http://dx.doi.org/10.3745/KIPSTB.2011.18B.4.241

종속격 정보를 적용한 동사 의미 중의성 해소

박 요 셉[†] · 신 준 철[††] · 옥 철 영[†††] · 박 혁 로[††††]

요 약

동형이의어는 여러 가지 의미를 가진 단어를 의미한다. 문장의 의미를 이해하기 위해서는 필수적으로 문장에 포함된 동형이의어의 의미를 결정해야 한다. 기존의 단어 의미 중의성 연구들은 공기 빈도를 기반으로 해결하였다. 하지만, 동사의 경우에는 정확도 향상을 위해서 격 정보가 중요하다. 왜냐하면, 동사 동형이의어의 의미는 행위의 주체나 객체에 따라 결정되어서 종속격(목적격, 부사격, 보격) 정보가 필요하며, 동사 동형이의어 의미마다 서로 다른 격 정보가 필요하기 때문이다. 본 논문에서는 한국어 격 정보를 적용한 동사 의미 중의성 해소를 제안한다. 격 정보는 표준국어대사전에 명시된 조사 정보를 이용하였다. 실험은 고빈도 동형이의어 12개를 대상으로 하였으며, 실험결과 정확도가 기존의 97.3%에서 98.7%로 1.34% 향상되었다. 이는 원래의 오류율을 2.7%에서 1.3%으로 절반정도 줄였다.

키워드 : 의미 분석, 단어 의미 중의성, 종속격 정보

Verb Sense Disambiguation using Subordinating Case Information

Yo-Sep Park[†] · Joon-Choul Shin[††] · Cheol-Young Ock[†††] · Hyuk-Ro Park[††††]

ABSTRACT

Homographs can have multiple senses. In order to understand the meaning of a sentence, it is necessary to identify which sense is used for each word in the sentence. Previous researches on this problem heavily relied on the word co-occurrence information. However, we noticed that in case of verbs, information about subordinating cases of verbs can be utilized to further improve the performance of word sense disambiguation. Different senses require different sets of subordinating cases. In this paper, we propose the verb sense disambiguation using subordinating case information. The case information acquire postposition features in Standard Korean Dictionary. Our experiment on 12 high-frequency verb homographs shows that adding case information can improve the performance of word sense disambiguation by 1.34%, from 97.3% to 98.7%. The amount of improvement may seem marginal, we think it is meaningful because the error ratio reduced to less than a half, from 2.7% to 1.3%.

Keywords : Semantic Analysis, Word Sense Disambiguation, Subordinating Case Information

1. 서 론

자연언어처리는 형태소 분석, 통사 분석, 의미 분석 및 화용 분석으로 구성된다. 기존에 형태소 분석 및 통사 분석은 많은 연구들을 통해 발전하였다. 하지만, 의미 분석 및 화용 분석의 연구는 이에 비하여 부족하다.

의미 분석(Semantic Analysis)은 통사 분석 결과에 해석을 가하여 문장이 가진 의미를 분석하는 작업이다. 문장의 의미는 구성하는 각 형태소의 의미가 합성되는 단순한 유형부터 은유처럼 고도의 분석을 요구하는 유형까지 다양하다. 이를 처리하기 위해서는 각 어휘 혹은 형태소에 의미 표지를 부여하고, 하위 범주와 같은 정보를 이용하여 부분의 의미를 통합하여 전체 의미를 구성하는 방법을 이용하는 것이 일반적으로 사용되는 방법이다[1]. 의미 분석 시에는 단어 의미 중의성(Word Sense Disambiguation) 문제가 발생한다.

단어 의미 중의성은 의미적 중의성을 의미하며 문맥을 통해 단어의 의미를 결정하는 작업이다[2]. 즉, 동형이의어(Homograph)의 의미를 결정하는 것이다. 이는 기계번역, 정보검색, 음성인식 및 합성, 철자교정 등 응용 분야의 기반 기술에 유용하다.

통계적 기반의 단어 의미 중의성 해결 방법으로 크게 지도학습(Supervised Method), 비지도학습(Unsupervised Method) 그리고 사전 기반 방법(Dictionary-based Method)이 있다. 첫째, 지도학습은 의미 부착된 학습 집합에 기반한 방법으로 결정트리(Decision Trees)[3,4], 나이브 베이지안

† 준 회 원 : 전남대학교 전자컴퓨터공학과 석사과정
†† 정 회 원 : 울산대학교 컴퓨터정보통신공학과 박사
††† 종신회원 : 울산대학교 컴퓨터정보통신공학과 교수
†††† 종신회원 : 전남대학교 전자컴퓨터공학부 교수
논문접수 : 2011년 6월 14일
수 정 일 : 1차 2011년 6월 28일, 2차 2011년 6월 30일
심사완료 : 2011년 7월 4일

(Naive Bayes)[5] 등이 있다. 둘째, 비지도학습은 학습 시 의미 부착 집합을 사용하지 않는 방법으로 공기 관계 그래프(Co-occurrence Graphs)[6] 등이 있다. 셋째, 사전 기반 방법은 기계 가독형 사전(Machine-Readable Dictionaries)[7,8]이나 시소러스(Thesaurus)[9,10,11] 등의 지식 베이스를 이용하는 방법이다[12]. 최근에는 정확도 향상을 위해서 혼합하여 연구를 진행한다.

표준국어대사전[13]의 5만 개의 기초어휘를 살펴보면 명사는 44,419개(66%), 동사는 15,564(23%)로 구성되며 동형이의어도 명사가 차지하는 비율이 높다. 하지만 동사 의미 중의성 해결이 명사에 비해 어렵다. 왜냐하면, 명사는 공기 빈도에 기반한 의미 정보만으로도 대다수 해결이 가능하지만 동사는 타동사는 목적어, 자동사는 주어에 사용된 명사가 중요한 역할을 하기 때문이다[26]. 또한, 보격 정보도 동사 중의성 해결에 도움이 된다.

