韓國 컴퓨터情報學會 論文誌
第13卷 第3號, 2008. 5.

2008-13-3-1-3

웹문서에서의 출현빈도를 이용한 한국어 미등록어 사전 자동 구축

박 소 영*

Automatic Construction of Korean Unknown Word Dictionary using Occurrence Frequency in Web Documents

So-Young Park*

요 약

본 논문에서는 한국어 형태소 분석의 성능향상을 위해서, 어절에서 미등록어를 인식하여 자동으로 사전을 구축하는 방법을 제안한다. 제안하는 사전 구축 방법은 전문 분석 기반 사전 구축 방법과 웹 출현빈도 기반 사전 구축 방법으로 구성되어 있다. 전문 분석 기반 사전 구축 방법은 전체 문서에서 반복적으로 나타나는 문자열을 미등록어로 인식하고, 웹 출현빈도 기반 사전 구축 방법은 반복되지 않은 문자열을 웹 문서에서 검색하여 그 출현빈도를 바탕으로 미등록어를 인식한다. 실험결과 전문 분석만을 바탕으로 하는 기존 접근방법에 비해서 웹 문서에서의 출현빈도도 함께 고려하여 제안하는 사전 구축 방법은 32.39% 정도 재현율이 높게 나타났다.

Abstract

In this paper, we propose a method of automatically constructing a dictionary by extracting unknown words from given eojeols in order to improve the performance of a Korean morphological analyzer. The proposed method is composed of a dictionary construction phase based on full text analysis and a dictionary construction phase based on web document frequency. The first phase recognizes unknown words from strings repeatedly occurred in a given full text while the second phase recognizes unknown words based on frequency of retrieving each string, once occurred in the text, from web documents. Experimental results show that the proposed method improves 32.39% recall by utilizing web document frequency compared with a previous method.

▸ Keyword : 미등록어 인식(Unknown Word Recognition), 사전 구축(Dictionary Construction), 웹문서 기반 접근방법(Web Document-based Approach)

• 제1저자 : 박소영
• 접수일 : 2008. 3. 10, 심사일 : 2008. 3. 27, 심사완료일 : 2008. 5. 24.
* 상명대학교 디지털미디어학부 전임강사
※본 연구는 2008학년도 상명대학교 소프트웨어미디어연구소 연구비 지원으로 이루어졌음

Ⅰ. 서 론

한국어 처리 시스템의 가장 기본이 되는 형태소 분석기는 어절에서 의미를 갖는 최소 단위인 형태소를 추출하는 시스템으로 사전이나 규칙과 같은 언어지식을 수작업으로 구축하거나 말뭉치에서 학습하여 이용한다[1]. 이러한 언어지식은 시스템 개발 당시 확보한 자료를 바탕으로 구축되므로, 자료에 없던 미등록어를 포함하는 어절이나 문장은 정확하게 분석할 수 없다는 한계가 있다. 즉, 정확한 분석 결과 없이 오분석 결과만 제시하는 문제가 발생하거나, 미등록어를 추정한 분석결과를 제시할 때 가능한 경우의 수가 많아져 과분석 문제가 발생한다[2]. 게다가, 신문기사, SMS 메시지, 웹 문서 등에는 최신 유행 경향에 따라 새로운 단어가 자주 등장하므로 이러한 문제가 더욱 심각하게 나타날 수 있다.

따라서, 본 논문에서는 사전에 등록되지 않은 미등록어를 사전에 자동으로 등록하는 한국어 미등록어 사전 자동 구축 방법을 제안한다. 제안하는 방법은 전문 분석 기반 미등록어 사전 구축 방법과 웹 출현빈도 기반 미등록어 사전 구축 방법으로 구성되어 있다. 먼저, 전문 분석 기반 미등록어 사전 구축 방법은 한 문서에서 나타난 단어들은 한 가지 의미로 사용된다는 경향[3]을 바탕으로 문서에서 반복되는 문자열을 미등록어로 인식한다[2,4]. 또한, 웹 출현빈도 기반 미등록어 사전 구축 방법은 반복되지 않은 문자열에서도 미등록어를 인식하기 위해서 웹 문서에서의 출현빈도를 분석하여 미등록어를 인식한다.

앞으로, 2장에서는 한국어 미등록어 분석과 관련된 기존 접근 방법에 대해 살펴본다. 그리고, 3장에서는 제안하는 미등록어 사전 구축 방법에 대해 설명하고, 4장에서는 제안하는 방법의 성능을 평가한다. 마지막으로 5장에서는 결론을 맺는다.

