자질집합선택 기반의 기계학습을 통한 한국어 기본구 인식의 성능향상

(Improving the Performance of Korean Text Chunking by Machine Learning Approaches based on Feature Set Selection)

황 영 숙 † 정 후 중 †† 박 소 영 † 곽 용 재 † 임 해 창 †††

(Young-Sook Hwang) (Hoojung Chung) (So-Young Park) (Young-Jae Kwak) (Hae-Chang Rim)

요 약 본 연구에서는 기계학습을 이용하여 한국어 기본구(base phrase)인식의 성능을 향상시키고자 할 때, 학습집합으로부터 획득 가능한 자질집합들 중 최적의 자질집합이 무엇이며, 자료부족 문제를 어떻게 완화할 것인가에 대해 논한다. 먼저 최적의 자질집합 선택은 "점증적 유용성"이란 관점에서 자질의 적합성을 정의하고 이러한 정의에 따라 자질집합을 선택한다. 그리고, 자료부족 문제 완화의 해결점을 찾기 위해 한국어의 통사적 특성을 고려한 형태소 품사체계 사용 및 선택적 어휘자질의 사용이 성능에 미치는 영향을 분석하고 결과를 제시한다.

다양한 크기의 문맥 및 속성, 품사체계에 따라 자질 집합을 구성하고, 서로 다른 특성을 갖는 학습기법인 결정트리와 메모리기반 학습기법을 적용한 결과, 한국어 기본구 인식에 유용한 자질은 품사, 어휘, 그리고 기본구 태그로, 두 학습 알고리즘 모두 동일하였다. 또한 한국어의 특성을 고려한 일반화된 품사체계 및 선택적 어휘자질의 사용이 자료부족 문제를 완화시켜주면서 안정된 성능을 보여주었다. 선택된 최적의 자질집합을 사용하여 결정트리와 메모리 기반 학습을 수행한 결과, 전체 기본구에 대해 각각 93.39%/93.41%, 90.99%/92.52%의 정확률/재현율을 얻었다.

키워드 : 한국어 문장의 기본구 인식, 기계학습, 자질집합선택, 결정트리 학습, 메모리 기반 학습

Abstract In this paper, we present an empirical study for improving the Korean text chunking based on machine learning and feature set selection approaches. We focus on two issues: the problem of selecting feature set for Korean chunking, and the problem of alleviating the data sparseness. To select a proper feature set, we use a heuristic method of searching through the space of feature sets using the estimated performance from a machine learning algorithm as a measure of "incremental usefulness" of a particular feature set. Besides, for smoothing the data sparseness, we suggest a method of using a general part-of-speech tag set and selective lexical information under the consideration of Korean language characteristics. Experimental results showed that chunk tags and lexical information within a given context window are important features and spacing unit information is less important than others, which are independent on the machine learning techniques. Furthermore, using the selective lexical information gives not only a smoothing effect but also the reduction of the feature space than using all of lexical information. Korean text chunking based on the memory-based learning and the decision tree learning with the selected feature space showed the performance of precision/recall of *90.99%/92.52%*, and *93.39%/93.41%* respectively.

Key words : Korean Base Phrase Recognition, Machine Learning, Decision Tree, Feature Set Selection, Memory-based Learning

· 이 논문은 2000년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2000-V01452-E00305).

† 학생회원 : 고려대학교 컴퓨터학과
yshwang@nlp.korea.ac.kr
ssoya@nlp.korea.ac.kr
yjkwak@nlp.korea.ac.kr

†† 비 회 원 : 고려대학교 컴퓨터학과
hjchung@nlp.korea.ac.kr

††† 종신회원 : 고려대학교 컴퓨터학과 교수
rim@nlp.korea.ac.kr

논문접수 : 2001년 11월 16일
심사완료 : 2002년 7월 2일

1. 서 론

한국어 문장에서의 기본구(base phrase) 인식은 통사적으로 서로 밀접하게 연결되어 있는 비재귀적(non-recursive) 형태의 모든 구를 인식하는 것으로 부분 구문분석(partial parsing)이라고도 한다. 이러한 기본구 인식은 완전 구문분석(full parsing)의 계산 복잡도(computational complexity)를 줄이기 위한 전처리로써 뿐만 아니라, 간단한 수준의 구문정보들을 요구하는 자연어 처리 응용 프로그램들에서 효과적으로 활용될 수 있다. 특히 완전 구문분석의 처리 비용이 비싼 반면에 성능이 기대를 따라주지 못한다는 인식이 높아지면서 높은 처리효율과 안정된 성능을 보장할 수 있는 부분 구문분석에 대한 요구가 늘어가고 있다. 다음의 예들은 기본구 인식과 그 응용사례를 보여주는데, 이로부터 인식결과의 응용범위를 가늠해 볼 수 있을 것이다.

(1) [NP 1991년 1월 중순], [NP 세종기지]에서는 [NP [NP 서울]의 [NP 기지방문단]]을 [VP 기다리고 있]었다.

(2) (VP [VP 1991년 1월 중순], (VP [NP 세종기지]에서는 (VP [NP[NP 서울]의 [NP 기지방문단]]을 [VP 기다리고 있]었다))).

(3) (TIME [NP 1991년 1월 중순]), (LOC [NP 세종기지]에서는) (OBJ [NP[NP 서울]의 [NP 기지방문단]]을) (PRED[VP 기다리고 있]었다).

즉, 예 (1)은 한국어 문장에서 기본 명사구(NP)와 기본 동사구(VP)를 인식한 것을 보여주고, 예 (2)와 예 (3)은 기본구 인식결과가 완전구문분석의 전처리나, 얕은 수준의 의미분석에 효과적으로 사용될 수 있음을 보여준다. 또한 기본구 인식결과는 그 자체로써 혹은 (2), (3)과 같은 처리 단계를 거쳐 정보추출이나 질의응답 시스템 등에도 사용될 수 있을 것이다.

기본구는 수동규칙을 이용하거나 기계학습을 통해 인식할 수 있다. 수동규칙의 경우, 전문가의 지식에 의존하여 인식하고자 하는 기본구에 대한 규칙을 수동으로 작성하는데, 가능한 모든 경우를 고려하는 것이 어려울 뿐만 아니라 일관성을 유지하기가 어렵다는 문제가 있다. 이에 반해 기계학습을 이용하는 경우는 대량의 학습 말뭉치로부터 자동으로 규칙/목적함수를 획득하게 되므로 시간과 비용을 절약할 수 있고, 일관된 작업을 수행할 수 있다는 장점이 있다. 그리하여 기계학습을 이용하는 경향이 높아지고 있는 추세이다.

지금까지 기계학습을 이용해 기본구를 인식하고자 했던 많은 연구들을 살펴보면 주로 학습기법에 초점을 맞추어 왔음을 알 수 있다. 이는 기존의 학습 알고리즘을 기본구 인식에 적용하거나, 기본구 인식에 적합한 학습 알고리즘을 새롭게 개발함으로써 보다 나은 성능을 획득하고자 함이었다. 그러나, 성능에 영향을 미치는 요인은 단지 학습 알고리즘뿐만이 아니다. 어떤 속성들을 자질로 선택하여 학습할 것인가 역시 성능에 중요한 영향을 미치는 것이다.

일반적으로 기본구 인식을 위한 학습집합은 다양한 속성들을 포함하는데, 작업 대상에 따라 포함되는 속성들이 달라질 수 있다. 예를 들어, 영어를 대상으로 하는 경우 단어단위로 띄어쓰기를 하므로 띄어쓰기 속성이 고려되지 않는다. 이에 반해, 교착어에 속하는 한국어는 하나 이상의 형태소가 한 어절을 형성하고 어절 단위로 띄어쓰기를 한다는 특징이 있으므로, 띄어쓰기 속성은 유용한 자질이 될 수도 있다. 또한 품사정보는 기본구 인식에서 많이 사용되는 자질중의 하나인데, 사용되는 품사체계에 따라 학습효과가 달라질 수도 있다. 이렇듯 학습집합에 포함되는 속성들은 학습 자질로 선택될 수 있으며, 자질집합을 어떻게 구성하느냐에 따라 동일한 학습 알고리즘이라 할지라도 학습효과는 달라질 수 있을 것이다.

본 연구에서는 기계학습을 적용하여 비 재귀적 기본구를 인식하고자 할 때, 중요한 자질이 무엇인가를 알아보고자 한다. 이를 위해 "점증적 유용성"이란 관점에서 자질의 적합성을 정의하고, 서로 다른 특성을 갖는 학습기법들을 적용하여 각 자질집합에 대한 성능을 비교·분석한 뒤, 최적의 자질집합을 제시하고자 한다. 또한, 자료부족 문제를 완화하기 위해 한국어의 통사적 단어형성 특성을 고려한 형태소 품사체계 사용 및 선택적 어휘자질의 사용이 성능에 미치는 영향을 분석하고 결과를 제시한다.

기계학습 기법은 자질의 정보량에 따라 자질 우선순위를 결정하고 귀납적 결론을 유도하는 결정트리 학습기법과, 모든 학습예제를 그대로 메모리에 저장하였다가 새로운 자료가 들어오면 학습예제들과의 유사도를 평가하여 결과를 출력하는 메모리기반 학습기법을 사용한다. 그리고 기본구 인식을 위한 자질집합은 다음의 방법에 따라 선택한다: 첫째, 학습 말뭉치로부터 추출 가능한 모든 단위 자질집합들을 만든다. 둘째, 이들의 가능한 모든 조합으로 만들어진 자질집합들을 대상으로 기계학습을 수행하고 성능을 비교·분석한다. 셋째 분석결과 가장 좋은 자질집합을 선택하고 이들을 기계학습에 적용했을 때의 성능을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서 기본구 인식

과 관련한 국내외 연구를 살펴보고, 3장에서 한국어의 특성을 고려한 비 재귀적 한국어 기본구에 대해 정의하고 기본구 인식의 문제를 태그부착의 관점에서 정의한다. 그리고 4장에서 기본구 인식을 위한 자질집합의 선택 및 자료부족 문제 완화를 위한 방법을 논하고, 5장에서 기본구 인식을 위한 기계학습 기법에 대해 간략히 기술한다. 6장에서 제시된 기계학습 기법들과 자질집합을 적용하여 수행된 실험결과들을 비교 분석하고 한국어 기본구 인식을 위한 최적의 자질집합을 제시한다. 마지막으로 7장에서 결론을 맺는다.