본 논문에서는 동사 의미 중의성을 해결하기 위해서 종속격 정보(목적격, 부사격, 보격)를 이용한 중의성 해소를 제안한다. 제안된 방법은 HMM[24]을 이용하였으며, 표준국어대사전에 명시된 격 정보를 이용하였다.

본 논문의 구성은 다음과 같다. 2장은 기존의 단어 의미 중의성의 연구에 대해서 살펴본다. 3장은 격 정보를 이용하여 단어 의미 중의성을 해결하는 방안에 대해서 기술한다. 4장은 기존 연구와 비교 실험 및 결과이며, 5장에서는 결론을 내린다.

2. 관련 연구

단어 의미 중의성은 소수의 샘플 어휘(Lexical sample or Targeted WSD)를 대상으로 하거나 모든 단어(All-words WSD)를 대상으로 한다. 후자의 경우가 현실적으로 가치가 있지만 말뭉치 구축에 많은 비용과 시간이 소요된다. 동형이의어는 일반적으로 문장에서 문맥 정보를 이용하여 의미를 해결한다. 즉, 대상 단어가 포함된 문장의 의미 정보들을 토대로 결정한다.

[14]에서는 사전 뜻풀이 말에서 추출한 통계적 의미 정보에 기반한 동형이의어 중의성 해결 시스템을 제안하였다. 의미 정보는 동형이의어의 사전 뜻풀이에서 체언과 용언을 추출하여 구성하였다. 정확한 의미 정보 추출을 위하여 동형이의어와 표제어가 의미적으로 상-하위어 관계 경우와 동형이의어가 뜻풀이말의 중간에 사용된 경우로 나누어서 구성하였다. 9개의 명사 동형이의어를 대상으로 실험한 결과 학습된 코퍼스 경우는 평균 96.11%, 미학습된 코퍼스 경우는 평균 80.73%의 정확률이었다. 비교적 작은 말뭉치 사전만을 이용해서 단순한 의미 계층 구조를 유추하고 이용함으로써 대용량의 의미 계층 구조가 없는 경우에 적용 가능한 장점이 있다. 하지만, 학습 말뭉치가 부족해서 의미 정보가 부족하며, 복문 구조의 경우에 적용하기 어렵다.

[15]에서는 사전 뜻풀이 말뭉치에서 구축한 의미 정보와 이를 적용한 베이지안 분류 모델을 이용하여 동형이의어를 해결하였다. 또한, 새로운 동형이의어 사전 확률 가중치 및 인접 어절에 대한 거리 가중치를 제안하였다. 중의성이 높은 명사 30개와 동사 16개를 실험한 결과 평균 81.55%의 정확률을 얻었다. 그러나 정확도를 개선하기 위해서는 의미 정보 정제가 필요하다.

[16]에서는 기존의 의미 정보를 정제하기 위하여 확률 정보, 거리 정보 및 격 정보, 문장 분할 정보 등을 추가한 단어 의미 중의성 모델을 제안하였다. 자료 부족 문제를 해결하기 위해서 울산대학교 어휘 지능망(U-WIN)을 이용하였다. 세종 의미 말뭉치를 대상으로 각 문장에서 중의성 단어의 의미를 결정하기 위한 중요한 의미 정보가 주변 명사(일반명사, 고유명사), 용언(동사, 형용사), 일부의 부사임을 확인하고 이를 바탕으로 동형이의어를 해결하였다. 실험 결과 명사 15개의 경우 평균 73.43%, 어휘 망을 적용한 경우 78.72%이고 동사 10개의 경우 평균 69.70%, 어휘 망을 적용한 경우 76.34% 이었다. 격 정보를 명사 동형이의어에만 적용하였으며, 단순히 격 조사의 유무만을 이용해서 정확하게 반영하지 못하였다.

[17]에서는 상호 정보량과 기분석된 복합명사 의미 사전에 기반한 동음이의어 중의성 해소 방안을 제시하였다. 어휘들 간의 연관 계수인 상호 정보량을 이용하여 자료 부족 현상을 해결하고자 했다. 상호 정보량의 단점을 보완하고 사전에 구조적으로 내포되어 있는 의미 결정 단서를 반영하기 위해서 상호 정보량의 값을 가지는 어휘 쌍의 비율과 사전 뜻풀이의 길이 및 의미 부착된 말뭉치로부터 추출한 의미 사용 비율을 가중치로 활용하였다. 또한, 기 구축된 복합명사 의미사전을 복합명사 어휘 중의성 해소에 이용하였다. 실험결과 LESK 방법론[18]에 비해 8.5% 정확률 향상을 보였으며, 다양한 가중치 중에서 의미 사용 비율 가중치가 동음이의어 중의성 해소에 가장 크게 기여함을 확인하였다. 하지만 많은 계산량에 따른 속도 문제가 발생하며 다양한 연관 관계를 이용할 필요성이 제기되었다. 특히, 한국어 특징을 고려한 연관 관계 가중치 부여가 필요하다.

[19]에서는 품사 태깅과 동형이의어 태깅이 문맥 정보에 의존함에 착안하여 HMM을 적용하여 해결하는 시스템을 제안하였다. 세종 말뭉치에서 유니그램(unigram)과 바이그램(bigram)를 추출하여 생성확률사전과 전이확률사전을 구축하였다. 실험결과 명사 동형이의어 9개를 대상으로 95.28% 성능을 보였다.

