품사 태깅 시스템의 신뢰도 측정

김 재 훈†

요 약

본 논문에서는 품사 태깅 시스템에서 신뢰도 측정 방법에 대해서 기술한다. 품사 태깅 시스템의 신뢰도는 품사 태깅 결과에 오류가 포함되지 않을 확률이다. 일반적으로 신뢰도 측정은 오류확률에 기반한다. 정확한 오류확률을 추정하기 위해서는 일반적으로 품사 태깅 시스템에서 사용되는 말뭉치보다 훨씬 더 많은 양의 말뭉치가 필요하다. 이 문제를 다소 완화시키기 위해서, 본 논문에서는 좀더 정확한 오류확률 추정하기 위해 교차확인 방법을 이용한다. 본 논문에서 사용된 품사 태깅 시스템은 시험말뭉치에 대해서 61%의 신뢰도를 보였다. 이는 한국어 문장의 형태소 수가 평균 20개이고, 품사 태깅 시스템의 정확률이 97.5%일 때의 신뢰도에 해당한다. 본 논문에서 사용된 품사 태깅 시스템이 미등록어가 없을 경우에 97.68%의 정확률을 보이므로 제안된 신뢰도 측정 방법이 어느 정도 타당함을 알 수 있었다. 제안된 신뢰도 측정 방법은 구문분석, 정보검색 등 여러 분야에 응용이 가능할 것이며, 본 논문에서는 품사태깅의 오류검출에 적용해보았다.

Measuring Reliability of POS Tagging Systems

Jae-Hoon Kim†

ABSTRACT

This presents a method for measuring reliability of a part-of-speech (POS) tagging system. The reliability is the probability that there are not mis-tagged words in its results, which are POS tagged sentences. In general, reliability is estimated based on the reciprocal of error probabilities. In order to estimate the error probabilities, training corpus should be much large compared with that to calculate approximately probabilities for tagging POS's. To relax this problem, this paper also describes a method for estimating more reliable error probabilities using cross validation. In an experiment, the reliability of our POS tagging system is about 61% on the average, that is equivalent to the reliability when the number of morphemes in a sentence is 20 and the accuracy of the POS tagging system is 97.5%. We believe that this method for measuring reliability of POS tagging systems is valid because the accuracy of our POS tagging system without unknown words is 97.68%. We expect that this model can be applied to syntactic analyzers and information retrieval systems. In this paper, we applied this model to an error detection system in POS tagging.

키워드 : 품사 태깅(POS tagging), 신뢰도 측정(Measurement of reliability), 오류검출(Error detection)

1. 서 론

품사 태깅은 주어진 문장을 구성하는 각 단어에 가장 적합한 품사를 결정하는 것이며, 품사 태깅 방법은 확률기반 방법[1], 규칙기반 방법[2] 등이 있다. 대부분의 품사 태깅 시스템의 성능은 95% 이상의 성능을 보이고 있다[3]. 이 결과는 매 20개의 단어마다 적어도 하나 이상의 오류가 있음을 말하고 있으며, 이 오류는 품사 태깅 시스템을 이용하는 구문 분석 시스템, 정보 검색 시스템, 기계 번역 시스템 등에 그대로 전파되어, 또 다른 오류를 야기시킨다. 이 문제를 다소 완화시키기 위해서 많은 연구자들은 여러 가지의 방법을 이용하여 품사 태깅 시스템의 성능을 향상시키고 있다[4-5]. 그러나, 대부분의 분류 시스템(classifier)과 같이 품사 태깅 시스템도 100% 정확한 결과를 얻을 수는 없다. 즉, 품사 태깅 시스템의 결과에는 항상 어느 정도의 오류를 포함하고 있다는 가정을 해야만 한다. 그렇다면 품사 태깅 결과에 어느 정도의 오류를 포함하고 있는지, 또한 어떤 단어가 오류를 가지고 있는지를 알 수 있다면, 응용 시스템에서 이를 쉽게 대처할 수 있을 것이다. 본 논문에서는 품사 태깅 결과의 오류 정도, 즉, 신뢰도를 측정하기 위한 방법을 제시하고, 품사 태깅에서 오류를 검출하기 위해 이 방법을 적용해 보았다.

※ 본 연구는 한국과학재단으로부터 핵심선분 연구 과제(과제번호 : 981-1212-036-2)의 일환으로 부분적인 지원을 받았으며, 또한 첨단정보기술 연구센터를 통하여 한국 과학재단의 지원을 받았다.
† 정 회 원 : 한국해양대학교 컴퓨터공학과,
한국과학기술원, 첨단정보기술연구센타
논문접수 : 2000년 12월 27일, 심사완료 : 2001년 8월 13일

