통계 정보를 이용한 전치사 최적 번역어 결정 모델

심광섭*†
성신여자대학교

Kwangseob Shim. 2004. A Statistical Model for Choosing the Best Translation of Prepositions. *Language and Information 8.1*, 101–116. This paper proposes a statistical model for the translation of prepositions in English–Korean machine translation. In the proposed model, statistical information acquired from unlabeled Korean corpora is used to choose the best translation from several possible translations. Such information includes functional word–verb co-occurrence information, functional word–verb distance information, and noun–postposition co-occurrence information. The model was evaluated with 443 sentences, each of which has a prepositional phrase, and we attained 71.3% accuracy. (**Sungshin Women's University**)

Key words: 기계번역(Machine Translation), 통계정보(Statistical Information), 공기정보(Co-occurrence Information), 거리정보(Distance Information), 상호정보(Mutual Information), 말뭉치(Corpus)

1. 서론

전통적인 규칙 기반 기계 번역 방식에서 번역 사전은 주어진 단어를 어떻게 번역할 것인가에 대한 정보를 담고 있다. 그런데 동일한 단어라 하더라도 문맥에 따라 여러 가지로 번역될 수 있기 때문에 일반적으로 번역 사전에서는 표제어에 대한 번역어가 단순 나열되어 있는 것이 아니라 해당 번역어로 번역되기 위한 통사·문맥 정보도 함께 기술되어 있다. 수작업으로 통사·문맥 정보를 수집하는 것은 상당히 어렵고 시간도 많이 걸리기 때문에 사람들은 전자 사전(machine-readable dictionary)과 같은 기존 언어 자원으로부터 번역에 필요한 정보를 자동 또는 반자동으로 추출하는 방법에 대하여 많은 관심을 보였다 (Copestake et al., 1994).

1980년대 후반에는 번역 사전을 전혀 사용하지 않고 병렬 말뭉치(parallel corpus)로부터 수집한 통계 정보만으로 번역을 수행하는 통계 기반 기계 번역 방식(stat-

* 136-742 서울 성북구 동선동 3가 249-1 성신여자대학교 자연과학대학 컴퓨터정보학부,
E-mail: shim@sungshin.ac.kr, esqui@hanmail.net, FAX: 02-924-3891
† 이 논문은 2002년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.

istical machine translation)에 대한 연구가 시작되었다. 미국 IBM 연구소의 Brown 등은 번역 과정을 번역 모델(translation model)과 언어 모델(language model)이라는 두 가지 통계적 모델로 단순화하여 순수한 통계 정보만으로도 기계 번역이 가능함을 보여 주었다 (Brown et al., 1990, 1993). 이 방법에서는 영어나 불어처럼 어휘·통사적으로 유사한 언어 사이에서는 비교적 좋은 결과를 얻을 수 있지만 한국어나 영어처럼 언어 특성이 전혀 다른 언어 사이에서는 만족스러운 결과를 얻기가 곤란하다. 이러한 문제점을 극복하기 위한 방안으로 최근에는 통사 기반의 통계적 번역 방식이 제안되기도 하였다 (Yamada and Knight, 2001).

통계적 번역 방식을 적용하기 위해서는 대량의 병렬 말뭉치가 필요한데, 캐나다와 같이 정치적 특수성으로 인하여 이중 언어 문서가 의무적으로 요구되는 경우를 제외하면 일부러 이중 언어 문서를 작성하는 경우는 드물기 때문에 일반적으로 임의의 언어 쌍에 대한 병렬 말뭉치를 구하기란 매우 어렵다. 그러나 웹(World Wide Web)의 확산에 힘입어 단일 언어에 대한 대량의 문서를 구하는 것은 그다지 어렵지 않게 되었다. 또 웹을 통해 다른 언어로 된 문서로 접근하는 것도 용이해 졌기 때문에 병렬되지 않은 다중 언어의 미가공 말뭉치(raw corpus)도 쉽게 구할 수 있게 되었다. 이에 따라 미가공 말뭉치로부터 여러 가지 유용한 언어 정보를 추출하는 연구가 많이 진행되고 있다. 기계 번역과 관련해서 보면, 목표 언어(target language)의 미가공 말뭉치로부터 통계 정보를 수집한 후 이를 기반으로 단어 의미 중의성 해소(word sense *disambiguation*)를 하여 번역어 결정하는 방법과 (Dagan, Itai and Schwall, 1991; Dagan and Itai, 1994; Lee and Kim 2002), 두 언어에 대한 개별적인 미가공 말뭉치로부터 수집된 정보를 이용하여 번역 사전을 구축하는 방법 등이 제안되었다 (Koehn and Knight, 2002).

