가 연구된 바 있다.⁵⁾ 본 논문에서는 의미 호응 관계를 이용한 개선된 시소러스를 사용하였다. 사용된 시소러스의 총 노드 개수는 18,946개이다.

공기 어휘 집합에서 의미 분별 정보로 활용할 수 있는 단어들 없이 경우에 시소러스로 단어들을 확장하여 의미 분별을 시도하게 된다. 사용하는 시소러스의 각 노드에 있는 명사들이 의미 주석되어 있지 않으므로 명사에 대한 동형이의어 분별에는 시소러스를 사용하지 않고 용언류 동형이의어의 분별에만 시소러스를 이용한다. 시소러스를 이용하여 용언류 동형이의어를 분별하는 공식은 아래와 같다.

[수식 1] 시소러스와 공기어휘집합을 이용한 동형이의어 분별

$$HG's\ Sense = argMax \sum_{i=1}^n LCS_{hg, \square} Tsr_{co-word}$$

[수식 1]은 공기 어휘 집합의 부족한 어휘들을 시소러스를 이용하여 분별하기 위해 사용할 수식이다. 동형이의어(hg)가 가지는 의미의 가지수는 n개이다. LCS는 공기 어휘 집합이고 LCS_{hg}는 동형이의어(hg)를 포함하는 공기 어휘 집합이다. 그리고 Tsr_{co-word}는 동형이의어(hg)와 구문 패턴 관계를 이루는 공기 어휘가 포함되어 있는 형제 노드들의 집합이다. 즉, [수식 1]은 동형이의어(hg)의 각 의미에 대한 공기 어휘 집합이 어휘 정보 부족으로 해결하지 못하는 어휘에 대하여 형제 노드들과 가장 많이 일치하는 쪽으로 의미를 분별하게 된다.

예를 들어, “논문을 쓰다.”라는 문장에 대하여 동형이의어를 분석하는 과정은 다음과 같다. 먼저, ‘쓰다’가 동형이의어이고 6가지의 의미를 갖는다. Table 6은 ‘쓰다’의 의미이다. ‘논문’은 ‘쓰다’와 공기하는 격어휘이므로 LCS에서 ‘논문/NNG,JKO,쓰/VV’를 찾는다. LCS에 ‘논문/NNG,

Table 6. ‘쓰다’의 뜻풀이

의미	뜻풀이
1	펜,연필,붓 따위로 획을 그어 글자를 이룬다.
2	(모자 따위를) 머리에 얹다.
3	사용하다. 이용하다.
4	시체를 묻고 무덤을 만들다.
5	웃, 장기 등에서, 말을 옮기다.
6	맛이 소태와 같다.

JKO,쓰/VV’가 없어 시소러스를 이용하게 된다. 일단 시소러스에서 ‘논문’을 찾아 형제 노드들을 가져온다. ‘논문’의 형제 노드들과 ‘쓰다’ 각각의 의미가 가지는 공기 어휘들을 일치시켜서 가장 많이 일치하는 쪽으로 ‘쓰다’의 의미를 결정한다. Table 7에서 볼 수 있듯이 ‘쓰다_1’과 일치된 빈도는 70이고 ‘쓰다_3’과 일치된 빈도는 4이다. 그래서 ‘쓰다_1’로 의미가 결정된다.

그러나 ‘쓰다_3’에서 일치된 어휘중 ‘서문’과 ‘한문’은 ‘쓰다_1’의 의미와 공기하는 어휘들인데 학습 코퍼스의 의미 태깅 오류로 인하여 발생된 결과이다. 그리고 ‘쓰다_3’과 일치된 어휘중에 ‘전기’ 같은 경우는 동형이의어이기 때문에 발생한 문제이다. 시소러스상에 동형이의어에 대한 구분이 있다면 이런 매칭은 일어나지 않았을 것이다.

