제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

자율 학습에 의한 실질 형태소와 형식 형태소의 분리

차영태*, 조세형*
*명지대학교 정보통신공학과
e-mail : sl20@mju.ac.kr

A Korean Language Stemmer based on Unsupervised Learning

Yongtae Cha*, Sehyeong Cho *
*Dept. of information communication engineering, Myong-Ji University

요 약 자연어의 처리를 위해 반드시 필요한 형태소 분석에는 여러 가지 방법이 있으나 기본적으로 사전을 갖춘 상태에서 가장 가능성 있는 후보를 선택하는 방식을 선택한다. 이러한 방식으로는 사전이 없는 미지의 언어를 분석하기는 불가능하다. 기지의 언어라도 지속적으로 어휘가 변하는 경우나 매우 특별한 분야의 경우에는 필요로 하는 사전이 존재하지 않는다. 본 논문에서는 태그가 없는 단순 말뭉치만을 가지고 자율학습을 이용하여 한국어의 실질 형태소와 형식 형태소를 분리해내는 기법에 대하여 기술한다.

1. 서론

컴퓨터 프로그램으로 하여금 자연 언어를 인식하고 이해하게 하려는 많은 연구가 지금까지 진행되어 왔다. 이러한 연구 중 한 분야는 학습을 통한 자연 언어의 처리, 즉 언어학습으로, 언어 학습을 다시 나누자면 형태소 학습, 구문 학습, 의미 학습 등으로 구분할 수 있다. 본 논문에서는 그 첫 단계라고 할 수 있는 형태소 학습에 치중한다. 말뭉치(corpus)를 이용하여 형태소 학습을 하는 경우 해당 말뭉치에 대한 품사 분석 자료(tagged corpus)가 준비되어 있어서 이를 이용할 수 있다면 지도학습으로 분류하게 된다. 반면에 단순 말뭉치(plain corpus)만을 이용한다면 자율 학습으로 분류된다. 본 논문에서는 자율 학습을 선택하였다. 그 가장 큰 이유는 형태소 학습의 이론적, 실질적 목적이 미지의 언어, 또는 매우 많은 수의 미지 어휘를 포함한 언어를 컴퓨터로 하여금 학습하게 하는데 있으므로 사람의 간섭이나 미리 준비된 정답이 요구되는 지도 학습을 사용함이 적절치 않기 때문이다.

2. 관련 연구

한국어의 형태소 분석의 수준은 이미 상용 제품에 사용되는 수준에 이르러 있다. [1]에 의하면 한국어 형태소 분석의 정확도는 89~97%에 달하는 것으로 보고되어 있으나 이들은 사전에 의존하는 방식이다 [2]은 한국어의 음절 특성을 형태소 분석에 이용한바 있다. 그 외의 형태소 분석 연구는 효율적인 사전의 구축[3][4] 이라든가 특정 언어(한국어)에 의존적인 지식을 이용하여 형태소 분석을 향상하는 연구[2][5], 통계적 방식과 규칙기반 방법의 접목 방식[1] 등에 관한 연구들이 이루어져 왔다.

한국어와는 달리 굴절어에 대한 자동 형태소 학습 방법은 상당히 오래 전부터 시도되어 왔다[7]. 초기의 연구는 미리 준비된 접미사 사전과 어근에 관련된 규칙을 이용한 것이었지만 최근에 들어서 언어에 독립적인 형태소 학습 방법이 연구되고 있다[8][9]. Schone[8]은 Goldsmith[9], Gaussier[10]등의 방법에 더하여 행렬의 Singular Value Decomposition 을 이용하여 차원을 줄이는 방식으로 LSA(Latent Semantic Analysis) 기법을 이용하여 중의성 문제를 해결하는 학습 방법을 시도하였으며 이 방법은 본 논문에서 제안된 방법과 병합될 수 있다.