신뢰도는 어떤 시스템이 일정한 기간 동안 특정한 조건 하에서 오류 없이 정상적으로 가동될 확률이다[6-7]. 본 논문에서는 이와 같은 개념을 품사 태깅에 적용하고자 한다. 일반적인 신뢰도 측정에서 특정 시스템을 문장의 태깅 결과로 모델링하고, 시스템을 구성하는 각 장비나 부품을 품사 태깅 결과에 속하는 단어와 품사의 쌍(이하에서는 "단어/품사"로 표기)으로 모델링한다. 문장의 태깅 결과에 오류가 포함되어 있다면 품사 태깅 결과 중 하나 이상의 단어/품사가 오류임을 의미한다. 이 오류는 특별한 환경이 되었을 때만 발생하게 된다. 본 논문에서는 이러한 환경을 문맥으로 특정 단어/품사에 대한 오류확률을 이용해서 문장에 대한 신뢰도를 측정한다. 이 신뢰도 측정 모델을 품사 태깅 오류검출 시스템에 적용하여 좋은 결과를 보였다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구로 한국어 품사 태깅과 신뢰도에 대해서 살펴보고, 3장에서 품사 태깅을 위한 신뢰도 측정 모델을 기술한다. 4장에서는 신뢰도를 구하기 위해 요구되는 오류확률의 추정 방법에 대해서 기술하고, 5장에서 품사 태깅에서의 신뢰도 측정 모델에 대한 실험 및 평가를 수행한다. 6장에서는 신뢰도 측정 모델의 응용으로 오류검출 방법에 대해서 기술하고 7장에서 결론을 맺고자 한다.

2. 한국어 품사 태깅과 신뢰도

2.1 한국어 품사 태깅

일반적으로 영어에 대한 품사 태깅은 주어진 문장에 대한 가장 적절한 품사열을 결정하는 것이다. 그러나, 한국어 품사 태깅은 주어진 문장에 대한 형태소열과 품사열을 동시에 결정해야 한다[8]. 한국어 품사 태깅 모델은 일반적으로 영어 품사 태깅에서 널리 사용되고 있는 은닉 마르코프 모델을 주로 이용한다[8-9]. 한국어 품사 태깅은 어절 정보를 모델 내에 포함시키느냐에 따라 어절 단위의 품사 태깅[10]과 형태소 단위의 품사 태깅[4]로 크게 구별된다. 어절 단위의 품사 태깅에서 어절 태그는 그 어절을 구성하는 형태소의 품사열로 표현되는 경우가 일반적이다. 어절 단위의 품사 태깅은 어절 단위의 문맥을 고려할 수 있다는 장점을 가지고 있으며, 어절 태그 수를 쉽게 고정시킬 수 없고, 자료 부족 문제와 동품사 중의성 문제[11]를 야기시키는 단점을 가지고 있다. 이에 반해, 형태소 단위의 품사 태깅은 품사 태그의 결정이 비교적 쉽다는 장점을 가지고 있으나, 한국어의 특징 중 하나인 어절 단위의 문맥을 직접적으로 고려할 수 없다는 단점을 가지고 있다. 최근에 와서는 이들 두 방법을 결합하는 모델[9, 12]도 등장하고 있다. 이들 대부분의 모델들은 모두 은닉 마르코프 모델의 변형이라고 볼 수 있으며, 또 다른 형태의 변형으로는 은닉 마르코프 모델과 규칙과의 결합 방법[13]이 연구되고 있다. 이 방법은 주로 Brill에 의해서 제안된 변형 기반 품사 태깅 방법[2]과 은닉 마르코프 모델의 결합이다.

2.2 신뢰도

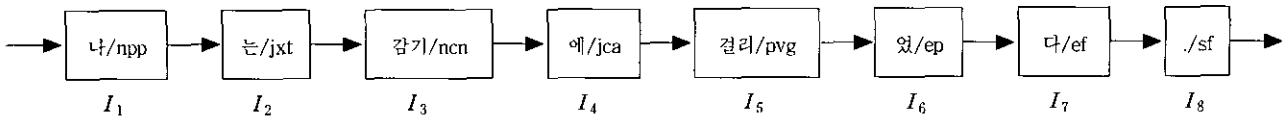
신뢰도란 어떤 장비가 주어진 조건 하에서 주어진 기간 동안에 의도된 일을 수행할 확률을 의미한다[6]. 다시 말해서 신뢰도는 어떤 시스템이 오류 없이 정상적으로 가동될 확률을 의미한다. 이 정의에서 "확률", "운용조건", "만족스럽게 작동", "사용기간"이라고 하는 네 가지의 요소가 포함되어 있다. 이 요소들이 장비나 부품의 신뢰도에 영향을 줄 수 있는 중요한 요소이다.

먼저 **확률**을 보면 신뢰도는 확률로 표시된다는 사실을 말하고 있다. 이것은 일반적으로 총 시행 회수 n 중에서 어떤 사건이 발생되는 백분율 %을 말한다. 즉, 어떤 장비가 50시간 동안 사용했을 때, 신뢰도가 70%라고 하는 것은 50시간 동안 사용하는 것을 100회로 가정하면 70회만 정상적으로 사용할 수 있다는 것을 의미한다. **운용조건**이란 각 부품이 정상적으로 작동할 수 있는 환경을 의미한다. 예를 들면, 반도체 칩 공정 과정에서 공기청정도, 온도, 습도 등이 공정조건에 적합해야만 한 공정이 정상적으로 이루어질 수 있다는 것이다. 이러한 환경적인 요인은 각 공정이나 각 부품이 정상적으로 작동하는데 크게 좌우된다. **만족스러운 작동**이란 마지막 공정에서 제품의 불량이 어느 정도 발생하지 않았다고 할 것인지를 분명히 밝히는 것으로 불량에 대한 판단의 기준이 되는 것이다. 일반적으로 제품의 부품이 불량을 발생시키는 조건으로 **사용 시간**이 있다. 즉, 신뢰도라고 하는 것은 시간함수로서 설명된다는 것이다. 일반적으로 부품은 10시간 동안 사용했을 때, 불량 발생률보다 20시간 동안 사용했을 때, 불량 발생률이 훨씬 높다는 것을 의미한다.

