工學碩士 學位論文

# 사례기반 학습을 이용한 한국어 어절 분류

Korean Eojeol Classification Using Instance-based Learning

指導敎授 金 載 熏

2002年 8月

韓國海洋大學校 大學院

컴 퓨 터 工 學 科

朴 浩 珍

本 論文을 朴浩珍의 工學碩士 學位論文으로 認准함

委員長　工學博士　朴　烋　讚　印

委　員　工學博士　辛　沃　根　印

委　員　工學博士　金　載　熏　印

2002年　8月

韓國海洋大學校 大學院

컴퓨터工學科　朴 浩 珍

# 목 차

# Korean Eojeol Classification Using Instance-based Learning

*Ho-Jin Park*

*Department of Computer Engineering*
*Korea Maritime University, Busan, Korea*

## Abstract

Generally, Internet users have exploited search engines to find the information that they need. Such search engines require fast processing and particularly morphological analysis in Korean. The notorious problem in Korean morphological analysis is over-generation, which is caused by the lack of morphotactics. This paper describes the eojeol classification in order to lighten the burden of the over-generation. In other word, we want to reduce the search space for morphological analysis using eojeol categories. In this paper, we propose a method for eojeol classification using an instance-based learning technique.

To evaluate our proposed system, we use two test corpora (KAIST and ETRI) that are part-of-speech tagged in Korean. In addition, we use the cross

validation method for training and evaluation since the test corpora are not enough. The average accuracies of the test corpora are 97% and 96.6% under 22 features, respectively, but the average accuracy is reduced into 95.5% even though the two corpora are combined. We believe that the tragedy results from the inconsistent tagging method in spite of the larger amount of training data. To select optimal features for our system, we employ backward sequential selection. As a result, we choose 16 features as the optimal features and the performance of our system is improved by about 0.2%. Furthermore the reduction rate is 35% on average when our system is applied to Korean morphological analysis.

# 1제 장 서 론

인터넷에서 사용자가 필요한 정보를 찾기 위해서 사용하는 가장 보편적인 방법은 검색엔진이다(엄재홍, 장병탁 2000). 하지만 검색엔진은 너무 많은 양의 정보를 제공하기 때문에 사용자에게 필요한 정보를 찾기란 그리 쉬운 일은 아니다. 이와 같은 문제를 해결하기 위해 최근 많은 연구자들은 웹기반 질의 응답 시스템(Buchholz and Daelemans 2001; Fujii and Ishikawa 2001)이나 정보 추출 시스템(TREC-10)을 개발하고 있다. 웹을 기반으로 하는 이들 시스템은 빠른 처리를 요구하고 있다.

한편 한국어를 대상으로 하는 웹기반 시스템은 필수적으로 형태소 분석과 같은 언어처리도구가 요구된다. 형태소 분석은 주어진 문장으로부터 의미의 기본 단위가 되는 형태소를 찾는 과정이며, 형태소 분리, 불규칙이나 굴절 현상에 대한 원형 복원 등의 작업이 필요하다. 한국어 형태소 분석의 가장 큰 문제는 형태소 분석의 결과가 너무 많다는 것이며, 이를 형태소 과잉분석이라고 한다. 예를 들면, 어절 “하나가”에 대한 형태소 분석 결과 수는 132개를 얻었다(김재훈, 서정연 외 1995). 형태소 과잉분석의 원인은 여러 가지가 있을 수 있으나, 가장 중요한 이유 중 하나는 형태소 배열규칙(morphotactics)의 제약조건이 부족하기 때문이다. 형태소 배열규칙의 제약조건을 강화하기 위해서 본 논문에서는 어절범주 정보를 사용한다. 즉, 형태소를 분석하기 전에, 한국어 어절에 대한 어절범주를 결정하여, 형태소 분석기의 탐색공간을 줄이고자 한다. 본 논문에서 어절에 대한 어절범주를 결정하는 것을 어절분류(eojeol classification)라고 하며, 본 논문에서는 사례기반 방법을 이용해서 어절을 분류한다. 사례기반 방법은 기존의 사례들을 학습하고, 학습된 사례들을 이용하여 새로운 사례를 분류하는 방법이다. 이러한 사례들을 벡터로 구성되어야 하며, 이를 자질벡터(feature vector)라고 한다. 본 논문에서는 오토마타 및 사전을 이용하여 자질벡터를 구성한다.

본 논문의 구성은 다음과 같다. 제2장에서는 관련연구로 여러 기계학습 방법을 이용한 분류 알고리즘에 대해서 살펴보고, 제3장에서는 본 논문과 관련된 한국어의 특징과 어절 범주를 정의한다. 제4장에서는 본 논문에서 제안한 한국어 어절 분류 시스템에 대하여 구체적으로 기술한다. 제5장에서는 제안된 시스템의 성능을 평가하고, 마지막으로 제6장에서는 결론과 향후 연구방향에 대해서 기술한다.

# 2제 장 관련 연구

분류란 주어진 자질벡터에 해당하는 범주를 결정하는 것이며, 일반적으로 학습을 통하여 시스템이 구성된다. 본 장에서는 분류의 일반적인 개념을 소개하고, 본 논문과 관련된 몇 가지의 기계학습 방법을 간단히 소개하고자 한다.

## 2.1 분류

분류란 주어진 자질벡터로부터 미리 정의된 하나 혹은 그 이상의 범주를 결정하는 것이다. 문서 분류의 예를 들어보자. 문서 분류는 주어진 문서를 단어로 분리하고, 분리된 단어를 자질로 하여 단어의 빈도수 등을 이용하여 자질벡터를 구성하고, 이 자질벡터를 이용하여 주어진 문서가 어떤 분야(예를 들면, 정치 분야, 스포츠 분야 등)에 속하는지를 결정하는 것이다. 최근에 개발되는 대부분의 분류 시스템은 학습 방법을 이용하는데, 일반적으로 널리 사용되는 학습 방법으로는 나이브 베이즈(naive Bayes)(Mitchell 1997), 결정 트리(decision tree)(Quinlan 1993), 신경망(neural network)(Haykin 1998), $k$-최근법(k-nearest neighbor)(Dasarathy 1991), 사례기반(instance-based) 방법(Daelemans and Bosch, et al 1997), 변형기반(transformation-based) 방법(Brill 1995) 등이 있다. 이하에서는 본 논문과 관련된 사례기반 방법과 결정트리 방법 그리고 변형기반 방법에 대해서 구체적으로 소개하고자 한다.

## 2.2 사례기반 학습

이미 경험한 사실로부터 어떤 규칙을 추출하여 그 규칙을 판단의 기준으로 간주하는 것 보다 이미 경험한 사실 중에서 새로운 상황에 가장 유사한 사실을 판단의 기준으로 간주하는 것이 효과적이라는 가설로부터 사례기반 학습이 시작되었다(Daelemans and Zavrel 2001). 이러한 학습 방법은 유사성기반(similarity-based) 혹은 예제기반

(example-based) 학습이라고도 불리며, 지도 학습(supervised learning) 방법이다. 또한 사례기반 학습은 $k$-최근법 알고리즘으로부터 나오게 되었다. 그림 1은 사례기반 학습의 흐름도이다.

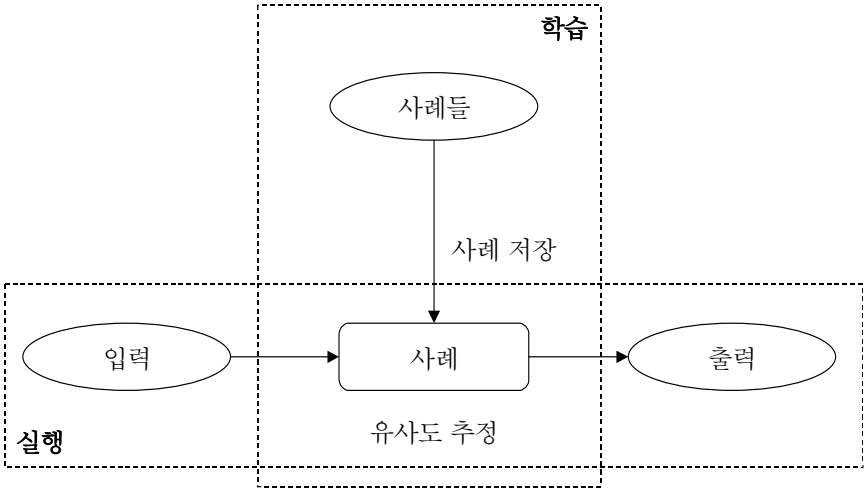


그림 1 사례기반 시스템 흐름도

Fig. 1 A flow diagram of an instance-based system

일반적인 사례기반 시스템은 학습부와 실행부로 구성된다. 학습부에서는 유사한 사례를 군집화하거나, 빠른 검색을 위해서 색인하여 적절한 형태로 사례를 저장한다. 실행부에서는 주어진 입력에 대해서 학습부에서 저장된 사례와 가장 비슷한 사례를 추출하고, 주어진 입력과 저장된 사례들의 유사도를 계산하여 범주를 결정한다. 사례기반 학습은 품사 태깅(part-of-speech tagging), 문서 분류(text classification), 기계 번역(machine translation) 등에 사용되었다(김영택 외 2001).

