# 품사 사전 자동 학습을 통한 중국어 단어 분할 및 품사 태깅

하주홍[0] 정 옥 이근배
포 항 공 과 대 학 교 컴 퓨 터 공 학 과
{miracle, zhengyu, gblee}@postech.ac.kr

# Chinese Segmentation and POS-Tagging by Automatic POS Dictionary Training

Ju-Hong Ha[0] Yu Zheng Gary G. Lee
Dept. of CSE, POSTECH

요 약

중국어의 품사 태깅(part-of-speech tagging)을 위해서는 중국어 문장들은 내부 단어 간의 명확한 분리가 없기 때문에 단어 분할(word segmentation)과 품사 태깅을 동시에 처리해야 한다. 본 논문은 규칙 기반(rule base)과 사전 기반(dictionary base) 기법을 혼합하여 구현한 단어 분할 시스템을 사용하여 입력 문장을 단어 단위로 분할하고, HMM(hidden Markov model) 기반 통계적 품사 태깅 기법을 사용한다. 특히, 본 논문에서는 주어진 말뭉치(corpus)로부터 자동 학습(automatic training)을 통해 품사 사전을 구축하여 구현된 시스템과 말뭉치간의 독립성을 유지한다. 말뭉치는 중국어 간체와 번체 모두를 대상으로 하고, 각 말뭉치로부터 자동 학습을 통해 얻어진 품사 사전으로 단어 분할과 품사 태깅을 한다. 실험 결과들은 간체, 번체 각각의 단어 분할 성능과 품사 태깅 성능을 보여준다.

## 1. 서론

단어 분할은 중국어 처리 연구에서 가장 기본이다. 중국어 문장들은 영어와 같이 각 단어들 사이를 구분하는 공백이 없고 단순히 문자열들로 표현하기 때문에 인접한 단어들을 분리해 내는 단어 분할이 선행되어야 한다. 중국어에서의 단어는 단일 문자이거나 복수 개의 문자들로 이루어져 있다.

단어 분할이 끝나면 각 단어들에 품사를 할당하게 된다. 품사는 형태소 (여기서는 단어) 해석이나 구문 해석을 하는데 중요한 정보이다. 하지만 어떤 단어들은 여러 개의 품사를 가질 수 있고, 따라서 역시 중국어에서도 애매성(ambiguity)이 발생하게 된다. 이런 애매성을 해결하기 위해서 중국어에서도 품사 태깅 과정이 필요하게 된다.

품사 태깅은 여러 가지 품사를 가지는 단어가 어떤 문장에서 나타날 때, 가장 적합한 품사를 단어에 부여(tag assignment)하는 과정을 말한다. 즉, 입력 문장 $W = w_1 w_2 \Lambda\ w_n$이 주어졌을 때, 문장 $W$에 가장 정확한 태그열 $T = t_1 t_2 \Lambda\ t_n$을 찾는 것이다.

본 논문에서는 중국어 간체와 번체 문장에 대해 단어 분리와 품사 태깅을 순차적으로 수행하는 통합 시스템을 구현하고 성능을 실험하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 중국어 단어 분할과 품사 태깅에 관해 현재까지 수행되어 왔던 관련 연구들을 살펴보고, 3장에서는 본 논문에서 제안한 품사 사전 자동 학습에 대하여 논한다. 4장에서는 구현된 단어 분할 및 품사 태깅 시스템에 대하여 설명한 후, 5장에서는 실험 및 분석을 하며, 마지막으로 6장에서 결론 및 고찰을 기술한다.

