한영 기계번역에서 결정 트리 학습에 의한 한국어 부사격 조사의 의미 중의성 해소

(Decision Tree based Disambiguation of Semantic Roles for Korean Adverbial Postpositions in Korean-English Machine Translation)

박 성 배 † 장 병 탁 †† 김 영 택 ††

(Seong-Bae Park) (Byoung-Tak Zhang) (Yung Taek Kim)

요 약 한국어는 격조사에 의해 구문 역할이 결정되고 하나의 조사가 여러 개의 의미를 가지는 특징이 있다. 특히, 부사격 조사는 그 의미의 다양성으로 인해서 한영 기계 번역에서의 조사 번역을 어렵게 만든다. 본 논문에서는 부사격 조사가 가질 수 있는 의미격을 24개의 클래스로 분류한 후, 50만 어절 크기의 말뭉치에서 추출한 학습 예제와 결정 트리 추론(decision tree induction)을 통해 부사격 조사의 의미격 결정 규칙을 학습하였다. 결정 트리 추론 시 나타날 수 있는 학습 예제의 부족 문제는 단어 클래스를 사용함으로써 해결하였다. 실험 결과, 6개의 부사격 조사에 대해서 평균적으로 76.2%의 정확도를 보였으며, 이는 가장 많이 나타나는 의미격을 부사격 조사의 의미격으로 결정하는 방법에 비해 26.0%의 정확도 향상을 의미한다.

Abstract Korean has the characteristics that case postpositions determine the syntactic roles of phrases and a postposition may have more than one meanings. In particular, the adverbial postpositions make translation from Korean to English difficult, because they can have various meanings. In this paper, we describe a method for resolving such semantic ambiguities of Korean adverbial postpositions using decision trees. The training examples for decision tree induction are extracted from a corpus consisting of 0.5 million words, and the semantic roles for adverbial postpositions are classified into 25 classes. The lack of training examples in decision tree induction is overcome by clustering words into classes using a greedy clustering algorithm. The cross validation results show that the presented method achieved 76.2% of precision on the average, which means 26.0% improvement over the method determining the semantic role of an adverbial postposition as the most frequently appearing role.

1. 서 론

한국어에서는 문장 내에 있는 각 단어의 구문적 역할이 격조사에 의해서 결정되므로, 주어진 한국어 문장을 이에 해당하는 정확한 영어 문장으로 번역하기 위해서 조사의 정확한 번역은 반드시 이루어져야 할 과정이다. 한국어 격조사는, 크게 주격, 서술격, 목적격, 보격, 관형격, 부사격, 호격의 7 가지로 나뉘어 질 수 있는데[1], 한영 기계 번역에서는 부사격 조사가 다른 나머지 격조사에 비해 의미 모호성을 보다 심하게 일으키므로 본 논문에서는 부사격 조사에만 관심을 둔다. 한영 기계 번역에서 부사격 조사를 적합한 영어의 전치사[1]로 표현하

† 학생회원 : 서울대학교 컴퓨터공학부
sbpark@nova.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
btzhang@cse.snu.ac.kr
ytkim@cse.snu.ac.kr
논문접수 : 1998년 7월 8일
심사완료 : 2000년 2월 17일

1) 부사격 조사가 영어로 번역될 때 여러 가지 품사로 번역될 수 있지만, 본 논문에서는 부사격 조사가 영어의 전치사로 번역되는 경우만 다루기로 한다. 다음 예문에서 부사격 조사 '에'가 사용되었지만, 이 문장은 '~에도 불구하고'를 하나의 인식 단위로 보아야지만 정확한 번역이 가능하다. 따라서, 본 논문에서는 영어의 전치사로 번역되기 힘든 부사격 조사가 나타나는 문장은 고려 대상에서 제외하였다.

려면 이 부사격 조사의 의미를 결정하여야 하는데, 조사가 표제어(head word)에 대한 그 성분의 의미격을 나타내므로 부사격 조사의 의미 해석을 부사격 조사의 의미격 결정 문제로 생각할 수 있다.

조사가 지닐 수 있는 의미를 구분하기 위해서 조정미는 한국어의 의미격을 30 가지로 구분한 후, 23 가지의 대표적인 조사를 그 의미격에 따라 분류하였다[2]. 하지만, 이 논문에서는 의미격이 한국어 의미 분석을 위해 주어진다고 보았기 때문에, 이런 의미격을 자동적으로 결정하는 방법을 제시하지 않았다. 한편, 영한 기계번역에서는 전치사의 의미를 결정할 때 비슷한 문제가 발생한다. 즉, 영어의 전치사구도 문맥에 따라서 여러 가지 의미를 가질 수 있기 때문에 심각한 의미 모호성을 지니고 있다. 이러한 의미 모호성을 해소하기 위한 규칙을 직관이나 시행 착오를 통해서 만들 수도 있으나[3], 이런 과정은 계산론적으로 비용이 너무 많이 들 뿐만 아니라 이 문제에 대한 지식이 충분한 전문가를 필요로 한다. 또한, 규칙이 기반한 영역(domain)이 바뀌면 규칙을 수동으로 다시 작성하여야 하는 문제를 지니고 있다. 더욱 문제가 되는 것은, 이렇게 해서 만들어진 규칙이 최적의 성능을 보인다는 보장을 할 수 없다는 것이다. 이런 한계를 극복하기 위해서, 강원석은 온-라인 시소러스인 WordNet과 역전파 신경망을 사용하여 전치사구의 의미를 결정하려고 시도하였다[4]. 그러나, 이 방식은 WordNet과 같은 기계 가독형 사전에 의존하였으므로, 신뢰할 수 있는 기계 가독형 사전이 존재하지 않을 때는 적용할 수조차 없는 제약을 보였다. 또한, 이 논문에서는 전치사의 의미격이 같더라도 한국어로 번역될 때 다른 조사로 번역될 수 있는 문제를 해결하기 위해서 비교적 많은 수인 63개의 의미격을 정의하여 사용하였다. 그러나, 이 방법은 한국어 어휘 특성 문제를 의미격 분류 문제에 포함시킴으로써 오히려 전체 문제를 더 복잡하게 만든 것으로 보인다.