## 2.3 결정트리

결정트리는 기호 학습 방법에서 대표적으로 사용되는 방법으로, 이산값(discrete value)을 가지는 함수를 추정하는 데 이용된다. 결정트리는 많은 학습자료를 효과적으로 검색하기 위해 학습자료를 트리형태로 표현한다. 일반적으로 트리의 노드는 질의에 해당하며, 질의의 답에 따라서 어떤 부트리가 결정된다. 그림 2는 결정트리의 한 예이다. 그림 2는 날씨와 주변 여건에 따라서 골프 경기를 진행할 것인지를 결정하기 위한 결정트리이다. 루트 노드는 라벨로 “날씨”를 가지고 있는데 이는 자질벡터의 한 요소이며, 각 에지의 라벨로 “맑음”, “흐림”, “비”를 가지는데 이는 자질 “날씨”에 대한 자질값이다. 그리고 말단 노드의 라벨은 분류 문제의 범주를 의미한다. 그림 2의 결정트리에 따르면 골프 경기가 진행되는 경우는 날씨가 맑고 습도가 75% 이하이거나, 날씨가 흐리거나 혹은 비가 오면서 바람이 불지 않을 때이다. 그와 반대로 골프 경기가 취소되는 경우는 경치가 맑고 습도가 75% 이상이거나 혹은 비가 오면서 바람이 불 때이다.

결정트리는 데이터의 분류와 일반화에 널리 이용되며, 생성이 용이하고, 학습을 통해 생성된 트리를 쉬운 규칙으로 구성할 수 있다는 장점이 있다. 또한 품사 태깅, 의미 중의성 해소, 문서 분류 등 다양한 자연언어처리 분야에 적용되어 왔다(김영택 외 2001).

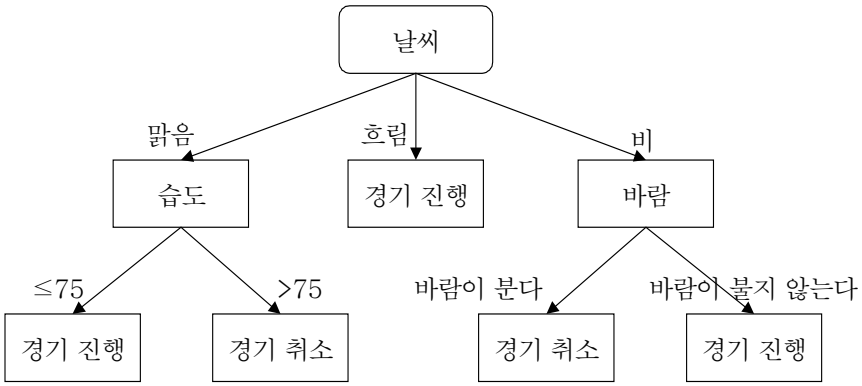


그림 2 결정트리의 예

Fig. 2 An example of a decision tree

## 2.4 변형기반 학습

변형기반 방법(Brill 1993)은 Eric Brill에 의해 품사 태깅을 위한 학습 방법으로 처음 소개되었다. 변형 기반 학습 방법은 빈도수를 이용하는 것과 같은 아주 간단한 방법으로 원시 말뭉치에 초기 태그를 부착하고, 학습 단계에서 초기 태그의 오류를 수정하기 위한 규칙을 추출한다. 오류 수정 규칙을 추출하기 위해서는 먼저 규칙틀(rule template)을 만들어 학습 말뭉치로부터 규칙틀에 맞는 규칙을 추출하고 추출된 규칙들 중에서 시스템의 오류를 가장 많이 수정하는 규칙을 시스템 규칙으로 등록한다. 이러한 과정을 반복적으로 수행하여 오류 규칙을 생성해 내는 방법이다. 그림 3은 변형 기반 학습 과정을 그림으로 표현하고 있다.

변형기반 방법은 품사 태깅, 전치사 접속 결정, 구문 분석, 철자 교정, 의미 중의성 해소 등 여러 자연언어처리 분야의 문제를 해결하는데 사용되었다(김영택 외 2001).

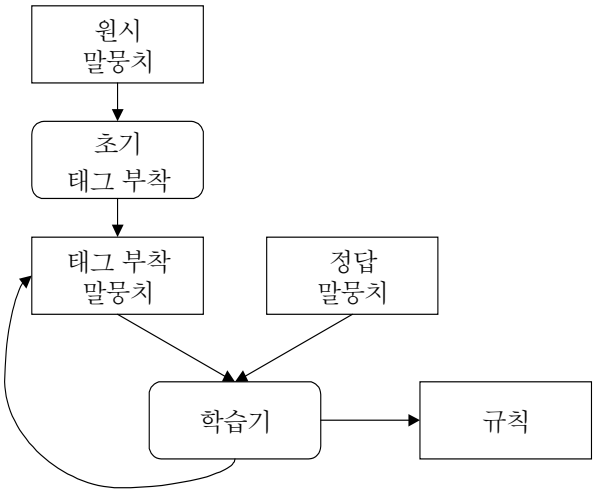


그림 3 변형 기반 학습 방법의 흐름도

Fig. 3 A flow diagram of the transformation-based learning

# 3제 장 한국어 어절범주

본 장에서는 한국어의 문법적 단위 및 어절에 대한 특징을 살펴보고, 어절범주에 대하여 정의한다. 어절범주는 5개의 범주로 나눌 수 있으며, 본 논문에서는 5개의 범주에 대하여 10개의 세부범주로 나누어 어절범주를 정의하였다.

## 3.1 한국어 어절

한국어는 몇 가지 문법적 단위를 지니고 있으며, 그 단위는 아래와 같다.

- 음절 : 자음과 모음이 합쳐 하나의 발음 단위
  예) 철/수/가 책/을 읽/었/다
- 형태소 : 최소의 의미 단위
  예) 철수/가 책/을 읽/었/다
- 어절 : 의미적 구성 단위, 문장 구성의 최소 단위
  예) 철수가/ 책을/ 읽었다
- 문장 : 의미상 하나의 완결된 사상, 감정을 나타내는 단위
  예) 철수가 책을 읽었다

본 논문의 대상은 어절이며, 어절은 아래와 같은 특성을 지니고 있다.

1. 어절과 어절은 띄어쓰기를 한다.
2. 어절은 의미상, 형태상으로 독립성이 있다.
3. 조사는 자립형태소와 어울려져야만 어절을 이룬다(철수+가).

4. 용언은 어간과 어미(읽+다)의 결합으로 어절을 이룬다.

## 3.2 한국어 어절범주

한국어 어절은 기능상으로 5개의 범주로 나눌 수 있다. 이러한 범주를 세분화하면 10개의 세부범주로 나눌 수 있다. 표 1은 한국어 어절의 범주를 나타낸다.

표 1 한국어 어절범주

Table 1 A list of korean eojeol categories

| 범주 | 세부 범주 | 어절 구성 | 예 |
|---|---|---|---|
| 체언 | Nc | (수사\| 명사)+ | 남북한/명사, 경제성장률/명사 |
| | Nj | (수사\| 명사) 접사* 조사+ | 중국/명사+의/조사,<br>산성비/명사+는/조사 |
| | Np | 대명사 접사* 조사+ | 이/대명사+들/접사+과/조사,<br>그/대명사+를/조사 |
| 용언 | Nv | 명사 접사* 어미+ | 우려/명사+되/접사+고/어미 |
| | V | (동사\| 형용사) 접사* 어미+ | 뒤집/동사+어/어미<br>있/보조용언+다/어미 |
| | P | 동사 | 달라/동사 |
| 수식언 | M | 관형사 | 우리/관형사, 아무/관형사,<br>이/관형사 |
| | A | 부사 조사* | 특히/부사, 거듭/부사,<br>너무/부사+도/조사 |
| 감탄사 | I | 감탄사 | 그래/감탄사, 자/감탄사,<br>말이다/감탄사 |
| 기호 | S | 기호 | “/기호, (/기호, )/기호 |

체언절은 3개의 세부범주, Nc(명사), Nj(명사 접사? 조사), Np(대명사 접사? 조사)로

나누었으며, 수사나 고유명사는 모두 명사 범주에 속한다. 용언절은 3개의 세부범주, Nv (명사 접사? 어미), V(동사/형용사 접사? 어미), P(동사/형용사)로 나누었으며, 수식언절은 2개의 세부범주, M(관형사)와 A(부사 조사?)로 나누었다. 그 외에도 I(감탄사)나 S (기호)로 나누었으며, 이들에 대해서는 세부범주를 나누지 않았다. 이하의 절에서는 각 세부범주에 대하여 자세히 기술한다.

