은닉 마르코프 모델을 이용한 음차표기된 외래어의 자동인식 및 추출 기법

Automatic Detection and Extraction of Transliterated Foreign Words Using Hidden Markov Model

오 종 훈* 최 기 선*
(Jong-Hoon Oh) (Key-Sun Choi)

요 약 본 논문에서는 한국어문서에서 음차표기된 외래어를 자동적으로 인식 및 추출하는 알고리즘을 제안한다. 제안된 방법에서는, 음차표기된 외래어 인식 및 추출 문제를 음절태깅문제로 변환한다. 음절태깅문제는 주어진 단어 내의 음절들에 대하여 순수 한국어를 구성하는 음절인지 또는 음차표기된 외래어를 구성하는 음절인지를 태깅하는 작업으로 정의된다. 이를 위하여, 주어진 어절 내의 음절의 나열을 순수 한국어 음절을 표현하는 상태와 외래어 음절을 표현하는 상태의 이진 상태(binary state)로 모델링한 은닉 마르코프 모델을 이용한다. 제안된 방법은 기존 연구에 비하여 높은 재현율과 정확률로 음차표기된 외래어를 인식 및 추출하였다.

주제어 외래어, 음차표기, 은닉 마르코프 모델, 음절태깅, 교차언어 정보검색, 자연언어처리

Abstract In this paper, we describe an algorithm for transliterated foreign word extraction in Korean language. In the proposed method, we reformulate the transliterated foreign word extraction problem as a syllable-tagging problem such that each syllable is tagged with a transliterated foreign syllable tag or a pure Korean syllable tag. Syllable sequences of Korean strings are modeled by Hidden Markov Model whose state represents a character with binary marking to indicate whether the character forms a Korean word or not. The proposed method extracts a transliterated foreign word with high recall rate and precision rate. Moreover, our method shows good performance even with small-sized training corpora.

1. 서론

한국어 문서에서 음차표기된 외래어[1]는 표준화된 외래어 표기법(외래어 표기법)이 있음에도 불구하고 사용자마다 다양한 형태로 표기된다. 예를 들어, 영어 단어인 'data'의 경우 한국어로 '데이터', '데이타' 등으로 음차표기된다[1, 5, 10]. 이러한 음차표기된 외래어의 다양한 형태로 인해 형태소 분석에 사용되는 사전에 등재되지 않는 경우가 많아, 한국어 형태소 분석에 있어 미등록어 문제의 원인이 된다. 일반적으로 한국어에 있어 하나의 어절은 내용어와 기능어로 구성된다. 예를 들어 '나는 학교에 그랑 간다.'라는 문장은 (1)과 같이 형태소 분석될 수 있다. (1)에서 밑줄친 형태소는 기능어를 나타내며, '는'은 주격 조사, '에'는 여격 조사, '랑'은 공동격 조사, 'ㄴ다'는 어미를 각각 나타낸다.

(1) 나+<u>는</u> 학교+<u>에</u> 그+<u>랑</u> 가+<u>ㄴ다</u>

* 한국과학기술원 전자전산학과/전문용어언어공학 연구센터
대전시 유성구 구성동 373-1, 우:305-701
{rovellia, kschoi}@world.kaist.ac.kr

Division of Computer Science, Department of EECS
Korea Advanced Institute of Science and Technology/KORTERM
373-1 Kusong-dong, Yusong-gu, Taejon, 305-701

전화: 042-869-5565
FAX: 042-867-8790
연구분야 : 자연언어처리, 정보검색, 한국어정보처리

1) 본 논문에서는 '음차표기된 외래어'와 '외래어'를 같은 의미로 사용한다.

한국어에 있어 기능어는 제한적인 형태로 나타나기 때문에, 내용어가 형태소 분석 사전에 등재되어 있지 않는 경우(내용어가 미등록어일 경우) 제한적인 형태의 기능어를 단순히 제거함으로써 기능어 앞에 나타나는 해당 내용어(나, 학교, 그, 가)와 내용어의 품사를 추정할 수 있다[2]. 하지만 이러한 단순한 경험적 규칙(heuristic)은 '오페라는'과 같은 음차표기된 외래어를 포함하는 어절의 경우 오류를 야기할 수 있다. '오페라'가 미등록어일 경우, [2]의 방법으로 처리하였을 때 '오페라는'을 '오페+라는'으로 잘못 분석하게 된다(올바른 분석은 '오페라+는'). '오페라는'에는 두 가지 가능한 기능어인 '는'과 '라는'이 존재하며, 기능어의 최장일치에 의해 '라는'이 기능어로 처리되어 '오페+라는'이라고 형태소 분석된다. 순수 한국어 어절의 경우 내용어와 기능어와의 결합에 있어 이러한 오류는 매우 드물게 발생하므로 대부분의 오류가 순수 한국어가 아닌 음차표기된 외래어에 의한 것으로 생각할 수 있다. 예를 들어 전 프랑스 대통령의 이름인 'Mitterang'의 경우 한국어로 '미테랑'으로 음차표기된다. 또한 '미테랑'이 미등록어일 경우 [2]의 방법에 의해 '미테+랑'으로 오분석될 수 있다.

이러한 문제를 해결하기 위해서는 주어진 어절에 음차표기된 외래어가 존재하는지를 인식하고 이를 추출하는 정교한 알고리즘이 필요하다. 본 논문에서는 음차표기된 외래어를 인식하고 추출하는 효율적인 알고리즘을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 다음 장에서는 먼저 관련연구로서 바이그램과 유니그램을 이용한 음차표기 인식 및 추출 방법에 대해서 살펴보고, 3장에서는 본 논문에서 제안하는 음차 표기 자동 인식 및 추출 알고리즘에 대하여 자세히 설명한다. 그리고 4장에서는 실험을 통해 본 논문에서 제시한 방법의 효용성을 보이며, 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

음차표기된 외래어를 추출하는 최근의 연구에서는 순수 한국어와 음차표기된 외래어간의 음절연결 상이성에 기반한 통계정보를 이용하였다[7, 11, 12]. 이들 연구에서는 음차표기된 외래어를 추출하기 위하여 '음차표기 외래어 인식'과 '음차표기 외래어 추출'[2]이라는 두 단계의 과정을 거쳐 음차표기된 외래어를 추출하였다.