Ⅱ. 기존연구

사전에 등록되지 않은 미등록어를 효과적으로 인식하기 위해서, 그동안 다양한 접근방법이 제안되었다. 이들은 크게 언어 지식 기반 접근방법, 주변 문맥 기반 접근방법, 전문 분석 기반 접근방법, 웹 정보를 활용한 접근방법이 있다.

첫째, 언어 지식 기반 접근방법은 형태소 패턴, 어절내 형태소 결합 정보와 같은 언어지식을 바탕으로 미등록어를 인식한다[5,6]. 그러나, 이러한 분석방법은 구축한 언어지식에 따라 성능이 크게 좌우될 수 있고, 언어지식의 구축 자체가 쉽지 않다는 문제가 있다[2].

둘째, 주변 문맥 기반 접근방법은 미등록어 주변에 나타나는 어휘의 통계정보를 바탕으로 미등록어를 인식한다[7,8,9]. 예를 들어, 영어에서는 단어 첫 글자의 대문자 여부, 하이픈 존재 여부, 'extra-'와 같은 접두사, '-tion'과 같은 접미사, '-ing'와 같은 어미 정보를 사용하여 미등록어를 인식할 수 있다[7,9]. 그러나 교착어인 한국어에서 단어는 매우 다양한 형태로 나타나므로 자료 부족 문제가 심각하게 나타날 수 있다.

셋째, 전문 분석 기반 접근방법은 문서에서 반복적으로 나타나는 문자열과 그 형태소 분석 결과를 바탕으로 미등록어를 인식한다[2,4]. 예를 들어, 문서에 나타난 어절 "출장가서"와 "출장가면"에 대해 각각 "출장/명사+가서/명사", "출장가서/미등록명사추정", "출장가/미등록동사추정+서/어미"와 "출장/명사+가면/명사", "출장가면/미등록명사추정", "출장가/미등록동사추정+면/어미"와 같이 형태소를 분석할 수 있다. 이때, 두 어절의 공통된 형태소 분석결과인 "출장가/미등록동사추정"을 바탕으로 "출장가/동사"를 미등록어로 인식한다[2]. 그러나 전문 분석 기반 접근방법은 미등록어가 문서에서 단 한 번만 나타난 경우 인식할 수 없다는 한계가 있다.

넷째, 웹 기반 접근방법은 "오바마"와 "Obama"와 같이 동일한 의미를 가지는 서로 다른 언어의 대응되는 단어 쌍을 찾기 위해 웹 검색 결과를 사용한다[10,11,12]. 예를 들어, 사전에 "오바마"가 등록되어 있지 않더라도 "오바마" 주변에 "Obama"가 함께 나타난 웹 검색결과를 분석하여 단어 쌍을 추출할 수 있다.

본 논문에서는 형태소 분석을 위한 미등록어 사전을 자동 구축하기 위해서 웹 문서에서의 출현빈도를 활용하는 방법을 제안한다. 제안하는 방법은 전문분석기반 접근방법을 보완하기 위해서, 문서에 한번만 나타나는 미등록어 후보에 대해 대량의 웹문서에서의 출현빈도를 바탕으로 검증하므로 자료 부족 문제를 완화할 수 있다. 또한, 제안하는 방법은 언어지식의 구축 부담을 줄이기 위해 미등록어의 주변 문맥 규칙을 형태소 분석 말뭉치에서 학습한다.

Ⅲ. 미등록어 사전 자동 구축

제안하는 미등록어 사전 자동 구축방법은 [그림1]과 같이 전문 분석 기반 미등록어 사전 구축과 웹 출현빈도 기반 미등록어 사전 구축의 2단계로 구성된다. 3.1에서는 전문 분석 기반 미등록어 사전 구축 방법에 대해 설명하고, 3.2에서는 웹 출현빈도 기반 미등록어 사전 구축 방법에 대해 기술한다. 그