양단희는 한국어 문장내 각 논항의 격 파악을 위한 직접적인 지식으로 격 원형성이라는 개념을 제안하였다[5]. 의미 분석에서는 의미 모호성을 해소하기 위해서 의미자질을 통한 선택제한(selectional restriction)을 사용하는 것이 일반적이라고 믿어져 왔으나, 일관성 있고 보편성 있는 의미소(semantic primitive)를 찾는 것이 어려운 일이었다. 따라서, 이 논문에서는 각 용언과 명사에 대해 격 원형성을 말뭉치로부터 미리 계산해 둔 후, 논항의 격을 이로부터 간단히 결정하는 방법을 제시하였다. 이 방법의 장점은 말뭉치로부터 기계학습을 통해 데이타를 구축하였기 때문에 은유나 환유 현상을 자연스럽게 다룰 수 있다는 점이다. 그러나 이런 현상을 반영하기 위해 어휘 수준에서 데이타를 학습했기 때문에 대량의 학습데이타를 필요로 하였다. 확률을 상대 빈도값으로 대신하기 위해서는 일정 빈도 이상인 경우만을 처리대상으로 삼아야 하므로 데이타 부족 문제가 심각하게 나타날 가능성이 높다. 또한, 여러 가지 이유로 인해 격조사가 표현할 수 있는 격 종류를 3 가지로 제한하였는데, 이 3 가지 격 종류로는 기계번역에 필요한 격의미를 충분히 반영하지 못한다.

본 논문에서는 결정 트리 학습에 기반하여 한국어 부사격 조사의 의미격을 결정하는 새로운 방법을 제시한다. 한국어 부사격 조사가 가질 수 있는 의미격을 정의하고 이를 바탕으로 한영 기계번역에서 부사격 조사의 의미해석을 하는 시스템에 대해서 설명한다. 이 시스템은 객관적으로 데이타로부터 의미격 결정 규칙을 생성하기 위해서 결정 트리를 이용한다. 결정 트리가 이산적 값을 갖는 함수에 대한 실용적인 학습 방법이고[6] 단어의 특성이 의미 표지나 다른 기타의 방법을 통해 이산적으로 표현될 수 있기 때문에, 결정 트리는 기계 가독형 사전 없이 문장에 나타나는 조사의 의미격을 결정하는 규칙을 학습하는 데 적합하다. 또한, 결정 트리로 학습된 규칙은 사람이 해석하기 쉬우므로 인지적인 면에서도 뛰어나다고 할 수 있다. 한편, 결정 트리 추론시 나타날 수 있는 학습 예제 부족 문제를 해결하기 위해서, 단어 자체가 아니라 단어의 클래스에 기반하여 결정 트리를 학습하는 방법을 사용하였다.

본 논문의 구성은 먼저 2장에서 문제를 명확히 정의하고, 3장에서 기본적인 개념을 설명한 후, 말뭉치로부터 정보를 얻는 방법에 대해서 설명한다. 4장에서 제시된 방법으로 학습된 규칙이 얼마나 정확히 부사격 조사의 의미격을 결정하는지에 대한 실험 결과를 보이고, 5장에서 결론 및 향후 과제에 대해서 토의한다.

"go"	
default	to
도달장소격	to
수단격	by
...	
"speak"	
default	in
발화전달격	to
수단격	in
...	

그림 1 전치사 번역을 위한 영어 동사 사전의 예

예문. 나의 충고*에*도 불구하고, 그는 여전히 태만하다.

2. 부사격 조사의 의미 모호성

영어는 구성성분의 위치와 전치사의 역할로 의미격이 결정되기 때문에, 한국어의 주격이나 목적격과 같은 격 조사는 영어로 번역될 때 대체로 사라지고, 부사격 조사는 주로 전치사로 번역된다. 그러므로, 번역된 영어 문장이 원시 한국어 문장과 동등한 의미를 가지기 위해서는 각 부사격 조사의 목표 전치사가 정확하게 선택되어야 한다. 그러나, 한국어에서 하나의 부사격 조사는 영어의 전치사처럼 여러 가지 의미를 가질 수 있어서 모호성이 발생한다.

예를 들어, 다음의 네 문장을 보자.

예문 1. 약수터*에* 도착해 보니, 이미 사람들이 많았다.
예문 2. 지하철 역*에* 많은 사람들이 있었다.
예문 3. 어제 아침*에* 나는 그와 만났다.
예문 4. 궁금한 마음*에* 그곳을 찾아갔다.

위의 네 문장 모두에서 부사격 조사 '에'가 쓰였지만, 네 문장 모두에서 서로 다른 뜻으로 사용되었다. 부사격 조사 '에'가 예문 1에서 '도달 장소격', 예문 2에서 '장소격', 예문 3에서 '시간격', 예문 4에서는 '이유격'으로 사용되었으므로, 영어로 번역될 때 각각 다른 전치사로 번역될 수 있다. 즉, 예문 1에서는 '에'가 '*at*'으로, 예문 2와 3에서는 '*in*'으로, 예문 4에서는 '*because of*' 등으로 번역된다. 따라서, 부사격 조사의 의미격에 따라 대역어인 전치사가 달라질 수 있으므로, 부사격 조사의 의미격을 밝히는 것은 부사어의 번역을 위해서 필수 불가결한 일이다.

또한, 부사격의 조사의 격이 결정된다고 해서 바로 그 조사의 전치사를 결정할 수 있는 것은 아니다. 아래의 예문 5와 예문 6을 보자.

예문 5. 나는 버스*로* 그곳에 갔다.
예문 6. 그 사람은 영어*로* 말했다.

위의 두 예문에서 부사격 조사 '로'는 모두 '수단격'으로 사용되었지만, '로'의 번역은 예문 5에서는 '*by*'가, 예문 6에서는 '*in*'이 되어야 한다. 예문 5의 '로'의 경우, 동사 '가다'가 영어 '*go*'로 번역되기 때문에 수단격은 교통 수단을 의미하는 '*by*'로 번역되어야 하고, 예문 6에서의 '로'의 경우는 동사 '말하다'가 영어 '*speak*'로 번역되기 때문에 '*in*'으로 번역되어야 한다. 즉, 같은 의미라고 할지라도 영어로 번역될 때 그 전치사를 지배하는 동사에 따라 선택되어지는 전치사가 달라질 수 있다. 그러므로, 의미격을 정확하게 결정했다고 하더라도 의미격 정보만으로는 정확한 번역 결과를 얻을 수 없다. 이 문제를 해결하기 위해서 [4]에서는 목표 언어(target language)의 미세한 차이까지도 표현할 수 있도록 의미격을 자세히 나눈 후, 이를 학습하는 방식을 취하였다. 그러나, 이 방법은 시스템이 구분해야 할 의미격의 수가 더욱 많이 늘어나므로, 학습에 필요한 데이타의 수가 많아지는 문제점을 지니고 있다.