첫 번째 단계에서는 통계정보를 이용하여 주어진 어절에 음차표기된 외래어가 포함되었는지를 결정한다. 이러한 통계정보는 음절의 유니그램(unigram)과 바이그램(bigram) 정보를 이용하여 식 (1)과 같이 사용된다.

$$D(W) = \frac{P(Foreign|W)}{P(Korean|W)} = \frac{P(W|Foreign) \times P(Foreign)}{P(W|Korean) \times P(Korean)} \qquad (1)$$

여기에서 $P(Foreign|W)$와 $P(Korean|W)$는 어절 W가 음차표기된 외래어인지 순수 한국어인지에 대한 조건부 확률을 나타낸다.

만약 $D(W)>1$일 경우, W가 음차표기된 외래어를 포함하는 것으로 판단한다. 식 (1)에서 $P(Foreign)$와 $P(Korean)$는 학습코퍼스에 나타난 한국어 단어와 음차표기된 외래어 단어의 빈도수를 이용하여 추정한다. 또한 $P(W|Foreign)$와 $P(W|Korean)$의 추정을 위하여 단어에 나타난 음절의 유니그램과 바이그램을 이용하였다.

두 번째 단계인 음차표기 외래어 추출 단계에서는 첫 번째 단계인 외래어 인식 단계에서 음차표기된 외래어를 포함한다고 판단된 어절에 대하여 순수 한국어와 접두사 및 접미사 등을 제거하여 음차표기된 외래어를 추출한다.

기존의 방법은 비교적 좋은 성능으로 음차표기된 외래어를 추출하였지만 몇 가지 한계를 가지고 있다. 첫 번째로, 기존의 방법은 두 단계의 과정으로 이루어져 있으며 첫 번째 단계인 외래어 인식의 성능이 낮아 두 번째 단계인 외래어 추출의 성능에 많은 영

2) '음차표기 외래어 인식'이란 "주어진 어절 \W에 대하여, \W에 음차표기된 외래어가 포함되어 있는지 없는지를 판단하는 문제"로 정의된다. 이에 반해 '음차표기 외래어 추출'은 "주어진 어절 \W에 대하여 \W에 포함된 외래어를 추출하는 문제"로 정의된다. 외래어 인식의 결과물이 '주어진 어절에 외래어가 포함되었나 아닌가'를 결정하는 이진값(binary value)이라면 외래어 추출의 결과물은 주어진 어절에 포함된 외래어 자체이다. 따라서 외래어 추출은 주어진 어절내에 포함된 외래어가 아닌 부분을 걸러내고 외래어가 포함된 부분만을 제시하여야 하기 때문에 외래어 인식보다 복잡한 과정이 필요하다. 예를 들어, '정보검색'과 '정보검색시스템'이라는 어절에 대하여, 외래어 인식단계에서는 '정보검색'은 외래어를 포함하지 않는다는 것을 인식하는 값 0을 할당하고 '정보검색시스템'은 외래어를 포함한다고 인식하는 값 1을 할당할 수 있다. 외래어 추출 단계에서는 외래어 인식 값이 1인 어절에 대하여 외래어를 추출한다. '정보검색시스템'이 외래어가 포함되었다고 인식되었기 때문에, '정보검색시스템'에서 외래어 '시스템'을 추출할 수 있다.
본 논문에서는 '음차표기 외래어 인식'과 '외래어 인식'을 같은 의미로, '음차표기 외래어 추출'과 '외래어 추출'을 같은 의미로 사용한다.

향을 끼친다. 이는 주어진 어절에 외래어가 포함되어 있다 할지라도 첫 번째 단계에서 외래어를 포함하지 않는 것으로 인식한 경우에는 두 번째 단계에서 외래어를 추출하지 않는다는 것이다. 두 번째로 기존 방법은 주어진 어절에 나타난 음차표기된 외래어를 구성하는 음절의 개수에 기반하여 음차표기된 외래어를 포함하는 어절을 인식한다. 이러한 특성으로 인해 기존의 방법은 음차표기된 외래어가 하나의 어절을 구성할 경우에는 좋은 결과를 나타낸다. 하지만 한국어에 있어 외래어는 순수 한국어와 함께 합성명사를 만들거나 기능어와 함께 하나의 어절을 구성하는 경우가 많기 때문에, [7, 11, 12]의 방법은 외래어를 구성하는 음절에 비해 순수 한국어를 구성하는 음절의 개수가 많은 어절의 경우 외래어를 올바르게 인식하거나 추출하지 못하게 된다. 예를 들어, '객체지향시스템에서'[3]의 경우에 외래어를 포함하지 않은 것으로 인식한다.

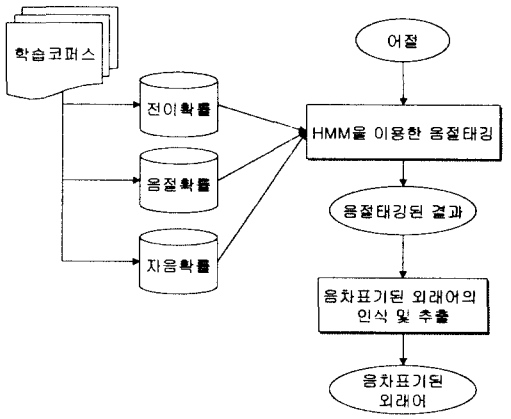
이러한 한계점들을 극복하기 위하여 본 논문에서는 은닉 마르코프 모델을 이용하여 주어진 어절에 나타나는 음차표기된 외래어를 자동적으로 인식 및 추출하는 알고리즘을 제안한다.

3. 음차표기된 외래어의 자동 인식 및 추출

3.1. 문제 정의

주어진 어절에 나타나는 외래어를 자동적으로 추출하기 위한 방법은 "외국 언어와 한국어는 음운학적으로 서로 체계가 다르기 때문에 순수 한국어를 표기할 때 자주 사용되는 음절과 음차표기된 외래어를 표기할 때 자주 사용되는 음절은 서로 다를 것이다"라는 전제에 기반한다. 특히, 영어의 경우 영어에서 자주 사용하는 자음인 'p', 't', 'c', 'f'는 순수 한국어를 표기할 때 자주 사용하지 않는 자음인 'ㅍ', 'ㅌ', 'ㅋ', 'ㅍ'로 각각 음차표기된다. 이러한 특성은 한국어 문서에서 외래어를 추출하는데 중요한 단서가 될 수 있다. 예를 들어, '오페라(opera)'에서 '페'는 순수 한국어를 표기할 때 자주 사용하지 않는 'ㅍ'이라는 자음을 사용하기 때문에 외래어를 구성하는 음절일 가능성이 높다고 유추할 수 있다. 하지만, 한국어의 음절은 자음과 모음의 조합으로 이루어져 있기 때문에 이러한 자음 정보만으로는 올바르게 외래어와 순수 한국어를 구성하는 음절인지를 판단하기 어렵다. 따라서 본 논문에서는 자음정보를 포함하는 음절정보와 전이정보만을 이용한 경우, 그리고 음절정보, 전이정보, 자음정보를 고려한 경우로 나누어 어절 내에 포함되어 있는 외래어를 인식 및 추출한다.