2. 관련연구

기본구(base phrase 혹은 chunk)의 개념은 1991년 Abney에 의해 처음 도입되었다[1]. Abney는 문장에서 끊어 읽는 운율적 휴지의 단위와 구문구조에 있어서의 단위가 상응한다고 보고 "하나의 중심어를 포함한 겹쳐지지 않는(non-overlapping) 단어들의 묶음"을 기본구로 정의하였다[1]. 이후 영어권에서는 [1]의 기본구에 대한 정의를 기반으로 기본구 인식에 대한 연구가 활발히 진행되었고, 최근에는 국내에서도 많은 연구가 시도되었다. 관련연구는 크게 수동규칙에 의한 방법과 기계학습에 의한 방법으로 구분되며, 국내외 연구를 살펴보면 다음과 같다.

2.1 수동규칙에 의한 방법

수동규칙으로 기본구를 인식하는 방법들은 주로 유한상태 오토마타나 경험적 규칙에 의존한다. Abney는 유한상태 오토마타를 사용한 다단계 구문분석 기법으로, 첫 번째 단계에서 서로 연관된 단어들의 덩어리인 기본구를 찾고, 이후 단계들에서 기본구들을 결합하여 더 큰 덩어리의 구를 만들면서 구문트리를 형성하고자 하였다[2]. 또한 Grefenstette는 단어열과 품사열을 입력받아 비재귀적 명사구와 동사구를 인식하고자 하였는데, 정규표현식에 의해 문법을 기술하고 유한상태 오토마타 툴을 사용하여 구문분석기를 구현하고자 하였다[3].

국내의 경우, 수동규칙을 사용한 연구로는 동사구 장벽 알고리즘을 이용한 연구[17]와 세 단계에 걸친 한국어 청크(Chunk) 인식에 대한 연구[18]가 있었다. 동사구 장벽 알고리즘을 사용한 연구는 정규표현식으로 작성된 규칙에 기반하여 명사구를 인식하고, 동사구 장벽 알고리즘을 적용하여 동사구를 묶고자 하였다. 한편 세 단계에 걸친 한국어의 청크 인식은 한국어의 띄어쓰기와 형식 형태소 특성을 이용하여 기본적인 구를 만들고, 그 다음에 간단한 CFG를 이용하여 명사구를 인식한 뒤, 세 번째 단계에서 대량의 말뭉치로부터 획득한 연어 패턴 정보를 이용하여 동사구를 인식하고자 하였다. 이러한 방법들은 비재귀적 기본구를 인식하기보다는 정의된 수준에서의 명사구 인식을 수행하고, 일부의 명사구 포함을 허용하면서 동사구를 인식하는데 초점이 맞추어졌다.

그러나 이러한 수동규칙에 의한 방법들의 경우, 모든 언어적 현상을 규칙으로 작성하기 어려울 뿐만 아니라, 규칙의 일관성을 유지하기가 어렵다는 문제가 있다. 그러므로 많은 성능향상을 기대하기 어렵다.

2.2 기계학습에 의한 방법

기계학습에 의한 기본구 인식은 기본구 인식 결과가 표기된 학습 말뭉치로부터 인식기를 학습한 후 실제 데이타에 적용하는 방법이다. 국외 연구로는 주로 은닉 마르코프 모델[4][5], 최대 엔트로피 모델[6][7]을 이용한 통계기반 접근법, 오류기반의 변형규칙 학습기법[8], 메모리 기반 학습방법[9][10][11][12][13], SVM을 이용한 학습기법[14]들이 있다. 또한 최근에는 단일 학습기법들의 한계를 극복하고자 여러 가지 학습기법을 복합적으로 사용하고자 하는 시도들[15][16]이 이루어지고 있다.

통계기반 방법의 초기 연구는 단어 사이에서 명사구가 시작할 것인지 아니면 끝날 것인지를 조건확률을 이용해 결정하는 것이었다[4]. 이후 제한된 수준의 구문범주를 인식하기 위해 HMM과 최대 엔트로피 모델을 결합한 방법들이 제시되었는데[6][7], 이들은 수동으로 자질 템플릿을 구성한 후 최대 엔트로피 개념에 따라 확률을 추정하여 기본구를 인식하고자 하였다. 한편, Ramshaw와 그의 동료들은 변형기반 학습기법을 적용하여 기본구를 인식하는 방법을 제시하였는데[8], 초기 예측이 표시되어 있는 학습 말뭉치로부터 초기 시스템을 생성하고, 올바른 태그가 명시되어 있는 목표 말뭉치와 규칙 템플릿을 사용하여 초기 시스템의 오류를 교정할 수 있는 규칙을 유도하고자 하였다. 이 방법은 기본구 인식을 품사부착과 동일한 관점에서 정의한 선구적 연구로서, 이후 많은 연구자들이 기본구 인식을 태그 부착의 문제로 바라보기 시작하였다.

그러나 통계기반이나 변형기반 방법들의 경우, 대부분 수동으로 자질 템플릿 및 규칙 템플릿을 구성했기 때문에 모든 언어현상을 포괄할 수 있는가에 대한 의문이 제기되고 있으며, 최적의 자질 템플릿(혹은 규칙 템플릿) 구성에 대해서는 거의 연구보고가 없는 실정이다.

메모리기반 학습방법은 기본구 인식에서 아주 활발히 연구되는 기계학습분야의 하나로 학습 말뭉치로부터 기본구 패턴을 추출하고 이들을 규칙화하여 적용하고자

하는 방법[9][10], 저장되어 있는 기본구 패턴들과 후보 기본구에 나타나는 품사 패턴들과의 유사도를 측정함으로써 가장 유사도가 높은 기본구를 결과로 제시하는 방법[11]등이 있다. 또한 학습자료를 품사나 어휘자질들의 벡터로 표현하고 자질들에 가중치를 부여하여 유사도를 계산하는 방법[12][13]들이 있다. 이러한 메모리 기반 학습기법은 자체적인 자질선택 기능이 없으므로 자질벡터를 어떤 자질들로 구성하는가에 따라 많은 성능차이가 발생할 수 있다.

최근 들어서는 자질선택 기능을 내포하고 있는 SVM (Support Vector Machine)을 이용하여 영어 문장의 기본구를 인식하고자 하는 연구[14]가 있었는데, 8개의 SVM을 결합한 결과 93.91($F_{\beta=1}$)이라는 높은 성능을 보고하였다. 이 방법은 학습예제를 중심어와 좌우 각 2단어의 품사, 어휘 그리고 기본구 태그로 표현하고 SVM으로 학습하였는데, 초기 자질집합의 선택이 성능에 미치는 영향에 대한 연구결과는 거의 발표된 바가 없다.

한편, 국내연구로는 변형규칙기반 학습에 의한 기본구 인식[19], 통계기반 접근법에 의한 연구[21][22][23] 등이 있다. [19]는 변형규칙기반 학습을 적용하여 비재귀 명사구를 인식하는 방법을 제안하였는데, 수동으로 작성된 200개의 템플릿을 사용하고, 처리 복잡도를 줄이기 위해 품사분류체계를 단순화시켜 처리하였다. 이 방법의 경우 한국어에 대한 전문적 지식을 바탕으로 세심한 주의를 기울여 템플릿을 작성하였지만, 학습집합에 존재하는 모든 언어현상을 포괄할 수 있는가에 대한 의문이 제기된다.

[20]은 결정트리 학습기법과 최대 엔트로피 개념을 결합하는 방법을 제시하였는데, 먼저 좌우 2 단어의 품사와 기본구 태그로 이루어진 자질집합을 사용하여 결정트리를 학습하고, 학습결과로부터 규칙을 자동으로 추출한 뒤, 최대 엔트로피 모델을 이용하여 규칙의 확률을 추정하고자 하였다. 이 연구결과는 결정트리 학습만을 사용했을 때보다 최대 엔트로피 모델과 결합했을 때, 성능이 향상되었음을 보여주었다. 그러나, 이 연구는 영어에 제한되었으며, 제한된 자질집합만을 대상으로 하고, 다른 가능한 자질집합에 대해서는 고려하지 않고 있다.

[22]는 가중적인 확률합(Weighted Probabilistic Sum)에 의한 기본구 인식방법을 제안하였는데, 현재 단어와 좌우 각 2 단어들의 품사를 가변적으로 사용하는 16개의 자질 템플릿을 만들고 각 자질 템플릿의 정보이득률을 가중치로 적용하여 기본구 태그를 인식하고자 하였다. 그러나, 이 방법 역시 자질집합이 품사로 한정되어 있고, 또한 제한된 문맥을 사용하고 있는데, 어휘나 기타 다른 자질을 포함한 적절한 자질집합을 선택한다면 더 나은 성능 향상을 기대할 수 있을 것이라 보여진다.

지금까지 기계학습 기법을 적용하여 기본구를 인식하고자 했던 국내외 연구들을 살펴보았다. 그런데 대부분 효과적인 학습기법의 개발이나 적용에는 많은 노력을 기울여 왔지만, 학습 알고리즘에 제시되는 자질집합의 선택에 대한 연구는 거의 전무한 상황이다. 자질선택 기능을 내포하지 않은 기계학습 방법이든 자질선택 기능을 내포한 기계학습 방법이든 학습 알고리즘에 주어지는 학습자료는 전문가의 경험적 지식에 의거하여 임의로 선택된 자질집합으로 구성되었을 뿐이었다.

그러나 자질선택 기능을 내포하지 않은 경우는 물론이고, 자체적인 자질선택 기능을 내포하고 있다 할지라도, 학습 알고리즘에 제시될 초기 자질집합의 선택은 성능에 많은 영향을 미칠 수 있다. 그러므로 학습 말뭉치의 구성과 학습자료 구축을 위한 자질집합의 선택은 성능향상 측면에서 매우 중요한 문제중의 하나가 된다.

또한 국내 연구의 경우, 비재귀적 명사구 인식에 대한 연구는 어느 정도 이루어지고 있으나, 동사구나 다른 비재귀적 기본구들의 인식에 대한 연구는 거의 없는 실정이다. 이에 본 논문에서는 한국어 문장에서의 비재귀적 명사구를 포함한 다른 기본구들을 정의하고, 한국어 기본구 인식에 유용한 자질집합을 어떻게 선택할 것인가에 대해 논의하고자 한다.

