은닉마르코프 모델(HMM)을 이용한 과학기술문서에서의 외래어 추출 모델

오종훈*○ 최기선*
* 한국과학기술원 전산학과
*전문용어 언어 공학 센터

Automatic Extraction of Technical Terminologies from Scientific Text based on Hidden Markov Model

Jong Hoon Oh* Key Sun Choi*
* Dept. of Computer Science, KAIST
* KORTERM

요 약

기술의 발달로 인해 수많은 용어들이 생성되고 있다. 이들은 대부분 전문용어이며 이는 비영어권 국가인 우리나라에 도입될 때, 외래어나 원어형태로 도입된다. 그런데 외래어나 원어형태의 전문용어는 형태소 분석기, 색인기 등의 시스템에서 오류의 원인이 되어, 이를 전처리기로 사용하는 자연언어처리 시스템의 성능을 저하시킨다. 따라서 본 논문에서는 외래어나 원어로 된 전문용어를 처리하기 위한 전단계로서 문서에서 자동적으로 외래어를 인식하고 추출하는 방법을 제시한다. 본 논문에서 제시하는 방법은 외래어 추출 문제를 태깅문제로 변환하여, 태깅 문제를 해결하는 기법 중의 하나인 은닉마르코프 모델 (Hidden Markov Model)을 이용하여 외래어 추출을 하였다. 그 결과 94.90%의 재현률과 95.41%의 정확도를 나타내었다.

1 서론

기술발달로 인해 새로운 개념을 나타내는 용어들이 생성되고 있다. 이런 용어들은 전문분야에서 쓰이는 용어가 대부분이며, 이러한 전문분야의 용어는 그 기술이 만들어진 나라의 언어로 쓰여진다. 과학기술분야의 새로운 전문 용어들은 대부분 영어로 만들어지며, 비영어권 국가에서 이러한 용어들을 도입할 때에는 그 나라의 언어로 '음역(*transliteration*)하여 만든 외래어' (이하 외래어)로 도입하거나 원어 (original language) 그대로 도입한다. [6] [10]

새로 생성된 '과학기술 분야의 전문용어' (이하 전문용어)들은 비영어권인 우리나라에서 한글문서에 나타날 때, 두 가지 문제점을 보인다. 첫째, 대부분의 전문용어는 외래어나 원어 그대로 나타나거나, 약어로도 나타난다. 예를 들어 '데이터 베이스'의 경우 한글 문서에 '데이터 베이스', 'Database', 'DB'의 형태로 나타난다. 둘째, 외래어로 된 전문용어는 '외래어 표기법'이 있음에도 불구하고, 사용하는 사람마다 다른 형태로 사용된다. 예를 들어 'data'의 경우 '데이터', '데이타', '데타' 등으로, 'digital'의 경우 '디지틀', '디지털', '디지탈' 등으로 쓰여지고 있다. 이는 우리의 음운 체계와 외국 언어의 음운 체계가 틀린 데에서 비롯되는데, 이러한 현상은 외래문물이 급속히 우리나라에 유입되면서 심화되고 있다. [11] [15] 위에 기술한 두가지 문제점으로 인해 외래어로 된 전문용어는 형태소 분석기나 색인기의 중요한 오류의 원인이 되어 이를 전처리기로 사용하는 자연언어처리 시스템의 성능을 저하시킨다. 따라서 외래어로 된 전문용어를 포함한 문서에 대한 효율적인 자연언어처리 시스템을 구성하기 위해서는 외래어로 된 전문용어에 대한 추출과 그에 대한 처리가 있어야 한다. 이를 위해서는 문서로부터 자동적으로 외래어를 추출하는 작업과 추출된 외래어가 전문용어인가를 결정하는 작업이 필요하다.

본 논문에서는 외래어로 된 전문용어를 효과적으로 처리하기 위해 필요한 외래어 추출과 전문용어 인식의 두 단계 중 첫 번째 단계인 문서에서 자동적으로 외래어 인식 및 추출하는 효과적인 방법을 제시하고자 한다.