본 논문에서는 이러한 외래어 인식 및 추출 문제를 음절태깅이라는 문제로 변환하여 문제를 해결한다. 음절태깅은 외래어를 구성하는 음절을 표현하는 태그와 순수 한국어를 구성하는 음절을 표현하는 태그를 정의하고 이들 태그에 기반하여 주어진 어절 내의 음절을 태깅하는 작업을 말한다. 이러한 음절태깅을 모델링하기 위하여 품사 태깅(POS-tagging)에 자주 사용되는 은닉 마르코프 모델(Hidden Markov Model)을 사용하였다. 음절태깅을 위한 은닉 마르코프 모델은 주어진 어절 내의 음절의 나열을 순수 한국어 음절을 표현하는 상태와 외래어 음절을 표현하는 상태의 이진 상태(binary state)로 모델링한다. 또한, 은닉 마르코프 모델에서 사용될 전이확률(transition probability), 음절확률(syllable probability), 자음확률(consonant probability)은 음절이 외래어를 구성하는 음절인지 순수 한국어를 구성하는 음절인지 태그된 학습 코퍼스로부터 추정된다.



(그림 1) 외래어 인식 및 추출 흐름도

이러한 모델링을 통하여, 주어진 어절에 나타나는 음절이 순수 한국어를 구성하는 음절일 경우에는 'K'라는 태그를, 외래어를 구성하는 음절일 경우에는 'F'라는 태그를 할당한다. 예를 들어, 어절 '오페라는'과 '미테랑'은 (2)와 같이 태깅될 수 있다.

3) 외래어를 구성하는 3음절인 '시스템'과 순수 한국어를 구성하는 '객체', '지향', '에서'의 6음절이 있다.

(2) 오페라는 => 오/F + 페/F + 라/F + 는/K.
미테랑 => 미/F + 테/F + 랑/F

음절태깅을 통한 외래어 추출은 태깅된 결과에서 'F'태그의 연속을 추출하는 작업으로 생각할 수 있으며, 외래어 인식은 태깅된 결과에서 'F'로 태그된 음절을 인식하는 작업으로 생각할 수 있다.

본 논문에서 제안한 음절태깅을 통한 외래어 인식 및 추출 알고리즘의 전체 과정은 (그림 1)과 같이 나타내어진다. 우선 음절태깅된 학습 코퍼스로부터 학습을 통하여 음절태깅에 필요한 정보를 추출한다. 추출된 정보를 이용하여 주어진 어절에 대하여 음절태깅 작업을 수행한 후 음절태깅된 결과에서 태그정보를 이용하여 외래어 인식 및 추출한다.

3.2. 은닉 마르코프 모델을 이용한 음차표기된 외래어의 자동 인식 및 추출

본 논문에서는 음차표기된 외래어 자동 인식 및 추출 문제를 음절태깅 문제로 정의한다. 음절태깅 문제는 식 (2)에서와 같이 $s_{1...n}$까지의 n개의 음절로 구성된 어절 S에 대하여, 확률 $P(t_{1...n}|s_{1...n})$를 최대화시키는 태그의 나열인 $t_{1...n}$을 찾는 문제로 정의된다. 식 (2)에서 $P(S|T)P(S)$는 조건부확률의 연속으로 표현할 수 있으므로 식 (3)과 같이 변환이 가능하다[13]. 또한 마르코프 독립 가정(Markov Independence Assumption)[8]에 의하여 식 (3)은 식 (4)로 변환 가능하다.

$$\phi(S) \stackrel{\text{def}}{=} \arg\max_T P(T|S) = \arg\max_T P(S|T)P(T) \quad (2)$$

$$P(S|T)P(T) = \prod_{i=1} P(t_i|t_{i-1},\ldots,t_1,s_{i-1},\ldots,s_1)\times P(t_i|t_{i-1},t_{i-2},\ldots,t_1) \quad (3)$$

$$P(S|T)P(T) = p(t_1|t_0)\times \prod_{i=2} p(t_i|t_{i-1},t_{i-2})\times \prod_{i=1} p(t_i|s_i,s_{i-1},t_{i-1}) \quad (4)$$

또한 자음정보를 추가하기 위하여 식 (4)는 식(5)와 같이 표현할 수 있다[4].

$$P(S|T)P(T) = p(t_1|t_0)\times \prod_{i=2} p(t_i|t_{i-1},t_{i-2})\times \prod_{i=1} p(t_i|s_i,s_{i-1},c_i,c_{i-1},t_{i-1}) \quad (5)$$

여기에서, s_i는 주어진 어절의 i번째 음절, c_i는 i번째 음절의 자음, t_0는 어절의 시작을 나타내는 태그, t_i는 어절의 i번째 음절에 대한 태그(F 또는 K)를 각각 나타낸다.