Semeval(Semantic Evaluations)는 단어 의미 중의성 경진대회이다. ACL-SIGLEX에서 주관하며 1998년부터 시작되어 3년마다 열린다. 처음에는 영어, 프랑스어, 이탈리아어를 대상으로 평가하였으며 한국어는 2001년도에 참여하였지만 이후에는 제외되었다. 2007년도에 영어를 대상으로 소수의 예제 샘플로 평가한 결과 NUS-ML 시스템이 최고의 성능인 88.7% 정확도를 달성하였다[20].

[21]에서는 지식 획득 병목을 해결하기 위해서 위키피디아를 이용하였다. WSD(Coarse-grained WSD)문제를 해결하기 위해서 두 개의 지식 기반 알고리즘을 실험한 결과 그

래프 기반의 접근 방식인 Degree Centrality 알고리즘[22]이 96.2%의 정확도를 달성하였다.

단어 의미 중의성의 기존 연구들은 주로 동형이의어 명사에 관심을 두었다. 그래서 동형이의어 동사를 명사와 동일시하여 같은 방법으로 처리해서 대체적으로 명사에 비해서 정확률이 낮다. 즉, 동사의 특성을 반영해서 처리하지 못하였다.

3. 종속격 정보를 이용한 동사 의미 중의성 해소

한국어는 조사와 어미가 발달하였다. 즉, 한국어를 연구하기 위해서는 필수적으로 이들을 중요하게 고려해야 되며 연구 방향 전체에 영향을 미치게 된다[23]. 기존 연구에서는 동형이의어 동사 처리 시 격 정보를 중요하게 고려하지 않았다. 그래서 본 논문에서는 조사에 의해 나타나는 격 정보를 이용해서 동사 동형이의어를 해결하고자 한다. 특히, 격 정보 중에서 목적격, 부사격, 보격을 이용한 동사 중의성 해소를 제안한다. 이는 표준국어대사전에 명시된 조사를 이용해서 격 정보를 적용하였으며 이를 HMM을 통해 반영하였다.

3.1 한국어의 동사와 격 정보

3.1.1 한국어의 동사

동사는 사람이나 사물의 움직임을 나타내는 품사이다. 한국어에서는 주로 문장의 끝에 위치하며 형용사와 함께 용언에 속한다. 동사는 목적어의 유무에 따라 타동사, 자동사로 구분되며 보어의 유무에 따라 불완전동사, 완전동사로 나눈다. 동사는 명사 다음으로 큰 비중을 차지하며, 동형이의어도 명사 다음으로 많다. 하지만 동형이의어 동사는 기존의 공기 빈도로 처리하는 것에 비하여 격 정보를 통해서 많은 경우에 정확하게 해결한다.

3.1.2 논항 구조

논항 구조(Argument Structure)는 술어(동사, 형용사)가 문장을 구성하면서 요구하는 논항들의 수와 그 의미역(Semantic Roles) 집합을 의미한다. 의미역에는 행동주(Agent), 피동주(Patient), 대상(Theme), 처소(Location), 도구(Instrument) 등이 있으며, 한국어의 논항 구조는 다음과 같다[15].

가다 : 논항구조 =[행동주$_A$+착점$_G$]
[진이가]$_A$[학교에]$_G$ 갔다.
차다 : 논항구조 = [행동주$_A$+피동주$_P$]
[연이가]$_A$[축구공을]$_P$ 힘껏 찼다.
지루하다 : 논항구조 = [경험주$_E$+자극$_S$]
[근이는]$_E$[그 영화가]$_S$ 무척 지루했다.

예문에서 술어는 공통적으로 두 개의 논항을 요구하지만 의미역 집합이 서로 다르다. 즉, '가다'는 [행동주+착점] 의미역 집합, '차다'는 [행동주+피동주] 의미역 집합, '지루하다'는 [경험주+자극] 의미역 집합이 필요하다. 이러한 논항들의 의미역은 격에 의해서 실현된다. '가다'는 두 논항이 조사 '가(주격)'와 '에(목적격)'로 실현되고, '차다'는 조사 '가(주격)'와 '에(목적격)'로 실현되며, '차다'는 두 논항이 조사 '가(주격)'와 '를(목적격)'로 실현된다. 결론적으로 논항의 의미역은 조사에 의해서 실현되는 격으로 표현된다.

3.1.3 종속격 정보(목적격, 부사격, 보격)

격은 문장 속에서 체언이나 체언 구실을 하는 말이 서술어에 대하여 가지는 자격을 말한다. 이를 실현하는 형태와 의미 기능으로 요약할 수 있다. 격은 명사나 명사 상당 어구에 조사가 결합하여 표시될 수 있으며, 위치로도 표현될 수 있고, 무표로도 나타낼 수 있다. 한국어는 대부분 격이 조사를 통해서 나타난다.

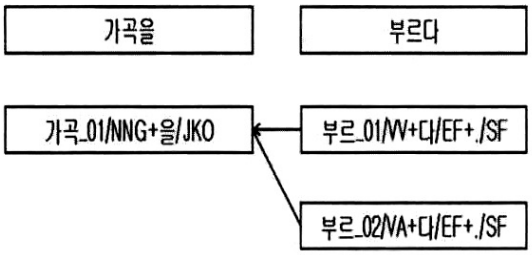
격 정보 중 목적격, 부사격, 보격이 동사 동형이의어 해결에 중요하다. 이는 동사의 의미가 격 정보를 통해서 나타나는 행위의 주체나 객체 등의 정보에 따라 결정되기 때문이다. 격 정보의 목적격은 문장에서 주로 '을/를(JKO)' 조사를 통해 나타나며, 부사격은 '에, 에서, 에게, 한테, (으)로, (으)로서, (으)로써(JKB)' 조사를 통해 나타난다. 한편, 보격은 문장에서 주로 '이/가(JKC)' 조사를 통해 나타나는 데 주격 조사와 중복이 된다. 이는 문장 성분이 주어인지 보어인지에 따라 구분하는 데 대부분 문장에 주격이 포함되기 때문에 동형이의어를 결정하는 데 중요한 정보가 아니어서 제외한다. 주격과 보격 조사는 동형이의어 태깅에서 품사 태깅 작업을 통해서 구분한다.