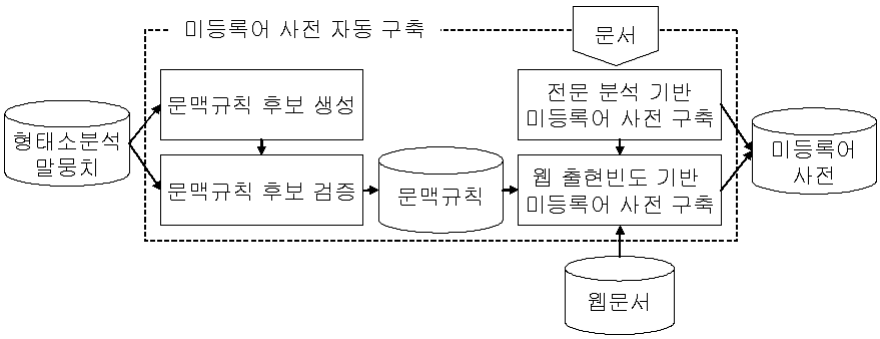

그림 1. 미등록어 사전 자동 구축 시스템
Fig. 1. Automatic Unknown Word Dictionary Construction System

리고 3.3에서는 웹 출현빈도 기반 미등록어 사전 구축 방법에서 사용하는 문맥규칙의 학습방법에 대해 자세히 설명한다.

3.1 전문 분석 기반 미등록어 사전 구축

전문 분석 기반 미등록어 사전 구축 방법은 기존 전문 분석 기반 접근방법의 기본 아이디어를 바탕으로 다음과 같이 구성된다. 먼저, 각 어절에 대해 확률적 형태소 분석[1]을 수행하고, 전체 형태소 분석 결과가 모두 낮은 확률로 분석되는 어절은 미등록어를 포함한다고 가정하여 미등록어 포함 어절 리스트에 등록할 수 있다. 그리고 [그림2]와 같이 문서 전체에서 나타나는 어절을 가나다순으로 정렬하고 앞뒤어절의 음절을 비교하여 2음절이상 동일한 음절열이 있으면 이를 최장 공통 부분문자열로 추출한다[4]. 즉, [그림2]와 같이 문서에 나타난 어절 "은희경", "은희경씨는", "은희경은", "은희경의"에서 최장 공통 부분문자열 "은희경"을 추출할 수 있다. 마지막으로, 미등록어 포함 어절리스트에 등록된 어절과 관련된 최장 공통 부분문자열은 명사로 추정하여 사전에 등록하고, 해당 어절은 어절리스트에서 제거한다. 즉, "은희경"을 미등록어 사전에 등록하고 관련된 어절 "은희경", "은희경씨는", "은희경은", "은희경의"를 미등록어 포함 어절 리스트에서 제거한다.

최장공통 부분문자열 (2음절이상)	문서에 나타난 어절의 가나다순 정렬		미등록어 포함 어절리스트
	일치부분	불일치	
은희경		유난히	
		유쾌하고	
	은희경		○
	은희경	씨는	○
	은희경	은	○
	은희경	의	○
		의심한다.	
		이름을	
		이산문학상을	○

그림 2. 전문 분석 기반 미등록어 인식 결과 일부
Fig. 2. Example of Some Results Recognized based on Full Text Analysis

3.2 웹 출현빈도 기반 미등록어 사전 구축

[그림2]의 어절 "이산문학상을"과 같이, 전문 분석 기반 미등록어 사전 구축 방법에서 미등록어 추출에 실패하여 미등록어 포함 어절리스트에 그대로 남아있는 어절에 대해서는 웹 출현빈도 기반 미등록어 사전 구축을 적용한다. 제안하는 웹 출현빈도 기반 미등록어 사전 구축 방법은 기존 사전에 등록되어 있지는 않지만 웹문서에서 매우 자주 나타나는 음절열은 새로운 단어라고 추정하고 미등록어로 인식한다. 이를 위해, [그림3]과 같이 주어진 각 어절에 대해 1음절씩 줄여가면서 문맥규칙과 조합하며, 조합한 문자열을 웹문서에서 검색한다. 그리고 모든 문맥규칙과 조합한 결과 그 출현빈도가 최소허용빈도보다 높은 부분 음절열은 미등록어로 인식하여 미등록어 사전에 명사로 등록한다.

