Journal of the Research Institute of Industrial Technology. Vol. 20, 2003

# 한국어에서 어절 범주의 정의 및 태깅

김재훈* · 박호진**

*한국해양대학교 기계 · 정보공학부 조교수

코리아오이즈닛 기술연구소 연구원

# Defining and Tagging Eojeol Tags in Korean

*Jae-Hoon Kim** · *Ho-Jin Park***

**Division of Mechanical and Information Engineering, National Korea Maritime University, Busan 606-791, Korea*

***Research Institute, Korea WISEnut, Inc., Seoul 137-130 Korea*

**요 약** : 본 논문은 어절에 대한 범주를 정의하고 정의된 범주를 이용한 한국어 어절 범주 태깅 방법을 제안한다. 본 논문에서는 기계학습의 일종인 사례기반 학습을 이용하며, 기계학습에 필요한 자질은 벡터로 표현되고, 자질은 사전과 오토마타를 사용하여 반자동으로 선정되고 자동으로 추출된다. 본 시스템의 성능을 좀더 객관적으로 평가하기 위하여, 공개된 두 가지 말뭉치(KAIST 말뭉치, ETRI 말뭉치)를 사용하였다. 평가 방법으로는 말뭉치의 부족 관계로 교차 검증 방법을 사용하였다. 본 시스템은 22개의 자질에 대해서 약 97%의 정확률을 보였다

**핵심용어** . 어절 범주, 태깅, 교차 검증, 사례기반 학습, 기계학습

**ABSTRACT** : *In this paper, we present the definition of the eojeol category and the method for eojeol category tagging. We use a machine learning technique call instance-based learning. We semi-automatically select feature for the machine leaning and automatically extract the features using dictionaries and finite-state automata. To evaluate our proposed system in an objective view, we use two publically available corpus, KAIST and ETRI, that are part-of-speech tagged in Korean and use the cross-validation evaluation to compensate it for data-sparseness of training corpus. The accuracy is about 97% under 22 features*

**KEY WORDS** : eojeol category, tagging, cross-validation, instance-based learning, machine learning

## 1. 서 론

한국어 정보처리를 위해서 필수적으로 요구되는 시스템은 형태소 분석기이다. 형태소 분석은 주어진 문장으로부터 의미의 기본 단위가 되는 형태소를 찾는 과정이며, 형태소 분리, 불규칙 현상이나 굴절 현상에 대한 원형 복원 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 얻는다[1]. 형태소 과잉분석의 원인은 여러 가지가 있으나, 가장 중요한 원인 중 하나는 형태소 배열규칙(morphotactics)에서 제약조건이 부족하기 때문이다. 본 논문에서는 형태소를 분석하기 전에 한국어 어절에 대한 어절범주를 결정하여, 형태소 분석기의 탐색공간을 줄이기 위해 형태소 배열규칙의 제약조건을 강화하기 위한 일환으로 어절 범주를 사용하고자 한다. 본 논문은 그 첫 번째 단계로서 어절의 범주를 결정하고 결정된 어절범주를 이용하여 어절범주를 태깅하는 방법을 제안하다. 본 논문에서는 기계학습의 한 종류인 사례기반 방법을 이용해서 어절을 태깅한다. 사례기반 방법은 기존의 사례 혹은 예제들을 학습하고, 학습된 사례 혹은 예제들을 이용하여 새로운 사례 혹은 예제를 분류하거나 태깅하는 방법이다. 이러한 사례들은 일반적으로 벡터로 표현되며, 이를 자질벡터(feature vector)라고 한다. 본 논문에서는 오토마타 및 사전을 이용하여 자질벡터를 구성한다.

본 논문의 구성은 다음과 같다. 제2장에서는 관련연구로 사례기반 학습 방법에 대해서 간단히 소개하고, 제3장에서는 본

* jhoon@mail.hhu.ac.kr 051)410-4574

** hanwool@wisenut.co.kr

논문과 관련된 한국어의 특징과 어절범주를 정의한다. 제4장에서는 본 논문에서 제안한 한국어 어절범주 태깅 시스템에 대하여 구체적으로 기술한다. 제5장에서는 제안된 시스템의 성능을 평가하고, 마지막으로 제6장에서는 결론과 향후 연구방향에 대해서 기술한다.