다음은 격 정보 중 목적격을 이용하여 동사 동형이의어 '부르다'를 해결하는 과정이다. (그림 1)은 HMM을 이용한 동형이의어 처리 과정이며, (그림 2)는 표준국어대사전 '부르다'의 뜻풀이이다. 부르다 뜻풀이 1번에는 목적격 조사 '을'이 명시되었고, 뜻풀이 2번에는 보격 조사 '이'가 명시되었다.

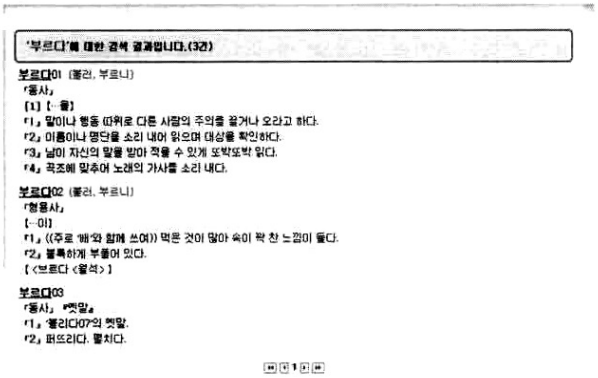
가곡을 부르다.
그녀는 자기를 부르는 소리를 듣고도 모른 척 하였다.
전화번호를 불러 줄 테니 꼭 전화해라.

공기 빈도를 기반으로 '부르다'의 의미를 결정하는 방법은 자주 등장하는 단어가 문장에 포함되었을 경우에 효과적이다. 하지만 공기 빈도 횟수만을 적용해서 처리하면 상대적으로 빈도 횟수는 적지만 의미 결정에 중요한 역할을 하는 단어를 정확하게 이용하지 못한다.

표준국어대사전에 명시된 조사 정보를 이용하면 이를 개선할 수 있다. 예문에서 '부르다' 앞에 조사가 목적격 조사 '을/를'이 있다. 조사를 통해서 격 정보를 적용하면 표준국어대사전 뜻풀이 1번 의미로 결정된다.



(그림 1) HMM을 이용한 처리 과정



(그림 2) 표준국어대사전 '부르다' 뜻풀이

결론적으로, 격 정보 중 목적격은 타동성을 지니는 동사의 의미 결정에 유용하며, 부사격은 따르다, 삼다, 주다, 여기다, 이르다 등과 같이 부사어를 꼭 필요로 하는 타동사에 유용하다. 보격의 경우 표준국어대사전 뜻풀이 조사가 목적격, 부사격 조사가 명시되어 있지 않을 경우에 유용하다.

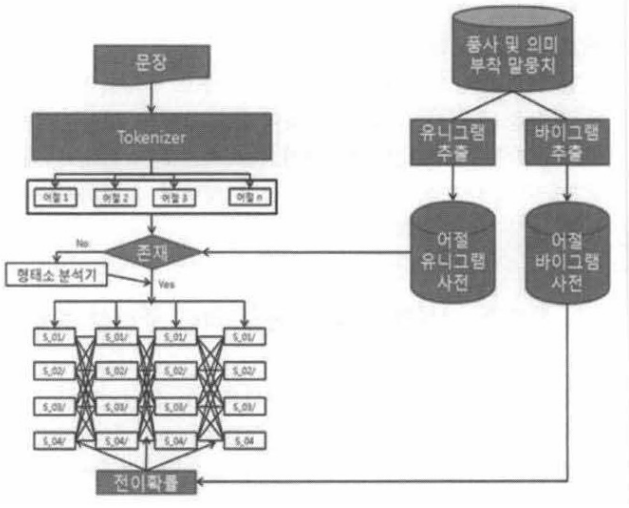
3.2 격 정보를 적용한 동사 의미 중의성 해소

격 정보를 이용한 동사 중의성 해결 방안은 표준국어대사전에 명시된 조사를 이용하여 처리한다. 이를 적용하기 위해서 HMM을 기반으로 한다. HMM은 주로 품사 태깅에 사용되는 모델이다. 이때 상태는 품사에 해당되는데 본 논문에서는 상태가 어절 상태에 해당한다. 즉, 한 어절의 모든 품사 및 모든 동형이의어에 해당한다.

3.2.1 HMM

HMM은 이중 통계적 모델로서 생성확률과 전이확률을 이용하여 최적의 상태 열을 찾는다. 이는 문맥 정보를 반영한 품사 태깅과 동형이의어 태깅에 적합한 모델이다.

생성확률은 어절이 형태소 분석의 발생확률로 적용하였고, 상태전이확률은 어절 간의 바이그램 통계를 이용한 어절의 전이확률을 사용하였다. 어절의 발생확률을 위한 유니그램 데이터는 어절의 품사 정보와 의미 정보들로 구성된다. 이는 기분석사전의 용도로 활용된다. 어절 유니그램 사전에는 품사패턴정보, 어근의 길이, 분석형태소개수 정보가 있으며, 어절 바이그램 사전에는 어절/품사-어절/품사, 어절-어근/품사, 끝형태소/품사-어근/품사 정보가 있다. 즉, 어절 바이그램 사전에는 한 어절내의 형태소 각각의 품사가 저장



(그림 3) 시스템 구조

된다. 예를 들면, 바이그램 어절 "수 있느냐는"은 "수_02/NNB 있/VV+느냐는/ETM" 형태로 저장된다.