### 3.2.1 체언절

체언은 문장의 주체적인 역할을 하며, 여기에는 명사, 대명사, 수사를 포함한다. 또한 본 논문에서는 외국어를 명사로 다루었다. 체언절은 3개의 세부범주, Nc, Nj, Np로 나누었으며 아래에서 각 세부범주에 대한 구체적인 예를 보일 것이다.

- Nc
  - '생산적인 **국회**' 라는 말이 있다.
  - 전경련은 **12일 21대** 새 회장으로 **최종현 선경그룹** 회장을 선임했다.
  - **작년 GNP총액** 약 **3천억** 달러가 적지는 않다.

- Nj
  - **이산화유황은** 바로 **산성비의 원인이** 된다.
  - 하루 내지 이틀이면 **중국의** 대기오염 **물질이 한반도와 일본에** 와 닿는다.
  - 그것이 **국제기준치보다 32배나** 초과해 들어 있다는 검역보고다.

• Np

- **그것은** 한국정치의 변동에 따른 국회의 대처자세다.
- 국회의 견제기능이 새삼 우려되는 이유도 **여기에** 있다.
- **이는** 또 고용의 축소로 이어진다.

### 3.2.2 용언절

용언은 문장의 서술적 역할을 하며, 여기에는 동사와 형용사를 포함한다. 용언절은 3개의 세부범주 Nv, V, P로 나누었으며, 아래에서 각 세부범주에 대한 구체적인 예를 보일 것이다.

• Nv

- 북한의 적극적 참여를 **요망한다**.
- 지나친 요구와 기대는 서로가 **자제해야** 할 **시기다**.
- 그것은 국민의 건강을 스스로 지키는 **수단이다**.
- **어딘가** 준비부족의 엉성한 대목이 있다는 인상을 갖게한다.

• V

- 매우 **바람직한** 일이다.
- 이제 **이러한** 우리의 요구가 실질적으로 관철된 셈이다.
- 통일원의 업무확대 주장도 **있었다**.
- 이에 **대한** 구체안을 빨리 제시해야 **한다**.

• P

- 붓글씨로 포장된 봉투 속에는 '잘 봐 **달라**' 는 사심들이 끼어 있음을 부인 못한다.

### 3.2.3 수식언절

수식언은 체언이나 용언을 수식하는 말을 의미하며, 2개의 세부범주, M과 A로 나누었으며, 아래에서 각 세부범주에 대한 구체적인 예를 보일 것이다.

• M

- 그러나 **이** 금액은 어느 의미에서 상징적인 숫자이다.
- 더욱이 **새** 정부는 국가안전기획부의 개혁을 약속하고 있다.
- 다 **자기** 나라의 경제 살리기가 초점이다.

• A

- 중소기업 **역시** 할 일이 많다.
- **그러기에는** **현재** 러시아가 직면하고 있는 상황이 너무나 심각하다.
- **그러나** **언제** **어떻게** 하느냐에 이르면 **항상** 말이 많다.

### 3.2.4 감탄사

감탄사는 문장 안에서 다른 단어와 독립적으로 쓰이며, 보통은 느낌이나 놀람, 부르고 대답하는 감정적인 언어 및 입버릇으로 내는 말을 말한다. 또한 감탄사는 기능에 따라서 감탄사가 되기도 하고 다른 품사가 되기도 한다.

• I

- **와**!
- **어디** 수은전지에 한할 일인가.
- 이리저리 다니며 짐승을 잡아먹고 열매 따 먹고 **말이다**.

### 3.2.5 기호

기호는 한글이 아닌 문자를 표현하는 범주로서 문장 종결 기호, 대등・접속 기호 등이 이 범주에 속한다. 또한 '%', '/', '-' 등 한글이 아닌 문자를 포함한다.

• S

- 그 결과가 마침내 92년 GNP(국민총생산) 4.7% 성장으로 나타났다.
- 썬 배를 설탕이나 시럽에 10-20분 정도 재워 둔다.
- 이런 물건들을 필통에 가지런히 넣어 두면, 다음에 사용하기에 편리하다.

# 4제 장 한국어 어절 분류 시스템

본 장에서는 한국어 어절 분류 시스템에 대하여 기술한다. 본 시스템은 두 가지 부분으로 구성되어 있는데 학습단계와 실행단계가 그것이다. 학습단계에서는 말뭉치로부터 자질을 추출하여 사례기반 학습을 통해서 최적화된 예제 색인을 얻으며, 실행단계에서는 최적화된 예제 색인을 이용하여 어절 분류를 한다.

## 4.1 시스템 구성

본 논문에서 제안하는 어절 분류 시스템은 사례기반 학습을 통하여 해당 어절을 범주로 분류하는 방법이다. 본 논문에서 사용하는 자질은 전문가의 경험을 통하여 선택하였다. 자질에 대한 내용은 4.2.3절에서 자세히 알아보도록 한다. 그림 4는 구현된 시스템의 개념도이다.

학습단계에서는 범주 부착 말뭉치를 학습하여 최적화된 예제 색인을 얻게된다. 실행단계에서는 학습단계에서 얻어진 최적화된 예제 색인을 사용하여 입력 문장의 어절을 분류하게 된다. 이하의 절에서 학습단계와 실행단계에 대한 구체적인 설명을 기술한다.

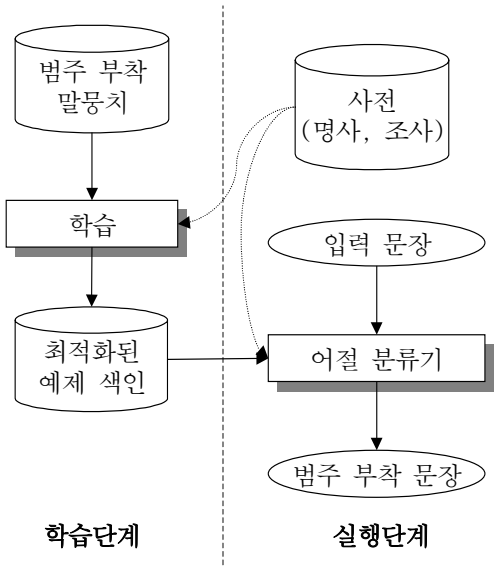


그림 4 어절 분류 시스템의 개념도

Fig. 4 Overview of the eojeol classifier

## 4.2 학습단계

학습단계에서는 품사 부착 말뭉치로부터 최적화된 예제 색인을 얻는 단계이다. 최적화된 예제 색인은 실행단계의 어절 분류기에 사용된다. 먼저, 품사 부착 말뭉치는 전처리기를 거쳐 기호가 분리된다. 분리된 말뭉치는 어절범주 부착기로 어절범주 부착 말뭉치로 변환되고, 자질 추출기로 학습에 필요한 자질을 추출하며, 사례기반 학습의 학습 데이터로 사용된다. 마지막으로 사례기반 학습은 최적화된 예제 색인을 저장한다. 그림 5는 학습단계의 흐름도이다.