4. 동형이의어 분별 과정

Fig. 4는 전체적인 동형이의어 분별 과정을 나타낸다. 먼저, 문장을 입력 받아 품사 태깅을 한 뒤 동형이의어가 있

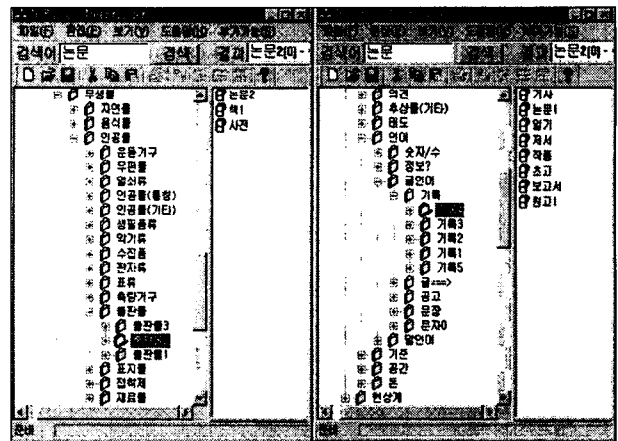


Fig. 3. 시소러스(논문의 형제 노드들).

Table 7. ‘논문’의 형제 노드와 ‘쓰다’의 공기 빈도

형제노드	의미번호	빈도	형제노드	의미번호	빈도
각본	1	3	서문	3	1
기사	1	1	전기	3	2
문장	1	4	한문	3	1
방문	1	1			
산문	1	2			
소설	1	1			
시	1	2			
원고	1	8			
이름	1	19			
편지	1	29			
총합	1	70	총합	3	4

을 경우 동형어의어와 관련된 품사 패턴을 찾는다. 품사 패턴에 의해 찾아진 어휘들이 LCS에 동일한 품사 패턴으로 일치하는 경우, LCS에 저장된 의미로 분석을 하게 된다. 품사 패턴의 어휘가 LCS에 일치되는 경우가 없을 경우, 시소러스를 이용하여 동형어의어 분별을 시도하는데, 이때 [수식 1]을 사용한다. 시소러스를 이용하여 동형어의어 분별에 실패할 경우, 앞뒤 5어절 이내의 어휘를 대상으로 Naive Bayesian정리를 이용한 계산 방법을 사용하여 동형어의어를 분별한다.⁹⁾ 끝까지 분석이 안될 경우 가장 높은 빈도의 동형어의어 의미를 선택한다.

실험 및 결과

실험에 사용된 동형어의어의 종류는 469개이다. 실험은 학습 말뭉치와 테스트 말뭉치를 달리하여 2가지 방향으로 진행하였다. 첫 번째는 ³⁾에서와 동일한 조건에서 테스트를 하였다. 즉, 국어사전의 의미태깅된 뜻풀이를 학습 말뭉치로 이용하고 세종계획의 의미 태깅된 150만 어절을 실험 대상으로 이용하였다. 첫 번째 실험에서 전체적인 동형어의어 분별과정을 통해 분석된 정확률은 90.05%이다. ³⁾에서보다 평균 정확률이 1%정도 낮은 수치를 보이고 있다. 1%

정도 낮은 이유는 학습 말뭉치로 사용한 의미 태깅된 사전의 뜻풀이에 의미 태깅의 오류가 있게 되면 LCS를 이용한 방법은 ³⁾의 naive bayesian모델과는 달리 치명적인 결과를 가져오게 된다. 또한 Table 8처럼 '같이'가 '쓰다_1'과 '쓰다_3'에 동일한 분별정보가 될 경우에는 약간의 오류가 발생할 수 있다. 두 번째는 세종계획의 의미태깅된 150만 어절중에 135만 어절을 학습 말뭉치로 이용하고 나머지 15만 어절을 실험 대상으로 이용하였다. 두 번째 실험에서 전체적인 동형어의어 분별 과정을 통해 분석된 정확률은 92.23%이다.

어휘공기집합의 데이터 부족 문제를 해결하는데 시소러스가 어느 정도의 도움을 주는지 Table 9를 통해 알 수 있다.