(그림 3)은 전체 시스템 구조이다.

최적 상태열은 Viterbi 알고리즘을 통해 얻는다. 즉, 관측열 $X=(X_1, X_2, X_3, ..., X_r)$가 일 때, 단일 최적 상태열 $q=\{q_1, q_2, q_3, ..., q_r\}$을 찾는다.

$$\delta_{t+1}(j) = \max_{1 \le i \le N} \delta_t(i) a_{ij} b_j(X_{t+1}) \quad (1)$$

$$\psi_{t+1}(j) = \arg\max_{1 \le i \le N} \delta_t(i) a_{ij} \quad (2)$$

식 (1)는 시간 t+1에서 j에 대하여 $\delta t+1(i)$를 최대화하는 상태의 트랙을 의미하며, 식 (2)는 배열 $\psi_{t+1}(j)$에 최적 상태 열을 탐색하기 위해서 저장한다. a_{ij}는 시간 t의 어떤 상태 i에서 시간 t+1의 어떤 상태 j로의 전이확률이다. $b_j(x_{t+1})$는 시간 t+1에서의 어떤 상태 j의 생성확률이다. 즉, 전이확률은 예를 들면 어절 "그는 그녀와"가 있을 때에 '그_01/NP+는/JX' 상태에서 '그녀/NP+와/JC' 또는 '그녀/NP+와/JKB' 상태로 이동할 확률이다. 다음은 상태 전이 확률에 관한 식이다.

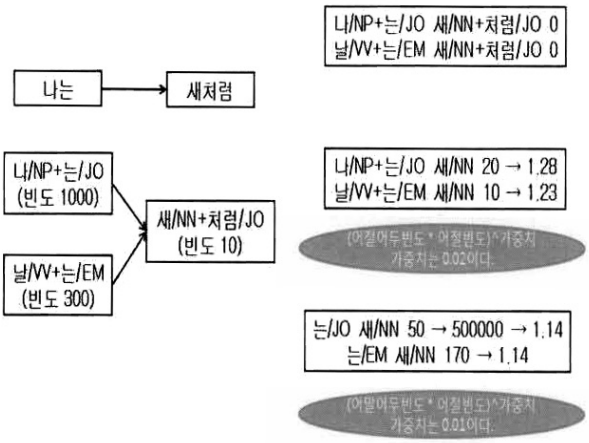
$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i), 1 \le i, j \le n \quad (3)$$
$$\text{이때 } \sum_{j=1}^{n} a_{ij} = 1 (n = \text{상태개수})$$

3.2.2 동형이의어 태깅

동형이의어 태깅은 문장 단위로 형태소 분석 과정을 통해서 어절당 후보들이 생성되어서 시작된다. 기본적으로 분석된 학습말뭉치에서 추출한 유니그램을 이용하여 분석한다. 이 과정에서 유니그램이 발견되지 않은 어절은 형태소 분석기를 이용하여 분석한 후, 유니그램의 품사 패턴 빈도, 어근의 길이 및 분석된 형태소 개수 등을 적용하여 최적의 형태소 분석 결과를 얻는다.

각 어절의 모든 후보가 준비되었다면 이후에 사용되는 의미 태깅은 크게 두 단계로 나누어진다. 첫 번째는 각 어절 사이에 후보들의 전이확률을 구하는 것이며, 두 번째는 전체 어절에 걸쳐 최적의 후보들을 하나씩 선택하는 것이다.

인접한 두 어절을 전체를 포함하는 정보가 없을 경우에는 전이확률을 어절의 어두(head)와 어말(function) 정보를 이용해서, 앞 어절 전체와 뒤 어절의 어두, 앞 어절의 어말과 뒤 어절 전체 등을 이용한다. 그래도 전이확률이 발생하지 않으면 인접한 두 어절의빈도수를 곱한다. 예를 들면, 어절 "나는 하늘을"이 앞 어절 전체와 뒷 어절 전체(나는 하늘을)의 바이그램 정보가 없을 경우에 앞 어절 전체와 뒤 어절의 어두의 정보(나는 하늘)를 찾는다. 만약에 발견되지 않으면 앞 어절의 어말과 뒷 어절의 어두의 정보(는 하늘)를 찾는다. 이러한 과정을 통해서 어두와 어말 정보를 반영한다[30].



(그림 4) 바이그램 어말과 어두 정보

격 정보는 동사 동형이의어의 빈도수에 가중치를 부여해서 적용한다. 표준국어대사전 뜻풀이의 조사 정보를 기반으로 다음과 같이 기준을 세울 수 있다.

기준 1> 목적격 정보가 자주 등장하므로 부사격이나 보격에 비해서 가중치가 낮다.

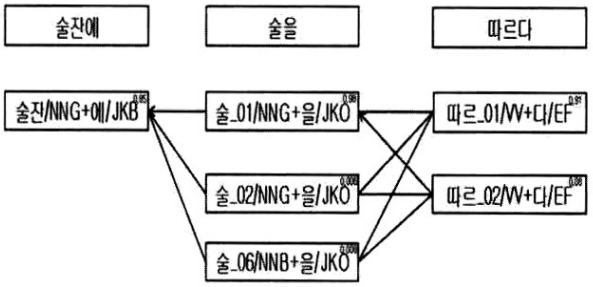
기준 2> 보격 정보는 자주 등장하지 않아서 목적격, 부사격 정보에 비해서 가중치가 높다.