## 2. 관련 연구

중국어 단어 분할 방법에는 여러 가지 방법들이 있다. 그 예로 통계적 기법[2], 사전 기반 기법[6, 5] 그리고 두 가지를 함께 사용하는 혼합형 기법[7]이 있다. 통계적 접근 방법 중 가장 흔히 사용하는 방법이 1단계(first order) HMM 이다. 인접한 두 문자들이 하나의 단어를 형성하는데 얼마나 관계를 가지는지를

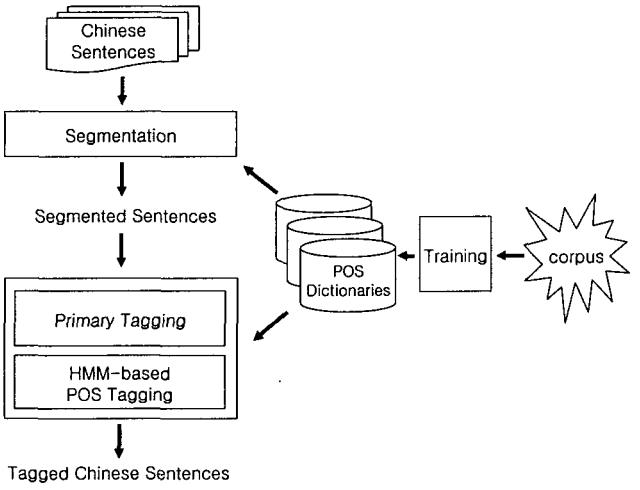


그림 1: 전체 시스템 구성도

인접된 문자들의 빈도로써 판단한다. 단점은 '데이터 빈약 문제(data sparseness problem)' 가 있다. 훈련 과정은 오직 학습 말뭉치에만 의존하는데 신문기사나 소설, 공문서 등은 서로 그 장르가 틀리기 때문이다. 또 문법적으로 당연한 부분에 대해서 정확성이 미치지 못하기도 한다. 사전 기반 기법에도 여러 가지가 있는데 크게 사전만을 사용한 기법과 언어적 지식과 결합한 방법이 있다. 여기서는 사전의 크기와 질이 성능에 중요한 영향을 미친다. 따라서 새로이 나타나는 단어들과 문서들의 장르에 따라 적절히 사전을 갱신해야 한다. 혼합형 기법은 두 기법을 적절히 섞어 사용하기 때문에 정확성이 매우 향상된다. 하지만 수행 속도가 느려진다는 단점이 있다.

본 논문에서는 사전 기반 기법과 통계적 기법을 순서적으로 적용하는 혼합적인 기법을 사용한다.

품사 태깅 방법에도 역시 규칙을 이용하는 방법[4], 확률 정보를 이용한 통계적 방법[1, 2, 3], 두 개의 방법을 혼합한 방법[10]이 있다. 규칙을 이용한 방법은 규칙에 해당하는 문장에 대해서는 높은 정확도를 가지는 장점이 있다. 하지만 정확한 규칙을 추출하기가 매우 힘들고, 제한된 문장들에 대해서만 사용된다는 단점을 가지고 있다. 확률 정보를 이용한 통계적 방법은 대량의 말뭉치에서 확률을 추출하기 때문에 대량의 데이터들을 처리할 수 있어 그 적용 범위가 매우 넓다. 그러나 규칙을 적용했을 때의 정확성을 유지하기가 힘들다. 이 두 방법을 보완하기 위해서 규칙과 통계적 정보를 혼합하여 사용하기도 한다.

본 논문에서는 HMM을 기반으로 한 통계적 방법을 사용한다. 특히, 새로운 말뭉치, 새로운 태그 집합에서 효율적으로 동작하기 위해서 통계적 방법에서 필요한 사전들을 주어진 말뭉치에서 자동 학습을 통하여 생성해 낸다. 따라서, 시스템과 사전 사이의 유연성과 적응성을 최대한 유지하게 된다.

## 3. 품사 사전 자동 구축

앞에서도 언급했듯이 본 논문에서는 시스템과 사전의 유연성과 적응성을 최대한 유지하기 위해서 품사 사전들을 말뭉치에서 자동학습을 통해 생성해낸다. 이를 위해서 사전 자동학습 모듈을 전체 시스템에서 따로 분리시켰다. 따라서 전체 시스템은 말뭉치에 관계없이 사전 자동학습 모듈을 통해 생성된 사전들로 단어 분리와 품사 태깅을 수행하게 된다.

