한국어 품사 태깅을 위한 다이내믹 링크 모델

황명진[1] 강미영[2] 권혁철[3]
부산대학교 컴퓨터공학과 한국어정보처리 연구실[1,3]
국립국어원 국어정보화팀[2]
{holgabun[1], hckwon[3]}@pusan.ac.kr
mykang@mct.go.kr[2]

A Dynamic Link Model for Korean POS-Tagging

Myeong-jin Hwang, Mi-young Kang, Hyuk-chul Kwon
Korean Language Processing Lab,
Dept. Computer Science and Engineering, Pusan National University

요 약

통계를 이용한 품사 태깅에서는 자료부족 문제가 이슈가 된다. 한국어나 터키어와 같은 교착어는 어절(word)이 다수 형태소로 구성되어 있어서 자료부족 문제가 더 심각하다. 이러한 문제를 극복하고자 교착어 문장을 어절 열이 아니라 형태소의 열이라 가정한 연구도 있었으나, 어절 특성이 사라지기 때문에 파생에 의한 어절의 문법 범주 변화 등의 통계정보와 어절 간의 통계정보를 구하기 어렵다.

본 논문은 효율적인 어절 간 전이확률 계산 방법론을 고안함으로써 어절 단위의 정보를 유지하면서도 자료부족문제를 해결할 수 있는 확률 모델을 제안한다. 즉, 한국어의 형태통사적인 특성을 고려하면 앞 어절의 마지막 형태소와 함께 뒤 어절의 처음 혹은 끝 형태소-즉 두 개의 어절 간 전이 링크만으로도 어절 간 전이확률 계산 시 필요한 대부분 정보를 얻을 수 있고, 문맥에 따라 두 링크 중 하나만 필요하다는 관찰을 토대로 규칙을 이용해 두 전이링크 중 하나를 선택해 전이확률 계산에 사용하는 '다이내믹 링크 모델'을 제안한다.

형태소 품사 bi-gram만을 사용하는 이 모델은 실험 말뭉치에 대해 96.60%의 정확도를 보인다. 이는 같은 말뭉치에 대해 형태소 품사 tri-gram 등의 더 많은 문맥 정보를 사용하는 다른 모델을 평가했을 때와 대등한 성능이다.

1. 개요

품사 태깅이란 문장이나 구 내에 나타나는 단어에 문법 범주를 부착하는 과정이다. 한국어에서 어절(word)은 보통, 띄어쓰기 단위를 사용하기 때문에 어절이라고 부른다. 한 어절을 형태소 분석해보면 다수의 분석 결과가 나오기도 하므로, 품사 태깅 과정에서 중의성 해소 과정은 필수적이다. 중의성은 해당 단어의 주위 문맥 정보를 이용하여 규칙이나 통계적인 방법으로 해소한다.

통계적인 방법을 사용할 때는, 통계 추출용 말뭉치의 균형성과 함께 자료부족문제가 이슈가 된다. 특히 한국어, 터키어를 비롯한 여러 교착어는, 어절이 다수 형태소로 구성되어 있어서 높은 생산성을 보이므로, 영어처럼 단순한 단어 구조를 가진 언어보다 자료부족문제가 특히 더 심하다.

이러한 자료부족 문제를 해결하고자 선행 연구자들은 어절 대신 형태소처럼 더 작은 단위를 이용하였다.[3] 그러나 문장을, 형태소의 나열로만 본다면 어절 단위의 특성 - 예를 들어 어절 내에서 이루어지는 파생형태소의 교착에 따른 어절 단위 문법범주 변화 정보를 잃게 된다. 따라서 자료부족 문제를 해결하면서, 어절 특성을 살릴 수 있는 적절한 방법이 필요하다.

본 논문에서는, 교착어의 품사 태깅에서 발생하는 이러한 문제점을 효율적으로 해결할 수 있는 '어절 간 전이 확률 계산 방법론'을 제안한다. 이것은 어절 간 전이 확률 계산에 형태소 품사 bi-gram만을 사용하고, 한 구간에서 여러 가지 가능한 bi-gram 중 하나만을 동적으로 선택해 사용하는 모델이다.

2장에서는, 기존의 교착어에 대한 통계기반 연구에서

어절 간 전이 확률을 어떻게 구했는지 알아본다. 3장에서는 어절 간 전이확률 계산 시 발생하는 문제점을 분석하고, 효율적인 전이확률 계산 모델을 제안한다. 4장에서는 관련 실험 후 제시한 모델의 성능을 분석하고, 5장에서 결론을 짓는다.