동형이의어 동사 뜻풀이가 v_1, v_2, ..., v_n 일 때 그 빈도수가 f_1, f_2, f_3, ..., f_n 이다. f_i' 는 격 정보 가중치 값을 적용한 값이며, f_i는 동사 동형이의어 빈도수이다.

$$V-sense=\begin{cases} v_1 \ldots . f_1 \\ . \quad \ldots . \\ . \quad \ldots . \\ v_n \ldots . f_n \end{cases} \tag{4}$$

$$f_i'=f_i^{(W_{obj}+W_{adv}+W_{com})} \tag{5}$$

표준국어대사전에 명시된 격 정보를 바탕으로 목적격 가중치(W_{obj})는 0.1승(빈도수^0.1), 부사격 가중치(W_{adv})는 0.4승, 보격 가중치(W_{com})는 0.5승을 부여한다. 이는 각 가중치에 0.1부터 1까지 적용해서 얻어낸 결과 값이다. 기준을 통해서 알 수 있듯이 목적격 정보는 자주 발생해서 목적격, 부사격 정보가 발생하는 경우에 더 높은 가중치를 부여해서 목적격 정보만을 가지는 경우로 결정되는 것을 피하도록 고려하였다.



(그림 5) HMM 모델 처리 과정

(그림 5)은 HMM을 적용하여서 "술잔에 술을 따르다" 문장을 처리하는 과정이다. 각 어절의 생성확률은 학습말뭉치를 통해서 얻어졌으며 어절의 발견 시에 총 빈도수 중에서의 각각의 어절의 확률에 해당한다. 어절 "술잔에"는 하나의 후보만 생성된다. 어절 "술"은 가장 높은 값인 술_01/NNG(마시면_취하는_음료)+을/JKO 이 선택된다. 하지만 "따르다"의 경우에는 따르_01/VV+다/EF(빈도수 103), 따르_02/VV+다/EF(빈도수 9)이다. 따르_01의 경우에는 표준국어대사전에 목적격 정보만 명시되어 있는데 문장에 목적격 정보를 확인할 수 있으므로 빈도수에 0.1승을 하면 1.58이 되며, 따르_02의 경우에는 표준국어대사전에 목적격과 부사격 정보가 명시되어 있으므로 빈도수에 0.5승을 하면 3.73이 된다. 이러한 과정을 통해서 동사 동형이의어를 해결하며 최적의 후보를 선택하게 된다.

4. 실험 및 결과

본 실험 목적은 동사 의미 중의성 해소에서 격 정보가 유용하다는 점에 있다. 실험을 통해 기존 HMM과 비교를 통해서 이를 확인하고자 한다. 실험은 21세기 세종계획 형태의미 분석 말뭉치 1,100만 어절을 대상으로 한다. 학습 말뭉치에서 유니그램과 바이그램을 추출하고, 추출한 유니그램과 바이그램에는 어절과 어절 품사별 빈도 정보가 있다. 형태소 분석기는 어절 사전 기반의 UTagger[25]를 이용한다. 기분석된 학습말뭉치의 유니그램을 이용하여 형태소 분석하며, 품사 태그는 세종 말뭉치 태그 셋으로 표시한다. 형태소 분석기 성능은 대체적으로 95%이상이며, 동사 동형이의어 12개를 대상으로는 오류가 거의 발생하지 않았다. 오류 발생 시에는 본 논문에서는 의미적 중의성 해소에 초점을 맞추어서 실험하기 때문에 제외한다.

기존 모델과 성능 비교를 하기 위해서 동형이의어 태깅 보고서[28, 29]를 바탕으로 고빈도 동형이의어 144개 중에서 12개를 선정하였다. 선정 기준은 빈도수를 기준으로 상위 12개의 동사를 선택하였다. 실험은 문장 단위로 진행하며, 10-폴드 교차 타당성(10-fold cross validation)를 적용하였다. 일반적으로 HMM는 품사 태깅에 자주 사용하는데, 본 논문에서는 HMM을 품사 태깅과 동형이의어 태깅이 문맥 정보를 통해서 해결되는 공통점을 고려해서 동시에 사용한다. 그 결과 각 어절마다 분석되어 발생하는 품사와 동형이의어들이 후보에 해당한다. 정확률은 동사 동형이의어를 포함하는 총 어절에서 정답 어절 수의 비율로 나타냈다.

<표 1>은 동사 동형이의어 12개의 기존 HMM과 비교하는 개별적 실험 결과이며, <표 2>는 동형이의어 12개를 대상으로 10-폴드 교차 타당성을 통해서 기존 HMM과 비교하는 총체적 실험결과이다.

〈표 1〉 동사 동형이의어 12개

	의미수	출현빈도	HMM	격정보 HMM
대비하다	4	3,169	94.86%	97.15%
따르다	2	113,512	96.54%	98.48%
끼이다	2	1,718	66.57%	77.27%
고르다	3	5,552	82.54%	90.56%
되다	5	443,287	99.06%	99.47%
부르다	3	31,305	95.83%	97.68%
열다	2	21,083	98.54%	99.20%
울리다	3	9,991	94.26%	97.96%
절다	4	1,006	45.48%	74.86%
젖히다	2	1,265	80.14%	95.82%
조르다	3	1,382	63.75%	81.43%
파다	2	4,092	60.86%	83.93%
평균			97.38%	98.72%

실험결과 전반적으로 HMM 정확률은 97.38%, 격 정보 HMM 정확률은 98.72%이다. 즉, 격 정보를 통해서 기존보다 1.34%가 향상되었다. 이를 통해서 동사 중의성에서 격 정보가 유용하다는 점을 확인할 수 있다.