3. 관심의 대상이 되는 한국어의 특징

교착어(agglutinative language)인 한국어는 단어에 조사, 어미 등의 접사가 붙는(agglutination) 경향이 있는데 이는 한국어에서 문법적으로 중요한 역할을 한다. 그 중 하나인 조사는 단어 또는 다른 어절의 뒤에 붙어서 하나의 어절을 구성하며 선행하는 단어/어절의 문법적인 위치를 결정하는 문법 형태소 또는 형식 형태소이다. 이러한 형식 형태소인 조사와 어미가 가지는 공통적인 특징은 어절의 끝에 위치한다는 것이다. 본 논문에서는 이러한 특징만을 사용하며 언어에 대한 다른 어떠한 지식도 사용하지 않고 실질 형태소와 형식 형태소를 분리해내는 방법을 기술한다.

4. 분석을 위한 초벌 알고리즘(baseline algorithm)의 비교

형식 형태소는 여러 종류의 실질 형태소와 결합되어 어절을 이룬다. 따라서 어절의 후반부에 자주 출현하게 된다. 이 지식을 역 이용하면 형식 형태소일 가능성이 많은 스트링을 찾아낼 수 있다. 여기서는 이러한 가능성을 찾아내는 초벌 알고리즘에 대해 설명하고자 한다.

4.1 빈도 방식

이 방식은 가장 분석이 용이한 방법이다. 이 방식은 조사나 어미와 같은 문법 형태소는 체언 및 용언과는 달리 그 종류가 많지 않고 반복적으로 쓰인다는 데서 착안을 한 것이다.

4.2 상대 빈도 방식

음절에는 형식 형태소의 여부와 상관 없이 빈도가 많은 음절과 적은 음절이 있다. 따라서 이러한 음절들은 당연히 끝 음절에도 많이 나타난다. 이는 평균 어절 길이가 3 정도인 점을 감안할 때, 특별히 어절 끝에 많이 쓰인 것이 아니라 본래 빈도가 많기 때문에 어절 끝에도 자연히 빈번하게 출현한 것이다. 따라서 이러한 단점을 보완하기 위해서 상대적 빈도(즉, 끝 음절 빈도 / 총 빈도)를 사용할 수 있다.

4.3 통계적 가설 검정 방식

언어의 생성을 각 음절에 대한 확률적 선택으로 볼 수 있다. 만일 임의의 지점에서 어느 음절이 선택될 확률을 p_0 라고 하고 이 음절이 어절 끝에 나타날 확률을 p 라고 하자. 또한 가정하기를 형식 형태소가 아닌 음절들은 무작위로(random) 선택된다고 가정해 보자. 이 가정이 사실이라면 형식 형태소는 어절 끝에 나타날 확률 p 가 p_0 보다 클 것이다. 한 음절을 기준으로 볼 때, 이 것은 성공/실패(출현/비출현)라는 베르누이 시행으로 볼 수 있으며 따라서 이항 검정의 방식으로 검정할 수 있다[8]. 즉, p 를 표본에서의 어절 끝 출현 확률, p_0 를 귀무 가설(null hypothesis)에 의한 음절의 출현 확률, N 을 음절 수(시행 횟수)로 볼 때,

$$z_0 = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}}$$

이며 예컨대 유의 수준 0.05 로 검정할 때, Z_0 값은 1.96 이상이 되어야 귀무 가설을 기각할 수 있다. 다시 말해서 Z_0 값이 1.96 이상이 된다면 95% 신뢰도를 가지고 당 음절은 형식 형태소라고 판단할 수 있다.

4.4 활용 빈도에 의한 방식

앞 절에서 본 바와 같이 "함께"라는 어절은 동일 어절이 34 회 출현함으로써 어절 "께"의 어절 끝 출현 확률을 높이게 되었다. 만일 동일한 어절이 동일한 형태소의 집합으로 구성되어 있다면 (일부의 예를 제외하고는 실제로 대부분의 경우에 그러하듯이) 음절이 얼마나 어절 끝에 자주 등장했느냐 보다는 얼마나 많은 종류와 조합을 하였느냐가 중요할 것으로 추측할 수 있다.

5. 복합 알고리즘

5.1 복합 알고리즘의 특징

본 논문에서 제시하는 알고리즘의 특징은 형식 형태소 후보의 파악을 위해 이항 검정(T-test)을 사용하였다는 것과 말뭉치에서의 어절의 출현 빈도를 따지지 않고 동일 어절의 다회 출현은 1 회로 간주하여 앞 절에서의 활용 빈도 방식과 통계적 가설 검정 방식을 조합하였다.