Table 9에 나와 있는 구문패턴의 재현률은 구문패턴이 얼마나 많은 동형어의어를 추출해서 분석에 이용할 수 있는가를 나타낸다. 그리고 LCS방법의 정확률은 구문패턴으로 뽑힌 정보중에서 LCS방법으로 분석된 것 중에 제대로 분석한 것들의 정확률이다. 시소러스 확장 방법의 재현률은 LCS방법으로 분석하지 못하는 구문패턴들 중에 시소러스로 확장 가능한 패턴의 개수를 말한다. 그리고 시소러스 확장 방법의 정확률은 어휘 공기 집합을 시소러스로 확장하여 분석된 결과가 어느 정도의 정확률을 갖는 지를 말한다.

Table 9의 분석결과에는 나와 있지 않지만 LCS방법으로 분석하지 못하는 경우가 실험1의 경우 약 69%이고 실험2의 경우 약 51%이다. 실험1의 경우에는 69%중에서 8%만이 용언류이고 나머지 92%가 명사이다. 그리고 실험

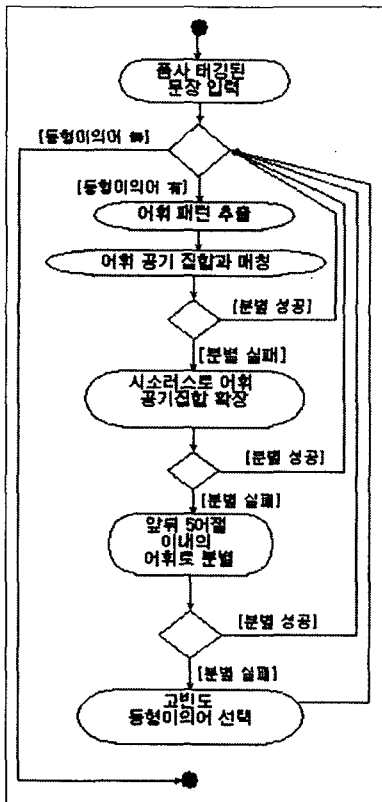


Fig. 4. 동형어의어 분별 과정.

Table 8. 오류의 가능성이 있는 어휘공기집합

fmpm	fpos	bmpm	bpos	sense	frq
같이	MAG	쓰	VV	1	1
같이	MAG	쓰	VV	3	1

Table 9. 실험1과 실험2의 재현률과 정확률

	실험1	실험2
구문패턴의 재현률	$100 \times \frac{237,135}{249,249} = 95.14\%$	$100 \times \frac{20,100}{21,999} = 91.37\%$
LCS방법의 정확률	$100 \times \frac{73,118}{74,886} = 97.64\%$	$100 \times \frac{9,768}{9,936} = 98.31\%$
시소러스 확장방법의 재현률	$100 \times \frac{10,372}{12,949} = 80.1\%$	$100 \times \frac{1,543}{1,604} = 96.2\%$
시소러스 확장방법의 정확률	$100 \times \frac{8,643}{10,372} = 83.33\%$	$100 \times \frac{1,404}{1,543} = 91.0\%$

2의 경우에는 51%중에서 15%만이 용언류이고 85%가 명사이므로 시소러스를 이용하여 동형이의어 명사를 분별할 수 있도록 시소러스의 각 노드에 의미 태깅을 해야 어휘 정보의 부족 문제를 해결할 수 있을 것이다.