표 1 부사격 조사가 가질 수 있는 의미격 집합

의미격	정의	예	의미격 표지
장소격	행위가 일어난 장소	그는 집*에서* 잔다.	L
시작 장소격	출발점이 되는 장소	그는 집*에서* 출발했다.	SL
도달 장소격	도달 장소	그는 학교*에* 도착했다.	AL
시간격	행위가 일어난 시간	나는 아침*에* 일어났다.	T
시작 시간격	출발점이 되는 시간	지금*부터* 시작하자.	ST
도달 시간격	도달 시간	내일*까지* 마치자.	AT
발화 전달격	발화(發話)의 상대자	그는 나*에게* 말했다.	U
반응격	행위에 대한 반응	그의 말*에* 놀랐다.	A
양태격	동작이 행해지는 방법	그는 큰 소리*로* 말했다.	M
여격	수여동사의 수용자	그는 그*에게* 선물을 주었다.	D
대항격	행위에 대한 저항	왜적*에* 맞서 싸웠다.	G
도구격	행위의 도구	연필*로* 글을 쓴다.	I
재료격	행위의 재료	벽돌*로* 집을 짓는다.	MT
제공자격	행위의 제공자	그*에게* 선물을 받았다.	SP
수익격	이익이나 손해를 받는 대상	나는 나라*에* 공헌했다.	B
수단격	행위의 수단이나 방법	그는 영어*로* 말했다.	ME
이유격	행동의 원인	그는 암*으로* 죽었다.	R
관계격	행위와 관계를 맺는 대상	그 옷이 너*에게* 어울린다.	REL
한정격	행위가 미치는 영역	이 규칙은 너*에게* 적용된다.	LIM
용도격	행동의 목적	꽃을 선물*로* 주었다.	P
역할격	역할, 자격, 지위	나는 선배*로서* 너에게 충고한다.	RO
행위자격	행위의 주체	나는 개*에게* 물렸다.	AG
근거격	판단의 근거	친구*로* 사람을 알 수 있다.	BS
변형격	변형의 결과	물이 얼음*으로* 변했다.	TR
근원격	사건의 시작	얼음은 물*로부터* 생긴다.	S

본 논문에서는 의미격으로 번역 어휘를 결정하는 문제를 의미격 결정 문제와 분리해서 처리한다. 예문 5와 예문 6에서 같은 의미격이 서로 다른 전치사로 번역되는 이유를 자세히 살펴보면, 그 전치사를 지배하는 동사가 다르기 때문임을 쉽게 알 수 있다. 즉, 부사격 조사의 번역 어휘인 전치사를 결정하는 것은 그 조사의 의미격과 동사의 영어 번역 어휘이다. 그러므로, 한영 기계번역 시스템에서 사용되는 영어 동사 사전에 의미격과 해당 전치사를 기록하여 둠으로써 전치사 선택 문제를 의미격 결정 문제와 분리할 수 있다(그림 1).

위와 같이 한영 기계번역에서 부사격 조사의 의미를 결정하기 위해서는 부사격 조사의 의미격 집합이 있어야 한다. 의미격이나 의미 자질에 대해 한국어를 위한 표준이 존재하지 않으므로 본 논문에서는 [3]에서 정한 의미격을 차용하였다(표 1).

3. 부사격 조사의 의미격 결정

3.1 결정 트리 학습

부사격 조사의 의미격을 결정하는 규칙을 학습하기 위해서, 본 논문에서는 *결정 트리*(decision tree)를 사용하였다. 결정 트리 학습은 널리 사용되며 귀납적 추론(inductive inference)에 매우 실용적인 방법 중 하나이다. 결정 트리가 노이즈에 강하고 논리합 표현을 학습하는 이산 함수를 유도하는 방법이고 본 논문에서는 이산값으로 표현되는 데이타를 사용하기 때문에, 결정 트리는 이 데이타로부터 규칙을 학습하는 데 적합하다고 할 수 있다.

결정 트리에서는 각 비단말 노드가 인스턴스의 어떤 속성(attribute) 검사를 뜻하고, 그 노드로부터의 가지(branch)는 이 속성이 가질 수 있는 가능한 여러 값 중에서 하나이다. 각 인스턴스는 트리의 루트 노드에서 시작해서 아래로 내려가면서 분류된다. 비단말 노드에서는 인스턴스의 속성값을 검사해서 해당 값에 따라 트리의 가지가 선택된다. 이러한 과정이 새로운 노드를 루트로 하는 서브 트리에서도 반복되어 마지막 단말 노드에까지 이르게 된다. 결정 트리 학습에서 가장 중요한 문제는 트리의 각 노드에서 어떤 속성을 검사할 것인지를 선택하는 것이다. 예제 인스턴스를 분류하는데 가장 유용한 속성을 선택하기 위해서, 정량적 측정 단위인 *정보 이득*(*information gain*)이 사용될 수 있다. 이 측정 단위는 결정 트리 학습 알고리듬으로 가장 널리 쓰이는 ID3 알고리듬과 그 후계자라고 할 수 있는 C4.5 알고리듬에서도 채택된 것이다. 목적 속성이 c개의 서로 다른 값을 가질 수 있을 때 예제 인스턴스 집합(S)의 엔트로피는

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

와 같이 정의된다. 그러면, 예제 인스턴스 집합 S에서의 속성 A에 대한 정보 이득 $Gain(S,A)$은 다음과 같이 정의된다.

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

여기서 $Value(A)$는 속성 A가 가질 수 있는 가능한 모든 값을 뜻하고, S_v는 A의 값이 v인 S의 부분 집합이다. 그러므로, $Gain(S,A)$는 속성 A에 따라 학습 예제들을 나눔으로써 얻어지는 엔트로피 감소의 평균으로 해석될 수 있다.

표 2는 본 논문의 결정 트리 학습에 사용된 속성과 그 속성이 가질 수 있는 값을 보인다. 속성 N과 V는 각각 부사격 조사의 표제 명사와 표제 동사를 뜻하며, 속성값은 이들이 속하는 다음 절에서 설명할 단어 클래스이다. 속성 A는 표제 명사 N이 보조사 '은/는'과 함께 쓰였는지를 지시하는 이진 속성이다. 보조사 '은/는'은 표제 명사 N이 가질 수 있는 의미격의 다양성을 제한하므로 유용한 속성이 된다. 속성 D는 표제 명사 N과 표제 동사 V 사이에 있는 단어의 수를 나타내며, d를 N과 V 사이의 거리라고 했을 때 속성값은 $d_1(d \leq 2)$, $d_2(2 < d \leq 5)$, $d_3(d > 5)$ 중의 하나이다. 속성 R은 문장의 전체 길이에 대한 속성 D의 상대 거리이다. 마지막으로, 속성 O는 표제 명사와 표제 동사 사이에 목적어가 있는지를 나타낸다.