기존의 방법 [2]은 어절에 나타나는 외래어의 개수에 의존하여 외래어를 추출하였기 때문에 어절에 외래어의 개수가 적을 경우, 외래어가 포함되어 있음에도 불구하고, 인식하지 못하는 경우가 발생한다. 본 논문에서는 이런 문제점을 해결하기 위해서, 어절에 나타나는 외래어의 개수와는 상관없이 외래어를 추출하는 방법을 제시한다. 또한 이를 위하여, 외래어추출 문제를 태깅문제로 변환하였다. 외래어 인식문제를 태깅문제로 변환하였을 경우 태그된 어절에 외래어를 나타내는 태그가 존재하면, 그 어절은 외래어가 있다고 인식하고, 외래어를 나타내는 태그가 존재하지 않으면 그 어절은 외래어를 포함하고 있지 않다고 인식한다. 외래어 추출문제를 태깅문제로 변환하였을 경우에는 외래어가 있다고 인식된 어절에 대하여 외래어를 나타내는 태그들에 대하여 해당 어휘를 추출함으로써 외래어를 추출할 수 있다. 본 논문에서는 외래어추출 문제를 태깅문제로 변환하고,

태깅문제를 해결하는 기법 중의 하나인 은닉마르코프 모델 (Hidden Markov Model)을 이용하여 외래어 추출문제를 해결하고자 한다.

본 논문의 구성은 2절에서 기존 연구와 외래어 인식,추출 문제의 중요성에 대하여 기술하고, 3절에서는 은닉 마르코프 모델을 이용한 외래어 인식, 추출 모델에 대하여, 4절에서는 실험 및 결과에 대하여, 그리고 5절에서는 결론 및 향후연구에 대하여 각각 기술한다.

2 기존 연구 및 외래어 인식 추출 문제의 중요성

2.1 외래어 인식 및 추출 문제의 중요성

외국언어는 한국어에서 외국어 형태 그대로 사용되거나 적절한 번역 절차나 표기과정을 거쳐 한국어로 사용된다. 전자를 외국어라 하고, 후자 중 번역이 아닌 원어의 발음이나 글자로부터 음차표기 (transliteration) 된 용어를 외래어라 한다.[11] [15] 외래어는 다음의 세 가지 특징을 가지고 있다. 첫째, 외래어가 한국어로 표기되기 때문에 순수 한국어와의 구별이 어렵다. 둘째, 외래어는 외국 언어의 음운체계를 가지면서, 한국어로 표기된다. 예를 들어 영어에서 자주 사용되는 'ph', 'p', 'f' 등은 한국어에서 'ㅍ'으로, 't' 등은 한국어에서 'ㅌ'으로 'c' 등은 한국어에서 'ㅋ'으로 음차표기되어 쓰이는데, 'ㅍ', 'ㅌ', 'ㅋ' 등의 자음은 순수 한글에서는 자주 쓰이지 않는 자음이다. 셋째,1절에서 기술하였듯이, 외래어는 사용하는 사람마다 다르게 표기되며, 원어, 약어 등이 혼용되어 쓰인다. 위에서 기술한 세번째 특성은 많은 경우 외래어가 사전에 등재되지 않는 중요한 요인이 된다. 따라서, 사전에 등재되지 않은 외래어를 포함한 문서에 대하여 이러한 사전을 이용하는 형태소 분석기나 색인기는 미등록어 문제를 야기시킨다. 이로 인해 형태소 분석기는 사전에 등재 되지 않은 단어 에 대하여 어절의 끝부터 조사나 어미를 제거하고 나머지 부분은 단순히 미등록어로 처리하는 방식을 취하는데,[8] 이 때 많은 오류가 발생한다. 예를 들어 '오페라는'과 '미테랑'은 다음과 같이 오분석 될 수 있다.