2. 관련 연구

2.1. 품사태깅 관련 일반 모델

품사 태깅은 단어열 $W=w_{1;n}=w_1, w_2, \cdots, w_n$가 주어졌을 때, 각 단어가 가질 수 있는 품사 후보 중 적당한 품사를 선택하여, 품사열 $T=t_{1;n}=t_1, t_2, \cdots, t_n$를 찾는 과정이라 할 수 있다. 대부분의 기존 통계 모델들은 Hidden Morkov model (HMM)에 기반을 두어 확률 $P(T|W)$를 최대화하는 변수 T를 찾는다.

$$\arg\max_T P(T|W)$$

$$= \arg\max_T \frac{P(T)\times P(W|T)}{P(W)} \quad (1)$$

$$= \arg\max_T P(T)\times P(W|T) \quad (2)$$

수식 (1)에서 (2)로 넘어가는 과정에서 $P(W)$는 모든 T에 대해서 상수이므로 제거했다. 그리고 이 수식에 다음 가정과 연쇄법칙(chain rule)을 적용한다.[6]

- 특정 단어는 그것의 품사에만 의존한다.
- 특정 단어의 품사는 그 이전 품사에만 의존한다.

$$P(W|T) = P(\prod_{i=1}^{n}(w_i|t_i)$$

$$P(T) = P(\prod_{i=1}^{n}(t_i|t_{i-1}) \quad (3)$$

그러므로 수식 (2)는 다음 식으로 나타낼 수 있다.

$$\arg\max_T P(T|W) =$$

$$\arg\max_T \prod_{i=1}^{n} P(t_i|t_{i-1})\times(w_i|t_i) \quad (4)$$

수식 (4)는 영어 같은 언어를 위한 품사 태깅 기본 수식이다.[4] 그런데 교착어에서는 $P(t_i|t_{i-1})$와 $P(w_i|t_i)$를 추출하기 쉽지 않다. 교착어에서의 단어, 즉, 어절은 하나 이상의 형태소로 구성되어 있기 때문이다.

$$t_i = mc_{i;1,n} = mc_{i;1}, mc_{i;2}, ..., mc_{i;k}$$

$$w_i = c_{i;1,n} = c_{i;1}, c_{i;2}, ..., c_{i;k} \quad (5)$$

따라서 교착어에서 어절 확률 $P(w_i|t_i)$과 전이확률 $P(t_i|t_{i-1})$을 어떻게 구해야 할 것인지에 대해 고민해야 한다. 이 중 어절 확률은 강미영이 [7]에서 제시한 방법론을 그대로 사용한다. 본 논문에서는 어절 간 전이확률을 어떻게 구할 것인지에 대해서 알아본다.

2.2. 교착어와 전이확률 계산의 복잡성

그림 1에서 볼 수 있는 것처럼 교착어에서 한 어절은 다수 형태소로 구성된다.

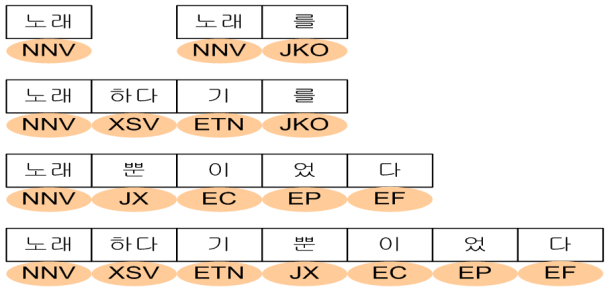

그림 1. 다수 형태소로 구성된 교착어 어절의 예
(약어는 부록 1을 참조할 것)

전이확률을 구하려는 방법으로 먼저 생각해 볼 수 있는 것은 수식(6)이다. 그러나 이 방법은 교착어의 특징인 어절의 높은 생산성 탓에 자료부족 문제가 생긴다.

$$P(t_i|t_{i-1}) = P(mc_{i;1,k}|mc_{i-1;1,l}) \quad (6)$$

기존 연구들([3])에서는 통계의 기본 단위를, 어절보다 더 작은 단위인 형태소나 음절 단위로 처리함으로써 자료부족 문제를 해결하려 하였다. 그러나 문장을 단순한 형태소 열로만 생각하면 어절의 문법적 특성(정보)을 읽게 되고(3장에서 구체적으로 설명), 수식 (7)처럼 하면 계산량이 많아지는 문제점이 있다.

$$P(t_i|t_{i-1}) = \sum_{j_1=1}^{k}\sum_{j_2=1}^{l} P(mc_{i;j_1}|mc_{i-1;j_2}) \quad (7)$$

교착어의 전이확률을 얻기 위한, 자료부족 문제도 없고 계산량도 적은 방법, 그러면서도 필요한 정보를 모두 얻을 수 있는 그러한 계산 방법을 찾아야 한다.

기존에 시도된 통계기반 품사 태깅 모델에서 한국어의 교착어적 특성에 맞는 전이확률을 제안한 대표적인 연구로 [2]에서 제안한 Twoply HMM을 들 수 있다.

이 모델에서 어절 내 확률은 형태소를 단위로 하는 일반적인 HMM을 사용하고, 어절 간 전이확률은 두 가지의 형태소 품사 bi-gram을 사용한다. 어절 간 전이확률 계산 방법을 좀 더 구체적으로 살펴보면, 앞 어절의 마지막 형태소와 뒤 어절의 첫 형태소 간 전이확률과 앞 어절의 첫 형태소에서 뒤 어절의 마지막 형태소 간 전이확률의 곱을 이용하였다.