3. 품사 태깅 시스템에 대한 신뢰도 측정

본 장에서는 품사 태깅 시스템에서 신뢰도 측정 방법에 대해서 기술한다. 먼저 일반적인 신뢰도 측정 모델을 품사 태깅 시스템에 어떻게 대응시키는지를 살펴보자. 일반적인 신뢰도 측정에서 시스템은 품사 태깅 결과로 대응시키고, 장비나 부품은 품사 태깅 결과에 포함된 단어/품사로 대응시킨다. 이제 품사 태깅 시스템의 신뢰도를 정의하기 위해서 먼저 품사 태깅 시스템에서 신뢰도 요소("확률", "운용조건", "만족스럽게 작동", "사용기간")를 살펴보자.

먼저 확률은 품사 태깅에서 어떤 단어가 특정 품사로 정확하게 태깅될 확률을 의미한다. 운용조건은 품사 태깅 결과에서 오류를 범할 수 있는 주변 환경을 말한다. 즉, 품사



(그림 1) 품사 태깅 시스템의 신뢰도를 측정하기 위한 개념도

태깅 오류가 발생될 오류 환경을 말한다. 품사 태깅에서의 운용조건은 상황에 따라서 다르게 정의할 수 있다. 다시 말해서 오류가 발생될 환경은 주변의 다양한 문맥에 의해서 결정될 수 있음을 의미한다. 다음은 만족스러운 작동에 대해서 살펴보자. 일반적인 신뢰도에서는 부품의 작동 특성을 연속 변수로 묘사할 수 있다. 그러나 품사 태깅에서 만족스러운 작동은 이진 변수로서 표현된다. 즉 오류가 발생했느냐 하지 않았느냐에 의해서 만족스러운 작동을 표현한다. 일반적인 신뢰도에서는 운용시간이 매우 중요한 요소로 작용한다. 그러나 품사 태깅에서 각 단어/품사에 대해서 시간이 경과함에 따라 불량률이 높아지는 것은 아니다. 따라서 품사 태깅의 신뢰도에서 운용시간은 중요한 요소가 되지 않는다.

이를 종합하면 품사 태깅 시스템의 신뢰도는 품사 태깅 결과나 각 단어가 주어진 문맥 하에서 정확히 태깅될 확률을 의미하며, (그림 1)과 같이 표현될 수 있다. 문장을 구성하는 어떤 단어에 품사 태깅 오류가 있다면, 그 문장의 품사 태깅 결과에 오류가 있는 것이다. 신뢰도 측정은 얼마나 많은 오류가 문장의 태깅 결과에 포함되어 있는지를 측정하는 것이다. 본 논문에서는 신뢰도 모델에서 가장 일반적인 모델인 직렬구조를 사용하며, 문장 S에 대한 신뢰도 $R(S)$는 식 (1)과 같이 표현된다[6].

$$\begin{aligned} R(S) &= R(I_1 I_2 \cdots I_n) \\ &= \Pr(I_1, I_2, \cdots, I_n) \end{aligned} \tag{1}$$

여기서, 문장 S의 품사 태깅 결과는 $I_1 I_2 \cdots I_n$이고, I_i는 i번째 단어/품사 (w_i/t_i)를 나타내며, n은 문장을 구성하는 단어의 수이다. 이를 연쇄규칙(chain rule)에 의해서 전개하면, 식 (2)와 같이 된다.

$$\begin{aligned} R(S) = \Pr(I_1)\Pr(I_2|I_1)\Pr(I_3|I_1, I_2)\cdots \\ \Pr(I_n|I_1, I_2, \cdots, I_{n-1}) \end{aligned} \tag{2}$$

식 (2)를 마코프 가정과 몇 가지 독립 가정을 이용해서 전개하면, 식 (3)을 얻을 수 있으며, 본 논문에서는 이를 품사 태깅의 신뢰도 측정에 대한 기본 모델로 간주한다.

$$\begin{aligned} R(S) &\simeq \Pr(I_1)\Pr(I_2|I_1)\Pr(I_3|I_2)\cdots\Pr(I_n|I_{n-1}) \\ &= \Pr(w_1, t_1)\Pr(w_2, t_2|w_1, t_1)\Pr(w_3, t_3|w_2, t_2)\cdots \\ &\quad \Pr(w_n, t_n|w_{n-1}, t_{n-1}) \\ &\simeq \Pr(t_1|h_1)\Pr(t_2|h_2)\Pr(t_3|h_3)\cdots\Pr(t_n|h_{n-1}) \\ &= \prod_{i=1}^{n}(1-P_e(t_i|h_i)) \end{aligned} \tag{3}$$

여기서 h_i는 문맥이고, $P_e(t_i|h_i)$는 품사 t_i가 h_i라는 문맥에 대해 오류가 발생될 확률이며, 식 (3)으로부터 문장 S에 포함된 i번째 단어/품사의 신뢰도 $R(w_i/t_i)$는 식 (4)와 같이 정의된다.

$$R(w_i/t_i) = 1 - P_e(t_i|h_i) \tag{4}$$

4. 교차확인 방법을 이용한 오류확률 추정

h_i를 결정하는 것은 분류 문제(classification)에서 특징(feature)을 결정하는 것과 같이 대단히 어려운 문제이기[14], 때문에 본 논문에서는 실험을 통하여 결정하도록 하고, 본 절에서는 오류확률 추정 방법에 대해서 논하고자 한다.