3. 한국어 문장의 기본구(Base Phrase) 인식

3.1 기본구의 정의

본 논문에서는 비 재귀적이며(non-recursive) 겹침이 없는(non-overlapping) 구(Phrase)를 기본구로 정의한다. 즉, 그 구조가 다른 하위 구조를 포함하지 않으면서 중심어가 최대로 투사되어 형성된 구를 기본구로 정의한다. 그리고 한국어 문장을 구성하는 기본구는 명사구(NP), 동사구(VP), 부사구(AP), 독립어구(IP)를 포함하며, 관형어구는 기본구에 포함하지 않는다. 비 재귀적이며 겹침이 없는 한국어 기본구를 정의할 때 고려한 사항은 다음과 같다.

첫째, 기본구 구성의 기본단위는 형태소로 한다. 이는 한국어의 교착어적 특성을 고려한 것으로 구를 형성하는 주체는 실질형태소들이고, 형식형태소는 주로 구와 구 사이의 문법적인 관계를 명시한다는 특성이 있으므로 이들을 활용하여 구를 명확히 구분하고자 함이다. 즉,

"도그마의 수정 변경이 가능하다는 사실" 이라는 문장에 대해 "(NP (VP (NP [NP 도그마] 의 [NP 수정 변경])이 [VP 가능하])다는 [NP 사실])"같이 구를 이루는 중심어들(도그마, 수정변경, 가능하 등)은 실질형태소들이 되고, '의', '이', '다는'등과 같은 형식형태소들로 구와 구 사이의 경계를 명확히 구분하면서 이들 사이의 구문적 관계를 연결하도록 한다.

둘째, 명사형 어미(-음, -기)나, 용언화 접미사(-하, -되)에 의한 파생어는 하나의 실질형태소로 취급하고 하나의 기본구로 묶는다. 즉, "~이(가) 가능+하+ㄴ"에서 "가능"을 명사구로 따로 묶기보다는 용언화 접미사 "-하"에 의해 파생된 "가능하"를 하나의 동사구로 묶는다.

셋째, 관용적 표현으로 굳어진 보조 용언구는 하나의 동사구로 묶는다. 예를 들면, "~을/라고 보+ㄹ 수 있+다"에서 "보+ㄹ 수 있"을 하나의 동사구로 묶는다.

넷째, 명사구가 관형격 조사 "-의"와 결합되어 관형어구를 이루고 이들이 다시 다른 명사구와 결합되어 명사구를 이루는 경우 각 명사구를 별도의 기본구로 취급한다. 예를 들어 "[NP [NP 도그마]의 [NP 수정변경]]"과 같은 경우 "[NP 도그마의 수정변경]는 "[NP 도그마]"를 하위구조로 갖기 때문에 "[NP 도그마]"과 "[NP 수정변경]"을 별도의 명사구로 인식한다.

다섯째, 단순하게 명사를 수식하는 관형어는 명사구로 묶어서 인식한다. 예를 들어 "[NP 그 사람]"이나 "[NP 대한 민국]의 [NP 뛰어난 운동선수들]"과 같은 경우가 이에 해당한다. 이때 "뛰어난"이 명사구로 함께 묶이는 것은 "뛰어난"이 명사를 수식하는 단순한 관형어로 사용되었기 때문이다.

여섯째, 관형절의 수식을 받는 명사구는 관형절로부터 분리하여 기본구를 형성한다. 예를 들어, "[[NP 도그마]의 [NP 수정 변경]이 [VP 가능하]]다는 [NP 사실]", 또는 "[[NP 기량]이 [VP 뛰어나]ㄴ [NP 운동 선수들]" 이라는 문장의 경우, "가능하"는 "+다는"과 "뛰어나"는 "+ㄴ"이라는 관형형어미와 연결되어 관형어를 형성하지만 이들은 "도그마의 수정 변경이 가능하+다는" 또는 "기량이 뛰어나+ㄴ" 이란 관형절의 하부구조이고, "가능하다는 사실", "뛰어난 운동선수들"은 겹쳐진 구조를 갖게 되므로 명사구로 묶일 수 없다. 그러므로 "사실", "운동 선수들"이 별도의 명사구로 분리된다.

일곱째, 구문적 표지 역할을 하는 조사와 연결/종결어미 및 기호는 어떠한 기본구에도 속하지 않는다.

3.2 기본구 인식의 정의

기본구 인식은 형태소 분석결과 품사태그가 부착된 문장의 각 형태소에 대해 적절한 기본구 태그를 할당하는 것으로 정의한다. 기본구 태그는 기본구에서의 각 형태소의 위치를 식별할 수 있도록 {B, I, O} 표기를 사용하며, 각 태그의 의미는 다음 [표 1]과 같다. [표 1]에서 XP는 명사구(NP), 동사구(VP), 부사구(AP), 독립어구(IP) 등을 포함한다.

표 1 기본구 태그의 표기법

B-XP	기본구 XP의 시작 위치에 있는 형태소
I-XP	기본구 XP의 내부와 마지막 위치에 있는 형태소
O	어떤 기본구에도 속하지 않는 형태소

이와 같은 기본구 인식에 대한 정의에 따라 학습집합은 문장을 구성하는 각 형태소마다 속성 값들과 목표값에 해당하는 기본구 태그를 갖도록 구성된다. 이때 속성은 형태소, 품사, 띄어쓰기 속성들이 포함되는데, 띄어쓰기 속성은 '0'(붙여쓰기)과 '1'(띄어쓰기)을 사용하여 속성 값을 표기한다. 다음의 예문 (4)~(5)는 기본구 인식에 대한 문제 정의에 따른 학습집합의 구성 예를 보여준다.

(4) "[AP 여기서] [NP 도그마]의 [NP 수정 변경]이 [VP 가능하]다는 [NP 사실], [AP 곧] [NP 비평기준]의 [NP 상대성]을 [VP 볼 수 있]다."

형태소	여기서	도그마	의	수정	변경	이	가능	하	다는	사실	,	곧
품사	AA	NNCG	PD	NNCV	NNCV	PS	NNCV	VX	EFD	NNCG	SS,	AA
띄어쓰기	1	0	1	1	0	1	0	0	1	0	1	1
기본구태그	B-AP	B-NP	O	B-NP	I-NP	O	B-VP	I-VP	O	B-NP	O	B-AP

형태소	비평	기준	의	상대성	을	보	ㄹ	수	있	다	.
품사	NNCV	NNCG	PO	NNCG	PO	VV	EFD	NNB	VJ	EFF	SS.
띄어쓰기	0	0	1	0	1	0	1	1	0	0	1
기본구태그	B-NP	I-NP	O	B-NP	O	B-VP	I-VP	I-VP	I-VP	O	O

(5) "[IP 그래서] [NP 칭기즈칸]은 [NP 뛰어난 기마 기술]로 [NP 인류 역사상 최대]의 [NP 제국]을 [VP 이루었]다."

형태소	그래서	칭기즈칸	은	뛰어나	ㄴ	기마	기술	로	인류	역사상	최대	의
품사	AC	NNP	PX	VV	EFD	NNCG	NNCG	PA	NNCG	NNCG	NNCG	PD
띄어쓰기	1	0	1	0	1	1	1	1	1	1	0	1
기본구태그	B-IP	B-NP	O	B-NP	I-NP	I-NP	I-NP	O	B-NP	I-NP	I-NP	O

형태소	제국	을	이루	었	다	.
품사	NNCG	PO	VV	EP	EFF	SS
띄어쓰기	0	1	0	0	0	1
기본구태그	B-NP	O	B-VP	I-VP	O	O

이제 한국어에서 기본구 인식은 각 형태소에 대한 기본구 태그의 부착 문제로 정의되었다. 그렇다면 문장을 구성하는 각 형태소에 적합한 기본구 태그를 어떻게 부착할 것인가? 이를 위해 본 연구에서는 기본구 태그가 부착된 대량의 학습 말뭉치로부터 기계학습을 수행하는데, 먼저 기본구 태그 결정에 영향을 미치는 자질들이 무엇인가를 분석하고, 선택된 자질들을 사용하여 학습을 수행함으로써 최대의 학습효과를 얻도록 할 것이다.

4. 한국어 기본구 인식을 위한 자질선택

자질선택 기능을 내포하지 않은 기계학습 방법이든 자질선택 기능을 내포한 기계학습 방법이든 학습 알고리즘에 제시될 초기 자질집합의 선택은 성능향상 측면에서 매우 중요한 문제중의 하나이다. 본 장에서는 학습 알고리즘에 최적인 자질집합을 선택하기 위해 학습집합으로부터 추출 가능한 모든 자질집합들을 대상으로 점증적 유용성(incremental usefulness)이란 관점에서 자질집합의 적합성(relevance of features)을 논의한다. 그리고 자료부족 문제를 완화하기 위해 한국어의 특성을 고려한 품사집합의 선택 및 선택적 어휘사용에 대해 논의한다.

4.1 점증적 유용성에 따른 자질집합의 선택

본 연구에서는 자질의 적합성을 "점증적 유용성(incremental usefulness)"이란 관점에서 정의하고, 이 정의에 따라 한국어 문장에서의 기본구 인식을 위한 자질집합을 선택한다. 그리고 자질의 적합성에 대한 정의를 좀더 구체화하기 위해 다음과 같은 표기법을 사용하기로 한다

지도 학습(supervised learning) 알고리즘에의 입력 자료는 N개의 학습 예제로 이루어진 집합 S라 한다. i번째 자질의 도메인을 F_i라 할 때 각 예제 X는 집합 $F_1 \times \cdots \times F_m$의 원소이고, 예제에 대한 목표 태그는 Y라 한다. 그러면, 학습예제는 <X, Y>의 쌍으로 표현된다. 이와 같은 표기와 함께 "점증적 유용성"이란 관점에서의 "적합한 자질"에 대한 정의는 다음과 같다.

Def. 학습예제집합 S, 학습 알고리즘 L, 그리고 자질공간 F가 주어졌을 때, 자질공간 $A \cup F$를 사용하여 L이 생성한 가정(Hypothesis)의 정확도가 F로부터 생성한 가정(Hypothesis)의 정확도보다 향상된다면, 자질공간 F에 대해 자질집합 A는 학습알고리즘 L에 점증적으로 유용하다고 말한다.

위에서의 "적합한 자질"에 대한 정의는 기본적 자질집합에서 출발하여 다른 자질집합을 추가 혹은 제거했을 때의 성능을 측정해 봄으로써 성능이 향상되는 방향으로 자질집합을 확장해 가는 방법이다. 이를 한국어 기본구 인식을 위한 자질집합 구성에 적용한다면 다음과 같다.