## 2. 관련 연구

분류(classification) 혹은 태깅(tagging)은 주어진 환경으로부터 태그 혹은 범주를 결정하는 것이며, 일반적으로 기계학습 방법을 많이 사용한다. 일반적으로 널리 사용되는 기계학습 방법으로는 나이브 베이즈(naive Bayes)[2], 결정 트리(decision tree)[3], 신경망(neural network)[4], $k$-최근법(k-nearest neighbor)[5], 사례기반(instance-based) 방법[6], 변형기반(transformation-based) 방법[7] 등이 있다. 본 논문에서는 사례기반 방법을 주로 이용하는데 본 절에서는 사례기반 학습 방법에 대해서 좀더 자세히 소개하고자 한다.

사례기반 학습은 유사도기반(similarity-based) 혹은 예제기반(example-based) 학습이라고도 불리며 지도 학습(supervised learning) 방법의 일종이다. 학습 과정은 빠른 검색을 위해서 유사한 예제를 군집화하거나 여러 가지 방법으로 색인하여 적절한 형식으로 사례를 저장한다. 분류 과정에서는 입력이 주어졌을 때, 저장된 사례와 가장 비슷한 사례를 추출하고, 추출된 사례의 부류를 입력의 부류로 결정하는 방법이다. 이 방법은 단어 발음변환, 품사 태깅, 전치사구 부착, 명사구 추출 등과 같은 자연언어 처리 분야[6]에 두루 사용되고 있다. 이 방법에서 입력 $X=(x_1, x_2, \ldots, x_n)$와 사례 $Y=(y_1, y_2, \ldots, y_n)$은 특별한 의미를 지닌 자질로 구성된 패턴이며, 두 패턴의 유사도 $\triangle(X, Y)$는 식 (1)과 (2)와 같이 정의된다.

$$\triangle(X, Y) = \sum_{i=0}^{n} w_i \, \delta(x_i, y_i) \tag{1}$$

$$\delta(x_i, y_i) = \begin{cases} \dfrac{x_i - y_i}{\max_i - \min_i} & \text{if } i\text{번째 자질} = \text{숫자, else} \\ 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases} \tag{2}$$

여기서 $\max_i$와 $\min_i$는 각각 $i$번째 자질이 가질 수 있는 최대값과 최소값을 의미하고, $w_i$는 $i$번째 자질의 가중치이다. 일반적으로 식 (3)과 같은 이득률(gain ratio)을 주로 사용하지만, 여러 가지 가능한 가중치 방법이 있다.

$$w_i = \frac{H(C) - \sum_{v \in V_i} P(v) \times H(C|v)}{-\sum_{v \in V_i} P(v) \log_2 P(v)} \tag{3}$$

$$H(C) = -\sum_{c \in C} P(c) \log_2 P(c) \tag{4}$$

여기서 $C$는 부류 집합이며, $V_i$는 $i$번째 자질에 대한 자질 값 집합이다. 식 (2)에서 기호 자질일 경우, 단순히 일치 여부만 유사도에 반영되고(일치도 측정법(overlap)), 어떤 자질이 특정 부류에 얼마나 기여하는지에 대해서는 유사도에 반영되지 않았다. 특정자질이 특정 부류에 기여하는 정도를 반영하는 유사도 측정법이 식 (5)와 같이 정의된다(기여도 차 측정법(value difference)).

$$\delta(x_i, y_i) = \sum_{k=1}^{n} |P(c_k|x_i) - P(c_k|y_i)| \tag{5}$$

이는 자질 내에 어떤 자질값이 특정 부류와 밀접한 관계를 가지고 있을 때, 좋은 결과를 가져오며, 모든 자질값이 모든 부류에 고른 영향을 줄 경우에는 일치도 측정법과 비슷한 결과를 가져온다.

사례기반 학습을 이용한 시스템의 개략적인 흐름도는 그림 1과 같다. 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.