일반적으로 오류확률을 추정하는 방법은 여러 가지가 가능하다[14]. 품사 태깅에서 오류확률을 구하는 방법으로 가장 간단한 방법은 품사 태깅 시스템을 이용하여 학습말뭉치를 태깅하고, 그 결과로부터 I_i에 대한 오류확률을 구하는 것이다. 그러나, 일반적으로는 학습말뭉치의 양이 충분하지 않기 때문에 다양한 환경(운용조건)을 오류확률에 충분히 반영할 수 없었다. 이 문제를 다소 완화시키기 위해서 본 연구에서는 교차확인(cross validation) 방법[15]을 이용하여 오류확률을 추정하였다.

교차확인 방법을 이용하기 위해서는 먼저 학습말뭉치를 n개로 나누고, 그 중 $n-1$개의 부학습말뭉치(sub-training corpus)를 이용해서 품사 태깅 시스템을 학습하고, 나머지 1개의 부학습말뭉치를 평가하여 오류확률을 추정한다. 그러나, n을 결정하는 것도 쉽지 않다. n이 너무 크면 1개의 부학습말뭉치의 크기가 너무 작아서 가능한 모든 오류를 추정할 수 없을 것이고, n이 너무 작으면 1개의 부학습말뭉치의 크기가 너무 커서 불필요한 오류에 대해서도 추정하게 될 것이다. 이를 해결하기 위해서 본 논문에서는 중심극한정리(central limit theorem)를 이용하여, n을 30으로 하였다[15]. 따라서 오류확률 $P_e(t_i|h_j)$을 추정하기 위해서 29개의 부학습말뭉치를 이용해서 학습하고, 나머지 하나(k번째 학습말뭉치)를 평가하여 모든 (t_i, h_j)에 대해서 식 (5)와 같이 $P_{e,k}(t_i|h_j)$를 추정한다.

$$P_{e,k}(t_i|h_j) = \frac{C_{e,k}(t_i, h_j)}{C_k(h_j)} \quad (5)$$

여기서 $C_{e,k}(t_i, h_j)$는 k번째 부학습말뭉치에서 문맥 h_j에 대해서 품사 t_i가 오류를 일으킨 횟수이고, $C_k(h_j)$는 k번째 부학습말뭉치에서 문맥 h_j가 발생한 횟수이다. 이 방법을 모든 30개의 부학습말뭉치에 적용하고, 각 부학습말뭉치에 대해서 구해진 $P_{e,k}(t_i|h_j)$를 평균하여 (t_i, h_j)에 대한 오류확률 $P_e(t_i|h_j)$를 식 (6)와 같이 구한다. 이를 종합하면 (그림 2)와 같은 알고리즘으로 정리된다.

알고리즘 : 오류확률 추정 알고리즘
입력 : 학습말뭉치 C
출력 : $P_e(t_i|h_j), \forall\ i,j$
방법 : 1. 학습말뭉치 C를 n개의 부학습말뭉치 C_k로 나눈다.
2. for k = 1, n do
C_k를 제외한 모든 C_i에 대해서 품사 태깅 시스템을 학습한다.
C_k를 품사 태깅하고, 오류를 표시한다.
C_k의 평가를 이용해서 $P_{e,k}(t_i|h_j)$를 구한다.
enddo

$$3.\ P_e(t_i|h_j) = \frac{\sum_{k=1}^{n} P_{e,k}(t_i|h_j)}{n}, \forall i,j \quad (6)$$

(그림 2) 교차확인 방법을 이용한 오류확률 추정 알고리즘

5. 실험 및 평가

5.1 실험 환경

교차확인 방법을 이용해 오류확률을 추정하기 위해 학습말뭉치를 <표 1>과 같이 30개의 부학습말뭉치로 나누었다. 말뭉치를 어떻게 나누느냐에 따라서 정확률에 큰 영향을 줄 수 있으므로 본 논문에서는 각 부학습말뭉치에 여러 장르의 문장이 골고루 포함되도록 하였다. 본 연구에서는 정확률을 높이는 것과는 상관없고 어떻게 오류확률을 정확하게 추출하느냐에 초점을 맞추고 있기 때문에 학습말뭉치를 나누는 데에는 노력을 기울이지 않았다. 부학습말뭉치의 평균 크기는 약 538개의 문장에 5,850개의 어절로 구성되었으며, 각 문장은 평균 10개의 어절로 구성되었고, 각 어절은 평균 2개의 형태소로 구성되었다.