'오페라는'(Opera+Josa) $\Longrightarrow$ '오페'+'라는'
'미테랑'(Mitterant) $\Longrightarrow$ '미테'+'랑'

이러한 오분석은 형태소 분석의 미등록어 처리시 외래어를 인식 및 추출할 수 있다면 해결될 수 있다. 즉, 형태소 분석의 미등록어 처리에서 미등록어중 외래어를 하나의 형태소 단위로 분리할 수 있다면, 외래어와 결합되는 다른 형태소도 올바르게 분석할 수 있다. 따라서 형태소 분석의 미등록어 처리 문제에 있어서 문서로부터 자동적으로 외래어를 인식 및 추출하는 작업은 중요하다.

외래어 인식 및 추출은 한국어 형태소 분석 뿐만 아니라 다국어 정보 검색에서도 중요한 작업이다. 다국어 정보 검색이란 '하나의 언어로 된 문서에 대해 다국어로 질의할 수 있는 것', '다국어 문서에 대해 다국어 질의어로 검색할 수 있는 시스템', '한 문서가 다국어로 작성되어 이를 검색하는 시스템' 등으로 정의 된다. [12] 한국어 전문 분야 문서의 경우, 한국어와 영어를 섞어 쓰는 경우도 있어서 실제 다국어 문서로도 볼 수 있다. 따라서 '한 문서가 다국어로 작성되어 이를 검색하는 시스템'의 정의에 의하여 한국어 전문 분야 문서에 대한 정보 검색은 다국어 정보 검색으로 볼 수 있다. 이러한 문서에는 외래어가 많이 쓰이기 때문에, 외래어를 인식 추출하여 그에 대한 원어를 찾을 수 있다면 한국어 질의나 영어 질의에 대하여 관련된 문서를 올바르게 찾을 수 있을 것이다.

따라서 외래어 인식 및 추출 문제는 전문용어를 인식하는 문제 뿐만 아니라 형태소 분석이나 다국어 정보 검색 등 자연 언어 처리에 있어서 중요한 작업이다.

2.2 기존 연구

기존의 외래어에 대한 연구는 원어로부터 외래어를 생성하는 문제나 외래어에 대한 원어를 찾는 문제가 주된 연구 대상이었다. [7] [11] [15] 최근 들어 과학기술문서에서 외래어를 자동적으로 추출하려는 연구가 있었다.[2] [14]

논문 [2]에서는 외래어를 추출하기 위하여 인식과 추출의 두 단계를 거친다. 논문 [2]의 외래어 인식단계에서는 문서상의 어절 중에서 외래어가 포함되어 있는 어절만을 인식하기 위하여 식 (1)을 사용하였다. 식 (1)은 주어진 단어 (또는 어절) W가 외래어를 포함하고 있는지를 결정한다.

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

$$P(W|Foreign) \approx \lambda_1 \times P(w_1|Foreign) \times \cdots \times P(w_n|Foreign) + \lambda_2 \times P(\phi w_1|Foreign) \times \cdots \times P(w_n\phi|Foreign) \quad (2)$$

식 (2)에서 w_i는 어절 W에서 i번째 어휘정보를 나타낸다. 식 (2)에서 $P(W|Foreign)$는 학습코퍼스로부터 바이그램과 유니그램 정보를 추출하여 $\lambda_1 + \lambda_2 = 1$ 인 λ_1, λ_2로 유니그램과 바이그램의 가중치를 조절하여 확률값으로 구한다. 식 (1)에서 D(W)가 1보다 크면 단어 W는 외래어가 포함된 것으로 인식하고 1보다 작으면 그 어절이 외래어가 포함되지 않은, 순수 한글로만 구성된 것으로 인식한다.

외래어 추출은 외래어 인식 단계에서 외래어가 포함되어 있다고 인식된 어절에 대하여 어절의 분리점을 기준으로 앞부분과 뒷부분에 대한 D(W)를 계산함으로써 외래어와 기능어를 나누고, 한국어복합명사 분석방법[13]를 이용하여, 순수한글과 조사, 어미를 제외한 외래어를 추출한다.