즉, 학습집합으로부터 획득 가능한 모든 경우의 단위 자질집합(하나의 속성으로 구성된 자질집합)을 생성하고 기본 자질집합에 차례로 단위 자질집합을 추가하면서 학습결과에 대한 성능을 분석해보는 것이다. 학습 말뭉치로부터 획득 가능한 자질집합들은 다음과 같고, 기본 자질집합은 현재 형태소의 품사 집합으로 한다.

- 현재 형태소의 품사(POS_0)
- 현재 형태소의 어휘(Lex_0)
- 현재 형태소의 띄어쓰기 정보(SP_0)
- 주변 형태소들의 어휘(Lex_{-m}, ... Lex_{-1}, Lex_{+1}, ... Lex_{+n})
- 주변 형태소들의 품사들(POS_{-m} ... POS_{-1}, POS_{+1}, ... POS_{+n})
- 주변 형태소들의 띄어쓰기 정보(SP_{-m}, ... SP_{-1}, SP_{+1}, ... SP_{+n})
- 주변 형태소들의 기본구 태그들(Cnk_{-m}, ... Cnk_{-1}, Cnk_{+1}, ... Cnk_{+n})

이때 함께 고려되어야 할 사항은 '어느 정도 크기의 문맥이 최적인가'이다. 본 논문에서는 현재 형태소를 중심으로 문맥을 왼쪽과 오른쪽으로 최소 0에서 최대 m(왼쪽)/n(오른쪽)까지를 고려하여 최적의 문맥크기를 결정하는데, 문맥을 최소 0에서 점차적으로 증가시켜 가면서 어느 문맥크기 이상에서 더 이상 성능 향상이 일어나지 않는다면 최대 성능 발휘 지점까지를 최적의 문맥크기로 결정한다.

4.2 품사체계와 선택적 어휘자질

학습 말뭉치를 사용하여 지도학습을 수행하는 경우 자료부족 문제는 피할 수 없는 문제중의 하나이다. 일반적으로 대부분의 연구들은 자료부족 문제를 완화하기 위해 어휘사용을 배제하거나[20, 21, 22] 선택적으로 어휘를 사용하고자 하였다[19]. 본 연구에선 선택적 어휘사용과 함께 품사체계의 선택이 자료부족 문제를 완화할 수 있다고 보고, 한국어의 통사적 특성과 의미적 특성을 고려할 때 품사체계 선택 및 선택적 어휘사용을 어떻게 할 것인가에 대해 알아본다.

먼저 품사체계의 경우, 품사의 세분화는 품사부착 시의 중의성을 가중시켜 학습말뭉치의 품사부착 일관성을 떨어뜨리거나 또는 자료부족 문제를 야기하기 쉽다는 문제

가 있다. 이는 품사를 자질로 사용하는 학습 알고리즘에 많은 영향을 미칠 수가 있다. 이에 다음의 사항을 고려하여 품사집합의 일반화와 세분화를 구분하고(부록 A참조), 품사집합의 세분화와 일반화에 따른 성능분석을 통해 기본구 인식에 적합한 품사집합을 선택하고자 한다.

첫째, 일반화와 세분화의 기준은 형태/통사적 특성이 다른 경우는 두 집합 모두에서 분명하게 분류하는 것으로 하고, 형태/통사적 특성이 동일하다 할지라도 의미적 특성이 다른 경우는 세분화한 것과 일반화한 것으로 나눈다.

둘째, 명사에서 자립명사의 경우, 통사적 특성이 일관적이지 않은 명사(예를 들면, '-하/-되' 등의 용언화 접미사와 결합하여 용언이 될 수도 있지만, 보통명사로도 사용되는 명사)에 대해서 하나의 품사로 일반화시킨 것과 세분화된 품사로 나눈 것, 두 집합으로 분류한다.

셋째, 용언의 경우, 동작이나 작용을 나타내는 것은 동사로, 성질이나 상태에 대해 서술하는 것은 형용사로 동일하게 나누고, 자립성이 결여된 보조용언에 대해서는 하나의 품사로 일반화한 것과 결합하는 본 용언에 따라 세분화시키는 것 두 집합으로 나눈다.

넷째, 수식언(관형사, 부사)은 통사적 특성이 분명히 드러나는 것은 두 집합 모두에서 분명하게 나누지만, 의미관계에 의해 세분화되는 것은 세분화된 품사집합에서만 나누는 것으로 한다.

다섯째, 형식형태소에서 어미의 경우, 부사전성 어휘는 대부분 표제어 사전에 등재되어 부사로 분석되는 경우가 많지만, 등재되지 않은 경우를 위해 세분화한 것과 세분화하지 않은 것으로 구분하기로 한다. 그 외의 경우는 일반화된 품사와 세분화된 품사체계가 동일하다.

여섯째, 접사의 경우는 명사형 접미사와 서술형 접미사의 경우 의미적 특성에 따라 세분한 것과 그렇지 않은 것으로 구분하고, 조사는 격의 특성에 따라 격조사를 세분화한 것과 일반화한 것으로 나눈다.

다음, 어휘는 의미적 기능을 담당하는 것으로 통사적 기능어휘와 밀착되어 문장을 구성한다. 그러므로 기본구 인식에 어휘를 사용할 수 있다면 정확도 향상에 도움이 될 것이며, 품사와 함께 사용한다면 의미적 혹은 통사적 세부 특성에 따라 품사를 세분화하는 작업도 필요하지 않을 것이다. 그러나 일반적으로 어휘사용은 자료부족 문제를 야기하기 쉽다.

다음의 [표 6]은 각 품사별로 자료부족 문제가 얼마나 심각한가를 보여준다. [표 6]에서 가장 자료부족 문제가 심각한 품사는 자립명사(12.6%)와 부사(4.6%), 그리고 용언(2.7%)이며, 형식형태소와 관형사, 의존명사 등은 상대적으로 자료부족문제가 덜 심각한 것으로 보인다. 이는 자료부족 문제가 심각한 자립명사, 부사 등에 대해서는 어휘를 사용하지 않고, 자료부족이 심하지 않은 품사들의 어휘를 사용하는 것이 어휘사용에 따른 자료부족 문제를 완화하는 방법이라는 것을 간접적으로 말해준다.

이에 본 연구에서는 실질형태소와 형식형태소의 어휘를 선택적으로 사용한 경우와 사용하지 않은 경우의 학습 알고리즘의 성능분석을 통해 최적의 선택적 어휘사용 방법을 선택하도록 한다.

표 6 테스트 집합의 자료부족 현상(학습집합의 형태소: 286,633개, 테스트 집합의 형태소 : 27,404개 자료부족률=학습집합에 없는 품사별 어휘의 수 * 100 / 테스트 집합의 품사별 어휘 수)

(단위:%)

	자립명사	의존명사	용언	조사	어미	접사	관형사	부사	기타
학습집합의 비율	29.4	3.0	10.6	18.7	18.8	6.3	1.5	3.0	8.8
테스트집합의 비율	27.2	3.1	12.4	18.8	19.8	6.0	1.5	3.1	8.0
자료부족률	12.6	1.4	2.7	0.4	0.2	0.8	0.2	4.6	0.8

5. 한국어 기본구 인식을 위한 기계학습 기법

한국어 기본구 인식을 위한 기계학습은 자질선택과 목표값 결정방법이 서로 다른 결정트리 학습(C4.5)[24]과 메모리기반 학습방법, 두 가지를 사용한다. 서로 다른 특성을 갖는 두 학습 알고리즘을 선택한 것은 한국어 기본구 인식을 위해 주어지는 자질집합과 학습 알고리즘과의 관계를 알아보고 초기 자질집합의 선택이 학습 알고리즘에 미치는 영향을 알아보기 위함이다.

5.1 결정 트리 학습(Decision Tree Learning)

결정 트리 학습은 귀납적 기계학습 방법 중의 하나로, 학습 자료들의 통계적 특성을 사용하여 오류 데이타에 견고한 결정함수를 만드는데, 주어진 자질집합을 바탕으로 뿌리노드에서부터 시작하여 생성 가능한 모든 결정트리들의 공간을 탐색하여 최적의 결정트리로 성장시켜

간다. 이때 주어진 자질집합에 따라 생성되는 결정트리의 학습 성능이 달라지게 된다.

즉, 첫 번째 탐색 단계에서 주어진 자질들 가운데 학습 자료들을 가장 잘 판별할 수 있는 최적 자질을 선택하고 이를 뿌리노드에 명시한다. 다음, 선택된 최적 자질의 가능한 값에 따라 자식 노드들을 만들고 학습 자료들을 분할하여 적절한 자식노드에 할당한다. 첫 번째와 두 번째 과정의 작업이 각 자식 노드에서 반복적으로 수행되며, 마지막 단말 노드에는 최적의 목표 값(즉, 기본구 태그)이 명시된다. 이때 각 노드에서의 최적 자질은 초기에 주어진 자질집합의 범위에서 결정되며, 각 자질이 제공하는 정보량을 측정한 결과 가장 많은 정보를 제공하는 자질로 선택한다. 정보량은 자질에 의한 정보이득량[1]을 자질에 의한 분할정보량[2]으로 나눈 정보이득률(Information Gain Ratio)을 사용하여 측정한다(식 5.1)[24]. 식 (5.1)에서 $H(S)$와 $H_{F_i}(S)$는 각각 자질 F_i가 주어지지 않았을 때와 주어졌을 때의 목표 값에 대한 학습집합 S의 엔트로피를[3] 의미하고, $SI_{F_i}(S)$는 자질 F_i에 의한 집합 S의 분할정보량을 의미한다.

$$IGR(S,F_i) = \frac{H(S)-H_{F_i}(S)}{SI_{F_i}(S)} = \frac{H(S)-\sum_{f_v \in F_i} \Pr(f_v)H(S_{f_v})}{-\sum_{f_v \in F_i} \Pr(f_v)\log \Pr(f_v)} \quad (5.1)$$

그런데, 이와 같은 결정트리 학습은 주어진 학습집합에 과적응(overfitting)되기 쉽다는 문제가 있다. 특히 언어처리의 특성상 자질 값들의 집합이 매우 큰 자질(예를 들면 어휘정보의 사용)이 사용될 수 있는데, 이 경우 학습집합이 자질 값에 의해 매우 작은 집합으로 분할되기 쉽고, 그 결과 분할된 학습집합에서 신뢰할 만한 결과를 얻기가 어렵게 된다. 또한 많은 자질을 사용하다 보면 결국에는 분할된 학습집합이 매우 작아져 학습 결과에 대한 신뢰도가 저하되게 된다.