한편, [5]는 또 다른 교착어인 터키어에서 효율적으로 어절 간 전이확률을 구하려고, 어절을 어근과, 파생경계(derivation boundary)로 분리되는 굴절그룹

(inflection group)의 열이라고 두어 구조를 단순화하였다. 또한, 어절 간 문법적 관계를 단순화하려고 두 가지 가정을 도입하였다. 현재 어절의 어근은 앞 두 어절의 어근에만 영향을 받고, 현재 어절의 각 굴절그룹은 앞 두 어절의 마지막 굴절그룹에만 영향을 받는다는 것이다.

[1]은 어절 간 전이 확률을 구하고자, 앞 어절의 마지막 형태소와 뒤 어절의 마지막 형태소 간의 전이확률을 이용하였다. 이때, 뒤 어절의 마지막 형태소는, 같은 어절 내의 다른 형태소들에 대한 정보를 부분적으로 제공하도록 하위범주화하였다. 예를 들어 수의존명사, 부사, 명사형 전성어미, 보조용언 등이 있었는지 없는지에 따라 어미와 조사 각각을 5종류로 하위범주화하였다.

기존 연구들에서는 어절 특징을 제대로 고려하지 못해 전이 정보가 빈약하거나,[2] 필요 이상의 계산을 많이 하며,[5] 이를 해결하고자 한 방법도 추가의 하위범주화 등으로 말미암아 그만큼의 통계 자료가 더 필요하다.[1] 따라서 이 논문은 교착어 품사 태깅에서 계산량을 줄이고 자료부족 문제를 해결하면서도 필요한 정보를 모두 얻을 수 있는 전이확률 계산 방법론을 제안한다.

3. 전이확률 계산의 최적화

3.1. 유용한 정보와 유용하지 않은 정보 분리

2장에서 교착어의 전이확률 계산 때 생기는 문제점 - 자료부족문제와 많은 계산량에 대해 살펴보았다. 이런 문제들은 한 어절이 다수 형태소로 구성되는 교착어의 특성 때문이다. 그러나 어절을 구성하는 모든 형태소들이 전이확률 계산에 필요한 정보를 제공하지는 않는다. 만약 어절 내 형태소들 중 유용한 몇 개만을 사용하여 전이확률을 계산할 수 있다면, 자료부족 문제를 해결하면서 계산량도 줄일 수 있다.

어절 간 전이확률 계산 시 유용한 정보가 무엇인지 알아보고자, 이 장에서는 우선 파생으로 경계 지어지는 문법그룹을 단위로 유용한 형태소 후보의 범위를 정의하고 유용한 형태소 단위를 분석한다.

3.1.1. 어절 내 문법그룹

한국어에서 한 어절을 이루는 형태소의 결합 순서는 그림 2의 상태도로 나타낼 수 있다. 이 상태도를 따르면 어절은 어근, 파생부, 굴절부로 구성된다. 굵은 화살표는 어근, 파생부, 굴절부 단위의 결합 순서를 나타내고 있고, 가는 화살표는 품사 단위의 세부적인 결합 순서를 나타내고 있다. 종료 상태는 그림의 복잡도를 줄이려고 괄호의 유무로만 구분했다. 괄호에 싸인 품사는 어절에서 가장 마지막에 있을 수 없다.

그림 2에 나타난 파생부는 이전 상태의 문법적 특성

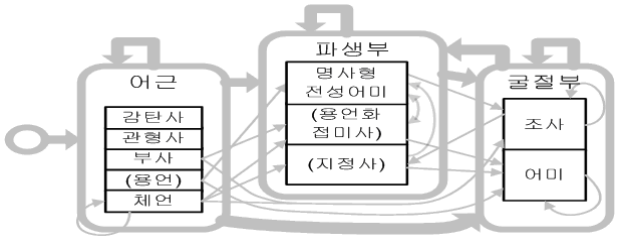

그림 2. 한국어 어절을 구성하는 형태소의 결합 상태도 (이 상태도는 보조사, 보조용언, 선어말어미 등과 관련하여 추가해야 할 부분이 있지만, 낮은 출현빈도, 그림의 복잡도, 논문 가독성을 고려해 생략했다. 그림에서 어미는, 명사형 전성어미를 제외한 나머지 어미만 해당된다.)

을 변화시킨다. 즉, 현재 어절의 문법 범주를 결정하는 것이다. 따라서 어절을 파생부를 이용해 문법적 특성이 다른 그룹으로 분리할 수 있으며, 본 논문에서는 이 그룹을 '문법그룹'이라 부른다. 그림 3은 그림 2를, 문법그룹을 이용해 간략화한 것이다. 각 문법그룹은 그들의 첫 형태소(어근이나 파생 형태소)의 문법 범주를 승계한다.

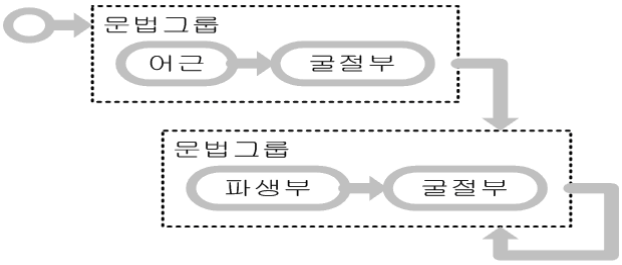

그림 3. 어절 내 문법그룹의 상태도

3.1.2. 전이확률 계산에서 유용한 문법그룹

문법그룹을 단위로 전이확률을 구할 때, 만들어질 수 있는 링크는 그림 4에 나타나 있다. 이렇게 여러 링크가 만들어질 수 있지만, 유용한 문법그룹만 골라낸다면 많은 링크를 제거할 수 있을 것이다. 그러면 유용한 문법그룹이 어떤 것인지 살펴보겠다.