논문 [2]의 방법은 인식단계의 결과가 추출단계의 입력으로 사용되기 때문에 인식단계에서의 성능은 전체 시스템의 성능에 중요한 요인으로 작용하게 된다. 그런데 논문 [2]에서는 식 (1)을 이용한 인식, 즉 어절에 나타난 외래어의 개수에 의존한 인식으로 인하여, 외래어의 개수가 적은 어절에 대하여는 외래어를 인식하지 못하는 경우가 발생한다. 예를 들어 '객체지향시스템에서'의 경우 '시스템'이라는 외래어가 3개 '객체지향'과 '에서'의 순수한글이 6개가 되어 외래어가 포함되지 않은 어절로 인식하게 된다. 또한 인식의 결과가 추출의 입력으로 사용되기 때문에 인식단계에서 인식되지 않은 어절에 대하여는 외래어를 추출하지 않기 때문에 전체 시스템의 성능의 저하를 가져온다. 이에 본 논문에서는 은닉마르코프 모델을 이용하여 어절에 나타난 외래어의 개수와 상관없이 효과적으로 문서에서 외래어를 인식, 추출하는 시스템을 제안하고자 한다.

3 은닉 마르코프 모델을 이용한 외래어 인식,추출 모델

3.1 문제 정의

외래어 인식,추출이라는 문제는 태깅 문제로 변환할 수 있다. 즉, 주어진 어절에 대하여 각 어절의 음절이 외래어로 쓰이는가 혹은 순수한글로 쓰이는가를 태깅하는 문제로 볼 수 있다. 외래어 인식의 경우 주어진 어절에 대하여 어절에 외래어가 포함되어 있는가를 검사하는 작업이다. 태깅문제로 변환하였을 경우, 주어진 어절에 대한 태깅 결과에서 외래어를 나타내는 태그가 있다면 해당어절이 외래어를 포함한다고 판단할 수 있고, 주어진 어절에 대한 대한 태깅 결과에서 외래어를 나타내는 태그가 없다면 그 어절에는 외래어가 포함되어 있지 않다고 판단할 수 있다. 예를 들어 하나의 음절이 외래어일 경우 'F'라는 태그를 할당하고, 하나의 음절이 한국어일 경우 'K'라는 태그를 할당한다고 할 때, 어절 '크로마토그라피'와 어절 '객체지향'은 다음과 같이 태깅될 수 있다.

크로마토라피 FFFFFFF $\Longrightarrow$ 외래어 포함
객체지향 KKKK $\Longrightarrow$ 외래어 포함하지 않음

외래어 추출의 경우 주어진 어절에 외래어가 있다고 인식되면, 해당 어절에서 순수한글과 조사 및 어미를 제외한 외래어만을 추출하는 작업이다. 이를 태깅 문제를 변환하였을 경우, 외래어가 있다고 인식된 태그된 어절에 대하여 외래어로 태그된 단어만을 뽑아내는 작업으로 변환이 가능하다. 예를 들어 '벨기에도'와 '오페라는'의 경우, 다음과 같은 처리가 가능하다.

오페라는('Opera'+는) FFFK $\Longrightarrow$ 오페라 FFF
벨기에는('Belgium'+는) FFFK $\Longrightarrow$ 벨기에 FFF

따라서 본 논문에서는 외래어를 인식, 추출하는 문제를 태깅문제로 변환하여, 태깅문제를 해결하기 위해 사용되는 계산 모델 중의 하나인 은닉마르코프모델을 이용하여 문제를 해결하고자 한다.[3][9]

3.2 은닉마르코프 모델

3.1절에서 기술하였듯이 외래어 인식, 추출문제를 태깅문제로 변환하였다. 본 절에서는 태깅문제를 해결하기 위해 사용되는 계산 모델 중 하나인 은닉마르코프 모델에 대한 기본적인 개념에 대하여 기술하겠다.