표 2 부사격 조사의 의미격을 결정하는데 사용되는 속성. d는 속성 N에서 V까지의 거리를 나타내며, r은 s를 전체 문장의 길이라고 했을 때 $r = d/s$로 정의된다.

속성	설 명	속 성 값
N	표제 명사	*N1, N2, ..., N200*
A	보조사 '은/는' 포함 유무	*yes* or *no*
V	표제 동사	*V1, V2, ..., V200*
D	N에서 V까지의 거리	$d_1(d \leq 2)$, $d_2(2 < d \leq 5)$, $d_3(d > 5)$
R	전체 문장 길이에 대한 D의 상대 거리	$s(r \leq 0.5)$ or $l(0.5 < r \leq 1)$
O	N과 V 사이의 목적어 존재 유무	*yes* or *no*

3.2 단어 클래스 생성

본 논문에서는 부사격 조사의 의미격 결정 규칙을 학

습을 위한 기초 정보로서 단어 클래스를 사용한다. 단어 자체를 사용하여 말뭉치에서 학습 정보를 추출하면, 많은 경우에 학습 예제의 빈도수가 낮고 이와 같이 빈도수가 낮은 학습 예제는 예제로서의 의미가 없다. 표 3은 말뭉치에서 추출한 동사와 부사격 조사가 첨가된 명사의 쌍을 추출한 결과를 보이고 있는데, 추출된 20,778개의 쌍 중에서 75%가 넘는 15,746개의 예제가 15번 이하의 빈도수를 가짐으로써 학습 예제로서의 기능을 하지 못한다. 이런 문제를 해결하기 위해서 본 논문에서는 단어 클래스에 기반하여 말뭉치에서 학습 예제를 추출하였다.

표 3 50만 단어 말뭉치에서 추출한 동사-명사 쌍의 빈도수에 따른 분포

빈도수	단어 쌍의 수	비 율
$n < 3$	9,574	46.0%
$3 \le n \le 15$	6,162	29.7%
$n > 15$	5,042	24.3%

Given n: the number of clusters we want
[step 1] Make a fully-connected weighted graph where
node = word
weight of the edge = augmented relative entropy between words connected by the edge
[step 2] $<i, j>$=two nodes with a minimum augmented relative entropy
[step 3] k = a new node resulting from merging i and j
[step 4] **for** all nodes l such that $l \neq i$ and $l \neq j$ and $l \neq k$
weight(l, k) = (weight(l, i) + weight(l, j)) / 2
[step 5] Remove node i and j from the graph.
[step 6] if(# of nodes in graph > n) **then** goto [step 2]
else print words in each node.

그림 2 확장 상대 엔트로피에 따라 단어를 클러스터링하는 욕심쟁이 알고리듬(Greedy Algorithm)

단어 클래스를 만들기 위해 단어 간의 유사도를 측정하는 단위로 상호 정보, 상대 엔트로피, WordNet 상의 거리, Tanimoto 측정식 등 여러 가지 단위가 있지만 [7,8,9,10], 본 논문에서는 단어를 확률분포로 나타내므로 단어의 유사도 단위로 확률 분포 사이의 거리인 상대 엔트로피(*relative entropy*)를 사용하였다. 상대 엔트로피는 확률 분포 $p(x)$와 $q(x)$에 대해서 아래와 같이 정의된다.

$$D(p \parallel q) = \sum_{x \in \chi} p(x) \log \frac{p(x)}{q(x)}$$

이 정의에서는 $0 \log \frac{0}{q} = 0$과 $p \log \frac{p}{0} = \infty$을 가정한다. 상대 엔트로피 $D(p \parallel q)$는 항상 0보다 크거나 같고 $p = q$일 때만 0이 되므로 이를 두 확률 분포 사이의 거리로 생각할 수 있다. 그러나, 일반적으로 교환 법칙이 성립하지 않고 ($D(p \parallel q) \neq D(q \parallel p)$) 삼각 부등식을 만족하지 않으므로, 이를 해결하기 위해서 본 논문에서는 *확장 상대 엔트로피*(augmented relative entropy) $D'(p \parallel q) = \frac{D(p \parallel q) + D(q \parallel p)}{2}$를 사용하였다.

말뭉치에서 사용된 동사 집합 $V = \{v_1, v_2, \cdots, v_m\}$과 명사 집합 $N = \{n_1, n_2, \cdots, n_k\}$에 대해서, p_{n_i}는 다음과 같이 표현될 수 있다.

$$p_{n_i} = \langle p(v_1|n_i), p(v_2|n_i), \cdots, p(v_m|n_i) \rangle$$

여기서 $p(v_j|n) = \frac{C(v_j, n)}{\sum_{v_i \in V} C(v_i, n)}$이고 $C(v_i, n)$는 동사 v_i와 명사 n이 동시에 나타난 횟수이다.

명사 집합으로부터 가능한 모든 명사 쌍에 대해서 확장 상대 엔트로피를 계산한 후, 명사들을 그림 2의 욕심쟁이 알고리듬(Greedy Algorithm)을 사용하여 명사 클래스로 분류한다. 이 알고리듬은 같은 방법으로 동사를 클러스터링하는 데에도 사용될 수 있다. 이 알고리듬에서는 클러스터링 하고자 하는 전체 단어를 노드로, 두 단어 사이의 확장 상대 엔트로피를 간선(edge)의 가중치(weight)로 삼은 완전 연결 그래프(fully-connected graph)를 만든 후, 노드들을 병합해 가면서 단어 클래스를 생성한다. 전체 그래프에서 가중치가 제일 낮은 간선으로 연결된 두 단어가 의미적으로 가장 비슷한 단어이기 때문에 step 3에서 이 두 단어를 묶어서 하나의 새로운 노드를 만들고 step 4에서 다른 노드와 이 새 노드 사이의 확장 상대 엔트로피 값을 변경한 후, step 5에서 기존의 두 단어를 그래프에서 제거한다. 이 과정을 한 번 수행할 때마다 두 개의 노드가 묶여서 하나의 노드가 되므로 노드의 수가 하나 줄게 된다. 그러므로, 전체 노드의 수가 원하는 수가 될 때까지 이 과정을 반복하면서 단어 클래스를 생성한다.