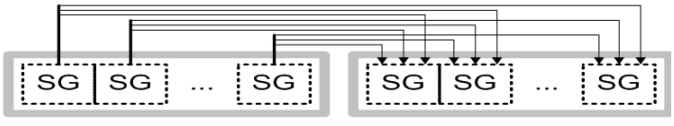

그림 4. 문법그룹(SG)을 사용한 어절 간 전이 링크

[5]에서 기술한 터키어의 특징 중에 "앞 어절에서는 가장 마지막 굴절그룹만이 뒤 어절과 문법적 관계를 형성하는 것으로 관찰된다."라고 언급하고 있다. 여기서 굴절그룹이라는 것은 본 논문에서 설명하는 문법그룹과 유사한 의미이다. 터키어의 이러한 특징은 한국어에서도 발견된다. 명사형 전성어미와 일부 보조사를 제외한

대부분 어미와 조사는 어절의 마지막에서 뒤 어절과 문법적 관계를 형성한다. 어미나 조사가 붙지 않은 관형사나 부사는 단독으로 어절을 구성(어절의 마지막 형태소)하면서 뒤 어절과 문법적 관계를 형성한다. 어절 마지막의 체언이나 명사형 전성어미도, 그 앞 형태소 중에 용언이 있었다 하더라도 뒤 어절과의 문법적 관계에서 앞 어절을 체언의 역할을 하게 한다. 따라서 어절 간 전이확률을 계산할 때 앞 어절에서는 가장 마지막 문법그룹만을 사용하여도 유용한 정보를 얻을 수 있음을 알 수 있다.

또한, 표 1을 보면, 한국어에서 앞 어절과의 관계에 영향을 끼치는 뒤 어절의 문법그룹은 99.62%의 어절에서 2개 이하로 나타나는 것을 알 수 있다. 따라서 뒤 어절에서는 첫 문법그룹과 마지막 문법그룹만 있다면 대부분의 어절 관련 정보를 얻을 수 있음을 알 수 있다.

표 1. 어절 내 문법그룹 수의 분포
(세종 말뭉치로부터 추출한 470만 어절 크기의 실험 자료로부터 관찰)

문법그룹 수	1	2	3	4
비율	90.76%	8.85%	0.38%	0.00%

이제 그림 5처럼 두 개의 링크만 남았다. 남은 링크는 모두 유용한 링크인지, 더 제거할 수 있는 링크는 없는지 살펴보겠다.

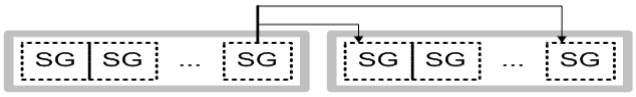

그림 5. 간결해진, 문법그룹(SG)을 사용한 어절 간 전이 링크

한국어에서 부사는 뒤에 오는 용언을 수식하므로 그림 6에서 부사 '헐레벌떡'은 동사 '달리다'를 수식한다. 그리고 관형사, 관형격조사, 관형형 전성어미는 체언을 수식하므로 그림 6에서 관형격조사는 '달리다'가 명사화된 '달리기' 혹은 명사형 전성어미 '기'를 수식한다.

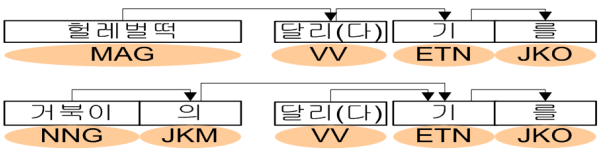

그림 6. 한국어 부분 문장의 의존 트리

그림 7은 그림 6을 문법그룹 간 링크로 표현한 것이다. 그림 7과 그림 6을 통해 뒤 어절의 여러 문법그룹 중 특정 하나의 그룹만이 앞 어절과 문법적 관계를 형성하는 것을 확인할 수 있다. 즉, 두 링크 중 하나는 사용하지 않아도 되는 링크이다. 그러나 어떤 링크가, 유용한 정보가 있는 링크인지는 문맥에 따라 달라지므로 일괄적으로 한 링크를 제거할 수는 없다.

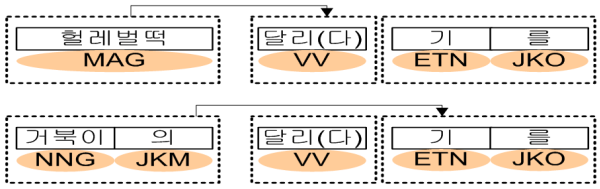

그림 7. 유용한 형태소 단위 전이 링크 예

3.1.3. 전이확률 계산에서 유용한 형태소

여기까지 본 논문은 어절 간 전이확률을 구할 때 앞 어절의 마지막 문법그룹과 뒤 어절의 처음과 마지막 문법그룹에 유용한 정보가 있음을 살펴보았다. 이제, 형태소 범위에서 살펴보겠다.