태깅 문제는 주어진 어절 W에 대하여, 식 (3)을 만족하는 품사열 T를 할당해주는 문제이다.

$$\phi(W) = argmax_T P(W|T)P(T) \tag{3}$$

식 (3)을 체인 룰 (chain rule)을 이용하여 전개하면, 식 (4)와 같은 조건부 확률의 연속적인 곱으로 나타낼 수 있다.

$$P(W|T)P(T) = \prod_{i=1}^{n} p(w_i|t_n,\cdots,t_1,w_{i-1},\cdots,w_1) \times p(t_i|t_{i-1},\cdots,t_1) \tag{4}$$

식 (4)에서 w_i는 어절 W에서 i번째 어휘정보를 나타내고, t_i는 어절 W에서 i번째 태그정보를 나타낸다.

식 (4)를 2차 마르코프 독립 가정(Markov Independence Assumption)을 사용하여 수식을 단순화 시키면, 식 (4)는 식 (5)로 단순화 된다.[3] [9]

$$P(T|W)P(W) = p(t_1) \times p(t_2|t_1) \times \prod_{i=1}^{n} p(w_i|t_i) \times \prod_{i=3}^{n} p(t_i|t_{i-1},t_{t-2}) \tag{5}$$

3.3 은닉마르코프 모델을 이용한 외래어 인식 및 추출 모델

본 논문에서 제안하는 외래어 추출 모델은 3.2절에서 기술한 은닉 마르코프 모델에 기반한다. 은닉 마르코프 모델에 기반한 외래어 추출 모델은 다음과 같은 3가지 정보를 이용한다.

- 어휘 정보
  - 바이그램 어휘정보, 유니그램 어휘정보
- 전이 정보
  - 트라이그램 전이정보, 바이그램 전이정보
- 초성 중성 어휘정보
  - 바이그램 초성 중성 어휘정보, 유니그램 초성 중성 어휘정보.

예를 들어, '객체지향시스템에서 KKKKFFFKK'의 경우에 전이정보는 바이그램의 경우 '(K K) (K K) $\cdots$ (F K) (K K)'로 표현되고, 트라이그램의 경우 '(K K K) (K K K) $\cdots$ (F F K) (F K K)'로 표현된다. 어휘 정보는 유니그램의 경우 '(객 K) (체 K) $\cdots$ (에 K) (서 K)' 로 표현되고, 바이그램의 경우 '(객체 KK) (체지 KK) $\cdots$ (템에 FK) (에서 KK)'로 표현된다. 또한 초성 중성 정보가 '(ㄱㄱ) (ㅊϕ) (ㅈϕ) (ㅎㅇ) (ㅅϕ) (ㅅϕ) (ㅌㅁ) (ㅇϕ) (ㅅϕ) KKKKFFFKK'으로 나타내어 질 수 있는데, 이에 대한 초성 중성 바이그램의 경우 '(ㄱㄱ) (ㅊϕ) KK' ..... '(ㅇϕ) (ㅅϕ) KK' 의 형식이 되며, 유니그램 초성 중성 어휘 정보의 경우, '(ㄱㄱ) K' ..... '(ㅅϕ) K'의 형식이 된다.

이러한 정보를 이용한 본 모델의 수식은 식 (6)으로 나타내어 진다.

$$\begin{aligned} P(T|W)P(W) = & \\ & p(t_1) \times p(t_2|t_1) \times \prod_{i=3}^{n} p(t_i|t_{i-1},t_{i-2}) \\ & \times \prod_{i=1}^{n} p(w_i|t_i) \times \prod_{i=2}^{n} p(w_i|t_{i-1},t_{i-2}) \\ & \times \prod_{i=1}^{n} p(ph_i|t_i) \times \prod_{i=2}^{n} p(ph_i|t_{i-1},t_{i-2}) \end{aligned} \tag{6}$$

식 (6)에서 t_i는 어절 W에서 i번째 태그 정보를 나타내고 ,w_i는 어절 W에서 i번째 어휘정보를, ph_i는 어절 W에서 i번째 초성 종성 어휘정보를 나타낸다.