결론 및 향후 연구

첫 번째 실험을 할 때 서로 다른 의미 태깅된 말뭉치를 사용하였기 때문에 의미 번호가 많이 달랐다. 그래서 469개의 동형이의어에 대해 수작업으로 의미 매핑 테이블을 작성하여야 정확률을 얻을 수 있었다. 품사 태깅은 말뭉치가 서로 다르더라도 비교적 쉽게 품사 변환을 할 수 있지만 동형이의어 태깅은 의미의 기준인 사전이 다를 경우에는 각 동형이의어마다 의미번호의 매핑을 해 주어야 하므로 서로 다른 말뭉치를 대상으로 전체 동형이의어를 테스트하기 위해서는 의미번호의 매핑작업을 자동으로 해주는 연구가 진행되어야 할 것이다. 또한 의미의 기술도 번호가 아닌 사람이 알아 보기 쉬운 기술 방법을 모색해야 할 것이다. 예를 들어, 명사인 경우에는 시소러스나 개념망에서 상위어를 이용하여 의미를 기술하면 알아 보기 쉬운 것이고, 동사인 경우에는 뜻풀이의 핵심 동사를 이용하거나 목적어 또는 주어에 대한 선택제약 정보로 동형이의어의 의미를 기술할 수 있을 것이다.

본 논문에서 제시한 LCS와 시소러스를 활용한 동형이의어 분별 방법이 정확률에서는 기존 연구보다 1%정도 떨어지지만 학습 코퍼스에 대한 오류가 수정되면 보다 좋은 결과가 예상된다. 또한 시소러스에 명사에 대한 의미 태깅이 되어 있다면 동형이의어 명사를 분별하는데 시소러스를 사용할 수 있어 더 높은 정확률이 예상된다.

어휘 공기 집합이 동형이의어뿐만 아니라 모든 어휘에 대해서 구축되어 있으므로 형태소 분석이나 구문 분석시에 어느 정도 도움을 주는 지를 실험할 계획이다.

REFERENCES

- 1) 김영관(2000) : “시소러스와 문맥정보를 이용한 색인어의 중의성 제거”, 석사 학위 논문
- 2) 김준수 외(2001) : “의미범주 및 거리 가중치를 고려한 통계 기반 동형이의어 분별 시스템”, 제 13 회 한글 및 한국어 정보처리 학술대회, pp487-493
- 3) 김준수 외(2002) : “한국어 동형이의어 태깅 시스템 구현”, 제 14 회 한글 및 한국어 정보처리 학술대회, pp24-30
- 4) 신사임(2001) : “공기정보 벡터를 이용한 한국어 명사의 의미 구분”, 제 13 회 한글 및 한국어 정보처리 학술대회, pp472-478
- 5) 양재형(1995) : “한국어 분석 모호성 해소를 위한 명사의 공기 유사성”, 박사 학위 논문
- 6) 옥철영 외(1998) : “우리말 개념망 명사 데이터 구축”, 한국전자동신연구원 최종 연구 보고서
- 7) 옥철영 외(2002) : “한국어정보처리에서 동형이의어 중의성 해결시스템”, 우수신기술 지정·지원사원 결과 보고서
- 8) 이승우(1999) : “국소 문맥과 공기 정보를 이용한 비교사 방식의 명사 의미 중의성 해소”, 석사 학위 논문
- 9) 이왕우(2002) : “구문패턴과 순환 뜻풀이망을 이용한 동형이의어 분별”, 제 29 회 한국정보과학회 춘계학술발표회
- 10) 이현아(1998) : “영-한 변환 시스템에서 한국어 구문 공기 정보를 이용한 역어 선택”, 석사 학위 논문
- 11) 임권록(1995) : “한국어 형태소 분석에서의 오분석 제거와 중의성 해결”, 박사 학위 논문
- 12) 조평옥(1996) : “한국어 명사의 의미 계층 구조 구축”, 울산대학교 석사학위 논문
- 13) 조형준(2000) : “한국어 병렬구문과 결합범주문법에서의 구문 분석”, 석사 학위 논문
- 14) 한상길(1999) : “시소러스 용어관계의 확장에 관한 연구”, 석사 학위 논문
- 15) Agirre E(1996) : “Word Sense Disambiguation Using Conceptual Density”, *Coling-96*
- 16) Xiaobin Li(1995) : “A WordNet-based Algorithm for Word Sense Disambiguation”
- 17) Yarowsky D(1992) : “Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora.” *Coling-92. Nantes, pp454-460*