이러한 문제를 해결하기 위해 다음의 두 가지 방법을 사용한다. 첫 번째 방법은 결정트리를 성장시켜 가는 과정에서는 분할된 학습집합의 크기가 항상 유효값 이상이 되도록 하면서 최적 속성을 선택하는 것이다. 두 번째 방법은 완전한 결정트리가 만들어지고 난 뒤 가지치기(pruning)를 통해 결정트리가 일반적인 예제들에 대해서도 잘 처리할 수 있도록 만드는 것이다. 이때 가지치기는 각 자질 값에 대해 가지치기 전후에서 발생 가능한 오류들을 추정한 결과, 가지치기를 수행한 결과가 오류를 적게 낸다면 가지치기를 수행하여 해당 속성을 고려하지 않도록 한다.

[그림 1]은 중심어와 좌우 각 3개의 형태소들에 대한 어휘(Lex), 품사(Pos), 기본구 태그(Cnk)를 자질로 하여 결정트리를 학습하고 가지치기를 수행한 결과를 보여준다. 뿌리노드에 있는 Cnk_{-1}(중심어 바로 왼쪽 형태소의 기본구 태그)가 최적 자질이며, 뿌리노드에서부터 차례로 자질값이 평가되면서 다음 자질 POS_0(중심어의 품사)가 선택되어 분기되고 종국엔 단말노드에 이르러 목표 값인 기본구 태그가 결정되도록 되어있다. 이때 초기 자질집합에 기본구 태그나 어휘가 포함되지 않는다면 최적 자질의 선택은 달라질 것이고, 그 결과 결정트리는 다른 모습으로 생성되어 있을 것이다.

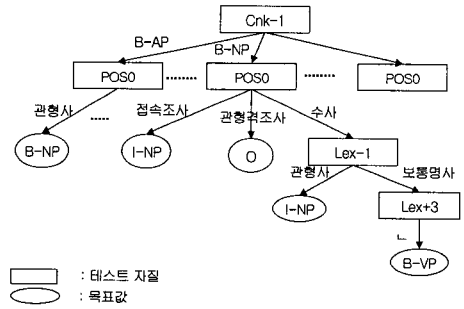


그림 1 결정트리의 학습결과: 사용자질은 중심어, 좌우 각 3개 형태소의 어휘(Lex), 품사(POS), 기본구 태그(Cnk)

이와 같이 결정트리 학습 알고리즘은 정보이득률이라는 평가함수를 이용하여 주어진 자질집합으로부터 기본구 인식에 적합한 자질들만을 선택하고, 오류 가능성이 높은 가지를 잘라냄으로써 오류 데이타와 미지의 데이타에 대해 견고하게 처리할 수 있다는 장점을 갖는다. 그러나, 이 모든 결과는 초기에 주어지는 자질집합에 종속되어 있다. 그러므로 초기 자질집합의 선택은 결정트리의 학습성능을 좌우하는 제일 요인이라고 할 수 있을 것이다.

5.2 메모리 기반 학습(Memory - based Learning)

메모리기반 학습은 학습할 때는 학습예제들을 메모리에 그대로 저장해 두었다가, 새로운 예제가 들어오면 학

1) 정보 이득량은 자질에 의해 학습집합의 혼잡도가 얼마나 감소했는가를 말한다.

2) 분할 정보량은 자질에 의해 학습집합이 얼마나 혼잡해지는가의 정도를 말한다.

3) $H(S_{f_v})$는 자질값 f_v에 의해 분할된 집합 S_{f_v}의 엔트로피를 의미한다.

습예제들과의 유사성을 평가하여 추론하는 학습방법으로, 전체 학습집합에 대해 한번에 결정함수(Target Function)를 추정하는 것이 아니라, 새로운 예제가 주어졌을 때만 국부적으로 목적함수를 추정하는 방법이다.

메모리기반 학습 알고리즘에서 가장 중요한 점은 새로운 예제 X_{new}와 학습집합 내의 모든 예제들 X_S와의 유사도를 측정하는 방법이고, 일반적으로 유사도는 식 (5.2)[25]와 같이 단순한 거리함수 $\triangle(X_{new}, X_S)$를 사용하여 측정된다. 그러나 본 연구에서는 패턴 X_{new}와 X_S 사이의 단순한 거리가 아니라, 자질의 적합성을 정보이득률(식 (5.1))로 구하고 이를 가중치로 사용하는 가중적 거리측정 함수를 사용한다(식 (5.3)[26]). 이는 자질마다 목표 값 결정에 공헌하는 정도의 차이가 있다고 보고, 이 공헌도를 가중치로 반영하기 위함이다.

$$\triangle(X_{new}, X_S) = \sum_{i=1}^{m} \delta(x_{new_i}, x_{s_i}) \quad (5.2)$$

$$where\ \delta(x_{new_i}, x_{s_i}) = \begin{cases} 0 & \text{if } x_{new_i} = x_{s_i} \\ 1 & \text{if } x_{new_i} \neq x_{s_i} \end{cases}$$

$$\triangle(X_{new}, X_S) = \sum_{i=1}^{m} w_i \delta(x_{new_i}, x_{s_i}) \quad (5.3)$$

$$= \sum_{i=1}^{m} IGR(S, F_i)\delta(x_{new_i}, x_{s_i})$$

[표 7]은 메모리 기반 학습이 어떻게 적용되는가를 보여 준다. 학습과정에서 측정된 자질들의 가중치(IGR)는 학습예제들과 함께 메모리에 저장되어 있고, 테스트 자료가 주어지면 학습예제들과의 유사도(△)를 측정하여 우선순위가 높은 k개의 후보들을 탐색한 뒤, 기본구 태그를 결정한다. [표 7]에서는 학습에 사용된 자질들의 가중치(IGR) 및 주어진 테스트 자료 "침묵/NC"에 대해 유사도를 평가한 결과, 우선 순위가 높은 순으로 찾아진 후보들을 차례로 보여주고 있다.

이러한 메모리 기반 학습 알고리즘은 결정트리 학습과는 달리 자체적인 자질선택 알고리즘을 갖지 않는다. 대신 주어진 자질들에 가중치를 부여하고, 이들을 최대한으로 이용한다는 특징을 갖는다. 또한 결정트리는 뿌리노드로부터 차례로 자질들을 검사하여 중간 노드의 모든 자질들이 만족하고 단말노드에 이르러야만 목표값을 결정할 수 있는 반면에, 메모리 기반 학습은 가중치가 높은 자질 중의 하나가 일치하지 않는다 할지라도 다른 자질들을 참조하여 목표 값을 결정할 수 있다는 차이가 있다. 그러나 이러한 차이에도 불구하고 이 방법 역시 결정트리와 마찬가지로 초기에 선택된 자질집합의 범위에서 학습자료가 구성되므로, 유사도 평가는 주어진 자질집합에 종속될 수밖에 없고, 자질집합을 어떻게 구성할 것인가가 성능 향상의 중요 요인으로 부각되게 된다.

6. 실험 및 평가

한국어 기본구 인식을 위한 실험은 크게 두 단계로 이루어졌다. 첫 번째 단계는 한국어 기본구 인식에 적합한 자질집합 중 주어진 학습 알고리즘에서 최고의 성능을 발휘할 수 있는 자질집합을 선택하기 위한 실험들이며, 두 번째 단계는 선택된 자질집합을 사용하여 한국어 기본구 인식을 학습하는 것이다.

6.1 실험환경

실험은 한국과학기술원에서 배포한 국어정보베이스[27]에 포함된 트리 태그 부착 말뭉치를 기본구 태그 부착 말뭉치로 변형하여 사용하였으며, 형용사구(ADJP)와 동사구(VP)는 모두 동사구로 통일하여 사용하였다. 또한 품사의 세분화 정도가 한국어 기본구 인식에 미치는 영향을 알아보기 위해 세분화된 품사집합과 일반화된 품사집합을 사용하여(부록 A.참조) 학습말뭉치를 별도로 구성하였다. 사용된 말뭉치는 모두 10만 여 개의 비재귀적 기본구로 구성되어 있는데, 학습을 위해 96,993개, 평가를 위해 9,385개의 비재귀적 기본구를 사용하였다. 각 비 재귀적 기본구의 분포는 다음 [표 8]과 같다.

표 7 메모리 기반 학습의 한 예

"그런데/AJ/B-IP 이러하/PA/B-NPㄴ/EM/I-NP 침묵/NC 이/JC 어느/MM 정도/NC..."

자질	POS$_0$	Le x$_0$	Le x$_{-3}$	POS$_{-3}$	Cn k$_{-3}$	Le x$_{-2}$	POS$_{-2}$	Cn k$_{-2}$	Le x$_{-1}$	POS$_{-1}$	Cn k$_{-1}$	Le x$_{+1}$	POS$_{+1}$	Le x$_{+2}$	POS$_{+2}$	Le x$_{+3}$	POS$_{+3}$	△
IGR	1	0.312	0.398	0.040	0.042	0.086	0.097	0.133	0.217	0.273	0.421	0.188	0.241	0.088	0.108	0.043	0.049	
자료	NC	NC	AJ	AJ	B-IP	PA	PA	B-NP	ㄴ	EM	I-NP	이	JC	MM	MM	NC	NC	
후보1	NC	NC	어	JM	O	PA	PA	B-NP	ㄴ	EM	I-NP	이	JC	MM	MM	NC	NC	.124
후보2	NC	NC	어	NB	I-NP	PA	PA	B-NP	ㄴ	EM	I-NP	이	JC	MM	MM	NC	NC	.124
후보3	NC	NC	AJ	AJ	B-IP	PA	PA	B-NP	ㄴ	EM	I-NP	어른	JC	MM	MM	NC	NC	.188
후보4	NC	NC	AJ	AJ	B-IP	PA	PA	B-NP	ㄴ	EM	I-NP	이	JC	NC	NC	풍	JC	.289

표 8 학습말뭉치와 평가말뭉치의 비재귀적 기본구의 분포

	명사구	동사구	부사구	독립언구	합계
학습말뭉치	56,554	32,833	5,501	2,015	96,993
평가말뭉치	5,254	3,370	525	236	9,385

평가는 정확도, 정확률, 재현률, F-평가를 사용하였으며, 각 평가 척도는 다음과 같다.

$$\text{태그정확도}(\%) = \frac{\text{정확한 기본구 태그의 수}}{\text{인식한 기본구 태그의 수}} \times 100(\%) \quad (6.1)$$

$$\text{정확률}(\%) = \frac{\text{정확한 비재귀 기본구의 수}}{\text{인식한 비재귀 기본구의 수}} \times 100(\%) \quad (6.2)$$

$$\text{재현률}(\%) = \frac{\text{정확한 비재귀 기본구의 수}}{\text{인식해야 할 비재귀 기본구의 수}} \times 100(\%) \quad (6.3)$$

$$F_{\beta=1} = \frac{(\beta^2+1)\times \text{재현률} \times \text{정확률}}{\beta^2 \times \text{재현률} + \text{정확률}} \quad (6.4)$$

6.2 결정트리의 가지치기를 위한 신뢰도 결정

최적의 자질을 선택하기 전에, C4.5 결정트리 학습에서 가지치기를 수행할 때 신뢰도를 얼마로 할 것인가에 대해 알아보았다. 예비 실험결과 가지치기에서 사용하는 신뢰도는 결정트리의 성능에 영향을 미치는 것으로 나타났다. [표 9]는 가지치기 수행시 사용된 신뢰도에 따른 성능을 보여주는데, 세분화된 품사집합(L)과 일반화된 품사집합(S)으로 구성된 학습집합으로부터 동일한 자질들을 선택하고 결정트리의 신뢰도만을 변경하며 실험한 결과이다. [표 9]에 따르면 신뢰도가 75%일 때 모두 가장 좋은 성능을 보여주고 있다. 그러므로 향후 결정트리 관련 실험에서는 75%의 신뢰도로 가지치기를 수행하고 평가하는 것으로 한다.