여기에서 초성 종성 어휘 정보가 필요한 이유는 초성 종성 정보를 통하여, 어휘 정보만을 이용하였을 경우 발생할 수 있는 데이터 부족 문제 (data sparseness problem) 를 해결하기 위함이다. 또한, 2.1에서 기술한 바와 같이 외래어의 경우 한국어에서 잘 쓰이지 않는 'ㅋ', 'ㅌ', 'ㅍ'의 자음이 많이 쓰인다는 특성이 있다. 따라서 모음은 외래어와 한국어를 구별하는데 있어 커다란 영향을 주지 않기 때문에, 자음 정보만을 사용하여, 외래어 인식 및 추출의 효율을 증대시키기 위함이다.

식 (6)에 의하여 태깅된 어절은 태깅된 외래어 품사에 의하여, 외래어를 인식하고 추출하게 된다. 예를 들어 '객체지향시스템에서'의 경우 'KKKKFFFKK'로 태깅될 수 있으며, '시스템'이 'FFF'로 태깅되기 때문에 '시스템'을 외래어로 인식, 추출할 수 있다.

본 논문에서 제안하는 외래어 추출방법은 각 음절에 대한 외래어 태깅으로 볼 수 있다. 하지만 이는 하나의 음절이 외래어로 자주 쓰인다고 해서 그 음절이 외래어가 되기 보다는, 바이그램 어휘정보와 바이그램 초성 종성 정보 그리고 바이그램,트라이그램 전이정보를 이용하여 하나의 음절을 태깅할 때 앞서 나온 연관된 음절정보를 이용하기 때문에, 음절보다는 어휘에 의존한 외래어 태깅이라고 볼수 있다. 예를 들어 '트리'라는 어절을 태깅할 때 '리'가 외래어 음절로 자주 쓰이기 때문에 외래어로 태깅되기 보다는 '트리'라는 어휘가 외래어로 자주 쓰이기 때문에 '리'라는 음절이 외래어로 태깅된다. 실제 '리'의 경우 실험데이터에서 '리'가 한국어로 쓰인 회수가 1043번 외래어로 쓰인 회수는 572번으로 단순히 '리'라는 음절만으로는 한국어가 될 확률이 높다.

따라서 본 논문에서 제안하는 외래어 추출 모델은 음절에 기반한 외래어 추출 모델이라기보다 어휘에 기반한 외래어 추출모델이라 할 수 있다.

4 실험

4.1 실험 데이터

실험 데이터는 KT SET 2.0 [5]의 키워드와 제목부분, KRIST SET [4]의 키워드와 제목부분, 그리고 컴퓨터분야 사전 [1]의 표제어를 어절별로 나누어 이용하였다. 예를 들어 '객체 데이터베이스에서는'는 '객체'와 '데이터베이스에서는' 으로 나누어 2개의 어휘로 실험에 사용하였다.

실험에 사용한 어휘수는 표 1과 같다

표 1: 각 실험 데이터의 어휘수

	어휘수	단일 어휘수
KRIST SET	60054	22805
KT SET	41257	11080
컴퓨터분야사전	22010	6590

표 1에서 어휘수는 중복된 어절을 포함한 어절의 어휘수를 나타내고, 단일 어휘수는 중복된 어절을 제외한 단일하게 나타나는 어절의 어휘수를 나타낸다.

표 2는 실험에 사용한 데이터에서 외래어가 차지하는 비율을 나타낸다.

표 2: 각 실험 데이터의 외래어 포함 비율

	한국어	외래어	외래어 비율
KRIST SET	52598	7456	12.42%
KT SET	29762	11495	27.86%
컴퓨터분야사전	14073	7937	36.06%

표 2는 각 실험 데이터 집합이 포함하는 외래어의 비율을 나타낸다. 그런데 표 2에서 특이한 점은 컴퓨터 분야인 KT SET과 컴퓨터분야사전이 여러 분야의 과학기술문서를 포함한 KRIST SET보다 외래어 포함 비율이 높다는 것이다. 이는 컴퓨터 분야처럼 과학기술의 발전에 민감한 분야일수록 외래어의 포함 비율이 높은 것으로 볼 수 있다. 즉, KRIST SET의 경우에는 여러 분야의 과학기술 문서를 포함하였으며, 컴퓨터 분야가 비교적 다른 과학기술 분야들보다 과학기술 발전에 민감하기 때문에, 컴퓨터 분야의 문서 집합인 KT SET이나 컴퓨터분야사전 보다는 외래어가 전체 어휘에서 차지하는 비중이 비교적 낮다고 볼 수 있다.