〈표 2〉 고빈도 동사 12개 동형이의어 실험

	실험1	실험2	실험3	실험4	실험5
총 어절수	1,112,281	1,109,193	1,114,223	1,113,459	1,109,953
동사 동형이의어 어절 수	6,440	6,353	6,272	6,773	6,627
HMM 동사 정답 어절 수	6,278	6,190	6,105	6,588	6,442
정확률	97.48%	97.43%	97.34%	97.27%	97.21%
격 정보 HMM 정답 어절 수	6,360	6,277	6,191	6,684	6,531
정확률	98.76%	98.80%	98.71%	98.69%	98.55%

	실험6	실험7	실험8	실험9	실험10
총 어절수	1,110,474	1,112,793	1,111,706	1,109,027	1,110,400
동사 동형이의어 어절 수	6,106	6,463	6,241	6,511	6,451
HMM 동사 정답 어절 수	5,988	6,315	6,078	6,318	6,280
정확률	97.58%	97.71%	97.39%	97.04%	97.35%
격 정보 HMM 정답 어절 수	6,040	6,387	6,152	6,423	6,369
정확률	98.92%	98.82%	98.57%	98.65%	98.73%

실험 결과 중에서 오류가 발생되는 경우는 다음과 같다. 첫 번째는 격 정보의 조사가 문장에 없는 경우이다.

> 제가 없어진 줄 알고 경찰까지 불렀던 어머니는
> 제가 관찰한 장면을 설명해 주자 차분히 이야기를
> ....
> ..... 경찰__04/NNG+까지/JX
> 부르__02/VA+었/EP+던/ETM .....

문장에서 '부르다'의 의미를 결정하기 위해서 필요한 격 정보 조사가 없다. 여기서는 조사 '까지'가 목격적 조사에 해당되지만, 표준국어대사전에 명시된 조사는 '을/를' 이기 때문이다. 이를 해결하기 위해서 목적격 조사 정보의 확장이 필요하다. 이는 다른 사전을 통해서 명시된 조사를 이용해서 개선해야 한다.

두 번째는 능동, 수동, 피동, 사동 유형의 문장이다.

> 예술가의 역할은 관객의 배를 부르게 하는 것이
> 아니라 배부른 관객을 취하게 하는 것이다.
> ..... 배__01/NNG+를/JKO
> 부르__01/VV+게/EC 하__01/VX+는/ETM .....

문장에서 '부르다' 목적격 조사 '를'이 있다. 여기서는 동사가 사동사이어서 격 정보 조사만을 이용하면 대상 단어가 정확하게 결정되지 못하다. 즉, 조사뿐만 아니라 문장의 유형까지 고려해서 동사 동형이의어를 처리해야 한다. 이는 구문 분석을 통해서 개선해야한다.

5. 결 론

본 논문에서는 조사를 통해서 한국어 격 정보를 적용한 동사 의미 중의성 해소를 제시하였다. 즉, 표준국어대사전에 명시된 목적격, 부사격, 보격 조사를 이용하여 동사 중의성

을 해결하였다. 격 정보를 적용하기 위해서 생성확률은 형태소 분석시 발생하는 확률로, 전이확률은 바이그램으로 처리해서 HMM에 적용하였다. 실험은 21세기 세종계획 형태의미 부착 말뭉치를 대상으로 하였다. 동사 동형이의어 12개를 대상으로 기존 연구와 비교 실험을 하였다. 실험 결과 동사 중의성 해소에서 격 정보를 유용하다는 점을 확인하였다. 하지만 동사 중에서 표준국어대사전에 격 정보가 명시되어 있지 않거나 문장의 유형을 고려하지 않아서 오류가 발생하였다. 이를 개선하기 위해서 다른 사전 자원을 이용해서 적용되지 않은 격 정보를 확장하거나 구문 분석을 적용해야한다.

단어 의미 중의성은 자연언어처리에서 해결하기 어려운 문제이다. 하지만, 자연언어처리 응용 기술의 기반이 되기 때문에 중요하다. 한국어의 경우에는 한자어에서 유래된 단어가 많아서 이로 인해 발생되는 의미 중의성이 있다. 특히, 한자어로 된 한 음절 관형사의 경우에는 HMM를 적용하면 많은 후보들이 생성되어 처리하기가 어렵다. 추후에 이를 해결하기 위해서 연구가 필요하다.

참 고 문 헌

[1] 김영택, '자연언어처리', 생능출판사, 2001.

[2] Roberto Navigli,"Word Sense Disambiguation : A Survey", ACM Computing Survey, Vol.41, No.2, 2009.

[3] J.R. Quinlan, "Induction of decision trees", Machine Learning Vol.1, No.1, pp.81-106, 1986.

[4] J.R. Quinlan, 'Programs for Machine Learning", Morgan Kaufmann', 1993.

[5] Gerand Escudero, "Naive Bayes and Exemplar-Based approaches to Word Sense Disambiguation Revisited", In Proceedings of the 14th European Conference on Artificial Intelligence, pp.421-425, 2000.

[6] Tae-Gil Noh, Seong-Bae Park, Sang-Jo Lee, "Unsupervised word sense disambiguation in biomedical texts with co-occurrence network and graph kernel", ACM fourth international workshop on data and text mining in biomedical informatics(DTMBIO '10), pp.61-64, 2010.

[7] Yorick A. Wilks, Brian M. Slator, Louise M. Gutherie, 'Electric Words: Dictionaries, Computers and Meanings', MIT Press, 1996.

[8] Andrew Harley, Dominic Glennon, "Sense Tagging in Action : Combining Different Tests with Additive Weightings", Proceedings of the SIGLEX Workshop on tagging text with lexical semantics, pp.74-78, 1997.

[9] Roget P.M, 'Roget's International Thesaurus', Bebook, 1991.

[10] John R.L. Bernard, "Macquarie Thesaurus", The Macquarie Library, 1986.