### 3.1. 말뭉치 구성

본 논문에서는 중국어 간체, 번체 모두에 대한 자동 품사 태깅을 위해서 두 가지의 말뭉치를 사용한다.

중국 본토에서 사용되는 간체의 품사 태깅을 위해서 북경 인민일보 1998년 1월부터 6월까지 6개월 분의 기사를 말뭉치[9]로 사용하였다. 약 650만개의 단어로 구성되어 있다. 말뭉치는 수작업으로 단어 형태소 분석과 태깅이 미리 되어 있으며, 여기에서 사용된 품사들은 총 43개로 이루어져 있으며 표 1과 같다. 대만에서 사용되는 중국어 번체의 품사 태깅을 위해서 Academia Sinica의 CKIP의 말뭉치[11]가 사용되었다. 간체 말뭉치와 같이 이미 단어 분할과 태깅이 되어 있고, 약 657만 단어로 구성되어 있다. 사용된 품사들은 총 56개이다(표 2). 특별히 번체 말뭉치의 품사 분류에 있어 동사를 16가지로 세분한 점이 주목할 만한 특징이다.

| 표기 | 내용 | 표기 | 내용 | 표기 | 내용 | 표기 | 내용 | 표기 | 내용 |
|---|---|---|---|---|---|---|---|---|---|
| Ag | 형용사어소 | e | 감탄사 | m | 수사 | p | 전치사 | v | 동사 |
| a | 형용사 | f | 방위사 | Ng | 명사어소 | q | 양사 | vd | 부사성동사 |
| ad | 부사성형용사 | g | 어소 | n | 명사 | Rg | 대명사어소 | vn | 명사성동사 |
| an | 명사성형용사 | h | 접두사 | nr | 인명 | r | 대명사 | w | 부호 |
| Bg | 구별사어소 | i | 성어 | ns | 지명 | s | 위치사 | x | 비어소 |
| b | 구별사 | j | 약어 | nt | 기관명 | Tg | 시간어소 | y | 어기사 |
| c | 연결사 | k | 접미사 | nx | 외래어 | t | 시간사 | z | 상태사 |
| Dg | 부사어소 | l | 습관용어 | nz | 기타전업명 | u | 조사 | | |
| d | 부사 | Mg | 수사어소 | o | 이성이태어 | Vg | 동사어소 | | |

표 1: 간체(북경어) 말뭉치 품사

| 표기 | 내용 | 표기 | 내용 | 표기 | 내용 | 표기 | 내용 | 표기 | 내용 |
|---|---|---|---|---|---|---|---|---|---|
| A | 형용사 | D | 일반부사 | Neqb | 후치수량 관형사 | VC | (p+빈어) 동작타동사 | VK | 상태구목적 어동사 |
| Caa | 병열연결사 | Na | 보통명사 | Nf | 양사 | VCL | 동작지방 빈어동사 | VL | 상태동사 목적어동사 |
| Cab | 열거연결사 | Nb | 전업명칭 | Ng | 접두사 | VD | 쌍빈어동사 | V_2 | "有" 동사 |
| Cba | 구미관련 연결사 | Nc | 지방명사 | Nh | 접미사 | VE | 동작구빈어 동사 | DE | 특수조사 |
| Cbb | 관련연결사 | Ncd | 위치사 | I | 감탄사 | VF | 동작동사 빈어동사 | SHI | "是" 동사 |
| Da | 수량부사 | Nd | 시간사 | P | 전치사 | VG | 분류동사 | FW | 외래어 |
| Dfa | 동사앞 정도부사 | Neu | 수사 | T | 어기사 | VH | 상태자동사 | 기호 | 10가지 |
| Dfb | 동사뒤 정도부사 | Nes | 부정관사 | VA | 동작자동사 | VHC | 상태 사역동사 | | |
| Di | 시태사 | Nep | 지시대명사 | VAC | 동작사역 동사 | VI | 상태타동사 | | |
| Dk | 구부사 | Neqa | 수량사 | VB | 동작타동사 | VJ | (p+빈어) 상태타동사 | | |