예를 들어, 미등록어 포함 어절 리스트에 어절 "이산문학상을"이 있으면, 문맥규칙과 조합하여 문자열 "이산문학상을이", "이산문학상을을", "이산문학상을에"를 웹 검색엔진에서 검색하고 그 출현빈도를 분석한다. 출현빈도가 최소허용빈도보다 작으면, 주어진 음절열은 새로운 단어가 아니라고 판단한다. 그리고, 어절 "이산문학상을"에서 오른쪽 한 음절을 제거하여 음절열 "이산문학상"을 만들고, 문맥규칙과 다시 조합하여 문자열 "이산문학상이", "이산문학상을", "이산문학상에"를 웹 검색엔진에서 검색하는 과정을 반복한다. 전체 출현빈도가 최소허용빈도보다 크면, 새로운 단어라고 판단하여 "이산문학상"을 미등록어 사전에 명사로 등록한다.

```
웹 출현빈도 기반 미등록어 사전 구축( 어절리스트, 문맥규칙리스트, 미등록어 사전 )
{
  // 어절 리스트에 미등록어를 추출해야하는 모든 어절에 대해서 while문 수행
  while ( 어절리스트에 분석할 어절 존재 )
  {
    // 어절에서 우측 음절을 하나씩 줄여나가면서 for문 수행
    for( 음절열 = 어절; 음절열크기 〉 1음절; 새 음절열 = 기존 음절열에서 우측1음절 제거 )
    {
      // 문맥규칙 리스트의 모든 문맥규칙에 대해서 음절열을 조합하여 웹검색엔진에서 검색
      while( 문맥규칙 리스트에 적용할 문맥규칙 존재 )
      {
        문자열 = 음절열 + 문맥규칙;
        웹출현빈도 =search( 문자열, 웹검색엔진 Google );
        // 음절열과 문맥규칙을 조합한 결과 웹에 거의 나타나지 않으면 미등록어사전에 등록 배제
        if ( 웹출현빈도 〈 최소허용빈도 )
        {
          break;  // 해당 음절열과 문맥규칙 조합을 중단하고 다음 음절열 처리
        }
      } // while( 문맥규칙 ) 종료
      // 음절열과 모든 문맥규칙을 조합한 결과 항상 웹에 자주 나타나면 해당음절열을 미등록어사전에 등록
      if ( 웹출현빈도 ≥ 최소허용빈도 )
      {
        insert( 음절열, 미등록어 사전 );
        break;  // 다음 어절 처리
      }
    } // for ( 음절열 ) 종료
  } // while ( 어절 ) 종료
} // 웹 출현빈도 기반 미등록어 사전 구축 함수 종료
```

그림 3. 웹 출현빈도기반 미등록어 사전 구축 알고리즘
Fig. 3. Unknown Word Dictionary Construction Algorithm based on Web Document Frequency

3.3 문맥규칙 학습

문맥규칙은 "명사#뒤음절열"로 규칙 템플릿을 구성한다. 이러한 규칙 템플릿을 바탕으로 어절에서 명사에 뒤따르는 뒤음절열을 형태소 분석 말뭉치에서 추출한다. 그리고 부정확한 규칙을 배제하고 정확한 규칙만을 선택하기 위해서 수식 (1)과 같이 규칙의 뒤음절열이 주어졌을 때 앞에 명사가 나타날 확률을 바탕으로 규칙정확도를 평가한다. 수식 (1)에서 확률값은 추출한 규칙을 형태소 분석 말뭉치에 재적용하여, 규칙의 뒤음절열이 나타난 빈도수와 규칙의 뒤음절열 바로 앞에 명사가 나타날 빈도수를 이용하여 계산한다. 마지막으로, 상위 n개의 규칙후보를 미등록어 인식을 위한 주변 문맥 규칙으로 선택한다.

$$\text{규칙정확도} = P(\text{명사} \mid \text{뒤음절}) \quad \cdots\cdots (1)$$

$$= \frac{freq(\text{명사}, \text{뒤음절})}{freq(\text{뒤음절})}$$

예를 들어, 형태소 분석 말뭉치의 "국내에서도 국내/명사+에서도/조사"에서 규칙 "명사#에서도"를 추출한다. 그리고 말뭉치에서 "에서도"가 나타난 빈도와 "에서도"의 바로 앞에 명사가 나타난 빈도를 바탕으로 규칙정확도를 평가한다. "에서도"는 거의 조사로 사용되므로 규칙정확도가 높게 평가되어 최종 규칙으로 선택한다. 반면에, "당국은 당국/명사+은/보조사"에서 규칙 "명사#은"을 추출할 수 있지만, "은"은 보조사뿐만 아니라 "많은", "높은", "씻은"과 같이 관형형 전성어미로도 많이 사용되므로 규칙정확도가 낮게 평가되어 최종규칙에서 배제된다.