각 문법그룹은 다수 형태소를 포함할 수 있다. 이것은 그림 2, 3의 상태도와 그림 7을 통해 알 수 있다. 만약 문법그룹의 여러 형태소 중 한 형태소에 대한 문법 특성을 안다면 문법그룹 내 나머지 형태소에 대한 문법 특성도 간접적으로 알 수 있다. 예를 들어, 그림 8에서 볼 수 있듯이, 어떤 문법그룹의 끝 형태소가 어미이면 그 앞 형태소는 용언, 지정사, 용언화 접미사 등만 가능하다. 즉 용언의 특성이 있는 문법그룹임을 알 수 있다. 마찬가지로 문법그룹의 첫 형태소가 체언이나 명사형 전성어미이면 나머지 형태소는 체언이나 조사일 것이고, 첫 형태소가 용언이나 용언화 접미사면 뒤 형태소는 어미일 것이다. 따라서 그림 6과 7에서 '의'와 실제 문법적 관계가 있는 것은 '기'이지만, '기' 대신 '를'을 사용해도 관형격조사 '의' 다음 어절에 체언의 성질이 있음을 알 수 있다. 이러한 특징으로 말미암아, 문법그룹에 속한 하나의 형태소만으로도 문법그룹 전체의 통사적 특성을 간접적으로 얻을 수 있다.

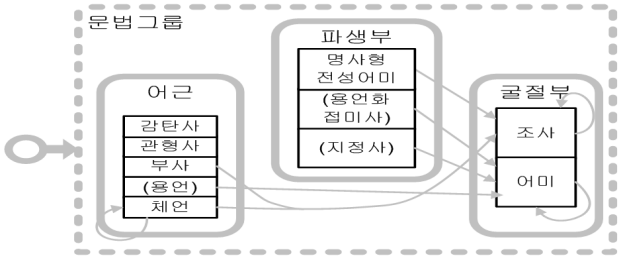

그림 8 문법그룹을 구성하는 형태소의 결합 상태도

따라서 본 논문은 교착어의 전이확률을 구하기 위한 최적의 형태소 조합으로 앞 어절의 마지막 형태소와 뒤 어절의 처음과 마지막 형태소를 사용한다. 이런 특징과 함께 그림 7을 고려하면 그림 9처럼 두 링크로 표현할 수 있고, 각각 링크1, 링크2로 부른다.

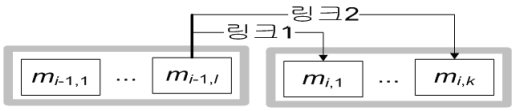

그림 9. 어절 간 형태소 단위 최적 링크

3.2. 전이확률 계산 방법

3.1절에서, 어절 간 전이확률을 구할 때 링크1과 2의 역할과 그 중요성에 대해 살펴보았다. 이 절에서는, 링크1과 2를 어떻게 이용하여 어절 간 전이확률을 구할 것인지에 대해 살펴보겠다. 표 2는, 이 절에서 살펴볼 6가지 모델을 비교하고 있다.

표 2. 어절 간 전이확률 추출을 위한 모델

모델	선택한 링크		하위범주화	n-gram
	링크1	링크2	사용여부	
모델0	0	0		3
모델1	0			2
모델2		0		"
모델3	0		0	"
모델4		0	0	"
모델5	0	0		"

모델0은 링크1과 링크2를 동시에 고려하는 모델로, 품사 tri-gram에 back-off smoothing model을 사용한 것과 유사한 모델이다. 이 모델은 파생으로 말미암은 품사의 변형 정보를 거의 다 얻을 수 있으면서도, 수식(6)과 같은 자료부족 문제는 많이 개선한 모델이다. 그러나 이 모델은 여전히 품사 tri-gram을 사용하고 있으므로 자료부족 문제에서 벗어나지 못한다.

모델1~5는, tri-gram model보다 자료부족 문제에서 자유로운 bi-gram에 기반을 둔 모델들이다.

모델1은 링크1만을 사용하는 모델이다. 이 모델은 어절에 파생 그룹이 포함되어 있을 때 (링크2가 필요한 경우) 그 정보를 얻을 수 없는 모델이다. 그러나 앞, 뒤 어절에서 각각 하나의 형태소만 사용하여 전이확률을 구할 때 가장 좋은 태깅 정확도를 얻을 수 있는 모델이다.[2]

모델2는 링크2만을 사용하는 모델이다. 이 모델도 어절에 파생 그룹이 포함되어 있을 때 (링크1이 필요한 경우) 파생되기 전의 정보(예를 들어 어근의 품사)를 얻을 수 없다.

모델3은 링크1을 사용한다. 모델1과의 차이점은 파생 유무에 대한 정보를 사용한다는 점이다. 실험에서는 뒤 어절의 첫 형태소 품사를, 파생이 된 것과 안 된 것으로 하위범주화하는 방식으로, 파생 유무에 대한 정보를 제공하였다. 이 모델은 하위범주화를 통해 추가의 통계 자료를 사용하였고, 대신 모델1의 단점을 보완하였다.