본 논문에서는 $p(t_i|t_{i-1},t_{i-2})$, $p(t_i|s_i,s_{i-1},t_{i-1})$, $p(t_i|s_i,s_{i-1},c_i,c_{i-1},t_{i-1})$를 추정하기 위하여 각 음절이 순수 한국어를 구성하는 음절인지 음차표기된 외래어를 구성하는 음절인지를 수작업으로 태그한 코퍼스를 사용하였다. 음절 태그된 학습코퍼스를 이용하여 전이확률 $p(t_i|t_{i-1},t_{i-2})$와 음절확률 $p(t_i|s_i,s_{i-1},t_{i-1})$, 그리고 자음 및 음절확률인 $p(t_i|s_i,s_{i-1},c_i,c_{i-1},t_{i-1})$는 식 (10)과 같이 추정된다[9]. 식 (10)에서 음절확률과 자음 및 음절확률은 자료희귀문제를 방지하기 위하여 선형결합을 통하여 보완된다. 음절은 자음과 모음으로 구성되므로 i^{th} 음절 s_i의 자음을 c_i, 모음을 v_i라고 정의하면, s_i는 c_i, vi으로 나타낼 수 있으므로, $p(t_i|s_i,c_i)$와 $p(t_i|s_i,s_{i-1},c_i,c_{i-1},t_{i-1})$를 식 (8)과 (9)와 같이 추정할 수 있다.

$$p(t_i|t_{i-1},t_{i-2}) = \frac{C(t_i,t_{i-1},t_{i-2})}{C(t_{i-1},t_{i-2})},\quad p(t_i|t_{i-1}) = \frac{C(t_i,t_{i-1})}{C(t_{i-1})} \quad (6)$$

$$p'(t_i|s_i) = \frac{C(t_i,s_i)}{C(s_i)},\quad p'(t_i|s_i,s_{i-1},t_{i-1}) = \frac{C(t_i,s_i,s_{i-1},t_{i-1})}{C(s_i,s_{i-1},t_{i-1})}$$
$$p(t_i|s_i,s_{i-1},t_{i-1}) = \lambda_1\times p'(t_i|s_i)+(1-\lambda_1)\times p'(t_i|s_i,s_{i-1},t_{i-1}) \quad (7)$$

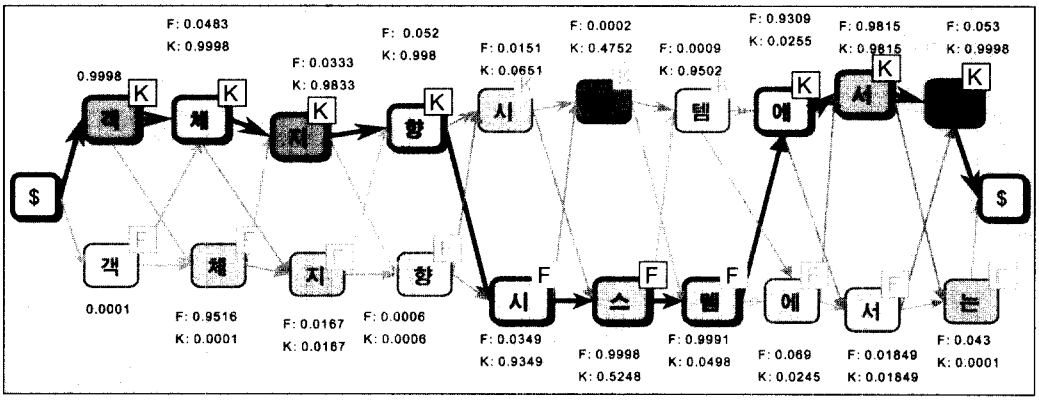
$$p(t_i|s_i,c_i) = \lambda_2\times\frac{C(s_i,t_i)}{C(s_i)}+(1-\lambda_2)\times\frac{C(c_i,t_i)}{C(c_i)}$$
$$= \lambda_2\times\frac{C(c_i,v_i,t_i)}{C(c_i,v_i)}+(1-\lambda_2)\times\frac{C(c_i,t_i)}{C(c_i)} \quad (8)$$

$$p(t_i|s_i,s_{i-1},c_i,c_{i-1},t_{i-1})$$
$$= \lambda_2\times\frac{C(s_i,s_{i-1},t_i,t_{i-1})}{C(s_i,s_{i-1},t_{i-1})}+(1-\lambda_2)\times\frac{C(c_i,c_{i-1},t_i,t_{i-1})}{C(c_i,c_{i-1},t_{i-1})}$$
$$= \lambda_2\times\frac{C(c_i,c_{i-1},v_i,v_{i-1},t_i,t_{i-1})}{C(c_i,c_{i-1},v_i,v_{i-1},t_{i-1})}+(1-\lambda_2)\times\frac{C(c_i,c_{i-1},t_i,t_{i-1})}{C(c_i,c_{i-1},t_{i-1})} \quad (9)$$

$$p(t_i|s_i,s_{i-1},c_i,c_{i-1},t_{i-1}) =$$
$$\lambda_1\times p'(t_i|s_i,c_i)+(1-\lambda_1)\times p'(t_i|s_i,s_{i-1},c_i,c_{i-1},t_{i-1}) \quad (10)$$

여기에서, $C(T)$는 음절 태그된 학습 코퍼스에 나타나는 T의 빈도수를 나타낸다.

t_{i-2}	t_{i-1}	t_i	$p(t_i \mid t_{i-1},t_{i-2})$	t_{i-2}	t_{i-1}	t_i	$p(t_i \mid t_{i-1},t_{i-2})$
	$	F	0.191102	F	F	K	0.092871
	$	K	0.808898	F	K	$	0.707317
$	F	$	0.039602	F	K	F	0.00238
$	F	F	0.956061	F	K	K	0.290303
$	F	K	0.004337	K	F	$	0.020217
$	K	$	0.021474	K	F	F	0.97288
$	K	f	0.002862	K	F	K	0.006903
$	K	K	0.975663	K	K	$	0.468705
F	F	$	0.325715	K	K	F	0.009444
F	F	F	0.581415	K	K	K	0.521851



(그림 2) '객체지향시스템에서는'의 음절태깅 도식화

식 (4), (5), (10)을 이용하여 주어진 어절에 대하여 음절태깅을 수행한다. 음절태깅결과에서 'F' 태그가 존재하면 주어진 어절 내에 음차표기된 외래어가 존재한다고 판단하며, 'F' 태그의 연속을 음차표기된 외래어로 추출한다. 예를 들어, 어절 '객체지향시스템에서는'는 〈표 1〉과 같이 음절태깅된다.

〈표 1〉 '객체지향시스템에서는'의 음절태깅 예

음절	객	체	지	향	시	스	템	에	서	는
음절태깅 결과	K	K	K	K	F	F	F	K	K	K

〈표 1〉의 결과에서 음절태깅결과 내에 'F' 태그가 존재하므로, '객체지향시스템에서는'은 음차표기된 외래어를 포함한다고 판단되며, 'F' 태그의 연속인 '시스템'은 음차표기된 외래어로 추출된다. 따라서 '객체지향시스템에서는'내의 순수 한국어는 '객체지향'과 '에서는'이 되고 '시스템'은 음차표기된 외래어가 된다. (그림 2)는 〈표 1〉의 음절태깅 과정을 도식화한 것이다.