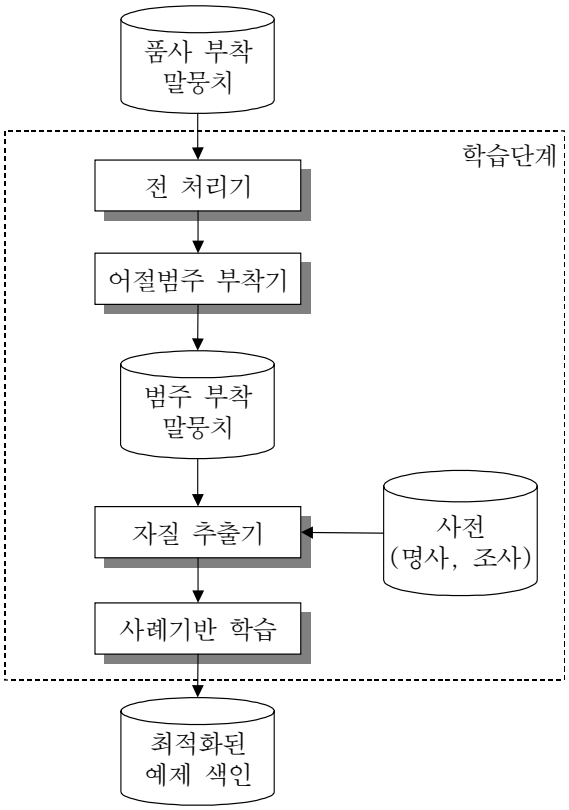


그림 5 학습단계 흐름도

Fig. 5 A flow diagram of learning steps

### 4.2.1 전처리기

전처리기는 어절에서 음절과 기호를 분리한다. 문장에서 기호는 부가 설명이나 단위, 측도 등을 나타내기 위하여 사용된다. 이러한 기호들이 포함되어 있는 어절은 두 개 이상의 의미로 이루어져 있다. 예를 들면, 어절 “이산화유황(SO2)의”는 “이산화유황의”와 이산화유황의 화학식 ″SO2″의 두 가지 의미로 이루어져 있다. 이러한 어절은 어절 범주를 결정하는데 문제가 있기 때문에 “이산화유황”, “(”, “SO2”, “)”, “의”로 분리하여 각각에

대한 어절범주를 결정한다. 분리된 “의”는 3장에서 설명한 어절범주에 속하지 않음으로 전처리기에서 분류된 조사, 어미 등을 위해서 본 논문에서는 표 2와 같은 내부 범주를 이용한다.

표 2 내부 범주

Table 2 A list of internal eojeol categories

| 범주 | 세부범주 | 어절 구성 | 예 |
|---|---|---|---|
| 조사 | Je | 조사+ 어미+ | 이란 |
| | Jj | 조사+ | 에서, 의, 가, 을 |
| 어미 | Ee | 어미+ | 는, 고, 라는 |
| | Ej | 어미+ 조사+ | 기보다는, 지요 |
| 지정사 | Ce | 지정사 어미+ | 이라는, 이라며, 이었다 |
| | Cj | 지정사 조사+ | 이라고까지, 이었음에도, 이라고 |
| 접사 | Xe | 접사+ 어미+ | 하면서, 들입니다, 시키려는 |
| | Xj | 접사+ 조사+ | 들이, 씩을, 씨가 |

또한 전처리기는 기호를 분리하는 것 외에, 분리된 기호를 후처리기에서 복원하기 위한 결합규칙을 생성하게 된다. 그림 6은 전처리기의 기호 분리 과정을 나타낸 그림이다. 결합규칙에서 ‘⊕’는 앞의 어절과 현재 어절과 분리되었다는 의미이다.

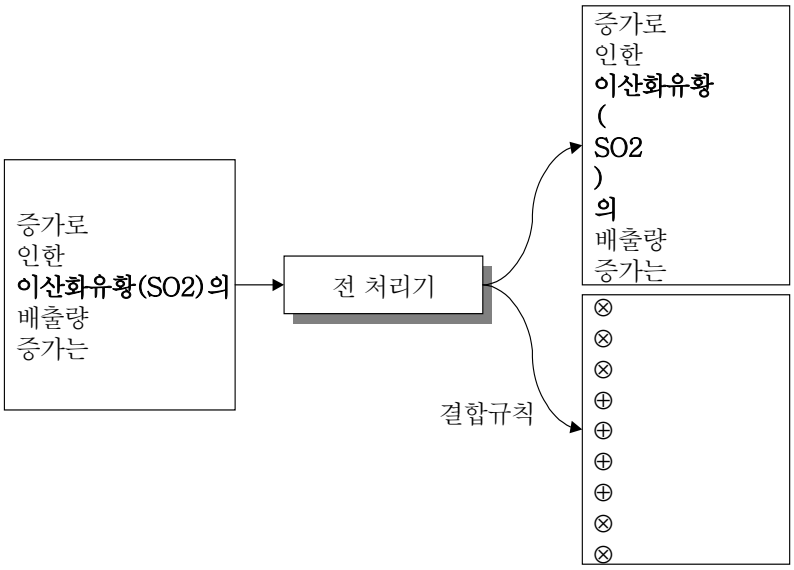


그림 6 전처리기

Fig. 6 A preprocessor for eojeol classifier

### 4.2.2 어절범주 부착기

학습에 필요한 말뭉치를 생성하기 위하여 어절범주 부착기는 품사 부착 말뭉치에 어절범주를 부착한다. 어절범주는 정규 표현식과 품사 정보를 이용하여 부착하였다. 어절의 범주를 부착하기 위하여 사용한 정규 표현식은 그림 7과 같이 유한 오토마타의 전이도로 표현할 수 있으며, Nj 범주를 인식하기 위한 정규 표현식의 일부만 나타낸 것이다.

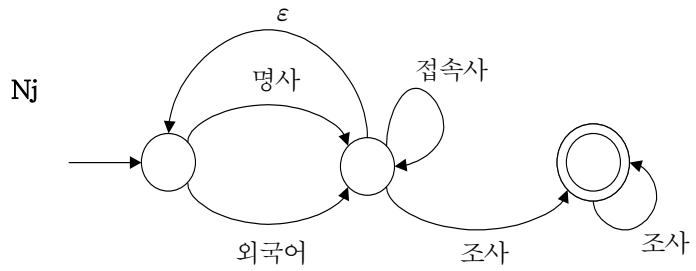


그림 7 상태 전이도로 표현한 Nj 범주 인식

Fig. 7 A finite-state automata for recognizing the category Nj

그림 7의 각 노드는 상태이며, 각 에지는 값에 따른 상태 전이를 나타낸다. 그리고, ε는 널(*null*) 문자열에 대한 상태 전이를 나타낸다. 예를 들어, “전문가/명사+들/접속사+의/조사”라는 어절이 입력으로 있을 때, “전문가”는 명사로 인식되고, “들”은 접속사로, “의”는 조사로 인식되어 Nj 범주가 부착된다.

범주를 부착하기 위한 정규 표현식은 여러 개로 구성되어 있고, 각각은 단계형으로 구성하여 각 단계에서 범주를 인식해 나아가는 방법으로 구현하였다.

### 4.2.3 자질 추출기

어절을 분류하기 위한 자질은 전문가의 경험을 통하여 22개의 자질을 선택하였다(표 3).

표 3 어절 자질표

Table 3 A list of eojeol features

| 자질번호 | 의미 | 자질값 |
|---|---|---|
| 1 | Nv 정규 표현식 | Nv, = |
| 2 | 이전 어절의 마지막 두 음절 | 두 음절, S |
| 3 | 이전 어절의 범주 | 범주, Bos |
| 4 | 현재 어절의 처음 두 음절 | 두 음절 |
| 5 | 현재 어절의 명사 위치 | 숫자, = |
| 6 | 현재 어절의 명사 포함 | N, = |
| 7 | 현재 어절의  마지막 두 음절 | 두 음절 |
| 8 | 현재 어절의 조사 | 조사 |
| 9 | 현재 어절의 어절 길이 | 1~4, >4 |
| 10 | 현재 어절의 심볼 | 심볼 |
| 11 | 현재 어절이 모두 영어 | +, − |
| 12 | 현재 어절의 숫자 포함 | +, − |
| 13 | 현재 어절의 과거형 | +, − |
| 14 | 현재 어절의 '하/되' 포함 | HD, = |
| 15 | 현재 어절의 '이다' 포함 | IDA, = |
| 16 | 현재 어절의 첫 번째 음절 | 한 음절, = |
| 17 | 현재 어절의 두 번째 음절 | 한 음절, = |
| 18 | 현재 어절의 세 번째 음절 | 한 음절, = |
| 19 | 현재 어절의 네 번째 음절 | 한 음절, = |
| 20 | 현재 어절의 첫 번째, 두 번째 음절 | 두 음절 |
| 21 | 현재 어절의 두 번째, 세 번째 음절 | 두 음절 |
| 22 | 현재 어절의 세 번재, 네 번째 음절 | 두 음절 |

표 3에서 1번 자질은 입력 어절이 Nv 정규 표현식에 인식되면 'Nv'를 자질값으로 사용하고, 그 외의 경우는 '='를 자질값으로 사용한다. Nv 정규 표현식은 명사에 "하다/되다"의 변형이 붙어 있는 어절을 인식하도록 구성되어 있다. 예를 들어, 어절 "사랑하다",

“사랑한다”, “결정되다” 등의 어절이 인식된다. 이와 같은 정규 표현식은 한국어의 “하다/되다”가 포함되어 있는 어절의 모호성 때문에 Nv 범주에 대해서만 정규 표현식을 자질로서 선택하였다. 2번과 3번 자질은 이전 어절의 음절 및 범주를 자질로서 사용하는데, 현재 어절이 문장의 시작이면 각각 'S'와 'Bos'를 자질값으로 사용한다. 4번과 7번 자질은 현재 어절의 처음 두 음절과 조사 및 어절에 포함되는 마지막 두 음절을 자질값으로 사용한다.