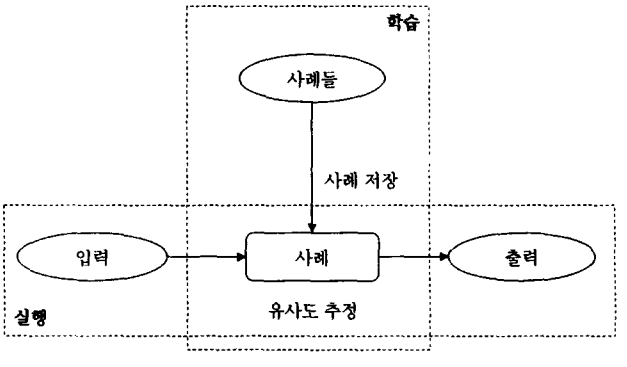


그림 1. 사례기반 시스템 흐름도

## 3. 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[8].

- 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[9].

체언절은 3개의 세부범주, Nc(명사), Nj(명사 접사?[1] 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.

표 1 한국어 어절범주

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

#### 3.2.1. 체언절

체언은 문장의 주체적인 역할을 하며, 여기에는 명사, 대명사, 수사를 포함한다. 한국어 텍스트에는 영어, 한자 등을 포함한 외국어들이 많이 포함되는데 이도 명사로 간주한다. 체언절은 3개의 세부범주, Nc, Nj, Np로 나뉘어지며, 아래에서 각 세부범주에 대한 구체적인 예를 보일 것이다.

- Nc : '생산적인 **국회**' 라는 말이 있다.
- Nj : **이산화유황은** 바로 **산성비의 원인이** 된다.
- Np : **그것은** 한국정치의 변동에 따른 국회의 대처자세다.

#### 3.2.2. 용언절

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

- Nv : 북한의 적극적 참여를 **요망한다**.
- V : 매우 **바람직한** 일이다.
- 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 나뉘어지며, 아래에서 각 세부범주에 대한 구체적인 예를 보일 것이다.

- M : 그러나 **이** 금액은 어느 의미에서 상징적인 숫자이다.
- A : 중소기업 **역시** 할 일이 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**!

#### 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.

## 4. 한국어 어절범주 태깅 시스템

### 4.1. 시스템 구성

본 논문에서 제안하는 어절범주 태깅 시스템은 사례기반 학

1) 기호 ?는 정규표현식을 의미하는 것으로 앞에 오는 객체가 0번 혹은 1번 출현할 수 있음을 의미한다 또한 이하에서 기호 *도 정규표현식을 의미하며, 객체가 0번 이상 나타남을 의미한다.

습을 통하여 주어진 어절에 어절범주를 할당하는 시스템이다. 본 논문에 사용되는 기본 자질은 전문가의 경험에 의해서 결정되었으며, 자질에 대한 구체적인 설명은 4.2.3절에서 기술될 것이다. 그림 2는 구현된 시스템의 개념도이며, 크게 학습단계와 실행단계로 구분된다. 학습단계는 주어진 자질에 대한 최적화된 예제 색인을 구하고 실행부는 학습부에서 구해진 최적화된 예제 색인을 사용하여 입력 문장의 어절에 대한 어절의 범주를 결정한다.

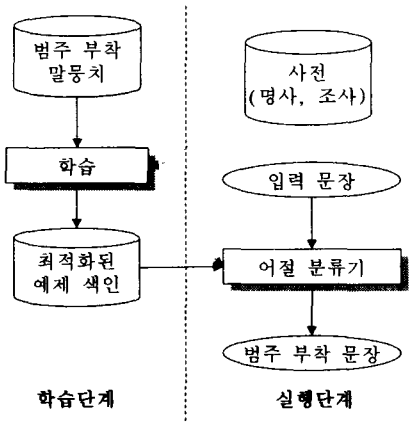


그림 2 어절범주 태깅 시스템의 개념도

## 4.2. 학습단계

학습단계에서는 품사 부착 말뭉치로부터 최적화된 예제 색인을 얻는 단계이며 좀더 구체적으로 설명하면 다음과 같은 절차로 이루어진다.

1) **전처리**: 주어진 어절에 기호가 포함될 경우 분리한다.
2) **어절범주 할당**: 분리된 어절에 대한 어절 범주를 할당한다. 이 과정을 통해서 그림 2의 범주 부착 말뭉치가 생성된다.
3) **자질 추출**: 주어진 어절에 대한 자질을 추출한다.
4) **학습**: 추출된 자질을 이용해서 사례기반 학습을 수행한다.