표 2: 번체(대만어) 말뭉치 품사

위의 말뭉치들을 사용하여 단어 분할과 품사 태깅에 사용할 품사 사전들을 자동으로 학습하게 되는데, 어휘 사전, 단어품사 확률 사전, 이진(bigram) 품사 확률 사전들을 생성하게 된다. 특히, 어휘 사전 자체를 자동으로 학습함으로써 언어나 도메인에 관계없이 단어 분할 이식률을 높일 수 있다. 자동 학습된 간체 어휘사전의 단어는 약 18만개이고, 번체 어휘사전의 단어는 약 15만개로 이루어져 있다.

### 3.2. 사전 구성

본 논문에서 구현한 시스템의 특징 중에 하나가 시스템과 말뭉치와의 유연성 및 적응성을 유지하고 있다는 것이다. 간체와 번체의 문자를 표현하는 코드 방식의 차이[1] 때문에 처리 모듈간의 차이는 있지만, 말뭉치에서 자동 학습을 통해 사전을 전부 생성함으로써 최대한 시스템이 특정 언어나 도메인 유형에 의존되지 않도록 유연성과 적응성을 유지하였다.

[1] http://www.haiyan.com/steelk/navigator/ref/

본 시스템에서 사용되는 사전들은 문자 빈도 사전, 단어 품사 확률 사전, 어휘 사전, 그리고 이진 품사 확률 사전이 있다. 문자 빈도 사전은 중국어 단일 문자가 가지는 출현 빈도를 값으로 가진다. 단어 품사 확률 사전은 특정 단어가 특정 품사를 가질 확률을 값으로 가진다. 어휘 사전은 미리 생성되어 있는 단어 품사 확률 사전으로부터 생성한다. 단어 품사 사전에서 특정 단어는 하나 이상의 품사와 출현 확률을 가지고 있다. 어휘 사전은 특정 한 단어가 가질 수 있는 모든 품사들을 각 품사의 출현 확률의 내림차순으로 결합된 문자열을 값으로 가진다. 예로 간체 단어 过는 어휘 사전에서 다음과 같은 형태를 취한다.

key : 过

value :

=₩u|=₩n|=₩nr|=₩Ng|=₩v|=₩d|=₩vn

값에서 간체 말뭉치로부터 자동 학습을 통해 过가 가질 수 있는 품사 중 출현 확률이 가장 높은 것이 품사 u이고, 가장 낮은 것이 품사 vn임을 나타내고 있다.

마지막으로 이진 품사 확률 사전은 말뭉치에서 특정 두 품사가 연속적으로 나타나는 확률을 값으로 가진다. 단어 품사 확률 사전과 이진 품사 확률 사전은 모두 자연로그를 취한 확률을 값으로 가진다.

## 4. 전체 시스템

본 논문에서 구현한 단어 분할과 품사 태깅 통합 시스템은 그림 1에서 그 전체적인 구성도를 보여주고 있다. 입력으로 들어오는 문장들이 중국어 간체 문장들인지, 번체 문장들인지에 따라 각각의 말뭉치에서 자동 학습시킨 사전들을 사용하여 단어 분할과 품사 태깅을 거쳐 최종으로 단어들이 분할되고 각 단어 마다 품사가 태깅된 문장들을 출력하게 된다. 지금부터 세부적인 시스템 구성을 살펴보기로 한다.

### 4.1. 단어 분할

본 논문에서 구현한 단어 분할 시스템은 그림 2에서 보는 것과 같다.