5.2 알고리즘.

1 단계: 이항 검정 (t-test)

1 단계에서는 모든 어절의 모든 어말에(rear substring) 대해 이항 검정을 실시하여 후보를 선정한다. 중복된 어말에 대한 중복 분석을 배제하고 효율적인 분석을 하기 위하여 역방향 트라이를 사용한다. 이항 검증을 위한 Z_0 값의 계산을 위해서는 우선 후보 형태소 (즉 역방향 트라이의 모든 노드에 대응) w 에 대해 스트링 w 의 본래 출현 확률을 구해야 한다. 이러한 확률은 이론적인 선험 확률(prior)이 존재하지 않으므로 표본을 분석함으로써 최우추정(Maximum Likelihood Estimation)에 의해 $p_0 = \#(w, C') / (w$ 가 나타날 수 있었던 위치의 수효$)$ 로 추정한다. 여기서 C' 은 원시 말뭉치 C 에서 중복 출현 어절을 제외한 새 (가상적) 말뭉치이며 $\#(x, C)$ 는 스트링 x 가 말뭉치 C 에 포함된 빈도를 말한다. 받침하나로 구성된 후보(예: "ㄴ")나 한 음절짜리 후보(예: "는")의 출현가능 위치 수는 음절 수와 같다. 1 음절보다 길고 2 음절 미만인 후보(예: "ㄴ다")은 2 음절 이상인 어절에서만 출현 가능하다. 그런데 2 음절 길이의 후보가 3 음절에서 나타날 수 있는 확률은 약간 복잡하다. 즉 2 음절 후보가 나타날 가능성이 있는 위치는 2 곳이지만 (1,2 음절 또는 2,3 음절) 후보의 앞뒤음절이 같지 않다면 두 군데에 동시에 나타날 수는 없다. 따라서 3 음절 어절에서 2 음절 후보가 출현하는 것은 엄밀한 의미에서 베르누이 시행이 아니다. 그러나 출현 확률이 1 보다 매우 작다는 가정에서 이 점은 무시하고 베르누이 시행인 것으로 근사한다. 따라서 길이가 l 음절인 (음절을 이루지 않는 음소는 1 음절로 계산) 후보가 출현할 수

있는 총 위치수는 $\sum_{i=l}^{l_{max}}(i-l+1)\times\#(i,C')$ 이며 여기서 l_{max} 는 가장 긴 어절의 음절 수이며 $\#(i,C')$ 는 길이 i 음절인 어절의 C' 내의 수효이다.

어절 끝 확률은 $p=\#(w\$,C')/(C'$ 내에서 w 를 포함할 수 있는 어절수) 로 계산하며 $\$$ 는 어절의 끝을 나타내는 가상적인 길이 0 의 음소이다. 즉 w 의 실제 어절 끝 출현횟수에 의한 최우추정치이다. 위의 식을 이용하여 Z_0 를 계산한 뒤 유의 수준 0.1 로 추정하여 후보 목록을 만들었다. 비교적 작은 2 만 5 천 어절(중복 제외)의 말뭉치를 분석한 결과 4 만여개의 어말 중에서 유의 수준 0.1 로 517 종의 후보가 선택되었다.

2 단계: 우연성 후보의 배제

이항 검정에서 어절 끝 출현 확률이 임의 확률보다 충분히 높은 후보들이 1 단계에서 선택되었으나 이들 중에서는 실제로는 후보형태소 자신 때문이 아니라 후보의 일부분 (rear substring) 때문에 선택되는 경우가 흔히 있다. 즉, w 와 δ 를 중복 없는 말뭉치 C'내의 음소열이라 할 때,

$Z_0(w)>T, Z_0(\delta w)>T,$ 인 경우

$p=\frac{\#(\delta w\$\in C')}{\#(w\$\in C')}$, $p_0=\#(\delta\in C')/(\delta$ 가 나타날 수 있었던 위치의 수효) 를 이용하여 다시 이항 검정을 하면 "ㅁ이다"의 경우 처럼 실제로는 의미 없는 경우를 제외시킬 수 있을 것이다. 실제 실험에서 1 단계에서 선택되었던 517 종의 후보 가운데 308 종이 탈락하고 219 종만이 남았다.