5번, 6번, 8번 자질은 사전을 이용한다. 5번 자질은 어절의 시작부터 명사 사전에 등록되어 있는 음절의 위치를 자질값으로 사용하며, 최장일치를 우선으로 한다. 명사 사전에 등록되어 있지 않을 경우에는 '='를 자질값으로 사용한다. 6번 자질은 어절에 명사가 포함되어 있으면 'N'을 자질값으로 사용하며, 그 외의 경우는 '='를 사용한다. 8번 자질은 어절의 뒤부터 조사 사전에 등록되어 있는 음절을 자질값으로 사용한다. 조사로 등록되어 있지 않으면 'S'를 자질값으로 사용한다. 그림 8은 5번과 8번 자질값을 추출하는 예를 보이고 있다.

9번 자질은 어절에 대한 음절 개수를 자질값으로 사용하며, 5음절 이상이면 '>4'를 자질값으로 사용한다. 10번 자질은 기호를 자질값으로 사용한다. 기호가 없을 경우 'S'를 자질값으로 사용한다. 11번부터 15번까지의 자질은 표 3에서의 의미에 따라서, 어절이 의미에 해당되면 '+'를, 그 외의 경우는 '-'를 자질값으로 사용한다. 16번부터 19번까지의 자질은 각각의 음절을 자질값으로 사용하며, 음절이 없을 경우 '='를 자질값으로 사용한다. 20번부터 22번까지의 자질은 두 음절을 자질값으로 사용한다. 예를 들어, 어절 “하나가”를 20번부터 22번까지의 자질값으로 표현하면, “하나 나가 가=”이 된다.

| 어절 | 명사 사전 | 자질값 |
|---|---|---|
| 산성비는 | : 존재 | 1 |
| 산성비는 | : 존재 | 2 |
| 산성비는 | : 존재 | 3 |
| 산성비는 | : 존재 하지 않음 | 3 |
| | **최종 자질값 :** | **3** |

**5번 자질 추출 과정**

| 어절 | 조사 사전 | 자질값 |
|---|---|---|
| 석재까지도 | : 존재 | 도 |
| 석재까지도 | : 존재 하지 않음 | 도 |
| 석재까지도 | : 존재 | 까지도 |
| 석재까지도 | : 존재 | 까지도 |
| 석재까지도 | : 존재 하지 않음 | 까지도 |
| | **최종 자질값 :** | **까지도** |

**8번 자질 추출 과정**

그림 8 5번과 8번의 자질값 추출 과정

Fig. 8 Extraction steps for the fifth and eighth feature value

### 4.2.4 사례기반 학습

사례기반 학습은 기존의 사례들을 이용하여 새로운 사례를 분류하는 방법이다. 본 논문에서의 사례기반 학습은 TiMBL[1] 도구를 사용하였다.

1) TiMBL: Tilburg Memory-Based Learner

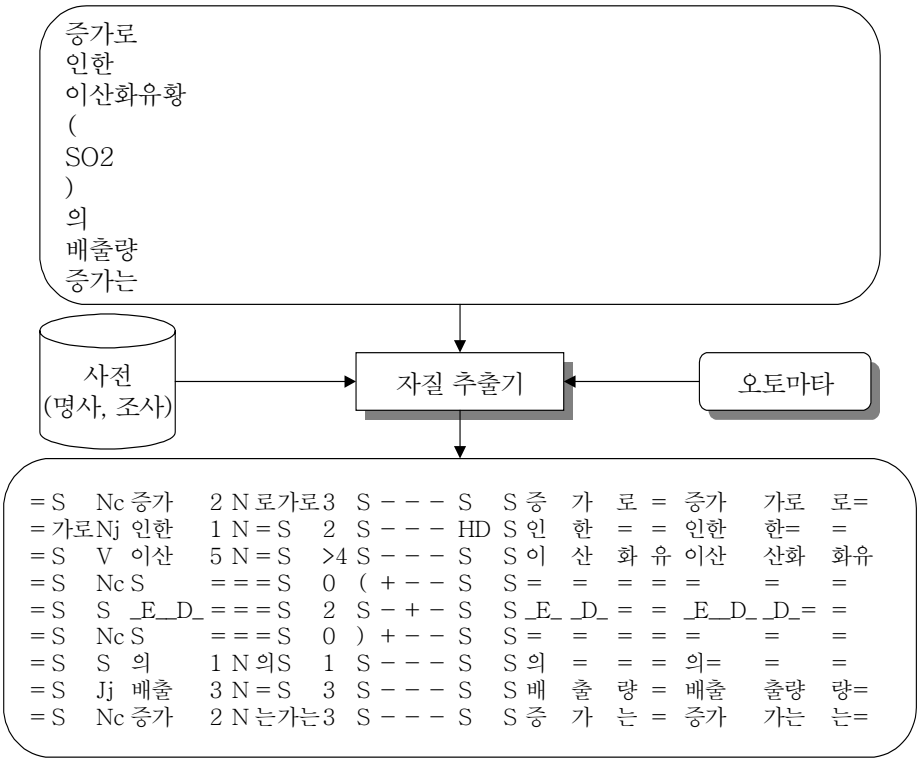


그림 9 어절 벡터로의 변환 과정

Fig. 9 A transformation step of eojeol vectors

TiMBL의 입력은 자질 벡터이다. 자질 추출기로 어절에 대한 자질을 벡터로 구성하고 어절범주 부착기로 해당 어절에 대한 범주를 부착하여 학습하였다. 그림 9는 어절을 TiMBL 도구의 학습 벡터로 변환하는 과정이다. 학습 벡터는 결정트리와 비슷한 구조로 저장된다.

## 4.3 실행단계

학습단계의 최적화된 예제 색인을 이용하여 실행단계에서는 입력 문장의 어절들을 어절범주로 분류한다. 각각의 분류된 어절은 범주 정보를 부착한 후, 후처리를 통하여 원래 어절로 복원하여 범주 부착 문장으로 출력한다. 그림 10은 실행단계의 흐름도이다.

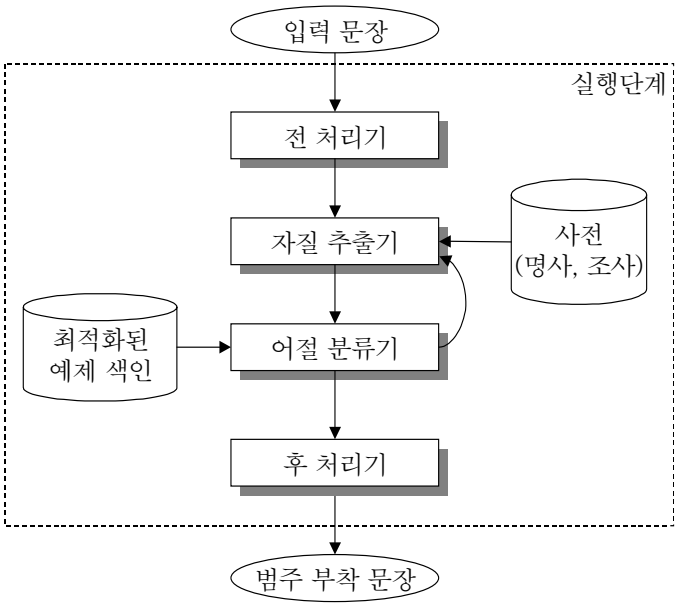


그림 10 실행단계 흐름도

Fig. 10 A flow diagram of processing steps

실행단계의 전처리기는 학습단계에서의 전처리기와 기능이 같다. 하지만 자질 추출기는 학습단계와 다르게 어절 분류기와 같이 동작되도록 되어있다. 자질의 2번째는 이전 어절의 범주를 자질로서 사용하는데, 이 자질은 어절 분류기의 결과를 이용하게 된다.

### 4.3.1 자질 추출기 및 어절 분류기

자질 추출기의 기본 동작은 학습단계와 같다. 하지만 2번 자질은 어절 분류를 해야만 얻을 수 있는 결과이다. 입력 어절이 있을 때 자질 추출기로 자질을 추출한 후 어절 분류기를 통하여 어절을 분류한다. 이렇게 분류된 어절범주는 다음 어절의 2번째 자질로 사용되게 된다.