입력된 중국어 문장은 전처리 분할(pre segment)[8]을 거치게 된다. 전처리 분할은 특수 단어들을 이용하여 문장을 짧은 문자열로 분할하는 모듈이다. 특수 단어는 표기, 숫자, 외국어, 기타 비 중국어 부호와 출현 확률이 높고 단어 조합 능력이 낮은 한자(e.g. 的)를 말한다. 예들 들어,

他是中国人，他的生日是5月13日。

과 같은 문장이 입력되어 전처리 분할을 거친 결과는

他是中国人/，他的生日是/5/月/13/日。

로 /에 의해 분할된다.

이렇게 전처리 분할 과정을 거친 후에는 말뭉치에서 자동 학습된 어휘사전을 기반으로 하여 전처리 분할 결과의 짧은 문자열에서 가능한 모든 단어를 추출한 후 Chih-Hao Tsai가 구현한 단어 분할 시스템[5]에서의 4가지 휴리스틱 규칙들을 따라 단어 분할 과정이 진행되게 된다. 먼저 위에서 추출한 단어들에서 오른쪽으로 최장 일치되는 3개의 단어들을 하나의 묶음(chunck) 단위로 묶고, 최장 길이의 묶음에서 처음 단어를 선택한다. 최장 길이의 묶음이 여러 개이면 다음 규칙으로 넘어간다. 두 번째 규칙은 평균 단어 길이가 최장인 묶음의 첫 단어를 선택하고, 평균 단어 길이가 최장인 묶음이 여러 개이면 다음 규칙을 적용한다. 세 번째 규칙은 단어 길이의 변화가 가장 작은 묶음에서 첫 단어를 추출한다. 이러한 묶음도 여러 개가 존재한다면 마지막 네 번째 규칙을 적용하게 된다. 여기에서는 자동 학습을 통해 생성된 중국어 문자들의 출현 빈도 사전을 이용하여 각 묶음 내의 각각의 문자들의 빈도 합이 가장 큰 묶음의 첫 단어를 추출하게 된다. 좀 더 자세한 내용을 위해서는 [5]을 참고하길 바란다.

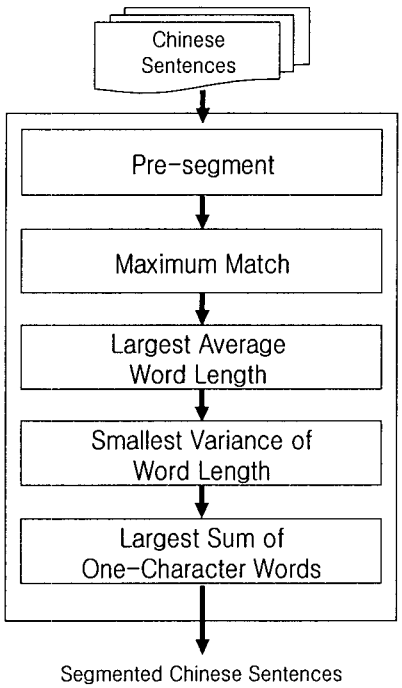


그림 **2**: 단어 분할 시스템

### 4.2. HMM 기반 품사 태깅

단어 분할 시스템의 결과로 얻어진 중국어 문장들은 어휘 사전을 이용해 문장 내에 단어들에 부여될 수 있는 모든 품사들을 할당한다. 이렇게 초기의 태깅된 문