표 9 결정트리에서 가지치기 시 사용한 신뢰도별 성능

신뢰도(%)		25	50	60	65	70	75	80	85	90
정확도 (%)	S	96.00	96.39	96.39	96.40	96.55	96.54	96.52	96.51	96.52
	L	96.00	96.20	96.25	96.38	96.45	96.60	96.50	96.47	96.45
$F_{\beta=1}$	S	92.59	93.06	93.07	93.12	93.24	93.28	93.14	93.07	93.04
	L	92.51	92.99	93.12	93.26	93.31	93.40	93.26	93.14	93.12

6.3 자질 집합의 선택

한국어 기본구 인식을 위한 자질집합 선택 실험은 다음과 같이 이루어졌다.

첫째, 가능한 자질들의 조합으로 만들어진 자질집합을 사용하여 학습한 결과들의 성능을 비교 분석하고 가장 좋은 성능을 내는 자질들의 조합을 선택한다.

둘째, 문맥의 크기를 얼마로 해야 성능이 최적이 되는가를 판단하기 위해 문맥 크기의 변화에 따른 학습결과들의 성능을 분석하고, 또한 각 자질들 사이의 우선순위를 자질들의 문맥위치와 관련하여 분석한다.

셋째, 한국어 기본구 인식에서 품사분류 체계의 세분화 정도가 성능에 영향을 미치는지를 분석한다.

6.3.1 최적의 자질유형 조합

하나의 형태소에 대해, 가능한 자질들(품사, 어휘, 띄어쓰기, 기본구 태그)의 조합을 만들면 총 15 가지(2^4-1)가 가능하다. 이 가운데, 품사는 기반자질(base feature)이라고 가정하고 항상 포함되도록 한다면 자질조합 결과 생성된 자질집합은 8($=2^3$)가지로 축소된다.

[그림 2]와 [그림 3]은 특정한 문맥크기(좌3, 우3)에서 학습 알고리즘들에 모든 가능한 자질집합들을 적용했을 때의 성능을 보여준다. 그림들에서 볼 때, 한국어 기본구 인식에서 유용한 자질은 특정 학습알고리즘과 무관하고, 중요자질은 품사와, 어휘 그리고 기본구 태그이며, 띄어쓰기 자질은 다른 자질에 비해 유용성이 낮은 것으로 보여진다.

주어진 크기의 문맥에서 상대적으로 높은 성능을 보여주는 결정트리 학습결과를 자세히 살펴보면 다음과 같다. 결정트리에서 띄어쓰기 자질은 품사자질만을 사용했을 때에 비해 약간의 성능 향상을 가져다 주지만, 기본구 태그나 어휘자질에 의한 성능향상에 비하면 상당

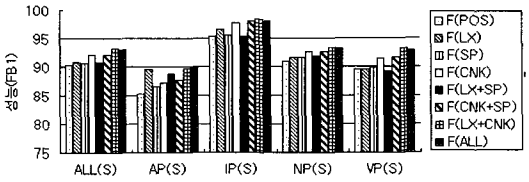


그림 2 사용된 자질집합에 따른 성능 비교(결정트리 학습기법): POS는 품사를, LX는 어휘를, SP는 띄어쓰기 정보를, CNK는 이미 결정된 기본구 태그를 의미하며, 사용된 문맥은 중심어와 좌우 각 3개의 형태소들임

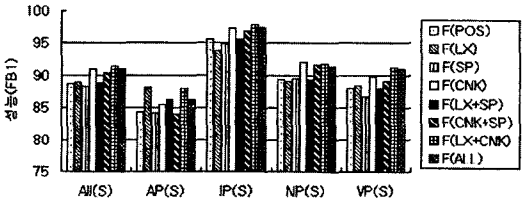


그림 3 사용된 자질집합에 따른 성능 비교(메모리기반 학습기법)

히 미미하다. 뿐만 아니라 품사, 어휘 그리고 띄어쓰기 자질을 함께 사용하는 것은 품사와 어휘자질만을 사용하는 것보다 낮거나 유사한 성능을 보여주는데, 이는 띄어쓰기 자질이 한국어 기본구 인식에 있어서 유용한 자질이 아니라는 것을 다시 한번 확인시켜준다.

이에 반해 이미 결정된 기본구 태그는 성능향상에 기여하는 바가 큰데, 특히 동사구와 명사구에 대해 각각 $F_{\beta=1}$ 값이 +1.88(18.2%)/+1.60(17.96%)의 성능향상을 보였다. 또한 어휘자질의 사용은 독립어구와 부사구에 대해 +1.25(27.2%)/+4.32(29.2%)라는 높은 성능 향상을 보여 주었다.

더 나아가 어휘와 기본구 태그 자질을 품사자질과 결합하여 사용한 경우 가장 좋은 성능을 보여주었는데, 품사자질만 사용하는 경우에 비해 전체적으로는 +2.96(30.6%)의 성능 향상이, 명사구와 동사구에 대해서는 각각 +2.32(26%), +3.67(35.5%)의 성능 향상이 있었다. 이러한 결과로부터 부사구의 인식에는 어휘자질이 중요한 역할을 하고 있고, 명사구나 동사구 인식에는 기본구 태그가 중요한 역할을 하는데, 이는 기본구 태그가 문형이나 동사의 하위범주 정보를 내포하고 있기 때문인 것으로 생각된다.

한편, 어휘자질은 한국어 기본구 인식에 있어서 유용한 자질로 사용될 수 있지만, 또한 심각한 자료부족 문제를 야기하기 쉽다. 그러므로 자료부족 문제를 완화시키면서 어휘를 적절히 사용할 수 있는 방안이 모색되어야 한다.

[그림 4]는 어휘자질을 사용하지 않았을 때, 선택적으로 사용했을 때, 모든 어휘를 사용했을 때의 성능을 보여주는데, 선택적 어휘사용에서 어휘사용이 배제된 품사들의 어휘 자질값은 품사로 대체하여 사용한 결과이다. 그림에서 확실한 것은 어휘자질을 사용하는 것이 사용하지 않는 것보다 성능 향상에 더 많은 도움이 된다는 것이다. 그러나, 실험결과는 또한 모든 어휘를 사용하는 것이 선택적으로 어휘를 사용하는 것보다 훨씬 더 결과가 좋은 것은 아니라는 것도 보여준다. 이는 모든 어휘

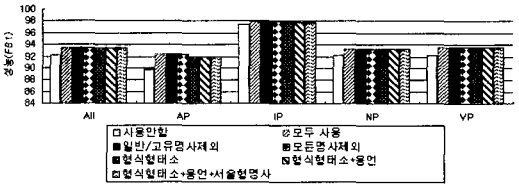


그림 4 선택적으로 어휘를 사용함에 따른 성능의 비교 (결정트리): 문맥은 중심어와 좌우 각 3개 형태소의 품사, 어휘, 기본구 태그를 사용

를 사용하거나, 자료부족 문제가 심각한 품사의 어휘를 사용하는 것이 자료부족 문제를 야기했기 때문이라고도 해석할 수 있다. 그러므로 어휘 사용에 따른 자료부족 문제를 완화하는 방안으로 어휘는 형식형태소나 자료부족이 심각하지 않은 품사의 어휘에 대해서만 사용하고 그 외는 품사로 대체하여 사용하는 것이 의미있다고 할 수 있을 것이다.

6.3.2 문맥의 크기

하나의 형태소를 기준으로 볼 때, 품사, 어휘 및 기본구 태그 자질들이 기본구 인식에 있어서 유용하다는 것이 실험적 결과로 밝혀졌다. 다음의 [표 10]은 품사, 어휘, 및 기본구 태그를 자질로 고정한 상태에서 문맥의 크기 변화에 따른 성능을 보여준다.

[표 10]에 따르면 문맥은 중심어의 어느 한쪽 문맥만을 보는 것보다는 좌우 문맥을 모두 보는 것이 성능 향상에 훨씬 더 좋은 영향을 미치고, 성능에 영향을 미치는 문맥의 범위는 제한적이라는 것을 알 수 있다. 즉, 결정트리로 학습한 결과는 좌우 각 3개의 형태소를 포함했을 때 성능이 가장 좋고, 문맥을 더 확장하면 성능은 오히려 저하되는 경향이 있다. 메모리기반 학습방법의 경우도, 좌우 각 2개의 형태소를 보는 경우 성능이 가장 좋고, 문맥을 확장하면 성능이 저하되는 경향이 있음을 볼 수 있다. 이는 더 많은 문맥을 보는 것이 때로는 혼잡도를 가중시켜 판별능력을 떨어뜨렸기 때문인 것으로 판단된다.