<표 2> 교차확인을 위한 말뭉치의 구성

말뭉치	개수			평균 문장당 어절수	평균 어절당 형태소수
	문장	어절	형태소		
0	559	6,026	13,047	10.77996	2.16512
1	557	6,359	13,692	11.41652	2.15317
2	555	5,979	12,892	10.77297	2.15621
3	553	5,990	12,785	10.83183	2.13439
4	552	6,045	12,974	10.95109	2.14624
...					
27	522	5,634	12,109	10.79310	2.14927
28	522	5,842	12,660	11.19157	2.16707
39	521	5,711	12,360	10.96161	2.16424
평균	538.7	5,850	12,637.8	10.86019	2.16013

5.2 미등록어를 고려한 품사 태깅 시스템의 성능

미등록어는 자연언어처리에서 피할 수 없는 현상이기 때문에 본 논문에서는 미등록어가 있다는 가정 하에서 성능을 평가하였으며, <표 2>가 미등록어를 고려한 품사 태깅 시스템의 성능이다. <표 2>에서는 어절과 형태소 단위의 정확률을 모두 보여주고 있다. 미등록어가 있을 경우에는 일반적으로 약 2~5% 이상의 차이가 있다. 그 만큼 미등록어를 정확하게 예측할 수 있다면 좋은 결과를 얻을 수 있을 것이다. 미등록어는 형태소 해석에서 커다란 문제가 되고 있다. 한국어의 경우, 미등록어 처리의 가장 큰 문제는 미등록어를 추정하는 것이 문제이기는 하지만, 형태소 해석에서 과분석으로 미등록어가 있는지 없는지를 알 수 없기 때문에 발생되는 문제가 더 클 것이다. 품사 태깅 시스템의 성능은 어절에 대해서 약 92%의 정확률을 보이며 형태소에 대해서는 약 95%의 정확률을 보인다. 이는 미등록어를 고려하지 않는 경우, 97.68%[17]보다 약 2%정도 낮은 편이다.

<표 3> 미등록어를 고려한 품사태깅 시스템의 성능

말뭉치	어절 단위			형태소 단위		
	오류수	오류율	정확률	오류수	오류율	정확률
0	465	0.077166	0.922834	631	0.048364	0.951636
1	498	0.078314	0.921686	654	0.047765	0.952235
2	495	0.082790	0.917210	674	0.052280	0.94772
3	527	0.087980	0.912020	686	0.053657	0.946343
4	525	0.086849	0.913151	675	0.052027	0.947973
...						
27	470	0.083422	0.916578	625	0.051615	0.948385
28	461	0.078911	0.921089	648	0.051185	0.948815
29	433	0.075819	0.924181	585	0.047330	0.952670
평균	470.7	0.074322	0.925678	628	0.045243	0.954757

5.3 교차확인 방법에서의 미등록어의 특성

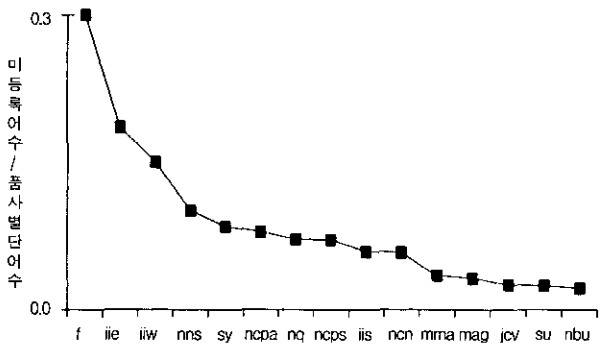
미등록어는 오류확률이 매우 높기 때문에 미등록어는 신뢰도 측정에 많은 영향을 준다. 본 절에서는 교차확인 과정에서 발생되는 미등록어의 특성에 대해서 살펴보고자 한다. 학습말뭉치가 충분히 크다고 하더라도 교차확인을 위해서 나뉘어진 부학습말뭉치의 크기는 충분하지 못할 수 있다. 교차확인 과정에서 발생되는 미등록어의 특성을

일반적으로 사전의 크기가 충분하다는 가정 하에서 생각하는 미등록어의 특성과는 매우 차이가 있다. 본 논문에서 미등록어는 학습말뭉치에 나타나지 않는 단어를 의미한다. 또한 미등록어는 형태소가 처음 나타났을 경우에도 미등록어이지만, 해당하는 형태소가 사전에 있었다고 하더라도 해당하는 품사가 사전에 없을 경우에도 미등록어로 간주한다. <표 9>는 품사별 미등록어 수를 보이고 있다(상위 10개인 품사).

〈표 4〉 품사별 미등록어 갯수

순 위	품 사	빈 도 수	확 률
1	ncn	4233	0.547465
2	nq	808	0.104500
3	ncpa	651	0.084195
4	pvg	540	0.069839
5	mag	330	0.042679
6	nnn	269	0.034790
7	paa	211	0.027289
8	ncps	189	0.024443
9	ef	73	0.009441
10	nbu	57	0.007371

일반적으로 대부분의 자연언어처리 시스템에서는 고유명사(nq)를 미등록어로 간주한다. 이 경우에 대해서는 일반적으로 모든 보통명사가 사전에 포함되어 있다고 가정하기 때문이나, 본 연구에서는 학습말뭉치에 나오는 단어 이외는 사용하지 않았다. 그 결과로 보통명사(ncn)에 대한 미등록어가 전체 미등록어의 약 55%를 차지하고 있다. 이는 전체 단어 중에서 보통명사가 약 70% 이상을 차지하는 것과 무관하지 않다. (그림 3)은 해당 품사가 나타났을 때, 미등록어가 될 확률을 보인 것이다. 시스템에 의해서 미등록어를 추정하는 것이 아니라 학습말뭉치에 포함되어 있지 않으면 모두 미등록어로 간주한다. (그림 3)에서 외국어(f)가 미등록어로 발생할 확률이 가장 높다. 이는 한국어 문장에서 외국어는 자주 사용되지 않을 뿐 아니라 영역의 변화에 따라 매우 다른 외국어가 사용됨을 암시적으로 말해주고 있다.