### 4.2.1. 전처리

표 2 내부 범주

| 범주 | 세부범주 | 어절구성 | 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 |

전처리기는 주어진 어절을 음절과 기호로 분리한다. 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.

### 4.2.2. 어절범주 할당

전처리 과정을 통해 분리된 조사, 어미 등은 3.2절에서 정의한 어절범주를 할당할 수 없기 때문에 시스템 내부에서 추가로 시스템 내부 범주를 표 2와 같이 정의한다. 전처리 과정을 통해서 분리된 어절에 간단한 오토마타를 이용해서 표 1과 표 2의 어절범주를 할당한다[9]. 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.

### 4.2.3. 자질 추출

어떤 자질을 선택하느냐는 시스템의 성능에 큰 영향을 준다. 본 논문에서는 전문가의 경험에 의해서 22개의 자질을 선택하였다(표 3). 표 3에서 1번 자질은 입력 어절의 어절 범주가 Nv이면 자질값으로 'Nv'를 가지고, 그 외의 경우는 '='를 가진다. 2번과 3번 자질은 이전 어절의 음절 및 범주를 자질로서 사용

표 3 어절 자질표

| 자질번호 | 의미 | 자질값 |
|---|---|---|
| 1 | Nv 정규표현식 | 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 어절이 명사 포함 | N, = |
| 7 | 현재 어절의 마지막 두 음절 | 두 음절 |
| 7 | 현재 어절의 조사 | 조사 |
| 9 | 현재 어절의 길이 | 1-4, >4 |
| 10 | 현재 어절이 심볼 | 심볼 |
| 11 | 현재 어절이 모두 영어 | +, - |
| 12 | 현재 어절이 숫자 포함 | +, - |
| 13 | 현재 어절이 과거형 | +, - |
| 14 | 현재 어절이 '하/되' 포함 | HD, - |
| 15 | 현재 어절이 '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 번째, 네 번째 음절 | 두 음절 |

하는데, 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 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 3은 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

#### 4.2.4. 학습

본 논문에서는 사례기반 학습 도구로 TiMBL(Tilburg Memory-Based Learner) 도구[6]를 사용하였다. TiMBL의 입력은 자질 벡터이며, 출력으로 최적화된 예제(자질 벡터)에 대한 색인이다. 본 논문에서 자질벡터는 4.2.3에서 설명한 자질 추출 과정에 의해서 생성된다 사례기반 학습에 대한 자세한 설명은 2장에 기술되었다.

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

5번 자질 추출 과정

| 어절 | 명사 사전 | 자질값 |
|---|---|---|
| 석재까지도 | 존재 | 도 |
| 석재까지도 | . 존재하지 않음 | 도 |
| 석재까지도 | : 존재 | 까지도 |
| 석재까지도 | : 존재하지 않음 | 까지도 |
| 석재까지도 | : 존재하지 않음 | 까지도 |
| | **최종 자질값** | 까지도 |

8번 자질 추출 과정

그림 3 5번과 8번의 자질값 추출 과정

### 4.3. 실행단계

학습단계의 최적화된 예제 색인을 이용하여 실행단계에서는 입력 어절에 어절범주를 할당하며, 할당된 어절은 범주 정보를 부착한 후, 후처리를 통하여 원래 어절로 복원하여 범주 부착 문장으로 출력한다. 다음과 같은 절차로 이루어진다

1) **전처리**: 주어진 어절에 기호가 포함될 경우 분리한다.
2) **자질 추출**: 주어진 어절에 대한 자질을 추출한다 이 때, 이전 어절의 어절범주가 사용된다
3) **어절범주 태깅**: 현재 주어진 어절의 범주를 결정한다
4) **후처리**: 전처리 과정에서 분리된 어절을 복원한다

실행단계의 전처리 과정은 학습단계에서의 전처리 과정과 같으며, 자질추출 과정의 경우에도 태깅된 자질이 피드백되는 것 외에는 같다. 어절범주 태깅 과정은 자질추출 과정에서 생성된 자질 벡터를 이용해서 TiMBL에 입력함으로써 주어진 어절의 범주를 결정할 수 있다