3 단계: 실질 형태소 후보의 생성 단계

어느 어절의 형식 형태소를 안다면 실질 형태소를 알 수 있고 거꾸로 실질 형태소를 안다면 형식 형태소를 알 수 있다. 완벽하지는 않지만 단계 1,2 를 통하여 많은 형식 형태소들을 찾아내었으므로 이제는 이를 이용하여 다시 실질 형태소를 구분해 낼 수 있다. 그러나 주어진 어절에서 형식 형태소의 집합만을 가지고 (만일 완벽한 집합이라고 하여도) 실질 형태소를 분리해 내기는 어렵다. 그 이유는 첫째, 어느 어절의 어미 부분이 우연히 어떤 실질 형태소와 같을 수가 있기 때문이다. 예를 들어 "ㄴ"은 많은 경우에 형식 형태소로 쓰이지만 그렇지 않은 경우, 예를 들어 "대문"의 경우가 그러하다. 둘째 이유로는 형식 형태소들끼리는 상호 포함 관계(rear substring)에 있을 수가 있으며 이 경우 두가지 이상으로 형태소 분리가 가능한 애매성의 문제(morphological ambiguity)가 존재하기 때문이며 마지막으로는 주어진 형식 형태소 후보에 여전히 오류가 있을 수 있기 때문이다. 여기에서 "실질 형태소는 단독으로서 어절을 구성하거나 형식 형태소와 결합하여 어절을 구성한다"라는 사실을 이용할 수 있다. 실질 형태소의 후보 w 는 $\#(w\delta\in C')$, 즉 형식 형태소후보 δ 와 결합된 어절이 말뭉치 C' 내에서 출현한 빈도에 의해 결정된다. 이론적으로는 이 또한 T-test 에 의하여 판정하는 것이 바람직한 듯이 보인다. 그러나 일반적으로 실질 형태소는 그 빈도가 형식 형태소와 비교할 때 매우 낮다는 특징 때문에 통계적 검정의 의미가 적어진다. 즉, 최우추정을 하기에 통계량이 매우 부족하다. 이러한 이유로 본 논문에서는 2 단계에서 선택한 형식 형태소 중 일정한 수 (threshold) n 개 이상 결합 사용된 어두를 실질 형태소 후보로 선정하였으며 n=2 로 하였다. 말뭉치의 크기가 큰 경우는 더 큰 n 값을 적용하는 것이 가능하다.

4 단계: 실질 형태소 후보와의 결합 형태에 의한 형식 형태소 후보의 배제

앞의 단계에서 형식 형태소와의 결합에 의해 실질 형태소를 찾아내었듯이 여기서는 역으로 실질 형태소와 몇 가지 형태로 결합하느냐에 의해 형식 형태소 후보를 배제한다. 예를 들어 "수록" 등에 쓰이는 "록" 은 높은 Z_0 값을 가지고 있어서 자칫 독립된 형식 형태소로 인식되기 쉽다. 그러나 "록"의 앞에 결합된 어두들은 실질 형태소로 구분되는 것이 없다. 본래 "형식 형태소는 여러 종류의 실질 형태소와 결합하여 사용되기 때문에 어절 말에 자주 출현한다"라는 가정에서 출발하였으므로 비록 어절 말에 자주 출현하더라도 실질 형태소와의 결합이 빈번하지 않으면 형식 형태소로 인정할 수 없다.

5 단계: 말뭉치의 1 차 형태소 분석 및 형식 형태소간 연어 확률(bigram)의 추출

1 차 형태소 분석은 어두가 실질 형태소일 확률과 어말이 형식 형태소일 확률의 곱으로 선택한다. 어미가 형식 형태소일 확률은 다음과 같다. 어느 substring w 가 나타날 확률을 p 라고 하고 이 것이 어미에서 나타날 확률을 p_s 라고 하자. a 는 어절말에 나타난 w 의 수효, b 는 어절말이 아닌 다른 위치에 나타난 수효, a'은 어절 말에 나타난 w 중 형식 형태소로 사용된 수효라고 하자. 또, A 는 어절의 수, B 는 어절말 이외에 w 가 나타날 수 있는 위치의 총합이라고 하자. 최우추정에 의하면 w 의 확률은 $p=\frac{a+b}{A+B}$ 이며 $p_s=\frac{a}{A}$ 이다. w 가 어절 말에 있을 때 이 것이 형식 형태소일 확률(MLE)은

$p_{GM}=\frac{a'}{a}=\frac{a-\frac{A}{B}b}{a}=1-\frac{Ab}{Ba}$ 이 된다.