표 10 일반화된 품사집합에서 문맥크기의 변화에 따른 성능($F_{\beta=1}$) (결정트리 학습/메모리기반 학습)

우 \ 좌	0	1	2	3	4	5
0		63.64/63.81	66.20/66.09	66.23/65.17	67.00/66.85	67.55/67.35
1	76.83/75.37	88.28/88.14	92.35/91.70	92.85/91.28	92.76/90.61	92.81/90.19
2	76.84/76.14	88.59/88.18	92.75/91.75	93.01/91.67	93.03/91.21	92.96/90.92
3	77.01/75.77	88.52/87.50	93.01/91.20	93.28/91.51	93.16/91.35	93.01/91.10
4	76.66/74.89	88.34/87.00	92.74/90.78	93.16/91.20	93.12/90.94	93.02/90.86
5	76.63/72.99	88.20/86.44	92.71/90.19	92.96/90.74	93.07/90.51	93.06/90.74

표 11 정보이득률과 카이제곱분포에 따른 각 자질의 우선순위

순위	정보이득률	x^2-분포	순위	정보이득률	x^2-분포	순위	정보이득률	x^2-분포
1	cnk-1	pos0	10	pos-2	lex+2	19	lex+4	lex+5
2	pos0	lex0	11	lex-2	pos-2	20	cnk-4	pos+4
3	lex0	cnk-1	12	lex+2	pos+2	21	pos+5	lex-4
4	pos-1	lex-1	13	pos+3	lex-3	22	lex+5	pos+5
5	pos+1	pos-1	14	cnk-3	lex+3	23	cnk-5	pos-4
6	lex-1	pos+1	15	lex+3	cnk-3	24	pos-4	lex-5
7	lex+1	lex+1	16	pos-3	lex+4	25	lex-4	cnk-4
8	cnk-2	lex-2	17	lex-3	pos-3	26	pos-5	pos-5
9	pos+2	cnk-2	18	pos+4	pos3	27	lex-5	cnk-5

또한, 문맥의 각 위치에 있는 자질들이 기본구 태그를 결정하는데 제공하는 정보이득률과 카이제곱 분포를 측정하고 우선순위를 평가한 결과 [표 11]과 같았다. [표 11]에서 정보이득률이 가장 높은 자질은 중심어 바로 왼쪽 형태소의 기본구 태그, 중심어의 품사, 중심어의 어휘 등이며, 중심어에서 멀어질수록 정보이득률이 낮게 측정되었다. 카이제곱 분포의 경우도 중심어에 대한 자질들이 가장 높고, 중심어에서 멀어질수록 자질들의 우선순위가 낮은 것으로 나타나, 정보이득률이나 카이제곱 분포 모두 유사한 자질 우선순위를 보여주었다.

[표 10]의 문맥크기에 따른 성능과 [표 11]의 자질들의 우선순위에 대한 실험결과로부터 중심어에 가까이 있는 자질일수록 많은 정보를 제공하고 목표값과 상관관계가 높고, 그 결과는 성능에 직결된다고 결론지을 수 있을 것이다.

6.3.3 품사체계의 세분화

본 실험에서는 한국어 기본구 인식에서 품사분류 체계의 세분화 정도가 성능에 어느 정도 영향을 미치는지를 분석하였다. 결정트리 학습과 메모리 기반 학습기법을 적용하여 실험한 결과는 각각 [그림 5] [그림 6]과 같다. 그림에서 어휘자질을 사용한 경우는 자료부족 문제가 심각한 자립명사는 해당 품사로 대체하고 그 외는 어휘를 그대로 사용한 결과이다.

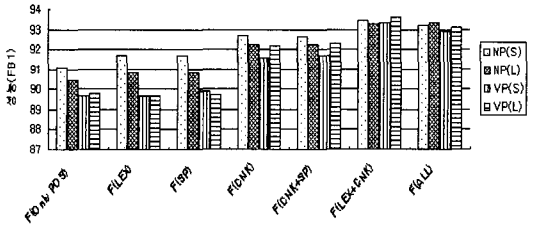


그림 5 품사집합의 크기에 따른 성능(결정트리 학습)

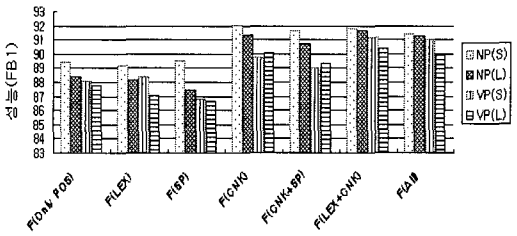


그림 6 품사집합의 크기에 따른 성능(메모리기반 학습)

실험결과를 자세히 살펴보면, 결정트리 학습결과는 명사구와 동사구가 서로 다른 특징을 보여주고 있고, 메모리기반 학습 결과는 명사구나 동사구 모두 품사를 세분화하는 것보다는 일반화한 것이 더 좋은 성능을 보여주고 있다.

결정트리에서 명사구와 동사구가 품사세분화에 대해 서로 다른 결과를 보여주는 것은 통계정보를 사용하여 자질선택을 한다는 점과, 한국어의 파생어 생산성과 관련된 것으로 분석되었다. 즉, 명사구에선 일반화된 품사가 자료부족 문제를 완화시켜줌으로써 안정된 통계정보를 획득하도록 해주었기 때문에 더 좋은 성능을 얻은 것으로 보인다.

이에 반해 동사구는 명사의 세분화(일반명사와 동사성/형용사성 명사로 세분)가 '-하'/'-되' 등의 용언화 접미사와의 결합가능성을 잘 분별할 수 있도록 했고, 또한 부사의 세분화가 동사구를 식별하는데 많은 도움을 주었기 때문에 세분화된 품사가 더 좋은 성능을 보여주었다. 이러한 차이는 기본구 태그를 사용하는 경우 완화되는 경향을 보였는데, 이는 자질값 집합의 크기가 작아 안정된 통계정보를 제공하는 기본구 태그자질이 품사보다 더 중요한 자질로 선택되고 기본구 태그 값을 중심으로 그 다음 자질이 선택되었기 때문이었다.

또한 자질선택을 하지 않는 메모리기반 학습에서 일반

화된 품사집합이 전반적으로 더 좋은 성능을 보이는 것은 품사의 세분화가 기본구 인식에 많은 영향을 미치지 않으며, 품사의 세분화가 필요한 경우는 어휘가 그 역할을 담당해 주기 때문이었다. 단 메모리기반 학습에 있어서도 동사구는 어휘를 사용하지 않았을 때 세분화된 품사집합이 때 더 좋은 성능을 보였는데, 이는 결정트리에서와 마찬가지로 명사로부터 파생된 용언들 때문이었다.

그러므로 파생어에 대해 전처리를 수행하고 파생되는 예를 없앤다면 일반화된 품사를 사용하는 것이 비용과 성능 면에서 더 좋은 결과를 얻을 수 있을 것이라 예측된다.

6.4 선택된 자질집합을 사용한 한국어 기본구 인식 결과

최적의 자질집합을 선택하기 위한 예비실험 결과, 자질은 품사, 선택적 어휘, 그리고, 기본구 태그를 사용하면서, 문맥 크기는 결정트리의 경우 중심어와 좌/우 각각 3개의 형태소를, 메모리기반 학습의 경우 문맥의 크기를 좌/우 각각 2개의 형태소를 보는 것이 좋은 것으로 밝혀졌다. 또한 품사집합은 너무 세분되지 않은 일반화된 품사집합을 사용하되 파생용언에 대해 전처리를 수행하는 것이, 어휘는 모든 어휘를 사용하기보다는 자료부족 문제가 심각한 명사를 제외하고 선택적으로 어휘를 사용하는 것이 더 좋은 것으로 평가되었다.

선택된 자질집합과 함께 결정트리 학습 및 메모리기반 학습을 적용한 결과, 한국어 기본구 인식결과는 [표 12]와 같다. 이때 어휘는 자립명사의 어휘를 배제한 어휘들을 사용하였으며, 품사는 일반화된 품사집합을 사용하였다. [표 12]에 따르면, 결정트리 학습결과가 메모리기반 학습결과보다 더 좋은 성능을 나타내고 있음을 볼 수 있다. 이는 결정트리 학습의 자질선택 능력이 잉여의 자질을 걸러내고 최적의 자질만을 선택해 냈기 때문인 것으로 판단된다.

또한 [표 13]은 자질집합선택을 거쳐 기계학습을 수행한 결과 중 한국어의 명사구 인식결과만을 기존연구와 비교한 결과를 보여주는데, 자질집합선택을 통해 메모리기반학습 기법을 사용하거나 결정트리 학습기법을 사용한 본 연구결과가 기존의 변형기반 학습기법이나 통계적 방법을 사용한 것에 비해 더 좋은 성능을 보이는 것을 볼 수 있다.

표 12 선택된 자질집합과 함께 학습한 한국어 기본구 인식결과

	메모리기반 (태그정확도:97.86%)			결정트리 (태그정확도:96.54%)		
	정확률	재현율	$F_{\beta=1}$	정확률	재현율	$F_{\beta=1}$
All	90.99%	92.52%	91.75	93.39%	93.41%	93.40
AP	87.20%	90.86%	88.99	90.26%	93.52%	91.86
IP	97.10%	99.15%	98.11	96.30%	99.15%	97.70
NP	91.88%	92.84%	92.36	93.58%	92.92%	93.25
VP	89.82%	91.87%	90.83	93.41%	93.80%	93.60

표 13 기존 연구와의 비교(명사구 인식)

	정확률	재현율	$F_{\beta=1}$
양재형(2000)	91.80%	90.70%	91.25
이신목(2001)	92.55%	90.90%	91.71
본 연구(메모리기반)	91.88%	92.84%	92.36
본 연구(결정트리)	93.58%	92.92%	93.25

7. 결 론

본 논문에서는 한국어 기본구 인식에 유용한 자질들이 무엇인가를 알아보기 위해 자질의 적합성을 "점증적 유용성"이란 관점에서 정의하였다. 또한 비재귀적 한국어 기본구 인식에 유용한 자질들을 밝혀내고, 각 자질이 특정한 방법에만 유용한 것이 아니라 기계학습을 사용하는 다른 방법들에서도 유용한가 여부를 밝히기 위해 서로 다른 특성을 갖는 기계 학습기법들을 선택하여 체계적인 실험을 수행하였다.