모델4는 링크2를 사용한다. 모델2와의 차이점은 링크2에 연결되고 나서 어절의 형태소가 굴절 형태소일 경우, 그 앞에 있는 어근이나 파생 형태소의 정보를 부분적으로 사용한다는 점이다. 실험에서는 어근이나 파생에 수의존명사, 부사, 명사형 전성어미, 보조용언 등이 있었는지 없는지에 따라 어미와 조사 각각을 5종류로 하위범주화하였다. 이 모델 역시 추가의 통계 자료를 이용하여 모델2의 단점을 보완하였다. 이 모델은 [1]에서 사용한 어절 간 전이모델과 같은 모델이다.

모델1과 2는 파생으로 말미암은 문제를 해결할 수 없고, 모델3과 4는 추가의 통계자료가 필요하다. 또한, 이 네 모델은 그림 10과 같은 분별력 문제가 발생한다. 즉, 전이확률 계산에 사용된 형태소는, 후보들 간에 모두 동일($m_{i,k,1}=m_{i,k,2}$)한 반면, 계산에 사용되지 않은 형태소의 후보들은 서로 상이($m_{i,1,1}\neq m_{i,1,2}$)한 경우 분별력을 읽게 된다.

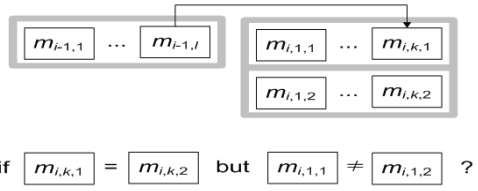

그림 10. 분별력 문제의 예

이러한 모델1~4와 모델0의 문제를 해결한 모델로, 본 논문에서는 모델5, '다이내믹 링크 모델'을 제안한다. 모델5는 문맥에 따라 링크1을 사용하기도 하고, 링크2를 사용하기도 한다. 모델5는 20여 개의, 링크 선택 규칙을 사용하여 문맥에 따라 링크1이 적당한지, 링크2가 적당한지를 동적으로 선택하여 계산하는 모델이다. 링크1, 2의 정보를 문맥에 맞게 선택하므로, 교착어에서 전이확률 구할 때 생기는 문제점인 높은 생산성으로 말미암은 자료부족문제나, 파생으로 말미암은 어절의 품사 특성 변화 문제 등을 대부분 해결할 수 있으면서, 추가의 통계 자료는 필요로 하지 않는 모델이다.

기존의 모델 중 Twoply HMM[2]도 두 개의 링크를 사용하지만 다이내믹 링크 모델과는 차이가 있다. Twoply HMM에서는 앞 어절의 마지막 형태소와 뒤 어절의 마지막 형태소를 연결하는 링크 대신 앞 어절의 첫 형태소와 뒤 어절의 마지막 형태소를 연결하는 링크를 사용한다. 이것은 그림 6과 7에서 관찰한 관형격조사 '의'와 뒤 어절의 체언과의 관계를 연결하지 못하는 것과 같이 어절 내에 파생 형태소가 있을 때 문법 관계를 제대로 연결하지 못하는 약점이 있다. 두 링크를 모두 사용하므로 오히려 관형격 조사와 용언을 연결하는 것과 같이 잘못된 정보를 포함한 링크를 사용하는 문제점이 있다.

본 논문에서 제시한 모델0에서 5는 전이확률 계산에 강미영이 [1]과 [7]에서 사용한, uni-gram과 범주 패턴, 학습한 파라미터를 이용한 형태소 n-gram 확률 추정 방법을 사용한다.

4. 실험

이 장에서는 다이내믹 링크 모델에서 사용한 링크 선택 규칙을 소개하고, 다이내믹 링크 모델의 태깅 정확도를 다른 전이확률 모델과 비교해 본다. 다이내믹 링크 모델을 비롯하여 모델0~4에 대한 실험은 모두 어절 확률 계산을 위해 강미영의 [7]에서 사용한 어절확률 추정 모델을 사용하였다.

4.1. 다이내믹 링크 선택 규칙 생성

다이내믹 링크 모델에서 사용한 링크 선택 규칙은 문맥에 따라 링크1과 링크2 중 적절한 하나를 선택하는 규칙이다.

표 3. 다이내믹 링크 선택 규칙

번호	조건	판단 범위	선택 link	정확도	재현률
1	문장경계 # 문장 첫 어절	어절	링크2	95.1	6.0
2	MMQ # NNBU	후보	링크1	98.4	0.6
3	* # 모두 단독형태소	어절	링크1	99.7	30.7
4	* # (MMQ & !MMQ)	어절	링크2	75.9	0.6
5	(JKG | EC[게 도록]) # 술어	어절	링크1	95.2	2.5
6	ETM # NNBG	어절	링크1	81.1	1.0
7	* # NNBG | NNG	어절	링크2	79.4	2.9
8	EC # VX	어절	링크1	93.4	2.1
9	* # (단독 형태소 후보 없음 & NNBU로 끝남 & NNBU로 끝나지 않음)	어절	링크2	97.8	0.7
10	* # (단독 형태소 후보 없음)	어절	링크1	98.0	21.1
11	MAG # (VV | VA)	후보	링크1	81.8	0.0
12	JKO # 술어	후보	링크1	91.8	0.2
13	다른 후보에 (MMQ # NNBU) 없음 & (JKM | ETM | MMG) # 체언	후보	링크1	90.9	0.5
14	MAG # *	후보	링크2	90.9	0.0
15	다른 후보에 (MMQ # NNBU) 없음 & (JKM | ETM | MMG) # *	후보	링크1	41.7	0.0
16	다른 후보에 (MMQ # NNBU) 있음 & (JKM | ETM | MMG) # *	후보	링크2	62.3	0.2
17	* # NNBU	후보	링크2	100	0.1
18	적용된 규칙 없음	어절	링크2	83.6	2.7