[11] David Yarowsky, "Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", Proceedings of the 14th COLING, pp.454-460, 1992.

[12] 옥철영, 김준수, 옥은주, 이왕우, 이재홍, 최호섭, '한국어정보처리에서 동형이의어 중의성 해결 시스템 기술', 정보통신부, 2002.

[13] http://www.stdweb2.korean.go.kr/(국립국어원 표준국어대사전)

[14] 허정, 옥철영, "사전의 뜻풀이 말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템", 정보과학회논문지(소프트웨어 및 응용), 제28권, 제9호, pp.688-698, 2001.

[15] 김준수, 최호섭, 옥철영, "가중치를 이용한 통계 기반 한국어 동형이의어 분별 모델", 정보과학회논문지(소프트웨어 및 응용), 제30권, 제11호, pp.1112-1123, 2001.

[16] 김준수, 옥철영, "정제된 의미정보와 시소러스를 이용한 동형이의어 분별 시스템", 정보처리학회논문지(B), 제12권, 제7호, pp.829-840, 2005.

[17] 허정, 서희철, 장명길, "상호 정보량과 복합명사 의미사전에 기반한 동음이의어 중의성 해소", 정보과학회논문지(소프트웨어 및 응용), 제33권, 제12호, pp.1073-1089, 2006.

[18] Michael Lesk, "Automatic Sense Disambiguation Using Machine Readable Dictionaries : How to Tell a Pine Cone from an Ice Cream Cone", SIGDOC-86 Proceedings of the 5th annual international conference on Systems Documentation, pp.24-26, 1986.

[19] 김동명, 배영준, 옥철영, 최호섭, 김창환, "HMM을 이용한 한국어 품사 및 동형이의어 태깅 시스템", 제20회 한글 및 한국어 정보처리 학술대회 발표논문, pp.12-16, 2008.

[20] Sammer S. Pradhan, Edward Loper, Dmitriy Dligach, Martha Palmer, "SemEval-2007 Tasks 17 : Enlglish Lexical Sample, SRL and All Words", Proceedings of the 4th International Workshop on Semantic Evaluations, pp.87-92, 2007.

[21] Simone Paolo Ponzetto, Roberto Navigli, "Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems", Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp.1522-1531, 2010.

[22] Roberto Navigli, Mirella Lapata, "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation", IEEE Transanctions on Parttern Analysis and Machine Intelligence, Vol.32, No.4, pp.678-692, 2010.

[23] 김원경, '한국어의 격', 박문사, 2009.

[24] Lawrence Rabiner(1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, Vol.77, No.2, 1989.

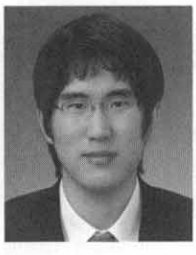
[25] 김재한, 옥철영, "어절 사전을 이용한 한국어 형태소 분석", 1994년 정보과학회 봄 학술발표논문집, 제21권, 제1호, pp.813-816, 1994.

[26] 임수종 박영자, 송만석, "가중치 정보를 이용한 한국어 동사의 의미 중의성 해소", 제10회 한글 및 한국어 정보 처리 학술대회 발표논문, pp.425-429, 1998.

[27] 남승호, '한국어 술어의 사건 구조와 논항 구조', 서울대학교 출판사, 2007.

[28] 옥철영, '세종 형태의미말뭉치 : 형태분석 오류 수정 및 동형이의어 태깅', 국어정보처리시스템 경진대회 사용자 설명서, 2010.

[29] 옥철영, 김혜영, 배영준, 신준철, 이용훈, 김홍순, 정성린, 김윤정, 최효식, 조희산, 최종원, 조미옥, 이민정, '어휘의미 관계 데이터베이스 확장', 국립국어원, 2010.
[30] 옥철영, 안미정, 김창환, 김혜영, Antangerel Changnaa, 배영준, 신준철, 이용훈, Nguyen Kiem Hieu, Vo Duc Thuan, 김홍순, 김지언, 이민정, 조미옥, '오픈 웹QA를 위한 어휘의미부착 기술개발', 한국전자통신연구원, 2010.



박 요 셉

e-mail : pys1249@ejnu.net
2009년 전남대학교 전자컴퓨터공학부(학사)
2009년~현 재 전남대학교 전자컴퓨터공학과 석사과정
관심분야 : 자연언어처리, 단어 의미 중의성

신 준 철

e-mail : dustsjc@nate.com
2007년 울산대학교 컴퓨터정보통신공학과 (학사)
2009년 울산대학교 컴퓨터정보통신공학과 (공학석사)
2009년~2011년 울산대학교 컴퓨터정보통신공학과(박사수료)
관심분야 : 문서 분류, 한국어 정보 처리

옥 철 영

e-mail : okcy@ulsan.ac.kr
1982년 서울대학교 컴퓨터공학과(학사)
1984년 서울대학교 컴퓨터공학과 (공학석사)
1993년 서울대학교 컴퓨터공학과 (공학박사)
1994년 러시아 TOMSK 공과대학 교환교수
1996년 영국 GLASGOW대학교 객원교수
1984년~현 재 울산대학교 컴퓨터정보통신공학과 교수
관심분야 : 한국어 정보처리, 지식베이스, 기계학습, 온톨로지

박 혁 로

e-mail : hyukro@chonnam.ac.kr
1987년 서울대학교 컴퓨터공학과(학사)
1989년 한국과학기술원 전산학과 (전산학석사)
1997년 한국과학기술원 전산학과 (전산학박사)
1994년~1998년 연구개발정보센터 선임연구원
1999년~현 재 전남대학교 전자컴퓨터공학부 교수
관심분야 : 정보검색, 자연언어처리, 데이터베이스, 인공지능