3.3 의미 모호성 해소

단어 의미 모호성 해소는 한 단어가 여러 가지 의미를 가질 수 있을 때 문장 내에서 의도하는 의미를 결정하는 것을 말한다. 한국어에서는 99.9%의 단어가 2개 이상의 의미를 가지는 것으로 알려져 있어서[8], 단어의 의미 모호성을 해소하지 않으면 앞 절에서 설명한 단어 클래스는 강성(hard) 클래스가 된다. 즉, 한 클래스 내의 각 단어의 의미는 그 단어가 가질 수 있는 여러 가

지 의미의 복합체가 되므로, 이 클래스에 기반한 자연언어 처리 모델이 오류를 일으키는 원인이 된다. 예를 들어, 금성 국어 사전에 따르면 단어 '배'가 6가지의 의미를 가지므로[11], 의미 모호성을 해소하지 않으면 단어 클래스에 속한 '배'라는 단어는 이 6가지 의미의 복합적인 의미를 가진다.

단어의 의미 모호성을 해소하기 위해서는 의미 부착 말뭉치(sense-tagged corpus)나 2개 국어 말뭉치(bilingual corpus)가 필요하지만 한국어에 대해서는 현재까지 신뢰할 만한 말뭉치가 존재하지 않으므로, 비감독 학습(unsupervised learning) 방식의 의미 모호성 해소 기법이 필요하다. 단어의 의미는 문맥에 의해 제약되는 특징이 있으므로 본 논문에서는 단어 의미를 기계 가독형 국어 사전에 기초해서 결정한다. 즉, 한 단어의 의미가 그 단어와 함께 한 문장에서 나타난 단어들과 사전에서 그 의미를 설명하기 위해 사용된 단어들 사이의 상관관계에 의해 결정된다고 가정하면, 그 단어의 정확한 의미는 사전에 기술된 여러 가지 의미 중 상관관계가 가장 높은 것으로 생각할 수 있다.

상관관계를 공식화하기 위해서, w를 의미가 모호한 단어, C를 w의 문맥, s_k를 사전에 나타난 w의 k번째 의미라고 하자. 그러면, C와 s_k는 다음과 같이 표현된다.

C : $w_1, \cdots, w_{i-1}, w, w_{i+1}, \cdots, w_n$

s_k : $w_{k1}, w_{k2}, \cdots, w_{kl}$

C와 s_k 사이의 상관관계는 아래와 같이 정의된다.

$$Cor(C, s_k) = \sum_{i=1}^{n} \frac{\sum_{j=1}^{l} D'(w_i, w_{kj})}{l}$$

여기서 $D'(w_i, w_{kj})$는 앞에서 정의된 w_i와 w_{kj} 사이의 유사도이다. 따라서, 단어 w의 정확한 의미 s는 다음과 같이 결정된다.

$$s = \arg\max_{s_k} Cor(C, s_k)$$

"배"
1. 사람이나 동물의 몸에서 위장 따위의 내장이 들어 있는 부분
2. 사람 또는 물건을 싣고 물 위로 떠다니는 물건. 선박.
3. 배 나무의 열매
4. 발생 초기의 생물체. 곧, 수정된 난세포가 자라서 된 극히 시초적인 것.
5. 갑절. 곱절
6. 술 또는 음료수의 잔 수를 세는 말

그림 3 금성 국어사전의 예

예문 7. 그는 빵으로 **배**를 채웠다.

예를 들어, 동아 국어사전에서는 명사 '배'를 그림 3과 같이 설명하고 있다. 따라서, '배'라는 단어가 포함된 예문 7의 '배'의 의미를 결정하기 위해서 사전에서 설명한 6개의 의미에 대해 각각 상관관계 값을 계산한다. 여기서 문맥 C는 예문 7이 되고, s_k는 '배'의 사전 의미 중 k번째 의미이다.

$$\begin{aligned}
Cor(C, s_1) &= \frac{D'(\text{그,사람}) + D'(\text{그,동물}) + \ldots + D'(\text{그,부분})}{9} \\
&+ \frac{D'(\text{빵,사람}) + \ldots + D'(\text{다스리,부분})}{9} \\
&+ \frac{D'(\text{배,사람}) + \ldots + D'(\text{배,부분})}{9} \\
&+ \frac{D'(\text{채우,사람}) + \ldots + D'(\text{채우,부분})}{9} \\
&\vdots \\
Cor(C, s_6) &= \frac{D'(\text{그,술}) + D'(\text{그,또는}) + \ldots + D'(\text{그,말})}{7} \\
&+ \frac{D'(\text{빵,술}) + \ldots + D'(\text{빵,말})}{7} \\
&+ \frac{D'(\text{배,술}) + \ldots + D'(\text{배,말})}{7} \\
&+ \frac{D'(\text{채우,술}) + \ldots + D'(\text{채우,말})}{7}
\end{aligned}$$

위의 수식에 의해 계산된 결과는 아래와 같다.

$Cor(C, s_1)$ = 7359.1 $Cor(C, s_2)$ = 4811.0

$Cor(C, s_3)$ = 104.4 $Cor(C, s_4)$ = 1058.5

$Cor(C, s_5)$ = 0.0 $Cor(C, s_6)$ = 5991.2

$Cor(C, s_1)$의 값이 다른 $Cor(C, s_i)$ $(2 \le i \le 6)$보다 크기 때문에, 예문 7의 '배'는 첫 번째 의미로 결정될 수 있다. 이렇게 해서 단어의 의미가 결정되면 의미가 포함된 단어는 원래 단어와는 독립된 단어로 인식하여 사용할 수 있다.

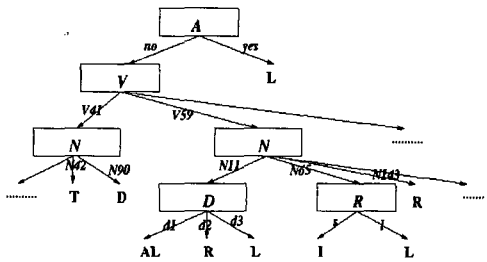


그림 4 조사 '에'에 대해서 학습된 결정트리의 예.