어절범주를 부착할 때에는 ':' 기호를 사용하여 어절과 범주를 구분하였다. 어절 분류기는 학습단계에서 사용한 TiMBL 도구를 사용하였다. TiMBL은 학습과 분류기의 기능을 가지고 있다.

### 4.3.2 후처리기

후처리기는 분류된 어절 및 전처리에서 분리된 기호를 복원하는 작업을 수행한다. 복원을 할 때에는 결합규칙을 사용한다. 또한 기호는 기호 범주 'S'를 사용하지 않고, 기호 자체로 변환하여 입력 어절의 형식을 그대로 사용하도록 하였다. 그림 11은 입력 문장에 대한 범주가 부착되는 과정이다. 그림 11에서 어절 "이산화유황(SO2)의"는 어절 분류를 거쳐 "이산화유황(SO2)의:Nc(Nc)Jj"로 범주가 부착된다.

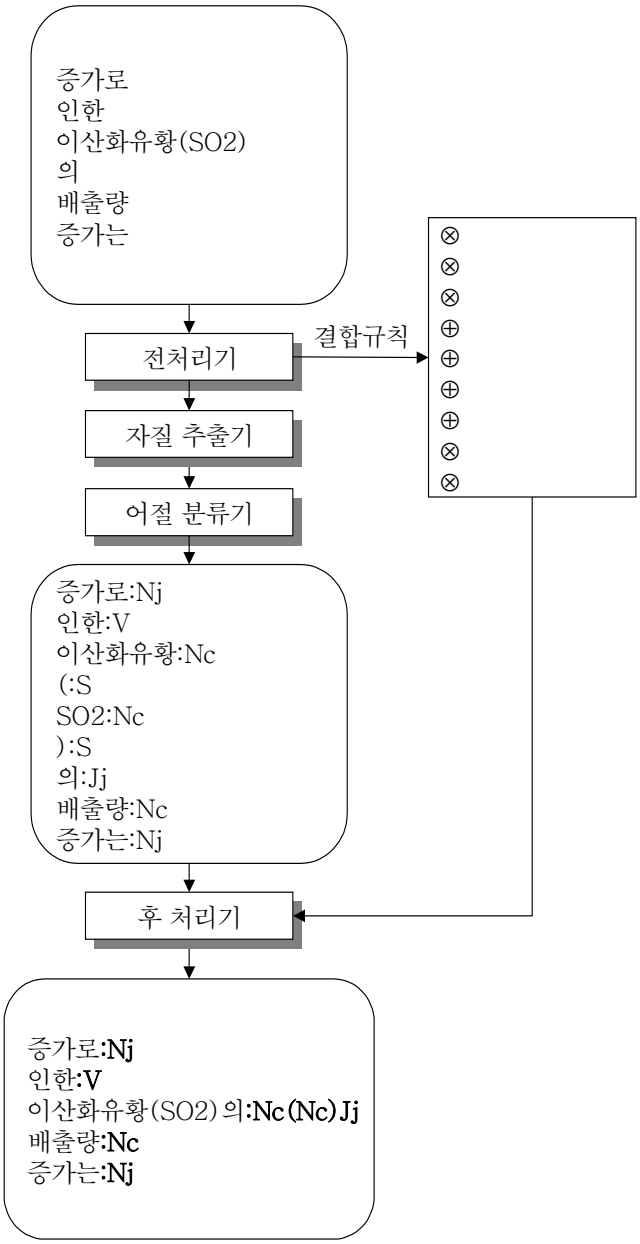


그림 11 어절 분류 과정

Fig. 11 A step of eojeol classification

# 5제 장 실험 및 평가

## 5.1 실험 말뭉치

본 실험에서는 두 종류의 평가용 말뭉치를 사용하였다. 이 평가용 말뭉치는 품사 부착되어 있는 말뭉치로서 KAIST 말뭉치(김재훈, 김길창 1995)는 약 20만 어절, ETRI 말뭉치(이현아, 이원일 외 1999)는 약 30만 어절 정도이다. 또한 성능을 평가하기 위하여 어절범주 부착기로 각 어절에 대하여 범주를 부착하였다.

## 5.2 성능 평가 방법

본 논문에서는 제안한 시스템의 성능을 평가하기 위하여 정확도(Accuracy)를 사용하였다(Manning and Schütze 1999). 정확도 $P$는 식 (1)로 정의된다.

$$P = \frac{A}{N} \tag{1}$$

여기에서 $N$은 분류된 어절의 총 개수이고, $A$는 시스템이 부착한 범주와 평가 말뭉치의 범주가 같은 어절의 개수이다.

또한 성능 평가 방법으로 교차 검증(cross validation) 방법을 사용하였다(Manning and Schütze 1999). 이 방법은 평가 말뭉치를 임의의 개수로 나누어 평가하는 방법이다. 예를 들어, 평가용 말뭉치를 A, B, C로 나누었을 경우, B, C를 학습 말뭉치로 사용하고, A를 평가 말뭉치로 사용하여 성능을 평가한다. B, C에 대해서도 이와 같이 성능을 평가하여 각각의 성능의 평균으로 시스템의 성능을 평가하는 방법이다. 이 방법은 말뭉치가

충분하지 않을 때 사용하는 방법으로, 본 논문에서는 10개의 말뭉치로 나누어 평가하였다.

## 5.3 어절 분류기 성능

22개의 자질에 대한 어절 분류기의 성능은 표 4에 나타나 있다. 두 종류의 말뭉치에 대하여 평가를 하였다. 또한 두 종류의 말뭉치를 합하여 평가하였는데, KAIST+ETRI 결과가 그것이다.

표 4 어절 분류 시스템의 성능

Table 4 Performance of our proposed system

| 말뭉치 번호 | KAIST | ETRI | KAIST+ETRI |
|---|---|---|---|
| 0 | 95.8% | 96.1% | 94.9% |
| 1 | 96.0% | 96.8% | 95.7% |
| 2 | 96.7% | 96.6% | 95.4% |
| 3 | 96.7% | 97.2% | 95.2% |
| 4 | 96.8% | 96.9% | 95.7% |
| 5 | 96.6% | 97.2% | 96.0% |
| 6 | 98.1% | 97.1% | 96.4% |
| 7 | 96.5% | 97.6% | 96.6% |
| 8 | 98.7% | 97.1% | 96.9% |
| 9 | 98.3% | 97.4% | 96.6% |
| **평균** | **97.0%** | **96.9%** | **95.9%** |

두 종류의 말뭉치는 평균 97%정도의 성능을 보였다. 하지만 두 종류의 말뭉치를 합쳤을 경우에는 평균 95.9%정도의 성능을 보였다. 이렇게 두 종류의 말뭉치를 합쳤을 경우에 성능이 저하되는 이유는 말뭉치의 장르가 다르기 때문이라고 생각된다. 또한 22개의

자질들의 학습 및 분류에 시간이 많이 소요되고 자질간의 간섭이나, 불필요한 자질이 포함되어 있을 것이라고 생각된다.

## 5.4 자질 최적화

본 논문에서 사용한 자질은 경험규칙으로 선택한 22개이다. 이런 자질들은 간섭이나 불필요한 정보가 포함되어 있다. 본 절에서는 자질에 대한 성능 평가를 함으로서, 분류 시스템의 성능을 향상시키고 분류에 필요한 자질을 선택하기 위한 실험이다. 그림 12는 자질의 성능을 평가하기 위한 흐름도이다.

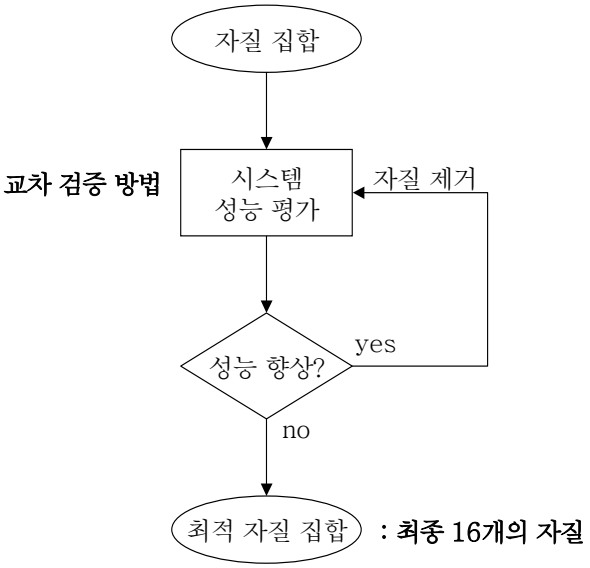


그림 12 최적 자질 선택 흐름도

Fig. 12 A flow diagram of the optimal features selection

22개의 자질에 대하여 성능을 평가하여 기본 성능이라 정의하고, 1번째 자질을 뺀 21개의 자질에 대하여 성능을 평가한다. 각각의 자질들을 뺀 후 성능을 평가하여 기본 성능과 비교하였을 때 성능이 저하되는 자질을 뺀 후, 다시 기본 성능을 평가하는 방법을 사용하였다. 최종적으로 시스템의 성능이 저하시키는 자질이 없을 때까지 반복하여, 최적 자질 집합을 선택한다. 표 5는 본 실험에서 선택되어진 최적화된 자질 집합이다.