= 어절 경계
* = 모든 경우를 허용
'단독 형태소' = 어절 분석 후보가 하나의 형태소로만 구성
'모두 단독 형태소' = 어절의 모든 분석가 '단독 형태소'
'단독 후보' = 어절의 형태소 분석 후보가 하나뿐임
! = 부정 연산, & = AND 연산, | = OR 연산
'판단 범위' = 규칙 판단에 사용하는 형태소 분석 후보의 범위. 만약, 후보 단위 적용이라면 전/후 어절에서 조건을 만족하는 후보 사이에만 규칙이 적용, 어절 단위 적용이라면 전/후 어절에서 조건을 만족하는 후보가 존재할 때 해당 전/후 어절의 모든 후보에 규칙이 적용

본 논문에서는 앞서 3.1.2절과 3.1.3절에서 살펴본 링크1과 링크2의 특성을 고려하여, 뒤 어절의 파생 유무에 따라 전 후 어절의 문법적 관계가 크게 달라지는 것을 살펴보았고, 이를 고려하여 몇 가지 문맥 규칙을 생성하였다. 또한, 링크1과 링크2를 각각 사용한 모델1과 모델2의 태깅 결과와 다이내믹 링크 모델의 태깅 결과를 비교하는 과정을 거쳐 규칙을 다듬고 확장하였다. 표 3에 링크 선택에 사용한 전체 규칙이 나와 있다. 전체 규칙의 수가 20여 개밖에 안 됨을 확인할 수 있다.

규칙 구조는 '#'를 기준으로 앞 어절과 뒤 어절의 조건, 규칙의 판단 범위, 규칙이 선택할 링크 등으로 구성된다. 각 형태소 품사 조건은, 특별히 언급하지 않았을 때 앞 어절에서는 마지막 형태소를, 뒤 어절에서는 첫 형태소를 의미한다. 각 규칙은 규칙 번호가 낮을수록 우선순위가 높다.

규칙은 기본적으로 3.1.2절과 3.1.3절에서 살펴본 링크1, 2의 특성을 고려하여 작성하였다. 그러나 규칙이 일반화 과정을 거치면서 숨은 규칙이 되어 직접적으로 보이지는 않는다. 예를 들어 규칙번호 11와 14에 '일반부사 # 용언' 문맥에 대한 규칙이 있다. 부사는 용언을 꾸미므로 규칙 11처럼 뒤 어절의 첫 형태소인 용언을 사용하도록 링크1을 선택한다. 만약 뒤 형태소가 '명사+용언화 접미사'의 구조를 가진다면 규칙 14가 적용되어 링크2가 선택될 것이다.

모델1과 모델2를 각각 적용하여 태깅한 결과에서 양쪽 모델의 태깅 결과가 모두 틀린 경우만 해당 어절에 대한 태깅이 틀렸다고 하는 가상의 모델 평가에서 품사 태깅 정확도가 97.43%였다. 이는 다이내믹 링크 모델이 얻을 수 있는 최고 성능으로 가정할 수 있다. 이 가정과 표 6의 실험 결과를 종합해 보면 링크 선택 규칙은 다이내믹 링크 모델의 정확도를, 이 모델이 얻을 수 있는 최고 정확도의 99.15%까지 끌어올렸음을 알 수 있다.

4.2. 통계 사전 구축과 정확도 평가

본 연구에서 실험한 모델들은 모두 통계에 기반을 둔 모델들이기 때문에 통계 사전 구축이 필수적이다. 통계 사전은 2006년 배포한 21세기 세종계획의 '연구 교육용 현대국어 균형 말뭉치'의 색인 말뭉치 10개 분야 1,000만 어절 중, Docu, General, Life, News, Science의 5개 분야 약 480만 어절 중 460여만 어절을 사용하였다. 세종 계획의 색인 말뭉치에서 사용한 품사 집합과 본 논문에서 사용하는 품사 집합이 서로 달라서, 본 논문의 품사 집합에 맞게 반자동으로 변환하여 사용하였다.

표 4. 학습 데이터

출처	크기(어절)	비고
세종 색인 말뭉치	약 460만	본 시스템에 맞게 품사 집합을 변환함

성능 평가에 사용한 말뭉치는 신문말뭉치와 중학교 교과서 말뭉치를 사용하였다. 신문말뭉치는 1년치 신문 말뭉치 약 940만 어절 중에서 임의로 3만 어절을 추출하였으며 전문가에 의해 수작업 정제하였다. 중학교 교과서 말뭉치는 소설, 수필, 시 등을 제외한 약 5만 어절이며, 역시 전문가에 의해 수작업 정제하였다.