본 논문에서는 표 1, 표 2의 특성을 가지는 실험데이터를 이용하여, 외래어 인식 및 추출 실험을 하였다.

4.2 실험 결과

실험은 세 종류의 각각의 실험데이터에 대한 실험, 그리고 세 종류의 실험데이터를 통합한 전체 실험데이터에 대한 실험로 나누어 수행하였다. 각 실험에서 90%는 학습데이터로 사용하고 나머지 10%는 시험데이터로 사용하였다.

평가기준은 재현율과 정확도로 나타내었으며 각각은 식 (7)의 방법으로 구한다.

$$\begin{aligned} \text{재현율} &= \frac{f_{correct}}{f_{extract}} \\ \text{정확도} &= \frac{f_{extract}}{f_{all}} \end{aligned} \tag{7}$$

식 (7)에서 $f_{correct}$는 올바르게 추출한 외래어의 개수를 의미하고, $f_{extract}$는 추출한 외래어의 개수를 의미한다. 또한 f_{all}은 실험데이터에 나타난 외래어의 총 개수를 의미한다.

표 3: 실험결과

	재현율	정확도
KRIST SET	92.05%	92.33%
KT SET	97.26%	95.72%
컴퓨터분야사전	97.03%	96.07%
통합	92.05%	92.33%

표 3는 각 실험 데이터에 대한 실험 결과를 나타낸다. 표 3의 결과는 전체적으로 높은 정확도와 재현율을 나타낸다. 또한 컴퓨터 분야

만의 문서를 포함하는 KT SET과 컴퓨터 분야의 표제어만을 포함하는 컴퓨터 분야 사전이 KRIST SET보다 정확도와 재현율면에서 좀 더 좋은 성능을 나타낸다. 이는 KRIST SET의 특성상 과학기술의 여러분야에 대한 실험집합이며, 표 2에서 나타난 바와 같이, KRIST SET의 외래어 비율이 낮음으로 인하여, 정형적인 외래어 패턴이 다른 실험집합보다 추출하기 어려웠기 때문으로 분석될 수 있다.

표 4은 논문 [2]에서 제시한 기법과 본 논문에서 제시한 은닉마르코프모델을 이용한 방법을 비교한 것이다.

표 4: 기존연구와의 비교

	실험 집합	재현율	정확도
논문[2]의 모델	KRIST SET	64.1%	82.04%
	KT SET	66.78%	77.85%
본 모델	KRIST SET	92.05%	92.33%
	KT SET	97.26%	96.07%

2절에서 기술하였듯이 논문 [2]의 기법은 외래어 인식과 추출의 두 단계로 이루어 지므로 표 4에서 통계적 기법에 대한 정확도는 '외래어 인식의 정확도 × 외래어 추출의 정확도'로 나타내었고, 재현율은 '외래어 인식의 재현율'로 나타내었다.

표 4의 결과로 볼 때 본 논문에서 제시하는 외래어 추출기법은 기존의 방법 [2]에 비해, KRIST SET의 경우 재현율이 27.95%, 정확도가 10.29%가 높으며, KT SET의 경우 재현율이 28.94% 정확도가 18.22% 높다. 특히 표 4에서 본 모델은 기존[2] 모델보다 재현율이 정확도보다 훨씬 좋은 성능을 나타내는데, 이는 본 모델이 어절에 나타난 외래어의 개수에 상관없이 외래어를 정확히 인식, 추출하기 때문에 외래어의 개수에 의존적인 기존의 모델에서 인식하지 못한 외래어를 인식하기 때문으로 분석될 수 있다.