(그림 4) 품사별 미등록어 확률 분포

감탄사(iie, iiw, iis)의 경우도 마찬가지인데, 감탄사는 일반적인 문장에는 잘 나타나지 않으나, 대화체나 소설 등에서는 매우 잘 나타나는 사실을 반영하고 있다.

5.4 신뢰도 측정

신뢰도 측정을 위해서 말뭉치를 학습말뭉치와 시험말뭉치로 나누었고, 각각에 대해서 신뢰도를 측정해 보았다. 시험말뭉치에는 미등록어가 포함되어 있으며, 미등록어에 대한 신뢰도는 1.0으로 하였다. 그 이유는 미등록어를 사용할 경우에는 100%의 오류가 있음을 알고 있기 때문에 이 미등록어를 오류로 간주해서는 안되므로 신뢰도를 1.0로 하였다. 신뢰도를 측정하기 위해서 h_i는 w_i와 $t_{i-1}w_i$를 사용했다. <표 4>은 실험결과이다.

〈표 4〉 학습말뭉치와 시험말뭉치의 신뢰도

h_i	학습말뭉치	시험말뭉치
w_i	61.64%	60.94%
$t_{i-1}w_i$	99.99%	99.90%

<표 4>에서 보아서 알 수 있듯이 신뢰도 측정은 h_i에 매우 민감하다. <표 4>에서 문맥정보를 많이 사용할수록 신뢰도를 정확하게 구할 수 있음을 말해주고 있다. 문맥이 크면 클수록 주어진 문맥에 대해서 정확한 오류를 찾을 수 있으나, 많은 문맥이 학습말뭉치에 나타나지 않을 수 있으며 이는 미등록어로서 처리된다. 미등록어에 대한 신뢰도는 1.0으로 하기 때문에 학습말뭉치와 큰 차이가 없었다. 문맥 $t_{i-1}w_i$에 대해서는 많은 미등록어로 말미암아 정확한 신뢰도로 보기는 어렵기 때문에 여기서는 문맥 w_i에 대해서 좀 더 분석해보고자 한다.

본 논문에서 사용한 품사 태깅 시스템[17]은 시험말뭉치에 대해서 61% 정도의 신뢰도를 가진다. 이 신뢰도는 61% 이상의 문장은 하나의 오류도 포함하지 않음을 의미한다. 이는 한국어 문장이 평균 20의 형태소로 구성되고, 품사 태깅 시스템이 97.5%의 정확률을 보일 때의 신뢰도에 해당한다 ($0.975^{20} = 0.603$). 본 연구에서 사용된 품사 태깅 시스템이 미등록어가 없을 경우에 97.68%이므로 신뢰도 측정 모델이 어느 정도는 타당한 모델임을 말해주고 있다[1].

6. 신뢰도 측정의 응용

6.1 신뢰도에 바탕을 둔 오류검출

품사 오류를 검출하기 위해서 본 논문에서 사용하는 방

1) 미등록어에 대한 신뢰도를 1.0으로 계산하기 때문에 미등록어가 없는 품사 태깅 시스템에 대한 정확률을 고려해야 한다.

법은 원하는 신뢰도에 따라서 오류를 검출할 수 있도록 한다. 오류검출 방법은 Greedy 알고리즘을 이용하며, 이를 요약하면 (그림 4)와 같다.

알고리즘 : 오류검출 알고리즘
입력 : 품사 태깅 결과(품사 태깅 시스템으로 출력)
출력 : 오류가 표시된 품사 태깅 결과
방법 : 1. 품사 태깅 결과의 각 형태소/품사 w_i/t_i의 $R(w_i/t_i)$를 적재한다.
2. 식(3)을 이용해서 문장에 대한 신뢰도 R_0를 계산한다
3. n = 1
4. 가장 오류를 발생하기 쉬운 형태소/품사 w_i/t_i를 찾아서 오류로 표시한다.
5. 식 (5)을 이용해서 다시 문장에 대한 신뢰도 R_n을 계산한다.

$$R_n = \frac{R_{n-1}}{R(w_i/t_i)} \quad (5)$$

6. if ($R_n < \alpha$) then { $n = n + 1$; goto 4 ; }
7. 오류검출 결과를 출력한다.

(그림 7) 신뢰도를 이용한 오류검출 알고리즘

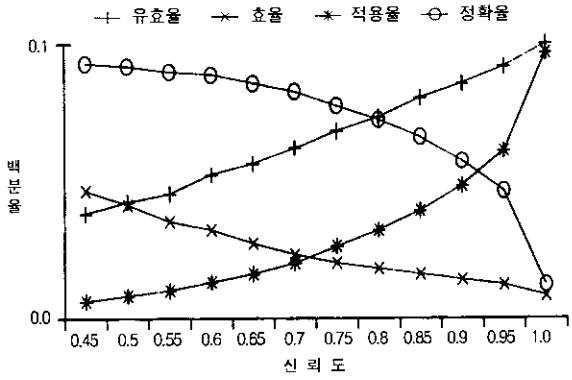
R_i는 i번째 구해진 문장에 대한 신뢰도를 의미하고, R_0는 품사 태깅 시스템의 품사 태깅 결과에 대한 신뢰도를 의미한다. 오류가 일어나기 쉬운 단어를 선택하는 방법은 Greedy 방법을 이용한다. 즉, 오류확률이 높은 것부터 낮은 순으로 정렬하고, 제일 높은 것을 차례로 선택하도록 한다. 이와 같은 방법으로 선택된 오류 형태소에 대해서는 신뢰도를 1.0로 바꾼다. 이 형태소에 대해서는 오류가 수정된 것으로 간주한다는 것이다. 그리고 나서 다시 식 (5)와 같은 방법으로 문장에 대한 신뢰도 R_n을 계산한다. 이렇게 구해진 R_n이 원하는 신뢰도 α보다 작다면 오류가 더 검출되어야 한다는 것을 의미하며, 그렇지 않다면 충분한 신뢰도를 갖추었기 때문에 더 이상 오류를 검출하지 않는다. 본 논문에서는 이 방법을 동적 오류 제어 방법이라고 한다.