만일 이 suffix 가 형식 형태소가 아니라면 중간에서 나타나는 것과 어미에서 나타나는 것은 같은 확률이라고 가정한다.

"나는 겨울 여행을 좋아한다"에서는 "나는"과 "여행을" 사이에 실질 형태소만으로 구성된 어절이 끼어있기 때문에 <는,을>의 연어의 확률을 이용할 수 없게 된다. 따라서 이 단계에서는 어미가 없는 실질 형태소만으로된 어절(ε -어절이라고 하자)을 제외한 나머지 어절만으로 연어 확률을 계산한다. 즉, 원거리 연어(long-distance bigram)를 이용한다.

6 단계: 연어 확률을 이용한 2 차 분석

N 어절의 문장에서 각 어절을 e_i, i=1, n, e_i 의 어미 후보를 s^i_j, j=1, c_i , c_i 는 어절 e_i 의 어미 후보 갯수라고 하자.

$$I = \arg\max_{I=<I_1,I_2,...,I_n>} p(s^1_{I_1})p(s^2_{I_2})...p(s^n_{I_n})$$

$$\prod_{k=1}^{n+1} p(<last(E,I,k-1),s^1_{I_k}>)$$

여기서 E 는 어절 벡터인 $e_1,e_2,...,e_n$ 이고 I 는 해 벡터 (solution vector) $< I_1,I_2,...,I_n >$, 그리고 함수 $last(E,I,k-1)$ 은 어절 벡터 E 에서 어미 후보 벡터를 I 로 잠정하였을 경우 $<e_1,...,e_{k-1}>$ 에서 ε 이 아닌 마지막 어미이다. 이 수식의 평가를 위하여 모든 조합을 계산 할 경우 시간적인 복잡도(time complexity) 는 $\prod_{i=1}^{n} c_i$ 로서 $c_i < m$ 이고 문장내 어절의 갯수가 n 일때 $O(m^n)$ 이며 이는 실시간에 계산해 내기 어려운 수치이므로 상당히 큰 n 값에 대해서는 보다 효율적인 계산 방법이 요구된다.

5.3 실험 결과

비교적 작은 말뭉치인 2 만 5 천 어절의 말뭉치에서 통계를 얻어 이 말뭉치에 있는 문장을 임의 추출하여 분석하였다. 본 연구에서는 아직 복합 어미에 대한 분석을 하지 않고 있으며 형태소에 대한 품사 붙이기 (tagging)도 하지 않으므로 성공과 실패의 표준을 올바른 실질 형태소의 분리로 삼았다. 이러한 기준으로 1 단계 분석에 의해 선택한 분리는 74% 의 성공률을 보였다. 2 단계 분석으로 오류 후보 상당수가 제외된 후 다시 분석한 결과 81.5%의 성공률을 보였으며 다시 4 단계 알고리즘에 의해 실질 형태소를 찾아낸 후재 분석 결과 85%의 성공률을 보였다.

또, 동일한 문장에 대해 학습용 말뭉치의 크기를 2 배로 하여 학습한 후 시도한 결과 성공률은 85%에서 87%로 향상되었다. 그러나 연어의 확률을 고려한 방식은 예상과 달리 가시적인 향상을 보이지 않았으며 오히려 일부 경우에 연어 확률을 고려하지 않았을 때 분석에 성공했던 어절에 대해 오답을 내는 결과를 가져왔다. 이러한 원인 중 가장 큰 것은 말뭉치의 크기에서 기인된 것으로 분석된다. 2 만 5 천 어절에서 약 200 개의 형식 형태소에 대해 bigram 을 구하였으므로 200x200 = 40,000 종류가 있는 반면 형식 형태소가 있는 어절은 대략 1 만 5 천 정도에 지나지 않으므로 1 쌍 당 평균 출현 횟수가 0.5 회에도 미치지 못하는 미미한 수치이다. 따라서 언어 정보를 활용하기 위해서는 획기적으로 큰 말뭉치를 이용하여야 할 것으로 생각된다.