장을 생성한 다음, 각각의 단어들에 문맥상 가장 적합한 품사를 부여하는 것이 얻고자 하는 최종 결과이다. 즉, 주어진 문장 $W = w_1 w_2 \Lambda\ w_n$에 대해 문맥상 가장 적합한 품사들 $T = t_1 t_2 \Lambda\ t_n$을 구하는 것이다. 수식으로 표현하면,

$$T^* = \arg\max_T P(T \mid W) \quad (1)$$

로 표현할 수 있다. (1)에서 $P(T \mid W)$는

$$P(T \mid W) = \frac{P(T)P(W \mid T)}{P(W)} \quad (2)$$

로 나타낼 수 있다. 식 (2)를 Markov 가정에 따라 아래 식 (3)과 (4)에 의해 단순화 되고,

$$P(T) \approx P(t_1)\prod_{i=2}^{n} P(t_i \mid t_{i-1}) \quad (3)$$

$$P(W \mid T) \approx \prod_{i=1}^{n} P(w_i \mid t_i) \quad (4)$$

$P(W)$는 상수이므로, 결국

$$T^* = \arg\max_T \prod_{i=1}^{n} P(t_i \mid t_{i-1})P(w_i \mid t_i) \quad (5)$$

와 같은 식 (5)를 유도할 수 있다. 식 (5)에 따라 Viterbi 검색 알고리즘(Viterbi search algorithm)을 거치게 되면 입력된 문장 속의 단어들에 가장 적합한 품사들로 태깅된 최종 문장을 출력하게 된다.

## 5. 실험 및 평가

구현된 전체 시스템은 그림 1에서 보여 주듯이 단어 분할 모듈, HMM 기반 품사 태깅 모듈, 말뭉치로부터의 사전 자동 학습 모듈 등은 리눅스 기반 C로 구현되었으며, 말뭉치에서 자동 학습한 사전들은 Berkeley DB[2]로 구성되어 있다.

### 5.1. 실험 결과

본 논문에서의 실험은 크게 세 가지로 나눌 수 있다. 첫 번째가 구현된 단어 분할 시스템이 중국어 간체와 번체 각각의 문장에 대해 얼마나 좋은 성능을 보여주는가를 실험한다. 두 번째는 단어로 분할된 문장들에 대한 품사 태깅 시스템에 대한 성능을 살펴보는 실험을 한다. 마지막으로 품사 태깅 시스템에서 $P(T \mid W)$와 $P(t_i \mid t_{i-1})$ 중 어느 확률이 시스템에 더 영향을 미치는가를 살펴보는 실험을 수행한다.

본 논문에서 사용된 말뭉치의 크기가 두 가지 모두 650만 단어로 매우 크다. 사용된 말뭉치가 매우 클 때에는 내부 데이터 실험과 외부 데이터 실험간의 차이가 없기 때문에, 사전 학습에 사용된 말뭉치의 일부를 실험 데이터(inner test)로 사용한다.

성능 평가 기준은 정보검색에서 가장 흔히 사용되는 정확율(precision)과 재현율(recall)을 사용한다.

단어 분할 시스템 성능 실험에서는 간체 50만 단어와 번체 40만 단어를 실험 데이터로 사용하였다. 실험 결과는 표 3과 같다.

| | 간체 (50만) | 번체 (40만) |
|---|---|---|
| 정확율(%) | 98.12 | 98.87 |
| 재현율(%) | 98.59 | 98.95 |

**표 3:** 단어 분할 시스템 실험 결과

실험 결과에서 보는 바와 같이 번체가 약 0.7% 좋은 결과를 보여주고 있다. 나타나는 오류들을 살펴보면 간체에서의 분할 오류 2% 중 50% 이상이 사용된 말뭉치에 오류가 있는 것을 확인할 수 있었고, 번체 말뭉치는 상대적으로 정제가 잘 되어 있음을 확인할 수 있었다. 나머지 오류들은 중국어 고유의 교차 오류(cross ambiguity)로 분석되었다.