6.2 오류검출 방법에 대한 평가

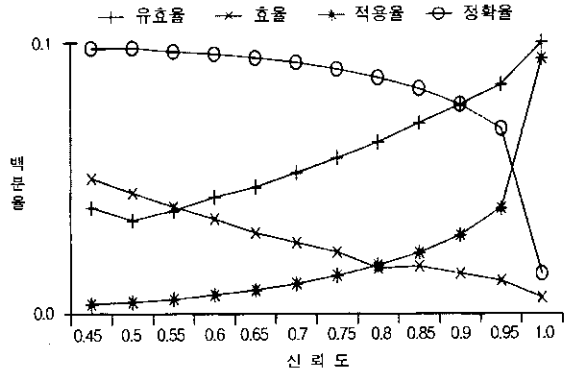
오류검출을 위해서 h_i는 w_i를 이용했다. 앞에서 보았듯이 $t_{i-1}w_i$를 이용할 경우 정확률은 대단히 높을 수 있으나, 견고성에 문제가 있기 때문에 오류검출을 위해서 문맥(h_i)으로 w_i를 사용한다. 동적 오류 제어를 위해서 본 논문에서는 아래와 같은 측도를 사용하였다. 여기서 시험말뭉치 내에 포함된 오류의 총 수를 E, 전체 단어 수를 T, 시스템이 찾은 오류 수를 F, 정확하게 찾은 오류 수를 H, 시스템이 오류와 그렇지 않은 단어를 정확히 구별한 수를 K라고 할 때, 각 측도를 아래와 같은 표현한다[18].

유효율(effectiveness) : 전체 오류 중에서는 올바르게 찾은 비율, $= H/E$
효 율(efficiency) : 시스템에서 찾은 오류 중에서 올바르게 찾은 비율, $= H/F$
적용율(coverage) : 전체 단어 중에서 오류를 찾은 비율, $= F/T$
정확률(accuracy) : 전체 단어 중에서 오류를 정확하게 인식한 비율, $= K/T$

유효율이 가능한 전체 오류를 얼마나 정확하게 찾았는지를 말하며 효율은 시스템이 오류로 검출한 것이 얼마나 정확한가를 말하고 있다. 따라서 유효율과 효율은 반비례한다. 유효율을 높이기 위해서는 극단적으로 모든 것으로 오류로 검출하면 100%가 되나 효율을 최소로 떨어지게 된다. 적용율은 전체 단어 중에서 오류를 검출된 단어 수이므로 유효율과 정비례한다. 정확률은 시스템이 오류인 것과 그렇지 못한 것을 구별할 수 할 수 있는 능력이다. 이들의 측도를 어절(그림 5)과 형태소(그림 6)로 나누어 시험말뭉치에 대해서 평가하였다.



(그림 7) 어절 단위의 오류검출 평가



(그림 8) 형태소의 오류검출 평가

(그림 5)와 (그림 6)는 신뢰도를 0.45에서 1.0까지 0.5씩 증가했을 때 오류검출에 대한 성능을 보이고 있다. 정확률은 신뢰도가 증가할수록 감소한다. 한국어 문장은 약 20개 정도의 형태소로 구성되고, 품사 태깅 시스템의 성능이 95%이므로 1개 정도의 오류를 포함하고 있는데 일반적으로 오류검출에서 1개 이상의 오류를 검출하므로 잘못 검출한 오류가 많이 포함되기 때문이다. 이것은 효율도 마찬가지 이유로 신뢰도가 증가할수록 감소한다. 반면에 적용률과 유효율은 신뢰도가 증가할수록 증가하게 된다. 평균적으로 문장당 1개의 오류를 포함하기 때문에 오류의 개수를 오류검출에 대한 임계값으로 사용할 수도 있으나, 이는 입력 문장의 길이가 다양하며, 문장의 길이가 짧더라도 많은 오류가 포함될 수 있기 때문에 오류검출에 임계값으로 사용하기에는 적당하지 않다.

7. 결 론

본 논문에서는 품사 태깅 시스템의 신뢰도 측정 방법을 기술하고 이 방법을 이용해서 품사 태깅에서의 오류검출 방법을 제시하였다. 품사 태깅의 신뢰도는 품사 태깅 시스템이 어느 정도의 오류를 포함하고 있는가를 수치적인 측도로 표현한 것이다. 본 논문에서 개발된 품사 태깅 시스템은 시험말뭉치에 대해서 61% 정도의 신뢰도를 가진다. 이는 한국어 문장이 평균 20의 형태소로 구성되고, 품사 태깅 시스템이 97.5%의 정확률을 보일 때의 신뢰도에 해당한다. 본 연구에서 사용된 품사 태깅 시스템[17]이 미등록어가 없을 경우에 97.68%이므로 신뢰도 측정 방법이 어느 정도는 타당한 방법임을 알 수 있었다.