3.4 부사격 조사의 의미격 결정 과정

부사격 조사의 의미격을 결정하는 규칙은 말뭉치에 기반한 결정트리 추론에 의해 얻어진다. 추론에 앞서, 학습 데이타 부족 문제를 해소하기 위해서 말뭉치에 나타난 명사 집합 N과 동사 집합 V를 클래스로 분류한다. 우선, 말뭉치에서 목적어-동사 관계를 이용하여 명사와 동사를 각각 200 개의 클래스로 분류한다(그림 2). 이 클래스는 단어의 의미 모호성이 해소되지 않은 것이므로, 3.3 절에서 설명한 과정을 통하여 N과 V를 의미

모호성이 해소된 집합 N'과 V'으로 확장한다. 의미 모호성이 해소된 명사 집합 N'과 동사 집합 V'을 다시 각각 200 개의 클래스로 분류하여 결정트리 추론에 쓰이는 표제 명사와 표제 동사 속성으로 사용한다.

표 4 의미 모호성이 있는 6개의 부사격 조사가 가질 수 있는 의미격의 종류

의미격	에게	부터	로	에	에서	까지
장소격				O	O	
시작 장소격		O			O	
도달 장소격	O		O			O
시간격				O		
시작 시간격		O				
도달 시간격						O
발화 전달격	O					
반응격				O		
양태격			O			
여격	O			O		
대항격				O		
도구격			O			
재료격		O	O	O		
제공자격	O				O	
수익격	O			O		
수단격			O			
이유격			O	O		
관계격	O			O		
한정격	O			O		
역할격			O		O	
행위자격	O		O	O	O	
근거격			O		O	
변형격			O			
근원격		O	O		O	

단어 클래스가 만들어지면, 간단한 파서를 이용하여 각 조사에 대해 표 2와 같은 형태의 튜플을 자동으로 추출한다. 예를 들어, 예문 1에서는 조사 '에'에 대해서, <*약수터*, *no*, *도착하다*, d_1, *s*, *no*>라는 튜플이 추출된다. 다시, 명사 '약수터'와 동사 '도착하다'가 각각 $N11$, $V59$ 클래스에 속하기 때문에 이 튜플은 t_i = <$N11$, *no*, $V59$, d_1, *s*, *no*>라는 튜플로 변경된다. 말뭉치에서 이런 튜플들의 집합 $T = \{t_i\}(1 \le i \le L)$를 자동으로 추출한 후, 각 튜플 t_i에 적합한 의미격을 수동으로 추가하여 학습 데이타를 구성한다. 각 조사에 대해서 이와 같이 구축된 학습 데이타를 사용하여 그림 4와 같은 결정트리를 추론한다. 그림 4에 의하면 부사격 조사 '에'에 대해서 <$N11$, *no*, $V59$, d_1, ?, ?>의 형태를 갖는 데이타는 '*도달 장소격*(AL)'의 의미격을 가지게 된다.

4. 실 험

4.1 실험 데이타

초등학교 읽기 교과서, 중학교 국어 교과서, 신문 기사로 구성된 50만 어절 크기의 말뭉치에서 15번 이상씩 사용된 1,564개의 명사와 993개의 동사를 학습 대상으로 하였다. 한국어 부사격 조사에는 많은 종류가 있지만, 본 논문에서는 번역을 할 때 부사격 조사 중 의미 모호성이 가장 심하게 발생하는 6개의 부사격 조사에 대해서 의미격과의 관계를 살펴보았다(표 4). 또한, 학습 데이타 부족 문제를 해결하기 위해서, 동사-목적어 관계를 이용하여 말뭉치에서 추출한 1,564 개의 명사와 993 개의 동사를 각각 200개의 클래스로 분류하였다.

4.2 단어 클러스터링 실험

학습 알고리듬의 성능을 평가하기 위해서 대표적인 클러스터링 방법인 *c*-means 알고리듬과 분포 유사도 [12]와 비교하였다. 분포 유사도를 계산할 때는 주격, 목적격, 부사격의 세 가지 문법 기능에 대해서 각각 유사도를 계산하여 더하여야 하지만, 본 논문에서는 실험의 편의성을 위해 목적격만 고려하였다. 성능을 평가하기 위한 단위로 cross-entropy를 사용하였는데, 이 값은 실제 확률 분포와 역관계에 있으므로 이 값이 작을수록 더 좋은 방법이라고 할 수 있다. 실험 결과에 따르면,

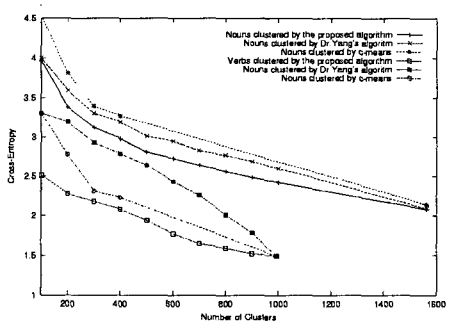
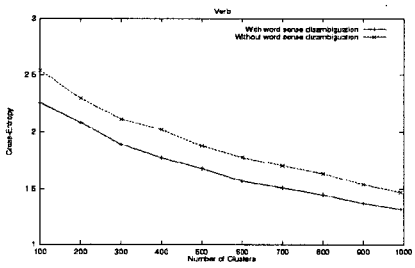


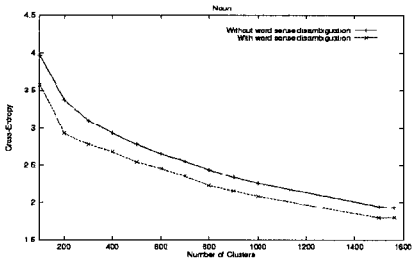
그림 5 제안된 클러스터링 방법과 c-means, 분포 유사도 클러스터링 방법과의 비교 실험 결과

명사에서는 분포 유사도가 c-means보다 나은 성능을 보이고, 동사에서는 c-means가 분포 유사도보다 나은 성능을 보이지만, 두 경우 모두 본 논문에서 제안한 방법이 가장 좋은 성능을 보이며(그림 5), 이는 다른 선행 연구 결과와 일치한다[8,9].

그림 6은 단어 의미 모호성 해소가 cross-entropy 감소에 얼마나 영향을 미치는지 보이고 있다. 이 결과에 따르면 명사와 동사 모두에서 단어의 의미 모호성을 해소했을 때가 하지 않을 때보다 좋은 성능을 보이며, 이는 일반적인 직관과 일치한다.