두 번째 실험인 품사 태깅 시스템 성능 평가에 사용된 데이터의 크기는 첫 번째 실험에서와 같다. 실험 결과는 아래 표 4와 같다.

| | 간체 (50만) | 번체 (40만) |
|---|---|---|
| 정확율(%) | 95.06 | 95.75 |
| 재현율(%) | 95.06 | 95.77 |

**표 4:** 품사 태깅 시스템 실험 결과

결과 표에서 보는 바와 같이 품사 태깅 시스템에서도 위에서 분석한 말뭉치 오류를 포함하더라도 간체, 번체 모두 95%의 성능을 나타낸다. 간체 결과에서 오류들을 살펴보면 동사(v)와 동명사(vn)간의 오류가 전체 오류의 40% 이상을 차지하는 것을 확인할 수 있었다.

시스템의 성능을 좀 더 향상시키기 위해 마지막으로 시스템이 $P(T \mid W)$와 $P(t_i \mid t_{i-1})$ 중 어느 확률에 영향을 받는지 여부를 확인하는 실험을 하였다. 각 확률에 가중치 $\alpha$와 $1-\alpha$를 곱하고 $\alpha$의 변화에 따라 성능의 변화를 살펴보았다. 실험 결과는 표 5와 같다.

두 가지 모두 $P(t_i \mid t_{i-1})$만 고려했을 때 약 72%의 정확도를 보였고, $P(T \mid W)$만을 고려했을 때는 각각 약 88%와 92%에 가까운 정확성을 보여주는 것으로 볼 때 $P(T \mid W)$이 품사 태깅 시스템 성능에 더 많은 영향을 미친다는 것을 확인할 수 있었다. 또, 가중치 적용 후의 결과에서도 최고의 정확도는 $P(t_i \mid t_{i-1})$의 가중치가 상대적으로 낮을 때 나타남을 볼 수 있었다. 번체는 무려 30%로 가중치를 낮출 때 최고의 정확도를 보임을

[2] http://sparcs.kaist.ac.kr/~neosado/docs/db-3.3.11/reftoc.html

실험에서 확인할 수 있다.

$$\alpha \times \log P(T|W) + (1-\alpha) \times \log P(t_i | t_{i-1})$$

| $\alpha$ | | 1 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 간체 | 정확율 | 88.344 | 90.376 | 92.522 | 93.780 | 94.617 | 95.020 | 94.770 | 93.476 | 90.055 | 82.769 | 72.439 |
| | 재현율 | 88.343 | 90.375 | 92.521 | 93.779 | 94.616 | 95.020 | 94.769 | 93.476 | 90.054 | 82.768 | 72.438 |
| 번체 | 정확율 | 91.710 | 93.902 | 95.094 | 95.753 | 95.328 | 94.328 | 92.030 | 87.656 | 80.460 | 75.056 | 72.172 |
| | 재현율 | 91.743 | 93.923 | 95.109 | 95.767 | 95.366 | 94.350 | 92.060 | 87.677 | 80.448 | 75.100 | 72.172 |

**표 5:** 가중치 부여에 따른 시스템 성능 변화

이전의 연구 결과들을 살펴보면 북경대학[10]에서 구현한 시스템은 본 논문에서 사용한 것과 같은 간체 말뭉치에서 단어 분할 정확율이 99.5%, 품사 태깅 정확율이 96.06%의 성능을 보여주고 있다. 번체의 경우 [2]에서 단어 분할 정확율이 99.80%, 품사 태깅 정확율이 96.06%임을 보여주고 있다. 이는 본 논문에서 구현한 시스템이 보여주는 것보다 조금 우수한 성능을 보여주고 있다. 하지만 본 논문에서는 말뭉치에 나타난 오류들을 거의 수정하지 않았고, 단지 구현된 시스템으로만 얻어진 결과이다. 실험을 통해 분석된 오류들을 수정한다면 좀 더 나은 성능을 보여줄 수 있을 것이다.