표 5 최적화된 자질 집합

Table 5 A list of optimized features

| 자질번호 | 의미 | 자질값 |
|---|---|---|
| 2 | 이전 어절의 마지막 두 음절 | 두 음절, S |
| 3 | 이전 어절의 범주 | 범주, Bos |
| 4 | 현재 어절의 처음 두 음절 | 두 음절 |
| 5 | 현재 어절의 명사 위치(전방향) | 숫자, = |
| 6 | 현재 어절의 명사 포함 | N, = |
| 7 | 현재 어절의  마지막 두 음절 | 두 음절 |
| 8 | 현재 어절의 조사(후방향) | 조사, S |
| 9 | 현재 어절의 어절 길이 | 1~4, >4 |
| 10 | 현재 어절의 기호 | 기호, S |
| 11 | 현재 어절이 모두 영어로 구성 | +, − |
| 12 | 현재 어절의 숫자 포함 | +, − |
| 13 | 현재 어절의 과거형 포함 | +, − |
| 14 | 현재 어절의 '하/되' 포함 | +, − |
| 16 | 현재 어절의 첫 번째 음절 | 한 음절, = |
| 17 | 현재 어절의 두 번째 음절 | 한 음절, = |
| 22 | 현재 어절의 세 번재, 네 번째 음절 | 두 음절 |

## 5.5 최적 성능

표 6은 5.4.절의 최적화된 자질 집합을 이용한 어절 분류 시스템의 성능이다.

표 6 시스템 최적 성능

Table 6 Performance under using optimized features

| 말뭉치 번호 | KAIST(이전) | ETRI(이전) | KAIST+ETRI(이전) |
|---|---|---|---|
| 0 | 96.9%(95.8%) | 96.8%(96.1%) | 95.6%(94.9%) |
| 1 | 96.6%(96.0%) | 97.2%(96.8%) | 95.8%(94.9%) |
| 2 | 96.7%(96.7%) | 96.9%(96.6%) | 95.7%(95.4%) |
| 3 | 96.7%(96.7%) | 97.5%(97.2%) | 95.2%(95.2%) |
| 4 | 96.7%(96.8%) | 97.0%(96.9%) | 96.3%(95.7%) |
| 5 | 96.8%(96.6%) | 97.2%(97.2%) | 96.2%(96.0%) |
| 6 | 98.2%(98.1%) | 97.2%(97.1%) | 96.5%(96.4%) |
| 7 | 96.8%(96.5%) | 97.9%(97.6%) | 96.6%(96.6%) |
| 8 | 98.7%(98.7%) | 97.3%(97.1%) | 97.0%(96.9%) |
| 9 | 98.3%(98.3%) | 97.4%(97.4%) | 96.7%(96.6%) |
| 평균 | 97.2%(97.0%) | 97.2%(96.9%) | 96.2%(95.9%) |

본 실험에서는 이전의 실험보다 평균 0.2% 정도의 성능 향상을 보인다. 22개의 자질집합에서 최적의 자질 집합을 추출하여 성능을 평가하였는데, 미비한 성능 향상을 보인 것은 몇 가지 문제점이 있다고 생각한다. 우선 어절 분류를 하는데 있어서의 자질이 별다른 검증 없이 선택되었다는 것이다. 이는 자질 중에서 아직 불필요한 자질이나 분류에 필요한 자질이 포함되어 있지 않을 수 도 있다고 생각한다. 또한 말뭉치를 분석해본 결과 품사 부착 말뭉치에서 오류가 발견되었다. 예로서, “일해야”는 “일하/pv+어야/ecs”로 분석되어 있고, “지향해야”는 “지향/nc+하/xsv+어야/ec”로 분석되어 있다. 어절범주 부착기

에서 “일해야”는 V 범주로 부착하고, “지향해야”는 Nv 범주로 부착한다. 같은 “명사+해야”에 대하여 품사 부착 오류 때문에 범주가 다르게 부착되는 것을 알 수 있다. 이는 학습에 영향을 주어 성능 저하의 원인일 될 수 있다.

## 5.6 오류 분석

오류 분석은 시스템의 성능을 개선하는데 많은 도움을 줄 수 있기 때문에 본 절에서는 구현된 시스템에서 발생되는 오류를 분석하고자 한다. 오류 분석을 하기 위하여 최적 자질 집합을 사용하였으며, KAIST 말뭉치와 ETRI 말뭉치를 합하여 분석하였다. 그림 13은 어절 길이에 대한 오류율을 보여주며, 그림 14는 범주에 대한 오류율을 보여주고 있다.

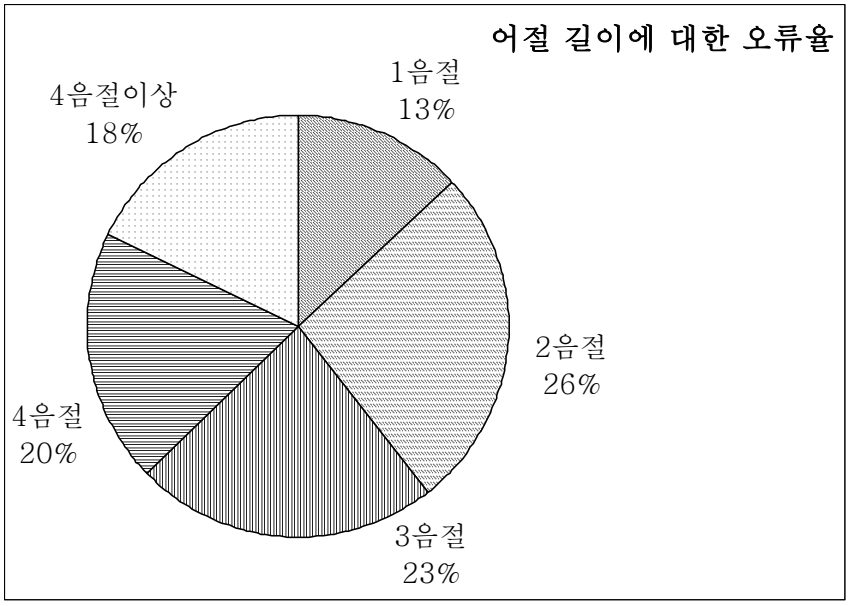


그림 13 어절 길이에 대한 오류율

Fig. 13 Error ratio for eojeol length

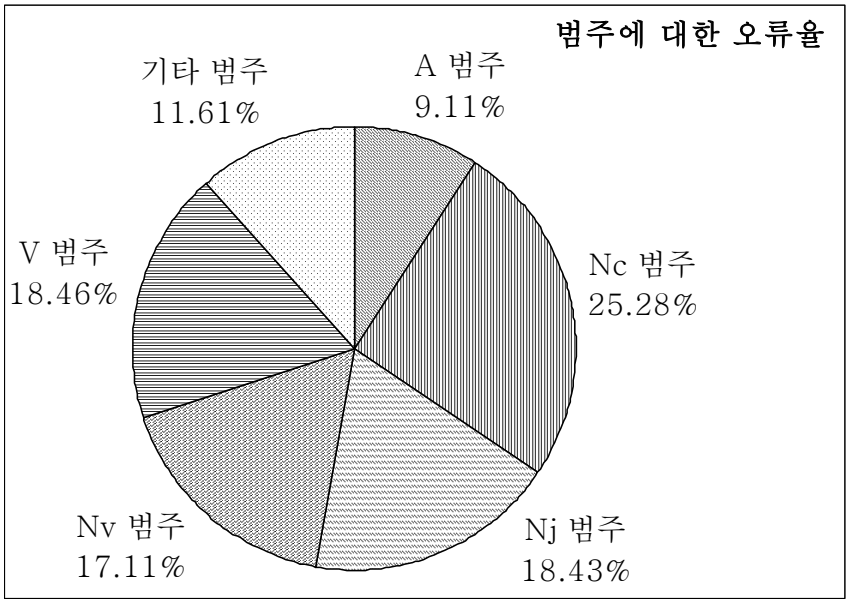


그림 14 범주에 대한 오류율

Fig. 14 Error ratio for eojeol category

그림 13에서 본 시스템의 오류 중에서 2음절과 3음절로 이루어져 있는 오류가 거의 50%에 가깝다. 이러한 짧은 어절들은 한국어의 특성상 모호한 경우가 많다. 예들 들면, 어절 “이는”은 수사와 조사가 결합된 “2는”과 대명사와 조사가 결합된 “이는”의 2가지 뜻을 가지고 있다. 이러한 문제점을 해결하기 위하여 이전 어절의 정보를 자질로서 사용하였지만 아직 오류가 많다. 어절에 대한 자질로서 이전 어절의 정보뿐만 아니라 다음 어절의 정보도 사용한다면 이러한 문제점을 해결할 수 있을 것으로 생각된다.