실험결과, 한국어 기본구 인식에 유용한 자질의 유형은 품사, 어휘, 기본구 태그로 학습 알고리즘에 관계없이 동일하였다. 그리고 문맥크기는 학습 알고리즘에 따라 약간 차이가 있지만, 좌우 문맥을 보되 중심어 주변의 제한된 범위의 문맥이 유용한 것으로 밝혀졌다. 또한 품사집합은 너무 세분화된 것보다는 일반화된 품사집합을 사용하고, 어휘는 자료부족문제가 심각하지 않은 어휘를 선택적으로 사용하는 것이 자료부족문제를 완화하면서 안정된 성능을 얻을 수 있는 것으로 평가되었다. 평가결과에 따라 선택된 최적의 자질집합을 사용하여 결정트리 학습과 메모리기반 학습을 수행하고, 한국어 기본구 인식에 적용한 결과, 각각 93.39%/93.41%과 90.99%/92.52%의 정확률/재현율을 얻었다. 또한, 이 중 명사구 인식결과만을 기존연구와 비교한 결과, 자질집합 선택을 통해 메모리 기반학습 기법을 사용하거나 결정트리 학습기법을 사용한 결과가 기존의 변형기반 학습기법이나 통계적 방법을 사용한 것에 비해 더 좋은 성능을 보여주었다.

향후 연구로는 본 연구를 통해 한국어 기본구 인식에 유용하다고 평가된 자질들이 SVM이나 다른 학습기법들에서도 동일하게 유용한지를 평가해 볼 예정이다. 그리고, 최적의 자질집합을 가장 성능이 우수한 방법에 적용하여 기본구 인식뿐만 아니라 어절간의 의존관계 분석 및 얕은 수준의 의미구조 분석으로까지 확장할 것이다.

참 고 문 헌

[1] S. Abney, "Parsing by Chunks," In R.C. Berwick, S.P. Abney and C. Tenny, editors, Principle-Based Parsing: Computation and Psycholinguistics, Kluwer, pp. 257-278, 1991.

[2] S. Abney, "Partial Parsing via. Finite-State Cascades," In Proc. of the ESSLLI '96 Robust Parsing Workshop, 1996.

[3] Gregory Grefenstette, "Light parsing as Finite State Filtering". In Proc. of the Workshop on Extended Finite State Models of Language, ECAI'96, 1996.

[4] K. W. Church, "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," In Proc. of the 2nd Conf. On Applied NLP, 1988.

[5] GuoDong ZHOU and Jian SU, "Error-Driven HMM-based Chunk Tagger with Context-Dependent Lexicon," In Proc. of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 2000.

[6] W. Skut and T. Brants, "A Maximum-Entropy Partial Parser for Unrestricted Text," In Proc. of the 6th Workshop on Very Large Corpora., 1998.

[7] Rob Koeling, "Chunking with Maximum Entropy Models," In Proc. of CoNLL-2000 and LLL-2000, pp. 139-141, 2000.

[8] L.A. Ramshaw and M.P. Marcus, "Text Chunking using Transformation-Based Learning," In Proc. of the 3rd ACL workshop on Very Large Corpora, 1995.

[9] Claire Cardie and David Pierce, "Error-Driven Pruning of Treebank Grammars for Base Noun Phrase Identification," In Proc. of COLING-ACL'98, pp. 218-224, 1998.

[10] Claire Cardie and David Pierce, "The Role of Lexicalization and Pruning for Base Noun Phrase Grammars," In Proc. of the 6th National Conference on Artificial Intelligence, 1999.

[11] Shlomo Argamon, Ido Dagan, and Yuval Krymolowski, "A Memory-Based Approach to Learning Shallow Natural Language Patterns," In Proc. of COLING-ACL'98, pp. 67-73, 1998.

[12] J. Veenstra, "Fast NP Chunking Using Memory-Based Learning Techniques," In Proc. of the 8th Belgian-Dutch Conference on Machine Learning, 1998.

[13] W. Daelemans, S. Buchholz, J. Veenstra, "Memory-Based Shallow Parsing," In Proc. of CoNLL, Bergen, Norway, 1999.

[14] Taku Kudo and Yuji Matsumoto, "Chunking with Support Vector Machines". In Proc. of NAACL-2001, 2001.

[15] Erik F. Tjong Kim Sang, W. Daelemans, H. Dejean, R. Koeling, Y. Krymolowski, V. Punyakanok, and D. Roth, "Applying system combination to base noun phrase identification," In Proc. of COLING. 2000.

[16] Hans van Halteren, "A Default First Order Family Weight Determination Procedure for WPDV Models," In Proc. of CoNLL-2000 and LLL-2000, pp. 119-12, 2000.

[17] 신효필, "최소자원 최대효과의 구문분석", 제11회 한글 및 한국어 정보처리 학술대회, pp. 242-248, 1999.

[18] Juntae Yoon, et. al. "Three Types of Chunking in Korean and Dependency Analysis based on Lexical Association," In Proc. of the 18th International Conference on Computer Processing Languages (ICCPOL'99), pp. 59-65, 1999.

[19] 양재형, "규칙기반 학습에 의한 한국어의 기반 명사구 인식", 정보과학회 논문지: 소프트웨어 및 응용, 제27권 제 10호, pp. 1062-1071, 2000.

[20] 박성배, 장병탁, "최대 엔트로피 모델을 이용한 텍스트 단위와 학습", 제 13회 한글 및 한국어 정보처리학술대회, pp. 130-137, 2001.

[21] 이신목, 강인호, 김길창, "방향성을 이용한 한국어 비재귀명사구 인식 모델", 제 13회 한글 및 한국어 정보처리학술대회, pp. 439-444, 2001.

[22] Young-Sook Hwang, Hoo-jung Chung, Yong-Jae Kwak, So-Young Park, "Shallow Parsing by Weighted Probabilistic Sum," In Proc. of the 19th International Conference on Computer Processing Languages(ICCPOL2001), 2001.

[23] Avrim L. Blum. (1997). "Selection of Relevant Features and Examples in Machine Learning," Journal of Artifical Intelligence, pp. 245-271.

[24] J. R. Quinlan. (1993). "C4.5: Programs for Machine Learning", Mateo: Morgan Kaufmann.

[25] D. W. Aha, D. Kibler and M. Albert. (1991). "Instance-based learning algorithms," Machine Learning, 6:37-66.

[26] Walter. Daelemans and Antal van den Bosch. (1992). "Generalisation performance of backpropagation learning on a syllabification task," In M. F. J. Drossaers and A. Nijholt, editors, Proc. of TWLT3: Connectionism and Natural Language Processing, pp. 27-37, Enschede. Twente University.

[27] 한국과학기술원, 국어정보베이스, v.1.0 (CD 배포판), 1997.

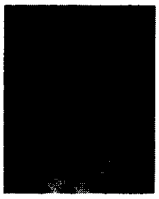
황 영 숙

1991년 고려대학교 전산과학과 학사. 1991년 ~ 1995년 쌍용정보통신 근무. 1998년 고려대학교 컴퓨터학과 석사. 1998년 ~ 현재 고려대학교 컴퓨터학과 박사과정. 관심분야는 자연어처리, 기계학습, 정보추출

[부록 A] 실험에 사용된 품사집합

대분류	Small POS Set(26)	Large POS Set(65)
체언	보통명사(NC)	일반명사(NNCG), 동사성명사(NNCV),형용사성명사(NNCJ), 고유명사(NNP)
	의존명사(NB)	의존명사(NNB), 단위성 의존명사(NNBU)
	대명사(NP)	인칭대명사(NPP), 지시대명사(NPI)
	수사(NN)	수사(NU)
접사	접두사(XP)	명사형 접두사(XPNN), 수사형 접두사(XPNU)
	접미사(SN)	명사형접미사(XSNN), 대명사형접미사(XSNP), 수사형 접미사(XSNU), 복수형 접미사(XSNPL)
	관형형접미사(SN)	관형사형 접미사(XSD)
	서술형접미사(SV)	동사화 접미사(XSVV), 형용사형 접미사(XSVJ)
		부사형접미사(XSA)
조사	격조사(JC)	주격조사(PS), 보격조사(PC), 목적격조사(PO),부사격조사(PA), 호격조사(PV)
	보조사(JX)	보조사(PX)
	접속조사(PN)	접속조사(PN)
	관형격조사(JM)	관형격조사(PD)
관형사	관형사(MM)	성상관형사(DA), 지시관형사(DI), 수관형사(DU)
부사	일반부사(MAG)	성상부사(AA), 서술부사(AP), 지시부사(AI), 접속부사(AC)
	서술성부사(MAJ)	동사성부사(AV), 형용사성부사(AJ)
감탄사	감탄사(II)	감탄사(C)
용언	동사(PV)	동사(VV)
	형용사(PA)	형용사(VJ)
	보조용언(PX)	보조동사(VVX), 보조형용사(VJX)
	서술격조사(CO)	서술격조사(I)
어미	종결어미(EF)	종결어미(EFF)
	연결어미(EC)	연결어미(EFC)
	명사형어미(EN)	명사형어미(EFN)
	관형형어미(EM)	관형형어미(EFD)
	선어말어미(EP)	선어말어미(EP)
		부사형어미(EFA)
기호	기호(SS)	온점(SS.), 물음표(SS?), 느낌표(SS!), 반점(SS,), 빗금(SS/), 쌍점(SS:), 반쌍점(SS;), 왼쪽따옴표(SS'), 오른쪽따옴표(SS'), 왼쪽괄호(SS(), 오른쪽괄호(SS)), 줄표(SS-), 줄임표(SSA), 기타(SSX), 외국어(SCF), 한자(SCH), 숫자(SCD)

정 후 중
1997년 고려대학교 컴퓨터학과 학사. 1999년 고려대학교 컴퓨터학과 석사. 1999년 ~ 현재 고려대학교 컴퓨터학과 박사과정. 관심분야는 자연어처리, 정보검색



곽 용 재
1997년 고려대학교 컴퓨터학과 학사. 1999년 고려대학교 컴퓨터학과 석사. 1999년 ~ 현재 고려대학교 컴퓨터학과 박사과정. 관심분야는 자연어처리, 구문분석, 정보검색, 기계학습

박 소 영
1997년 상명대학교 전자계산학과 학사. 1999년 고려대학교 컴퓨터학과 석사. 1999년 ~ 현재 고려대학교 컴퓨터학과 박사과정. 관심분야는 자연어처리, 기계번역, 한국어 정보처리

임 해 창
1991년 ~ 현재 고려대학교 컴퓨터학과 교수. 1993년 인지 과학회 이사. 1994년-998년 한국 정보과학회 편집위원. 1998년 5월 ~ 2000년 5월 한국정보과학회 한국어정보처리연구회 운영위원장. 1999년 3월 ~ 2000년 8월 고려대학교 컴퓨터과학기술연구소 연구소장. 관심분야는 자연어처리, 구문분석, 정보검색, 기계학습