일반적으로 확률이나 신뢰도를 추정하기 위해서는 많은 양의 학습 말뭉치가 필요하다. 본 연구에서는 이를 해결하기 위해서 교차 확인 방법을 이용하였으며, 이 방법은 주로 시스템을 평가하기 위해 사용되는 방법이다. 본 논문에서는 이를 확률함수 추정, 즉, 보여지지 않은 확률함수를 추정하는 용도로 사용되었으며, 실험을 통해서 매우 효과적임을 알 수 있었다.

본 논문에서 신뢰도 측정은 학습 말뭉치에서 각 단어가 유발시키는 오류에 근거를 두고 있다. 이 신뢰도는 오류를 고려한다는 점에서 확률과 크게 다르며, 근본적으로는 확률에 기반을 둔 연구이다. 본 연구는 이와 같은 신뢰도를 기반으로 오류 검출 시스템을 구현해 보임으로써 신뢰도의 유용성을 평가하였다.

앞으로 신뢰도 모델을 좀더 많은 응용 시스템에 적용하여 이 모델의 유용성이 평가되어야 할 것이다. 또한 오류확률을 추정하기 위한 동적 문맥에 좀더 깊은 연구가 요구된다.

참 고 문 헌

[1] Dermatas, E. and Kokkinakis, G., "Automatic stochastic tagging of natural language texts," *Computational Linguistics*, Vol.21, No.2, pp.137-163, 1995.

[2] Brill, E.(1995). "Tranformation-based error driven learning and natural langauge processing : a case study in part-of-speech tagging," *Computational Linguistics*, Vol. 21, No.4, pp.543-564.

[3] Church, K. W. and Mercer, R. L., "A introduction to the special issue on computational linguistics using large corpora," *Computational Linguistics*, Vol.19, No.1, pp. 1-24, 1993.

[4] 김재훈, "오류-보정 기법을 이용한 어휘 모호성 해소", 한국과학기술원, 전산학과, 박사학위 논문, 1996.

[5] Lin, Y.-C., Chiang, T.-H., and Su, K.-Y., "Automatic model refinement - with an application to tagging," *Proceedings of the International Conference on Computational Linguistics (COLING-94)*, Kyoto, Japan, pp. 148-153, 1994.

[6] 박경수, 신뢰도 공학 및 정비이론, 희중당, 1987.

[7] Ramakumar, R., *Engineering Reliability : Fundamental and Applications*, Prentice Hall, 1993.

[8] 김재훈, 임철수, 서정연, "은닉 마르코프 모델을 이용한 효율적인 한국어 품사 태깅", 정보과학회논문지, 제22권, 제1호, pp.136-146, 1995.

[9] 신중호, 한영석, 박영찬, 최기선, "어절구조를 반영한 은닉 마르코프 모델을 이용한 한국어 품사 태깅", 제6회 한글 및 한국어 정보처리 학술대회 발표논문집, 시스템공학연구소, 대전, pp.389-364, 1994.

[10] 이운재, "한국어 문서 태깅 시스템의 설계 및 구현", 한국과학기술원, 전산학과, 석사학위논문, 1993.

[11] 이상주, 임희석, 임해창, "은닉 마르코프 모델을 이용한 두 단계 한국어 품사 태깅," 제6회 한글 및 한국어 정보처리 학술대회발표논문집, 시스템공학연구소, 대전, pp.305-312, 1994.

[12] 김진동, 임희석, 임해창, "Twoply HMM : 한국어 특성을 고려한 형태소 단위의 품사 태깅 모델", 정보과학회논문지 (B), 제25권, 제1호, pp.183-192, 1998.

[13] 임희석, "언어 지식과 통계 정보를 이용한 한국어 품사 태깅 모델", 고려대학교, 컴퓨터학과, 박사학위논문, 1997.

[14] Nadler, M. and Smith, E. P., *Patterhn Recognition Engineering*, John Wiley & Sons, 1993.

[15] Helstrom, C. W., *Probability and Stochastic Process for Engineering*, Macmillan Publishing Company, 1991.

[16] Cohen, P. R., *Empirical Methods for Artificial Intelligence*, The MIT Press, 1995.

[17] 김재훈, "가중치 망을 이용한 한국어 품사 태깅", 정보과학회논문지(B), 제25권, 제6호, pp.951-959, 1998.

[18] Forster, G. F. *Stastical Lexical Disambiguation*, Master's Thesis, McGill University, School of Computer Science, 1991.

김 재 훈

e-mail : jhoon@hanara.kmaritime.ac.kr

1986년 계명대학교 전자계산학과(학사)

1988년 한국과학기술원 전산학과(공학석사)

1996년 한국과학기술원 전산학과(공학박사)

1988년~1997년 한국전자통신연구원, 선임연구원

1997년~1999년 한국해양대학교, 컴퓨터공학과, 전임강사

1999년~현재 한국해양대학교, 컴퓨터공학과, 조교수

2000년~현재 한국과학기술원 첨단정보기술연구센터, 연구원

2001년~현재 USC, Information Sciences Institute, 방문연구원

관심분야 : 자연언어처리, 한국어 정보처리, 정보검색, 음성언어처리