Ⅳ. 실험 및 평가

제안하는 미등록어 사전 자동 구축 방법의 성능을 평가하기 위해서, 웹에서 무작위로 44개의 신문기사를 추출하고, 이를 바탕으로 구축된 사전의 질을 평가하기 위해서 수식 (2), (3), (4), (5)와 같이 정확률, 재현율, F-measure, 정답포함률을 사용하였다. 〔표1〕과 같이 실험한 신문기사는 총 13,429 어절로 구성되어 있으며, 247개의 미등록어를 포함한다. 또한, 각 신문기사는 평균 319.74 어절로 문서의 길이가 최대 683 어절부터 최소 104 어절까지 길이분포가 다양하다. 그리고, 기사당 평균 5.93개의 미등록어를 포함하는데, 전체 513 어절 중 22개의 미등록어를 포함하는 경우와 전체 139 어절 중 약 11.5%에 해당하는 16개의 미등록어를 포함하는 경우도 있다.

$$\text{정확률} = \frac{\text{사전에 등록된 올바른 미등록어 수}}{\text{사전에 등록된 미등록어 수}} \quad \cdots\cdots (2)$$

$$\text{재현률} = \frac{\text{사전에 등록된 올바른 미등록어 수}}{\text{전체 올바른 미등록어 수}} \quad \cdots\cdots (3)$$

$$F-measure = \frac{2 \times \text{정확률} \times \text{재현율}}{\text{정확률} + \text{재현율}} \quad \cdots\cdots\cdots (4)$$

$$\text{정답포함율} = \frac{\text{올바른 형태소분석결과를 포함한 어절수}}{\text{전체 어절수}} \quad \cdots\cdots\cdots (5)$$

표 1. 실험 신문기사 분석
Table 1. News Paper Test Set Analysis

	총계	최대	평균	최소
어절	13,429	683	319.74	104
미등록어	247	22	5.93	0
미등록어비율	1.84%	11.51%	1.84%	0.0%

〔표2〕는 신문기사에 나타난 미등록어 247어절을 미등록어 포함 어절에 등록하여 구축한 사전의 질을 나타낸다. 기존 전문 분석 기반 미등록어 사전 구축 방법〔2,4〕는 〔표2〕와 같이 정확률은 높지만 재현율은 상대적으로 낮은 성능을 보인다. 이는 전문 분석 기반 미등록어 사전 구축 방법은 단어가 여러 형식형태소와 결합하여 다양한 형태로 나타날 때 유용한데, 실제 신문기사에서 미등록어의 45.75% 정도의 단어가 한번만 나타나므로 전문 분석 기반 미등록어 사전 구축 방법을 적용할 수 없었다.

따라서, 문서에서 한번만 나타난 어절에서도 미등록어를 추출할 수 있도록 제안하는 미등록어 사전 자동 구축 방법은 웹 출현빈도를 이용하여 미등록어 사전을 구축한다. 즉, 웹 출현빈도 기반 미등록어 사전 구축 방법은 Google 검색엔진을 사용하여 미등록어를 추출하므로, 문서에 한번만 나타나도 적용할 수 있다. 〔표2〕에 보는 바와 같이, 웹 출현빈도 기반 미등록어 사전 구축 방법의 자체 정확률은 다소 떨어지지만, 전문 분석 기반 미등록어 사전 구축 방법을 보완하여 재현율을 32.39%정도 향상 시킨다.

표 2. 미등록어 사전 자동 구축 결과
Table 2. Experimental Results of Automatic Unknown Word Dictionary Construction

	정확도	재현율	F-measure
전문 분석 기반 미등록어 사전 구축	97.01	52.63	68.24
웹 출현빈도 기반 미등록어 사전 구축	86.02	32.39	47.06
통합	92.51	85.02	88.61

그러나, 웹 문서에서도 거의 나타나지 않은 단어는 제안하는 방법에서도 제대로 추출되지 못하는 한계를 보인다. 예를 들어, 어절 "류인명"과 "쉬베이훙"에서는 미등록어가 전혀 추출되지 않았고, 어절 "리오펠의"와 "술라주의"에서는 잘못된 미등록어 "리오"와 "술라"가 추출되었다.

제안하는 방법으로 자동 구축된 미등록어 사전의 유용성을 평가하기 위해서, 미등록어 사전의 사용 여부에 따른 형태소 분석기의 정답포함률을 비교하였다. 〔표3〕과 같이 미등록어 사전을 사용하면 약 1.7% 정도 정답포함률이 향상되었고, 미등록어에 대해서만 형태소 분석을 수행한 결과 약 86.06%의

정답포함률을 보인다. 이는 자동으로 구축한 미등록어 사전도 형태소 분석기의 성능향상에 도움이 될 수 있음을 보인다.