표 5. 평가 데이터

출처	크기(어절)
1년치 신문	32,502
중학교 교과서	46,984
계	79,486

모델별 품사태깅 결과가 표 6에 있다.

표 6. 모델별 성능 비교

모델	정확도	비고
모델0	96.58%	링크1& 링크 2(Tri-gram)
모델1	95.84%	링크 1
모델2	93.84%	링크 2
모델3	96.41%	링크 1+하위범주
모델4	96.40%	링크 2+하위범주
모델5	**96.60%**	다이내믹 링크 모델

다이내믹 링크 모델은, 어절 간 전이 확률을 구기 위한 통계 데이터로 형태소 품사 bi-gram만을 사용한다. 그러나 이 모델은, 하위범주화한 품사 집합의 형태소 품사 bi-gram을 사용한 모델3, 4나, 형태소 품사 tri-gram을 사용한 모델0과 대등한 성능을 나타낸다. 이 실험 결과는, 적은 규칙을 이용해, 전이 확률 계산에 이용할 적당한 형태소를 선택하게 함으로써, 형태소 품사 bi-gram만을 사용해 계산된 전이확률의 가치(정보량)가 bi-gram 이상의 정보를 사용해 계산한 전이확률의 가치만큼 높아질 수 있음을 보여주고 있다.

5. 결론

본 연구에서는 교착어인 한국어의 어절 전이 확률을 계산하는 방법에 대해 논하였다. 본 논문에서는 먼저, 어절의 전이 확률을 계산하는 데는 어절의 모든 형태소가 필요하지 않음을 가정하였고, 논의를 통해 앞 어절의 마지막 형태소와 뒤 어절의 처음과 끝 형태소에 어절 간 전이확률 계산에 필요한 거의 모든 정보가 있다는 결론을 도출하였다. 그리고 이 세 형태소를 두 개의 전이 링크라고 보고, 이 두 링크를 문맥에 따라 한순간에 하나의 링크만 선택하게 하는 규칙을 이용해 다이내믹 링크 모델을 완성하였다. 링크 선택 규칙은 이 모델의 이론적 최고 정확도인 97.43% 대비 99.15%까지 끌어올렸으며, 약 8만 어절의 실험 자료에서 96.60%의 태깅 정확도를 얻었다. 이 모델은 통계 정보로 형태소 품사 bi-gram만을 사용하지만, 이보다 많은 문맥 정보를 이용하는 다른 모델과 대등한 정확도를 보였다.

<Acknowledgement>

이 논문은 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (No. R01-2007-000-20517-0)

<참고 문헌>

[1] 강미영, "한국어의 형태·통사적 특징을 고려한 범주 기반 가변 n-gram 품사 태깅 모델", 부산대학교 컴퓨터공학과 박사학위 논문, 2007.

[2] 김진동, 임희석, 임해창, "Twoply HMM:한국어 특성을 고려한 형태소 단위의 한국어 품사 태깅 모델”, 한국정보과학회 논문지(B), Vol.24, No.12, pp.1502-1512, 1997.

[3] 이상호, "미등록어를 고려한 한국어 품사 태깅 시스템 구현", 한국과학기술원 전산학과 석사학위논문, 1992.

[4] Charniak, E., C. Hendrickson, N. Jacobson and M. Perkowitz. "Equations for Part-of-Speech Tagging". Proceedings of the Eleventh National Conference on Artificial Intelligence, AAAI Press/MIT Press, pp. 784-789, 1993.

[5] Hakkani-Tür. D. G, Oflazer. K, Tür. G, "Statistical Morphological Disambiguation for Agglutinative Languages", Computers and the Humanities, Vol. 36 No.4, 2002.

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

[7] Mi-young Kang, Sung-won Jung, Kyung-soon Park and Hyuk-chul Kwon, "Part-of-Speech Tagging Using Word Probability Based on Category Patterns", Lecture Notes in Computer Science, Vol.4394, pp.119-130, 2007.

<부록>

1. 실험에 사용한 품사집합과 약어표

품사	약어	품사	약어
일반명사	NNG	호격조사	JKV
동작성 명사	NNV	인용격조사	JKQ
상태성 명사	NNA	보조사	JX
고유명사	NNP	접속조사	JC
일반의존명사	NNBG	종결어미	EF
단위성 의존명사	NNBU	선어말어미	EP
인칭대명사	NPP	연결어미	EC
지시대명사	NPD	관형형 전성어미	ETM
양수사	NRC	명사형 전성어미	ETN
서수사	NRO	인용형 어미	EQ
동사	VV	일반접두사	XPG
형용사	VA	수접두사	XPU
보조용언	VX	복수접미사	XSP
지정사	VC	일반접미사	XSG
일반관형사	MMG	수접미사	XSU
수관형사	MMQ	관형사화 접미사	XSM
일반부사	MAG	동사화 접미사	XSV
접속부사	MAJ	형용사화 접미사	XSA
감탄사	IC	외국어	SL
주격/보격조사	JKS	한자	SH
목적격조사	JKO	도량형단위	SU
관형격조사	JKM	화폐단위	SC
부사격조사	JKG	기호	PNT
		문장경계	SB