후처리 과정은 전처리 과정에서 분리된 어절을 복원한다. 이 때 어절범주를 어떻게 결합할 것인가가 문제이다 본 논문에서는 이를 해결하기 위해서 기호범주를 제외하고는 어절범주 단계에서 태깅된 어절범주를 그대로 사용한다 기호의 경우에 기호범주 'S' 대신에 기호 자체를 사용한다 예를 들면, 어절 "이산화유황(SO2)의"는 다섯 개의 어절 "이산화유황", "(", "SO2", ")", "의"로 분리되며 각각 어절범주 태깅 결과는 이산화유황:Nc", "(:S", "SO2:Nc", "):S", "의:Jj"이다 이 결과를 후처리 과정을 거치면, "이산화유황(SO2)의:Nc(Nc)Jj"이 되며, 이 결과가 최종 어절범주 태깅의 결과이다.

## 5. 실험 및 평가

### 5.1. 실험 말뭉치

본 논문에서는 성능을 평가하기 위해 두 종류의 평가용 말뭉치를 사용한다. 이 평가용 말뭉치는 품사가 부착되어 있는 말뭉치로서 KAIST 말뭉치[10]는 약 20만 어절, ETRI 말뭉치[11]는 약 30만 어절 정도이다.

### 5.2. 성능 평가 방법

본 논문에서는 제안한 시스템의 성능을 평가하기 위하여 정확도(Accuracy)를 사용하였다[12]. 정확도 $P$는 식 (6)으로 정의된다.

$$P = \frac{A}{N} \tag{6}$$

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

한 성능 평가 방법으로 교차검증(cross validation) 방법을 사용하였다[12] 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. 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

표 4 어절 분류 시스템의 성능

| 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%** |

표 5 최적화된 자질 집합

| 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 | 숫자, = |
| 6 | 현재 어절이 명사 포함 | N, = |
| 7 | 현재 어절의 마지막 두 음절 | 두 음절 |
| 7 | 현재 어절의 조사 | 조사 |
| 9 | 현재 어절의 길이 | 1-4, >4 |
| 10 | 현재 어절이 심볼 | 심볼 |
| 11 | 현재 어절이 모두 영어 | +, - |
| 12 | 현재 어절이 숫자 포함 | +, - |
| 13 | 현재 어절이 과거형 | +, - |
| 14 | 현재 어절이 '하/되' 포함 | HD, - |
| 16 | 현재 어절의 첫 번째 음절 | 한 음절, = |
| 17 | 현재 어절의 두 번째 음절 | 한 음절, = |
| 22 | 현재 어절의 세 번째, 네 번째 음절 | 두 음절 |

### 5.4. 자질 최적화

본 논문에서 사용한 자질은 경험규칙으로 선택한 22개이다 이런 자질들은 간섭이나 불필요한 정보가 포함되어 있다 본 절에서는 자질에 대한 성능 평가를 함으로써, 분류 시스템의 성능을 향상시키고 분류에 필요한 자질을 선택하기 위한 실험이다 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 성능을 저하시키는 자질이 없을 때까지 반복하여, 최적 자질 집합을 선택한다. 표 5는 본 실험에서 선택되어진 최적화된 자질 집합이다

### 5.5. 최적 성능

표 6은 54절의 최적화된 자질 집합을 이용한 어절 분류 시스템의 성능이다 본 실험에서는 이전의 실험보다 평균 02% 정도의 성능 향상을 보인다. 22개의 자질집합에서 최적의 자질집합을 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 부착 과정에서 "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

표 6. 시스템 최적 성능

| 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%)** |

### 5.6. 오류 분석

오류 분석은 시스템의 성능을 개선하는데 많은 도움을 줄 수 있기 때문에 본 절에서는 구현된 시스템에서 발생되는 오류를 분석하였다. 오류 분석을 하기 위하여 최적 자질 집합을 사용하였으며, KAIST 말뭉치와 ETRI 말뭉치를 합하여 분석하였다. 그림 4는 어절 길이에 대한 오류율을 보여주며, 그림 5는 범주에 대한 오류율을 보여주고 있다. 그림 4에서 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 5는 어절 범주에 대한 오류율로, 시스템의 오류 중에서 각각의 범주가 다른 범주로 잘못 부착된 경우를 보여준다. 가장 오류가 많은 범주는 Nc 범주이고, V 범주, Nj 범주, Nv 범주 순이다. 대부분의 오류에서 이 네 가지의 범주가 서로 잘못 부착된 경우가 많았다. 이러한 오류들은 자질을 다양하게 선택함으로써 해결할 수 있을 것으로 기대된다.