(그림 2)에서 각 노드에 할당되어 있는 확률값은 $p(t_i|s_i, s_{i-1}, t_{i-1})$을 나타내며, 'F:확률, K:확률'의 형식으로 되어 있다. 이전 태그가 'F'일 경우에는 'F:확률'의 값이 되고, 이전태그가 'K'일 경우에는 'K:확률'의 값이 된다. 표로 나타난 부분은 전이확률을 나타낸다. (그림 2)에서 "$"에서 "객:K"로 가는 경로는 하나밖에 존재하지 않으므로 $P(\$ \rightarrow 객{:}K)$는 식 (4)에 의해 $p('K'|\$) \times p('K'|객, \$)$로 표현된다. 또한 "$"에서 "객:F"로 가는 경로도 하나밖에 존재하지 않으므로 $P(\$ \rightarrow 객{:}F)$도 식 (4)에 의해 $p('F'|\$) \times p('F'|객, \$)$로 표현될 수 있다.

두 번째 노드인 "체:K"에 대하여 다음과 같은 조건부 확률을 구하고, 이 중에서 높은 값을 가지는 노드가 "체:K"에 연결될 수 있는 이전 노드 (previous node)가 될 수 있다.

여기서 $t_{i-2} = '\$'$는 이전 단계에서 "객:K"가 선택한 이전노드가 '$'로 선택되기 때문에 $t_{i-1} = 'K'$일 경우에는 $t_{i-2} = '\$'$가 선택된다.

$$p(객{:}K \rightarrow 체{:}K) = p(t_i = 'K' | t_{i-1} = 'K', t_{i-2} = '\$') \times$$
$$p(t_i = 'K' | s_i = 체, s_{i-1} = 객, t_{i-1} = 'K')$$

$$p(\text{객:}F\rightarrow\text{체:}K)=p(t_i='K'|t_{i-1}='F',t_{i-2}='\$')\times$$
$$p(t_i='K'|s_i=\text{체},s_{i-1}=\text{객},t_{i-1}='F')$$

마찬가지 방법으로 "체:F"에 대하여 다음과 같은 조건부 확률을 구하고 "체:F"에 연결될 수 있는 이전 노드를 결정한다. 여기서 $t_{i-2}='\$'$는 이전 단계에서 "객:F"가 선택한 이전노드가 '$'로 선택되기 때문에 $t_{i-1}='F'$일 경우에는 $t_{i-2}='\$'$가 선택된다.

$$p(\text{객:}K\rightarrow\text{체:}F)=p(t_i='F'|t_{i-1}='K',t_{i-2}='\$')\times$$
$$p(t_i='F'|s_i=\text{체},s_{i-1}=\text{객},t_{i-1}='K')$$
$$p(\text{객:}F\rightarrow\text{체:}F)=p(t_i='F'|t_{i-1}='F',t_{i-2}='\$')\times$$
$$p(t_i='F'|s_i=\text{체},s_{i-1}=\text{객},t_{i-1}='F')$$

이러한 과정을 모든 노드에 대하여 적용하며 각 노드의 이전노드가 선택한 현재노드의 상태를 알 수 있으며, 이 정보를 이용하여 비터비 알고리즘[8]을 적용하면 (그림 2)와 같은 굵은 실선의 경로를 얻게 된다. 비터비 알고리즘의 적용은 마지막 노드인 '$'로부터 시작하여 '$'의 이전 노드인 '는:K'를 선택하고, '는:K'에서는 '는:K'의 이전 노드인 '서:K'를 선택한다. 이러한 방법을 반복하면 최초 노드인 '$'까지의 경로를 결정할 수 있게 된다.

4. 실험 및 평가

4.1. 실험 데이터

본 논문에서는 제안된 외래어 인식 및 추출 방법에 대한 실험을 하기 위하여 전기 전자 및 컴퓨터 분야의 4,414문서를 포함하는 KT 실험집합[3]과 생물학, 물리학 등의 과학기술 분야의 13,515 문서를 포함하는 KRIST 실험집합[6]을 사용하였다. 또한 각 실험집합에 나타나는 어절에 대하여 수작업으로 'F'와 'K'로 음절태깅하였다. 〈표 2〉는 두 실험집합에 나타나는 어절에 대하여, 순수 한국어만으로 구성되었는지 외래어를 포함하고 있는지에 대한 분석결과를 나타낸다.

〈표 2〉 각 실험집합의 특성 및 분석 결과

	KW	TFW	전체
KRIST 실험집합	52,598 (87.58%)	7,465 (12.42%)	60,054
KT 실험집합	29,762 (72.24%)	11,495 (27.86%)	41,257

〈표 2〉에서 KW는 순수 한국어만으로 구성된 어절의 개수를 나타내며, TFW는 음차표기된 외래어를 포함하는 어절의 개수를 나타낸다. 분석 결과에서 KT 실험집합(약 27.9%의 어절이 외래어를 포함)은 KRIST 실험집합(약 12.4%의 어절이 외래어를 포함)보다 많은 수의 어절이 음차표기된 외래어를 포함하고 있다. 실험결과에서는 이러한 실험집합의 특성이 실험결과에 미치는 영향에 대하여도 기술한다.

4.2 실험방법

본 실험에서는 제안된 방법의 성능을 보이기 위하여 다섯 가지 종류의 실험을 수행한다.

1. 제안된 방법의 견고성(robustness)를 평가하기 위하여, 두 가지 경우에 대하여 실험하였다. 첫 번째로 학습 코퍼스와 실험코퍼스를 같은 실험집합에서 추출하여 실험을 수행한다. 본 논문에서는 이러한 실험방법을 동종간 실험(homogeneous test)이라 정의한다. 예를 들어, 학습코퍼스와 실험코퍼스 모두 KT 실험집합에서 추출하여 사용할 경우 동종간 실험이다. 두 번째로, 학습 코퍼스와 실험 코퍼스를 서로 다른 실험집합에서 추출하여 실험을 수행한다. 이를 이종간 실험(heterogeneous test)이라 정의한다. 예를 들어 학습코퍼스는 KT 실험집합에서 추출하고, 실험코퍼스는 KRIST 실험집합에서 추출할 경우 이종간 실험이다.
2. 제안한 방법과 기존 연구[7, 11, 12]와의 성능 비교를 위한 실험을 수행한다. 실험에서는 동일한 학습코퍼스와 실험코퍼스를 사용하여 성능을 비교한다.
3. 음절정보와 전이정보를 이용한 경우(식 (4))와 음절정보, 전이정보, 자음정보를 이용한 경우(식 (5))를 비교 실험한다.
4. 순수 한국어, 순수 외래어, 한국어와 외래어가 같이 사용된 경우를 나누어 유형별 성능을 비교 평가한다.
5. 제안한 방법과 학습코퍼스의 양과의 상관관계를 측정하기 위한 실험을 수행한다. 실험에서는 일정한 양의 실험코퍼스에 대하여 학습코퍼스의 양을 변화시켜 적용한다.