6. 결론 및 향후 연구 방향

본 논문에서는 미지의 언어나 또는 기지의 언어라도 지속적으로 어휘가 변하는 경우에 사전의 부재로 형태소 분석을 할 수 없는 경우의 문제를 해결할 수 있는 방법을 연구하였다. 연구된 기법의 주요 특징으로는 첫째, 사전 등의 언어 관련 지식이 필요하지 않고 오직 단순 말뭉치만이 필요하다는 것이고, 둘째, 자율학습을 이용함으로써 사람의 간섭이 필요하지 않아 학습에 필요한 시간과 노력이 거의 들지 않는다는 점이다. 또한 잘 확립된 통계적 방법론을 이용하기 때문에 일반적인 휴리스틱과는 달리 이론적인 기반이 확고하여 확장 및 발전이 용이하다.

현재로서는 형식 형태소 중 가장 확률이 높은 후보를 선택하게 되어 있다. 그러나 한국어에서는 실제로 형식 형태소 여러 개가 연속하여 출현하는 현상을 볼 수 있다. 본 알고리즘에 의한 분리 결과를 보면 형식 형태소를 분리하였음에도 불구하고 어두에 여전히 형식 형태소가 남아있는 현상을 볼 수 있다. 따라서 분리된 어두에 대한 반복적인 형태소 분석에 의해 추가적인 형태소를 분리함으로써 정확도를 높일 수 있을 것으로 생각된다. 또한, 본 연구의 결과를 유사한 특징을 가진 외국어에 적용함으로써 언어에의 비의존성을 시험해보는 것도 향후의 중요한 연구 방향의 하나이다.

참고문헌

[1] 신상현, 이근배, 이종혁, "통계와 규칙에 기반한 2 단계 한국어 품사 태깅 시스템", 정보과학회 논문지(B) 24 권 2 호, pp.160-169, 1997.2.

[2] 강승식, "음절특성을 이용한 한국어 불규칙 용언의 형태소 분석," 정보과학회논문지(B) 제 22 권 10 호, pp. 1480-1487, 1995.10

[3] 최재형, 이상조, "양방향 최장 일치법에 의한 한국어 형태소 분석기에서의 사전 검색 횟수 감소 방안," 한국정보과학회 논문지 vol.20 no.10, pp.1497-1507, 1993.10.

[4] 김철수, 배우정, 이용식, 青江純一, "이중배열 트라이 구조를 이용한 한국어 전자 사전의 구축," 정보과학회 논문지(B) 제 23 권 1 호, pp.85-94, 1996.1.

[5] 임희석, 윤보현, 임해창, "배제 정보를 이용한 효율적인 한국어 형태소 분석기," 한국정보과학회 논문지, 22 권 6 호, pp.957-964, 1995.6.

[7] Lovins, J.B., Development of stemming algorithms. Machine Translation and Computational Linguistics, 11, 1968

[8] Patrick Schone and Daniel Jurafsky, "Knowledge-free Induction of Morphology using Latent Semantic Analysis," in proceedings of the ACL99 workshop: Unsupervised learning in Natural Language Processing, University of Maryland.

[9] J. Goldsmith, "Unsupervised learning of the morphology of a natural language," University of Chicago, http://humanities.uchicago.edu/faculty/goldsmith.

[10] E. Gaussier, "Unsupervised learning of derivational morphology from inflectional lexicons," in proceedings of the ACL99 workshop: Unsupervised learning in Natural Language Processing, University of Maryland.

[11] 김홍규, 강범모, "한국어 형태소 및 어휘 사용 빈도의 분석", 고려대학교 민족문화연구원, 2000, 7.

[12] 21 세기 세종 계획, 문화관광부, http://sejong.or.kr/