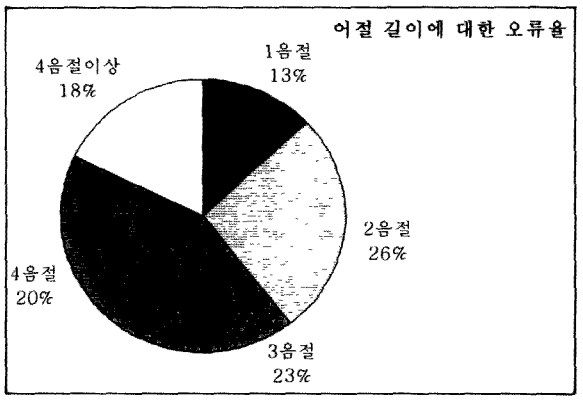


그림 4 어절 길이에 대한 오류율

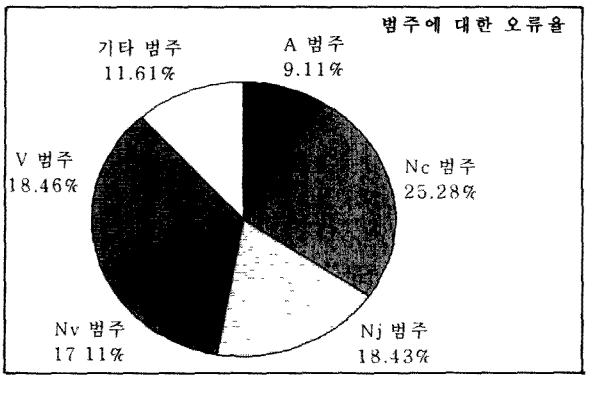


그림 5 범주에 대한 오류율

### 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.

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

여기서 $N$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, 어절에 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

| 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% |

## 6. 결과 및 향후 연구방향

본 논문에서는 기계학습 방법을 이용한 한국어 어절범주 태깅 시스템을 제안하였다. 사례기반 학습의 자질은 22개를 사용하였다. 자질들은 오토마타 및 명사, 조사 사전을 이용하여 추출하였고, 거의 대부분의 자질들은 음절을 추출한 것이다. 본 논문에서 제안한 어절범주 태깅 시스템은 단순한 자질을 사용함으로 어절범주를 태깅할 수 있다는 장점이 있다.

본 시스템의 성능을 평가하기 위하여 두 종류의 말뭉치(KAIST 말뭉치, ETRI 말뭉치)를 사용하였고, 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 보

었다 이는 두 종류의 말뭉치의 성으가 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, 어절범주를 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 구현할 수 있을 것으로 기대된다

## 참 고 문 헌


[1] 김재훈, 서정연, 김길창, 1995 *실용적인 한국어 형태소 해석*, 한국과학기술원, 전산학과, CS-TR-95-98

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

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

[4] Haykin, S., 1998 *Neural Networks, 2*nd edition, Prentice-Hall Inc.

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

[6] 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

[7] 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.

[8] 조규빈, 1986. *하이라이트 고교문법*, 지학사.

[9] 박호진, 사례기반 학습을 이용한 한국어 어절 분류, 한국해양대학교, 컴퓨터공학과, 석사학위 논문, 2002.

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

[11] 이현아, 이원일, 임선숙, 허은영, 이재성, 차건희, 박재득, 1999. "표준안에 따른 품사 부착 말뭉치 구축", *제 1회 형태소 분석기 및 품사태거 평가 워크숍 논문집*, pp.40-43.

[12] Manning, C D and Schutze, H, 1999 *Foundations of Statistical Natural Language Processing*, The MIT Press.


원고접수일 : 2002년 x월 x일
원고채택일 : 2003년 x월 x일