표 3. 형태소 분석기의 성능
Table 3. Accuracy of Morphological Analysis

	정답포함률 (미등록어사전×)	정답포함률 (미등록어사전○)
전체	98.16	99.85
미등록어	0.00	86.06

V. 결론

본 논문에서는 미등록어로 인한 한국어 형태소 분석의 성능 저하를 개선하기 위해서, 어절에서 미등록어를 추출하여 미등록어 사전을 자동으로 구축하는 방법을 제안하였다. 제안하는 방법은 전문 분석 기반 미등록어 사전 구축 방법과 웹 출현빈도 기반 미등록어 사전 구축 방법으로 구성되어 있다. 먼저, 전문 분석 기반 사전 구축 방법은 한 문서에서 나타난 단어들은 한 가지 의미로 사용된다는 경향을 바탕으로 전체 문서에서 반복적으로 나타나는 문자열을 미등록어로 인식한다. 또한, 웹 출현빈도 기반 사전 구축 방법은 반복되지 않은 문자열을 웹 문서에서 검색하여 그 출현빈도를 바탕으로 미등록어를 인식한다. 따라서, 제안하는 모델은 대량의 웹문서에 나타난 출현빈도를 사용하므로, 자료 부족 문제를 완화할 수 있고, 주어진 문서에서 한번만 나온 미등록어도 인식할 수 있다. 실험결과 전문 분석만을 바탕으로 하는 기존 접근방법에 비해서 제안하는 사전 구축 방법은 웹 문서에서의 출현빈도도 함께 고려하여 32.39% 정도 재현율이 높게 나타났다.

참고문헌

〔1〕 이도길, 한국어 형태소 분석과 품사부착을 위한 확률 모형, 고려대학교 박사학위 논문, 2005.

〔2〕 박봉래, 전문분석에 기반한 한국어 미등록어의 인식, 고려대학교 박사학위 논문, 1999.

〔3〕 David Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods", Proceeding on 33rd Annual Meeting of the Association for Computational Linguistics, pp.189 -196, 1995.

〔4〕 김선호, 윤준태, 송만석, "한국어 문서 처리를 위한 동적 생성 로컬 사전 기반 미등록어 분석", 정보과학회논문지: 소프트웨어 및 응용, 제29권 제6호, 407쪽-416쪽, 2002.

〔5〕 양장모, 김민정, 권혁철, "언어정보를 이용한 한국어 미등록어 추정", 한국정보과학회 봄 학술발표논문집, 제23권 제1호, 957쪽-960쪽, 1996.

〔6〕 차정원, 이원일, 이근배, 이종혁, "형태소 패턴 사전을 이용한 일반화된 미등록어 처리", 정보과학회 인공지능 연구회 춘계학술대회 논문집, 37쪽-42쪽, 1997.

〔7〕 Ralph Weishedel, Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeff Palmulcci, "Coping with Ambiguity and Unknown Words through Probabilistic Models", Computational Linguistics, Vol.19, No.2, pp.359-382, 1993.

〔8〕 Masaaki Nagata, "Automatic Extraction of New Words from Japanese Texts using Generalized Forward-Backward Search," Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.48-59, 1996.

〔9〕 Andrei Mikheev, "Unsupervised Learning of Word-Category Guessing Rules," Proceeding on 34rd Annual Meeting of the Association for Computational Linguistics, pp.327-334, 1996.

〔10〕 Pu-Jen Cheng, Jei-Wen Teng, Ruei-Cheng Chen, Jenq-Haur Wang, Wen-Hsiang Lu, Lee-Feng Chien, "Translating unknown queries with web corpora for cross-language information retrieval," Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 25-29, 2004.

〔11〕 Masaaki Nagata, Teruka Saito, Kenji Suzuki, "Using the web as a bilingual dictionary," Proceedings of the workshop on Data-driven methods in machine translation, p.1-8, 2001.

〔12〕 Qing Li, Sung Hyon Myaeng, Yun Jin, and Bo-Yeong Kang, "Translation of Unknown Terms Via Web Mining for Information Retrieval," LNCS Vol. 4182, pp.258-269, 2006.

저 자 소 개

박 소 영(So-Young Park)

1997년 2월: 상명대학교 전자계산학과(이학사)

1999년 8월: 고려대학교 컴퓨터학과(이학석사)

2005년 2월: 고려대학교 컴퓨터학과(이학박사)

2007년 3월 - 현재: 상명대학교 디지털미디어학부 전임강사

관심분야: 자연어처리, 기계학습, 정보검색