마지막으로 한가지 관심을 가지고 살펴볼 사항은 간체의 품사 종류가 43개이고 번체는 56개임에도 불구하고 성능은 오히려 번체가 조금 더 높음을 확인할 수 있었다. 일반적으로 품사의 종류가 많아지면 애매성(ambiguity)이 높아진다고 생각할 수 있다. 그 이유를 살펴보면 간체의 경우 약 18만 단어 중 두 개 이상의 다중 품사를 가지는 단어(multi-tagged word)가 15,662개로 어휘 사전 전체의 8.7%를 차지한다. 반면, 번체는 약 15만 단어 중 다중 품사를 가지는 단어가 6,286개로 4.2%만 차지하고 있다. 이는 번체가 동사를 16개로 세분한 것도 다중 품사 단어들을 줄이는데 영향을 주는 것으로 분석할 수 있다. 따라서, 번체가 품사의 종류가 많음에도 다중 품사 단어가 간체의 50% 수준이기 때문에 본래 내재된 애매성이 줄어 간체에 비해 더 낮은 성능을 보여준다고 분석할 수 있다.

## 7. 결론 및 고찰

본 연구에서는 중국어 처리를 위해서 단어 분할과 품사 태깅 시스템을 구현하고 그 성능을 실험하였다. 단어 분할 시스템에서는 사전 기반 기법과 통계적 기법을 모두 사용한 혼합형 기법을 사용하였고, 품사 태깅 시스템은 HMM 기반의 통계적 기법을 사용하여 지도 학습(supervised learning)을 실시하였다. 데이터에 대한 시스템의 유연성 및 적응성을 유지하기 위해 사전을 말뭉치로부터 자동 학습을 통해 생성하였다. 실험에서는 중국어의 두 가지 종류인 간체 문장과 번체 문장 모두에 대해 성능 테스트를 실시하였다. 두 가지 모두 최종 약 95%의 성능을 보여주고 있으며, 번체가 조금 더 좋은 성능을 보여주고 있다.

결과에서 나타난 오류를 살펴보면 간체의 경우 비슷한 품사간의 오류가 가장 많았으며, 두 가지 모두 말뭉치 내에 오류가 다수 포함된 것으로 확인되었다.

향후 결과 내의 오류의 유형들을 자세히 분석해 오류 후처리를 실시한다면 좀 더 좋은 결과를 얻을 수 있을 것으로 기대된다.

## 6. 참고 문헌


[1] S.H. Bai "THE METHODIC RESEACH OF GRAM MATICAL TAGGING CHINESE CORPUS." In 《机器翻译研究进展》, 408-418(1992)

[2] C.H. Chang and C.D. Chen "A Study on Integra ting Chinese Word Segmentation and Part-of-S peech Tagging", Communications of COLIPS, 3, 1 (1993), 69-77

[3] C.H. Chang and C.D. Chen "HMM-based part-of-speech tagging for Chinese corpora", In Proc. of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, pages 40-47, Columbus, Ohio, USA, June 1993

[4] S.H. Liu, K.J. Chen, L.P. Chang and Y.H. Chin "Automatic Part-of-speech tagging for Chinese corpora", Computer Processing of Chinese and Oriental Languages, Vol 9, No 1, pp31-47(1995)

[5] C.H. Tsai "MMSEG: A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm", http://www.geocities.com/hao510/mmseg/,(1998)

[6] J. Yu and S. Yu "Some Problems of Chinese Segmentation", in the first international workshop on Multi Media Annotation (MMA-2001)

[7] 刘挺,"串频统计和词形匹配相结合的汉语自动分词系统",中文信息学报 Vol12 No1,(1998)
[8] 王永成 "中文词的自动办理。",In Journal of Chinese Information Processing Vol.4 No.4,(1990)
[9] 俞士汶，现代汉语语料库加工——词语切分与词性标注规范与手册，北京大学计算语言学研究所，1999/4
[10] 周强 "□□和□□相□合的□□□□□注方法"，北京大□□算□言□□究所 http://chinese.pku.edu.cn/shisheng.htm (2001)
[11] 中央研究院平衡语料库的内容与说明, institute of information science academia sinica. 98-04