(가) 동사



(나) 명사

그림 6 단어 의미 모호성 해소 전과 후의 클러스터링 성능 비교. (가)는 동사에 대한 cross-entropy의 변화를 보이고, (나)는 명사에 대한 cross-entropy의 변화를 보인다.

4.3 의미격 결정 실험

의미격 결정 규칙을 학습하기 위해서 형태소 분석기와 간단한 파서를 이용하여[3] 각 부사격 조사에 대해서 표 2와 같은 형태의 튜플을 말뭉치에서 자동으로 추출하여 학습 데이타를 구축하였다. 구축된 데이타의 수와 가장 많이 쓰이는 의미격은 표 5와 같다.

표 5 50만 어절의 말뭉치에서 추출된 부사격 조사별 학습 예제의 수

부사격 조사	에게	부터	로	에	에서	까지
튜플의 수	144	43	695	1201	303	28
기본 의미	발화전달격	시작시간격	도달장소격	장소격	장소격	도달시간격
기본 의미의 성공률	35.4%	48.8%	40.9%	65.3%	57.4%	53.6%

수집된 튜플의 수가 많지 않으므로, 의미격 결정 학습의 정확도를 결정하기 위해서 *10-fold cross validation* 기법[2]을 사용하였다. 말뭉치에서 부사격 조사의 쓰임새를 조사해 본 결과, '에'는 *장소격*으로 65.3%, '로'는 *도달 장소격*으로 40.9% 등의 가장 많은 쓰임새를 보였다(표 5). 따라서, 이 퍼센트를 기준 정확도로 삼을 수 있다.

표 6 부사격 조사의 의미격 결정 규칙 10-fold cross validation 실험 결과

횟수	에게	부터	로	에	에서	까지
1	73.3%	70.0%	83.3%	83.3%	74.2%	66.7%
2	64.3%	80.0%	82.3%	79.5%	73.3%	100.0%
3	73.3%	70.0%	77.4%	75.8%	74.2%	100.0%
4	85.7%	70.0%	78.3%	78.3%	66.7%	100.0%
5	64.3%	60.0%	77.1%	81.6%	77.4%	66.7%
6	64.3%	75.0%	76.7%	84.0%	70.0%	66.7%
7	69.4%	80.0%	80.0%	84.1%	71.0%	66.7%
8	64.3%	80.0%	79.7%	78.3%	64.5%	100.0%
9	71.4%	80.0%	77.7%	84.0%	64.5%	50.0%
10	80.0%	70.0%	77.1%	80.2%	73.3%	100.0%
평 균	71.0%	75.4%	79.0%	80.9%	70.9%	81.7%
의미 모호성 해소를 하지 않을 때	64.3%	72.0%	76.1%	77.7%	67.2%	81.7%
기준 정확도	35.4%	48.8%	40.9%	65.3%	57.4%	53.6%

본 논문에서 제시한 방법의 실험 결과에 따르면, 평균적으로 '에게'에 대해서 71.0%, '부터'에 대해서 73.5%, '에'에 대해서 80.9%, '로'에 대해서 79.0%, '에

2) k-fold cross validation은 주어진 알고리듬을 k번 실행한 결과의 평균으로 평가하는 방법을 말한다. 초기의 학습 데이타를 k개로 나눈 후, k-1개의 부분 데이타로 학습하고 나머지로 알고리듬을 평가한다. 서로 다른 k개의 부분 데이타에 대해서 이 과정을 반복한 후, 결과의 평균으로 알고리듬을 평가한다.

서'에 대해서 70.9%, '까지'에 대해서 81.7%의 정확도를 나타낸다(표 6). 따라서, 기준 정확도보다 평균 약 76.2%의 정확도를 보이며, 이는 26.0%의 성능 향상을 의미한다. 실험 결과에 보이는 오류의 또 다른 원인은 수집된 데이타의 수가 일반적인 경우를 다룰 만큼 충분히 많지 않았고 결정트리의 크기를 확인 집합(validation set)으로 제어했음에도 불구하고 학습 데이타(training data)가 과도하게 학습(overfit)되었기 때문인 것으로 보인다.

표 7은 결정트리 학습에 사용된 속성 중 어느 것이 실제로 중요한지를 나타낸다. 이 표는 부사격 조사 '에'에 대해서 결정트리를 학습한 결과이며, 이 표에서 정확도 감소는 그 속성을 포함하지 않고 학습했을 때 의미격 결정 규칙의 정확도가 얼마나 감소했는지를 뜻한다. 실험 결과에 따르면, 여러 속성 중 가장 중요한 것은 표제 동사 V이다. 특히 속성 D는 오히려 성능을 떨어뜨리는 역할을 하는데, 그 이유는 더 연구해 보아야 하겠다.

표 7 결정트리 학습시 각 속성의 중요도.

속 성	정확도 감소 (%)
N	0.4
A	0.8
V	2.4
D	-0.1
R	0.4

5. 결론 및 향후 과제

본 논문에서는 결정트리 학습에 기반하여 한국어 부사격 조사의 의미 모호성을 해소하는 방법을 제시하였다. 말뭉치에서 추출된 데이타를 결정트리로 추론함으로써 부사격 조사의 의미격 결정 규칙을 얻은 후, 이 규칙을 사용하여 부사격 조사의 의미격을 결정하였다. 이 방법에 대한 실험 결과, 의미 모호성이 있는 한국어 부사격 조사에 대해서 평균적으로 약 76.2%의 비교적 높은 정확도를 보였다. 이는 가장 자주 쓰이는 의미격을 그 부사격 조사의 의미격으로 결정한 경우보다도 약 26.0% 정도의 정확도 향상을 보인 것이다. 따라서, 말뭉치로부터 추출한 예문을 결정트리로 추론하여 의미격 결정 규칙을 생성하는 것이 타당하다고 할 수 있다. 양단희의 격 원형성을 이용한 격조사의 의미결정 실험에서는 지도 및 비지도 학습 방식을 사용했을 때, 73.5%의 정확도를 얻었다[5]. 이 실험에서는 하나의 조사에 대해서 단 3개의 의미격만을 고려하였지만, 본 실험에서는 더 많은 의미격을 사용하였음에도 불구하고 더 좋은 결과를 얻었다. 그리고, 기계학습을 통해 규칙을 이끌어 냄으로써 은유와 환유 현상을 처리할 수 있었다. 또한, 본 논문에서 제시된 방법은 자동으로 데이타를 수집한 후 이를 학습하는 방식이었으므로 매우 범용적인 방법이다. 따라서, 이 방법을 영어 전치사의 의미격 결정과 같은 문제에도 그대로 적용할 수 있을 것이다.