4.3 실험 평가

평가 방법으로 정보검색에서 가장 보편적으로 사용되는 재현율과 정확률을 사용한다[14]. 재현율은 주

어진 실험집합에 나타나는 정답에 대하여 올바르게 찾아낸 정답의 비율을 나타내며, 정확률은 시스템이 제시한 정답에 대하여 올바르게 찾아낸 정답의 비율을 나타낸다. 이 개념을 본 실험의 평가를 위하여 식 (11)과 같이 정의한다.

$$\text{재현율} = \frac{\text{올바르게 추출된 외래어의 개수}}{\text{실험코퍼스에 나타나는 외래어의 개수}}$$
$$\text{정확률} = \frac{\text{올바르게 추출된 외래어의 개수}}{\text{제안한 방법이 추출한 외래어의 개수}} \quad (11)$$

정확률과 재현율이 모두 높을수록 성능이 좋은 것이다. 하지만 재현율과 정확률은 일반적으로 서로 반비례의 관계에 있어, 한 쪽을 높이면 다른 한 쪽이 내려가는 것이 보통이다.

4.4. 동종간 실험과 이종간 실험

동종간 실험과 이종간 실험을 위하여, KT실험집합과 KRIST 실험집합에서 학습코퍼스와 실험코퍼스를 각각 추출한다. KRIST 실험집합의 경우 KT실험집합보다 많은 양의 문서를 포함하여 실험집합의 크기에서 서로 차이를 나타낸다. 이러한 차이는 학습 코퍼스와 실험코퍼스의 양에 영향을 미치게 되어 올바른 실험 결과를 이끌어 낼 수 없다. 본 실험에서는 이러한 문제점을 해결하기 위하여 KT실험집합은 실험집합에 나타나는 모든 어절을 사용하며, KRIST 실험집합의 경우 KT실험집합과 같은 양을 임의로 추출하여 사용한다. 그리고 각 실험집합의 90%의 어절은 학습코퍼스로 사용하고 10%는 실험코퍼스로 사용한다.

〈표 3〉과 〈표 4〉는 동종간 실험과 이종간 실험 결과를 나타낸다. 〈표 3〉과 〈표 4〉에서 학습코퍼스의 실험집합과 실험코퍼스의 실험집합이 KRIST 실험집합일 경우 재현율은 90.69% 이며 정확률은 91.37%이다(동종간 실험결과). 또한, 학습코퍼스의 실험집합이 KRIST 실험집합이고, 실험코퍼스의 실험집합이 KT 실험집합일 경우 재현율은 82.94% 이며 정확률은 87.42%이다(이종간 실험결과). 실험결과에서 동종 실험집합 실험에서는 KRIST 실험집합과 KT실험집합을 사용한 경우 모두에서 좋은 성능을 보임을 알 수 있다. 이종간 실험에서는 KT실험집합을 학습코퍼스로 사용한 경우에 KRIST 실험집합을 학습코퍼스로 사용한 경우보다 보다 좋은 성능을 나타냄을 알 수 있다. 이는 KT실험집합의 경우 KRIST 실험집합보다 음차표기된 외래어가 많이 포함하고 있어 음차표기된 외래어에 대한 확률이 보다 정확하게 추출될 수 있었기 때문으로 추정된다.

〈표 3〉 외래어 추출 실험결과의 재현율

		학습코퍼스의 실험집합	
		KRIST 실험집합	KT 실험집합
실험코퍼스의 실험집합	KRIST 실험집합	90.69%	89.17%
	KT 실험집합	82.94%	96.24%

〈표 4〉 외래어 추출 실험결과의 정확률

		학습코퍼스의 실험집합	
		KRIST 실험집합	KT 실험집합
실험코퍼스의 실험집합	KRIST 실험집합	91.37%	90.18%
	KT 실험집합	87.42%	95.46%

4.5. 기존 연구와의 비교 실험

기존연구[7, 11, 12]와의 비교실험을 위하여 각 실험집합으로부터 90%를 학습코퍼스로 10%를 실험코퍼스로 추출한다. 〈표 5〉와 〈표 6〉은 외래어 인식과 추출 실험의 결과를 나타내고 있다. 실험 결과에서 제안된 방법은 외래어 인식과 추출 모두에서 높은 정확률과 재현율을 나타낸다. 실험결과, 외래어 인식에서 평균 24.87%의 재현율 향상을 나타내며, 외래어 추출에서는 평균 42.8% 재현율 향상과 평균 17% 정확률 향상을 보이고 있다. 외래어 인식의 경우 본 논문의 기법은 95%이상의 재현율과 97%이상의 정확률을 나타내는데 비해, 기존 방법은 본 논문의 기법에 비해 낮은 재현율을 나타낸다. 따라서 본 논문에서 제안한 외래어 추출의 성능이 외래어 인식의 성능에 기존 연구에 비하여 큰 영향을 받지 않음을 알 수 있다. 기존방법의 경우 KT 실험집합에 대한 외래어 인식에서 비교적 높은 성능을 나타내지만 외래어 추출에서 KRIST 실험집합과 비슷한 성능을 나타낸다. 하지만 기존 방법이 외래어 추출에 있어 순수 한국어와 접두사 및 접미사 등을 제거하여 음차표기된 외래어를 추출하기 때문에 이 과정에서 많은 오류를 나타내었던 것으로 분석된다. 특히 재현율의 성능향상이 정확률의 성능향상보다 크게 나타나는데 이러한 결과

의 원인은 다음과 같이 요약될 수 있다.