그림 14는 어절 범주에 대한 오류율로, 시스템의 오류 중에서 각각의 범주가 다른 범주로 잘못 부착된 경우를 보여준다. 가장 오류가 많은 범주는 Nc 범주이고, V 범주, Nj 범

주, Nv 범주 순이다. 대부분의 오류에서 이 네 가지의 범주가 서로 잘못 부착된 경우가 많았다.

이러한 오류들은 자질들은 다양하게 선택함으로서 해결할 수 있을 것으로 기대된다.

## 5.7 형태소 분석 축소율

본 절에서는 형태소 분석기의 축소율에 대하여 평가하고자 한다. 축소율이란 기존의 형태소 분석 개수와 제안한 시스템의 어절범주를 이용한 형태소 분석 개수의 비율로 정의하였다. 축소율 $R$은 수식 (2)와 같이 정의된다.

$$R = \frac{C}{N} \tag{2}$$

여기에서 $N$은 형태소 분석의 결과 개수이고, $C$는 제안한 시스템의 어절범주와 형태소 분석의 결과의 어절범주가 동일한 것에 대한 개수이다. 예를 들어, 어절 “상해”의 어절범주가 Nc이라고 할 때, 어절 대한 형태소 분석 결과 수는 12개이다. 이러한 분석 결과 중에서 Nc 범주를 가지는 분석 결과가 5개이면, 축소율은 약 40%정도가 된다.

표 7은 두 종류의 말뭉치에 대한 축소율을 보이고 있다. KAIST 말뭉치는 평균 35.6% 정도의 축소율을 보였고, ETRI 말뭉치는 평균 37.2%정도의 축소율을 보였다. 이는 형태소 분석기의 탐색공간과 시간을 축소율만큼 줄일 수 있을 것이다.

표 7 형태소 분석 축소율

Table 7 Reduction ratio of morphological analysis

| 말뭉치 번호 | KAIST | ETRI |
|---|---|---|
| 0 | 36.1% | 38.3% |
| 1 | 35.7% | 37.0% |
| 2 | 36.9% | 38.5% |
| 3 | 35.8% | 37.4% |
| 4 | 38.0% | 36.5% |
| 5 | 34.3% | 36.6% |
| 6 | 36.9% | 37.4% |
| 7 | 34.9% | 34.9% |
| 8 | 32.0% | 36.6% |
| 9 | 35.4% | 38.8% |
| **평균** | **35.6%** | **37.2%** |

# 6제 장 결과 및 향후 연구방향

본 논문에서는 사례기반 학습 방법을 이용한 한국어 어절 분류 시스템을 제안하였다. 사례기반 학습의 자질은 22개를 사용하였다. 자질들은 오토마타 및 명사, 조사 사전을 이용하여 추출하였고, 거의 대부분의 자질들을 음절을 추출한 것이다. 본 논문에서 제안한 어절 분류 시스템은 단순한 자질을 사용함으로 어절 분류를 할 수 있다는 장점이 있다. 또한 자질에 대한 학습을 통하여 보다 낳은 성능을 보였다.

본 시스템의 성능을 평가하기 위하여 두 종류의 말뭉치(KAIST 말뭉치, ETRI 말뭉치)를 사용하였다. 평가 방법으로서 정확도를 측정하였다. 또한 실험에 사용된 두 종류의 말뭉치는 학습에 필요한 사례를 충분히 만들지 못하여 교차 검증 방법을 사용하였다. 본 시스템은 22개의 자질을 사용하였을 경우, 각각 평균 97%와 평균 96.5%를 보였으며, 두 종류의 말뭉치를 합쳤을 경우, 평균 95.9%의 성능으로서 1%정도의 성능 차이를 보였다. 이는 두 종류의 말뭉치의 장르가 다르기 때문이라고 생각된다. 또한 최적 자질을 선정하기 위한 실험에서 16개의 자질을 선택하여 시스템의 성능을 평가했을 경우, 평균 0.2%정도의 성능 향상을 보였다. 또한 본 시스템을 형태소 분석기에 적용해 보았을 경우, 어절 범주를 사용하지 않은 분석결과보다 평균 35% 정도의 축소율을 보였다.

본 논문에서는 22개의 자질을 임의로 선택하여 사용하였다. 이는 자질간의 간섭을 가져올 수 있고, 중복된 자질을 포함하고 있다. 이를 개선하기 위해서는 한 어절 앞의 자질뿐만 아니라 한 어절 뒤의 자질도 포함을 해야할 것이다. 또한 최적 자질의 선택에서 자질 집합을 선택할 때 정보 이득 방법 등의 개선된 자질 선택 방법이 필요하다. 또한 성능 개선 방향으로서는 어절 분류 후 변형 기반의 오류에 의한 학습 방법을 사용함으로 성능 향상을 가져올 수 있을 것이다. 마지막으로 본 어절 분류 시스템을 형태소 분석 및 품사 태깅, 정보 추출, 정보 검색 등의 다양한 분야에 응용함으로서 더욱 향상된 시스템을 구

현할 수 있을 것으로 기대된다.

# 참고 문헌


김영택 외, 2001. *자연언어처리*, 생능출판사.

김재훈, 김길창, 1995. *한국어에서의 품사 부착 말뭉치의 작성요령: KAIST 말뭉치*, 한국과학기술원, 전산학과, CS-TR-95-99.

김재훈, 서정연, 김길창, 1995. *실용적인 한국어 형태소 해석*, 한국과학기술원, 전산학과, CS-TR-95-98. http://nlplab.kmaritime.ac.kr/demo/f_kma.html

엄재홍, 장병탁, 2000. "대규모 문서 데이터 집합에서 Q&A를 위한 질의문 분류 기법", *2000 봄 학술발표논문집*, B권, 한국정보과학회, pp. 253-255.

이현아, 이원일, 임선숙, 허은영, 이재성, 차건희, 박재득, 1999. "표준안에 따른 품사 부착 말뭉치 구축", *제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집*, pp.40-43.

조규빈, 1986. *하이라이트 고교문법*, 지학사.

Brill, E., 1995. "Transformation-based error-driven learning and natural language processing:A case study in part-of-speech tagging", *Computational Linguistics,* Vol. 21, No. 4, pp. 543-565.

Buchholz, S. and Daelemans, W., 2001. "Complex Answers: A Case Study using a WWW Question Answering System", *Natural Language Engineering*, Special Issue on Question Answering.

Daelemans, W. and van den Bosch, A. and Weijters, T., 1997. "IGTREE: Using Trees for Compression and Classification in Lazy Learning Algorithms", D. Aha(ed.), *Arificial Interlligence Review* 11, Special issue on Lazy Learning, Kluwer Academic Publishers.

Daelemans, W. and Zavrel, J. and van der Sloot, K., 2001. *TiMBL: Tilburg Memory-Based Learner*, version 4.1, Reference Guide, ILK Technical Report ILK-0104, Tilburg University, http://ilk.kub.nl

Dasarathy, B. V., 1991. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, McGraw-Hill Companies Inc.

Fujii, A. and Ishikawa T., 2001. "Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration", *Computers and the Humanities*, Vol. 35, No. 4, pp. 389-420.

Haykin, S., 1998. *Neural Networks,* second edition, Prentice-Hall Inc.

Manning, C. D. and Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*, The MIT Press.

Mitchell, T., 1997. *Machine Learning*, McGraw-Hill Companies, Inc.

Quinlan, J. R., 1986. "Induction of decesion tree", *Machine Learning*, Vol. 1, pp. 81-106.

Quinlan, J. R., 1993. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc.

TREC-10, 2001. *Proceedings of the Text REtrieval Conferences*. http://trec.nist.gov