본 논문에서는 학습 데이타의 부족 문제를 해결하기 위하여 '단어 클래스'를 사용하였는데, 단어 클래스는 '의미 자질'보다 약간의 정확성은 떨어지지만 자동으로 구축할 수 있기 때문에 비용이 적게 들고 객관성과 확장성을 확보할 수 있는 장점이 있다. 하지만, 본 논문에서는 단어 클러스터링을 위해서 동사-목적어 관계만 활용하였으므로, 동사 클래스는 타동사로만 제약되는 한계가 있다. 따라서, 목적어뿐만 아니라 주어도 약하기는 하지만 용언을 제약하는 성격을 가지고 있으므로, 다른 성분을 단어 클러스터링에 활용하는 방법에 대해서 연구할 계획이다. 또한, 본 논문에서는 단어 클래스를 명사와 동사에 대해서 각각 200 개씩으로 나누었는데, 이 클래스의 수에 대한 객관성이 결여되었다. 따라서, 단어 클래스를 어느 정도로 했을 때, 최적의 성능을 보이는지를 연구하여야 한다.

본 논문에서 사용한 단어 클러스터링 방법은 미경험 단어에 대해서 단어 클래스를 결정하기 위해서 단어의 분포 정보를 얻기 위해서 대량의 말뭉치를 필요로 하는 문제점이 있다. 이 문제는 비록 단어 클래스 자동 생성의 공통적인 문제점이기는 하지만, 이 문제를 해결하기 위해서 Self-Organizing Map과 같은 신경망 모델을 활용하여 해결할 계획이다.

참 고 문 헌

[1] 남기심, 고영근, "표준 국어문법론", 탑출판사, 1989.
[2] 조정미, 김길창, "한국어 의미 해석시 중의성 해소에 대한 연구", 정보과학회지, 제14권, 제7호, pp.71-83, 1996.
[3] 황호성, "한영 기계 번역에서 부사격 조사의 번역", 서울대학교 석사학위 논문, 1998.
[4] W. S. Kang, J. Y. Seo, K. S. Choi and G. C. Kim, "A Neural Network Method for the Semantic Analysis of Prepositional Phrases in English-Korean Machine Translation," *Computer Processing of Chinese and Oriental Languages*, Vol. 8, No. 2, pp. 163-176, 1994.
[5] 양단희, 송만석, "기계학습에 의한 단어의 격 원형성 자동 획득", 정보과학회논문지, 제25권, 제7호, pp.1116-1127, 1998.
[6] J. R. Quinlan, *C4.5: Programs for Machine*

Learning, Morgan Kaufmann Publication, 1993.
[7] 천성진, "의미 계층을 이용한 전치사구의 수식 위치와 의미 결정에 관한 연구", 서울대학교 석사학위 논문, 1995.
[8] 박영자, "사전을 이용한 단어 의미 자동 클러스터링 : 유전자 알고리즘 접근법", 연세대학교 박사학위 논문, 1998.
[9] Pereira F., Tishby N. and Lee L., "Distributional Clustering of English Words," In *Proceedings of Annual Meetings of the Association for Computational Linguistics(ACL)*, pp.183-190, 1993.
[10] Jun Gao and XiXian Chen, "Probabilistic Word Classification based on a Context-sensitive Binary Tree Method," *Computer Speech and Language*, Vol. 11, No. 2, pp.307-320, 1997.
[11] "뉴 에이스 국어 사전", 금성사, 1987.
[12] Jaehyung Yang, "Conjunction Identification in Korean Noun Phrase Coordination Using Cooccurence Similarity," *Computer Processing of Oriental Language*, Vol. 10, No. 4, pp.391-408, 1997.
[13] 강원석, 서정연, 김길창, "영한 기계 번역에서의 전치사구 처리를 위한 격의미 체계와 의미속성 집합", 제6회 한글 및 한국어 정보처리 학술대회 논문집, pp.177-180, 1994.
[14] 김나리, 김영택, "한국어 동사 패턴에 기반한 한국어 문장 분석과 한영 변환의 모호성 해결", 정보과학회논문지, 제23권, 제7호, pp.766-466, 1996.
[15] 이휘봉, 이종혁, 이근배, "구문의존구조에서 개념그래프 생성을 위한 한국어의 의미분석", 한국정보과학회 봄 학술발표 논문집, 제24권, 제1호, pp.463-466, 1997.
[16] 박성배, 김영택, "한국어 부사격 조사의 의미격 결정", 한국정보과학회 봄 학술발표 논문집, pp.399-401, 1998.
[17] Kuand-Hua Chen and Hsin-Hsi Chen, "Attachment and Transfer of Prepositional Phrase with Constraint Propogation," *Computer Processing of Chinese and Oriental Language*, Vol. 6, No. 2, pp. 123-142, 1992.
[18] E. V. Siegel and K. R. McKeown, "Emergent Linguistic Rules from Inducing Decision Trees: Disambiguating Discourse Clue Words," In *Proceedings of the 12th National Conference on Artificial Intelligence(AAAI)*, pp. 820-826, 1994.
[19] S. B. Park and Y. T. Kim, "Semantic Role Determination in Korean Relative Clauses Using Idiomatic Patterns," In *Proceedings of the 17th International Conference on Computer Processing of Oriental Languages(ICCPOL)*, pp. 1-6, 1997.
[20] D. Petitpierre, S. Krauwer, D. Arnold, and G. B. Varile, "A Model for Preference," In *Proceedings of Annual Meetings of the Association for Computational Linguistics(ACL)*, pp. 134-139, 1987.
[21] E. Charniak, *Statistical Language Learning*, MIT Press, 1993.
[22] Hindle D., "Noun Classification from Predicate-Argument Structures," In *Proceedings of Annual Meetings of the Association for Computational Linguistics(ACL)*, pp. 268-275, 1990.

박 성 배
1994년 한국과학기술원 학사. 1996년 서울대학교 컴퓨터공학과 석사. 1996년 ~ 현재 서울대학교 컴퓨터공학부 박사과정 재학. 관심분야는 한국어정보처리, 기계번역

장 병 탁
정보과학회논문지 : 소프트웨어 및 응용 제 27 권 제 5 호 참조

김 영 택
정보과학회논문지 : 소프트웨어 및 응용 제 27 권 제 5 호 참조