1. 기존의 방법이 어절 내에 음차표기된 외래어를 구성하는 음절보다 순수 한국어를 구성하는 음절이 많을 경우에는 음차표기된 외래어를 포함하지 않는다고 판단하기 때문에 추출하지 못하는 외래어가 발생하므로 재현율이 떨어진다.
2. 제안된 방법은 이러한 경우에도 음차표기된 외래어를 올바르게 추출하므로 재현율의 향상을 보인다.

〈표 5〉 기존 연구와의 비교 실험 결과 (외래어 인식)

	실험집합	재현율	정확률
기존방법 [7, 11, 12]	KT 실험집합	89.12%	95.09%
	KRIST 실험집합	69.27%	98.08%
제안한 방법	KT 실험집합	99.11%	98.16%
	KRIST 실험집합	95.95%	97.70%

〈표 6〉 기존 연구와의 비교 실험 결과 (외래어 추출)

	실험집합	재현율	정확률
기존방법 [7, 11, 12]	KT 실험집합	64.10%	82.04%
	KRIST 실험집합	66.78%	77.85%
제안한 방법	KT 실험집합	96.24%	95.46%
	KRIST 실험집합	91.18%	92.17%

4.6. 자음정보 추가 유무에 따른 외래어 추출 성능 비교 실험

자음정보 추가 유무에 따른 외래어 추출 성능 비교 실험을 위하여 각 실험집합으로부터 90%를 학습코퍼스로 10%를 실험코퍼스로 추출한다. 자음정보를 추가하지 않은 경우는 식 (4)를 이용하여 외래어를 추출하였으며, 자음정보를 추가한 경우에는 식 (5)를 이용하여 외래어를 추출한다. 〈표 7〉은 실험결과를 나타낸다. 실험결과에서 자음정보를 추가한 경우 자음정보를 추가하지 않은 경우에 비해 재현율과 정확률에서 성능 향상을 나타낸다. 이러한 성능향상은 외래어에서 자주 출현하는 'ㅍ', 'ㅋ', 'ㅌ'과 같은 자음정보와 순수한국어에서 자주 출현하는 'ㄱ', 'ㄴ' 등과 같은 자음정보가 외래어를 추출하는데 유용하게 사용되었다는 것을 보여준다.

〈표 7〉 자음정보 추가 유무에 따른 외래어 추출 성능 비교 실험 결과

	실험집합	재현율	정확률
음절정보, 전이정보 : 식 (4)	KT 실험집합	96.24%	95.46%
	KRIST 실험집합	91.18%	92.17%
음절정보, 전이정보, 자음정보 : 식 (5)	KT 실험집합	97.26%	96.07%
	KRIST 실험집합	92.05%	92.33%

4.7. 유형별 외래어 추출 성능 비교 실험

유형별 외래어 추출 성능 비교실험을 위하여 각 실험집합으로부터 90%를 학습코퍼스로 10%를 실험코퍼스로 추출한다. 〈표 8〉과 〈표 9〉는 이들 실험집합의 학습코퍼스와 실험코퍼스에 나타난 유형의 개수를 나타낸다. 〈표 8〉과 〈표 9〉에서 순수 한국어는 "논리함수"와 같이 어절이 모두 한국어로 이루어진 경우, 순수 외래어는 "라이브러리"와 같이 어절이 모두 외래어로 구성된 경우, 외래어와 한국어의 조합은 "객체지향시스템에서는"이나 "오페라는"과 같이 어절이 한국어의 기능어나 내용어와 외래어의 조합으로 이루어진 경우를 나타낸다. 〈표 8〉과 〈표 9〉에서 순수 한국어 유형의 비율이 가장 높으며, 외래어와 한국어의 조합 유형이 가장 낮은 비율을 나타내는 것을 알 수 있다. 특히 KT 실험집합의 경우 순수 외래어의 비율이 KRIST 실험집합보다 높게 나타나는 것을 알 수 있다. 〈표 10〉은 유형별 외래어 추출 실험 결과를 나타낸다. 실험 결과에서 순수 한국어나 순수 외래어의 경우는 매우 높은 성능을 나타내는 반면, 외래어와 한국어의 조합은 순수 한국어나 순수 외래어에 비해 비교적 낮은 성능을 나타낸다. 이는 학습코퍼스에 나타난 외래어와 한국어의 조합으로 구성된 어절수가 비교적 적어 자료 부족 문제로 인한 것으로 분석된다. 이러한 오류를 줄이기 위해서 순수 외래어와 한국어의 기능어와의 조합을 통한 학습데이터의 생성과 같은 데이터의 보강이 필요하다고 하겠다.

〈표 8〉 각 실험집합의 학습코퍼스에 포함된 유형별 개수

유형	실험집합	
	KT 실험집합	KRIST 실험집합
순수 한국어	24,941개 (69.28%)	47,261개 (87.52%)
순수 외래어	9,202개 (25.56%)	3,391개 (6.28%)
외래어와 한국어의 조합	1,857개 (5.16%)	3,348개 (6.20%)
총계	36,000개	54,000개

〈표 9〉 각 실험집합의 실험코퍼스에 포함된 유형별 개수

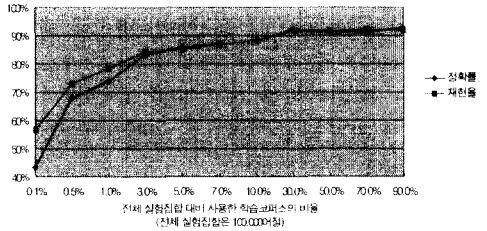
유형	실험집합	
	KT 실험집합	KRIST 실험집합
순수 한국어	2,772개 (69.3%)	5,292개 (88.20%)
순수 외래어	1,007개 (25.18%)	366개 (6.10%)
외래어와 한국어의 조합	221개 (5.52%)	342개 (5.70%)
총계	4,000개	6,000개

〈표 10〉 유형별 외래어 추출 실험 결과

유형	실험집합	재현율	정확률
순수 한국어	KT 실험집합	99.20%	99.39%
	KRIST 실험집합	99.62%	99.53%
순수 외래어	KT 실험집합	97.48%	98.15%
	KRIST 실험집합	92.29%	98.48%
외래어와 한국어의 조합	KT 실험집합	87.00%	86.62%
	KRIST 실험집합	89.00%	84.59%