따라서 본 논문에서 제시한 은닉마르코프모델을 이용한 외래어 추출은 일반적인 태깅문제로 변환한 간단한 방법으로 재현율과 정확도의 모든 면에서 좋은 성능을 내는 효율적인 외래어 추출방법이라고 하겠다.

5 결론 및 향후 연구

본 논문에서는 외래어 추출문제를 태깅 문제로 변환하여, 은닉 마르코프 모델을 이용한 과학 기술 문서에서의 외래어 추출기법에 대하여 기술하였다. 본 논문의 기법은 기존의 방법 [2]과는 달리 외래어 인식 및 추출에 있어 어절에 나타난 외래어의 개수에 독립적인 방법을 사용하여, 좋은 성능을 나타내었다. 하지만 본 논문이 제시한 방법은 KRIST 실험집합과 같은 여러 분야의 문서가 포함되어 있는 실험집합의 경우보다 KT SET과 컴퓨터 분야 사전처럼 하나의 분야의 문서로만 구성된 실험집합이 외래어 추출에 있어 더 좋은 성능을 나타내었다. 따라서 향후 분야에 독립적으로 외래어를 추출하는 연구가 지속적으로 필요하다.

본 논문의 기법은 전문용어를 문서에서 자동적으로 추출하기 위한 방법의 전단계이므로, 향후 이를 이용한 전문용어의 추출의 연구가 진행될 수 있을 것이다.

또한 형태소 분석의 미등록어 처리에 있어서 외래어에 대한 미등록어문제는 본 논문이 다루는 문제와 유사하기 때문에 본 논문의 기법이 유용하게 이용될 수 있을 것이다.

참고문헌

[1] 한국사전 연구사. 컴퓨터 정보 용어대사전. 한국사전연구사, 1995.

[2] 정길순, 권윤형, 맹성현. 외래어와 영어처리를 통한 검색 효과 향상. In *추계학술발표논문집 24권 2호*, pages 189–192, 1997.

[3] 강인호, 김재훈, 김길창. 최대 엔트로피 모델을 이용한 한국어 품사 태깅. In 한글 및 한국어정보처리 학술발표논문집, pages 9–14, 1998.

[4] 이준호 외. 정보 검색 연구를 위한 krist 테스트 컬렉션의 개발. 정보관리학회지, 12(2), 1991.

[5] 김성혁 외. 자동 색인기 성능 시험을 위한 testset의 개발. 정보관리학회지, 11(1):929–932, 1994.

[6] 강현화, 서상규. 경제학 전문용어 사전의 국어학적 분석. In 전문용어언어공학 심포지움, pages 55–66, 1998.

[7] 정길순, 맹성현. 외래어의 자동음역을 통한 영어 단어 생성. In *춘계학술발표논문집 25권 1호*, pages 429–431, 1998.

[8] 강승식. 한국어 자동 색인을 위한 형태소 분석 기능. In 춘계학술발표논문집 *22권 1호*, pages 930–932, 1995.

[9] 김재훈. 오류 보정 기법을 이용한 어휘 모호성 해소. PhD thesis, 한국과학기술원 전산학과, 1996.

[10] 송영빈. 전문용어의 위상과 사전 구축. In 전문용어언어공학 심포지움, pages 87–97, 1998.

[11] 이재성. 다국어 정보검색을 위한 영-한 음차 표기 및 복원 모델. PhD thesis, 한국과학기술원 전산학과, 1998.

[12] D.A Hull, G. Grefenstette. Quering across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of ACM SIGIR Conference on Information Retrieval*, 1996.

[13] S.H. Myaeng et al. On language dependency in indexing. In *Proceedings of the first International Workshop on Information Retrieval with Oriental Language*, 1996.

[14] Y.H. Kwon et al. Foreign word identification using statistical method for information retreival. In *ICCPOL*, pages 675–680, 1997.

[15] J.S. Lee, K.S. Jeong, S.H. Myaeng, K.S. Choi. Automatic identification and back-transliteration of foreign words for information retrieval. *Information Processing and Management*, 1999.