4.8. 학습코퍼스의 양에 따른 성능 변화 실험

(그림 3)은 학습코퍼스의 양에 따른 제안된 모델의 정확률과 재현율의 성능변화를 나타낸다. 실험을 위하여, KRIST 실험집합과 KT 실험집합을 하나의 실험집합으로 만들었다. 만들어진 전체 실험집합의 어절 수는 약 100,000개이며, 이 중 10%인 10,000개의 어절을 실험코퍼스로 고정하여 사용하고 나머지를 학습코퍼스로 사용한다. (그림 3)에서 x축은 전체 실험집합 중에서 사용한 학습코퍼스의 비율을 나타내며, y축은 재현율과 정확률을 나타낸다. (그림 3)에서 재현율과 정확률이 x축의 30% 지점에서 (학습코퍼스의 양이 30,000 어절일 때) 수렴하기 시작하는 것을 알 수 있다. 또한 학습코퍼스의 양이 매우 작을 때 (x축의 3% 지점인 학습코퍼스의 어절 수가 3,000개 일 경우에 정확률과 재현율은 각각 약 84% 정도를 나타낸다.)에도 제안된 방법은 좋은 성능을 나타냄을 알 수 있다.



(그림 3) 학습코퍼스의 양에 따른 음차표기 외래어 추출의 성능 변화

본 장에서의 실험결과를 요약하면 다음과 같다.

1. 본 논문의 기법은 동종간 실험과 이종간 실험에서 모두 좋은 성능을 나타내었다.

2. 본 논문의 기법은 기존 연구 방법에 비하여 재현율과 정확률 모두에서 성능향상을 나타내었다.

3. 본 논문은 적은 양의 학습코퍼스를 사용한 경우에도 외래어를 효율적으로 추출하였다(학습 코퍼스가 3,000어절 실험코퍼스가 10,000어절 일경우에 약 84% 정도의 재현율과 정확률을 나타내었다.)

5. 결론

본 논문에서는 한국어 문서에서 나타나는 외래어를 인식 및 추출하는 방법에 대하여 기술하였다. 제안한 방법은 외래어 인식 및 추출 문제를 주어진 어절이 순수 한국어를 구성하는 음절인지 외래어를 구성하는 음절인지를 태깅하는 음절태깅이라는 문제로 변환하여 해결하였다. 은닉 마르코프 모델에 기반한 본 논문의 기법은 실험집합이 같은 경우뿐만 아니라 실험집합이 다른 경우에도 좋은 성능을 나타내었다. 또한 기존 연구에 비하여 외래어 인식과 외래어 추출에서 모두 성능향상을 보여주었다.

자음정보의 유용성을 살펴보기 위하여 본 논문에서

는 음절정보와 전이정보만을 사용하여 외래어를 추출한 경우와 음절정보, 자음정보, 전이정보를 사용하여 외래어를 추출한 경우에 대하여 비교 실험하였으며, 실험결과 자음정보가 외래어 추출에 유용하다는 것을 보였다. 유형별 외래어 추출 실험에서는 모든 유형에 대하여 비교적 좋은 성능을 나타내었다. 학습코퍼스의 양에 따른 외래어 추출 실험에서는 적은 양의 학습코퍼스로도 좋은 성능을 나타냄을 보였다.

향후 외래어와 한국어의 조합으로 구성된 어절에서의 외래어 추출 성능을 향상시키기 위하여 순수외래어로 구성된 어절과 한국어 기능어의 결합을 통한 데이터의 보강에 대한 연구가 필요할 것이며, 음절, 전이, 자음정보를 효율적으로 통합하는 방법에 대한 연구도 진행되어야 할 것이다. 또한 은닉 마르코프 모델의 통계적 특성에 의해 본 논문의 기법은 다른 전문분야에서 나타나는 음차표기된 외래어의 추출에도 사용될 수 있을 것이다.

참고문헌

[1] 강병주, 이재성, 최기선(1999), 외국어 음차표기의 음성적 유사도 비교 알고리즘, 정보과학회 논문지 (B) 제26권 제10호, 1237-1246.

[2] 강승식(1995), 한국어 자동 색인을 위한 형태소 분석 기능, 한국정보과학회 춘계학술발표논문집 22권 1호, 930-932.

[3] 박영찬, 최기선, 김재군, 김영환(1996), 한국어 정보검색을 위한 시험용 데이터 모음 2.0 개발, 1996년도 한국정보과학회 인공지능 연구회 춘계학술 대회, 59-65.

[4] 오종훈, 최기선 (1999), "은닉 마르코프 모델을 이용한 과학기술문서에서의 외래어 추출 모델", 제11회 한글 및 한국어 정보처리 학술대회, pp. 137-141

[5] 이재성(1998), 다국어 정보검색을 위한 영-한 음차표기 및 복원 모델, 박사학위 학위논문, 한국과학기술원 전산학과.

[6] 이준호, 최광남, 한현숙, 김종원, 남성원 (1995), 정보 검색 연구를 위한 KRIST 테스트 컬렉션의 개발, 정보관리학회지, 12권 2호, pp. 225-232

[7] 정길순, 권윤형, 맹성헌(1997), 외래어와 영어처리를 통한 검색효과 향상, 한국정보과학회 추계학술발표논문집 24권 2호, 189-192.

[8] Allen James(1995), Natural Language Understanding, The Benjamin/Cummings Publishing Company.

[9] Huang, X. D., Y. Ariki, and M. A. Jack(1990), Hidden Markov Models for Speech Recognition , Edinburgh University press.

[10] Kang, B. J., and K. S. Choi(2000), Two Approaches for the Resolution of Word Mismatch Problem Caused by English Word and Various Korean Trasliterations in Korean Information Retrieval, In the Proceeding of the International Workshop on Information Retrieval with Asian Languages (IRAL '2000), 133-140.

[11] Kwon, Y. H., Jeong, K. S. & Myaeng, S. H.(1997), Foreign Word Identification Using a Statistical Method for Information Retrieval, Proc. of the 17th International Conference on Computer Processing of Oriental Languages, Hong Kong.

[12] Myaeng, S. H., Kwon, Y. H. & Jeong, K. S.(1997), The Effect of a Proper Handling of Foreign and English Words in Retrieving Korean Text, Proc. of International Workshop on Information Retrieval with Asian Languages (IRAL '97), Tsukuba, Japan.

[13] Rabiner, L. (1989), Tutorial on hidden Markov models and selected applications in speech recognition, In Proceedings of the IEEE, Vol 77.

[14] Salton, G. and McGill, M. J.(1983), Introduction to Modern Information Retrieval, McGraw-Hill.