전치사 접속(prepositional phrase attachment)은 자연 언어 처리에 있어서 구조적 모호성을 유발하는 주요 요인 중의 하나로, 접속 모호성 문제를 해결하기 위한 여러 가지 방안들이 제시되었다 (Collins and Brooks, 1995; Pantel and Lin, 2002; Volk 2002). 기계 번역의 경우 전치사가 어디에 접속되는가에 따라 번역 결과가 완전히 달라질 수 있기 때문에 전치사 접속 모호성 문제는 반드시 해결되어야 할 문제 중의 하나이다. 그런데 접속 모호성 해소란 전치사가 무엇을 수식하는가를 결정하는 것일 뿐 전치사의 의미를 결정하는 것은 아니다. 따라서 전치사의 접속 모호성이 해소되었다 하더라도 기계 번역을 하기 위해서는 문맥에 맞는 전치사의 번역어를 결정(target word selection)하는 과정이 필요한데, 본 논문에서는 접속 모호성 문제가 해소된 전치사에 대하여 한국어 미가공 말뭉치(raw corpus)로부터 수집된 통계 정보를 이용하여 최적 번역어를 결정하는 모델을 제시한다.

2. 전치사 번역어 결정 모델

2.1 전치사 번역 및 번역 사전

영어 전치사는 대개 우리말의 조사나 어미와 같은 기능어(function word)로 번역이

된다. 그런데 전치사는 그 용법이 매우 다양하여 같은 전치사라 하더라도 문맥에 따라 다른 의미로 사용되므로 적절한 번역어를 선택하는 것은 그다지 쉽지 않다. 다음은 하나의 전치사가 여러 가지 의미로 사용될 수 있음을 보여 주는 예이다.

(1) 가. The game ended in tie. 그 경기는 무승부로 끝났다.

나. Salt dissolves in water. 소금은 물에 녹는다.

다. They seek solace in religion. 그들은 종교에서 위안을 찾는다.

라. I got burnt in the hand. 나는 손을 데었다.

마. He lacks in administrative ability. 그는 행정 능력이 부족하다.

전치사에 대한 번역어를 잘못 선택하게 되면 다른 내용어(content word)에 대한 번역은 올바르게 했다 하더라도 전체 번역문이 부자연스럽게 되어 번역문으로부터 원문의 의미를 파악하기가 어렵게 되며, 경우에 따라서는 원문의 의미가 잘못 전달되기도 한다. 예를 들어 다음 문장에서,

(2) 가. He is gentle in voice. 그는 목소리가 부드럽다.

나. It is soluble in fat. 그것은 지방에 녹는다.

전치사 in은 각각 '-가', '-에'로 번역이 되었는데, 똑같은 문장에서 전치사 부분을 다음과 같이 잘못 번역했다고 하자.

(3) 가. 그는 목소리로 부드럽다.

나. 그것은 지방으로 녹는다.

3의 (가)는 주변 내용어로부터 문장 의미를 추측할 수는 있겠지만 문장이 부자연스럽기 때문에 이 문장을 읽고 의미를 파악하는 데에는 다소 시간이 걸릴 것이다. (나)의 경우에는 자연스러운 문장 같아 보이지만 원문과는 전혀 다른 의미를 지니고 있다.

위 예에서 보듯이 기계 번역에서 전치사의 번역은 내용어에 대한 번역 못지않게 중요한데, 과거에는 어떤 문맥에서 어떤 전치사가 어떻게 번역되는가에 대한 정보를 번역 사전에 사람의 손으로 일일이 코딩해 넣는 방식을 취하였다. 다음은 심광섭 등이 제안한 방법에 따라 2의 원문을 번역하는데 필요한 번역 사전을 구성한 예이다 (심광섭 외, 1992).

(4) 가. (ADJ "gentle"
...
((↓ PP (H "in") (T "가")) (T "부드럽다"))
...)

나. (ADJ "soluble"
...
((↓PP (H "in") (T "에")) (T "녹는다"))
...)

4의 (가)는 형용사 gentle의 직접 지배를 받는 전치사구(PP)의 머리(head)가 in인 경우 in은 '가'로 번역하고 gentle은 '부드럽다'로 번역한다는 것을 나타낸다. (나)도 같은 방법으로 해석할 수 있다. 이와 같이 과거에는 형용사나 동사가 특정 전치사와 함께 사용될 경우 이 전치사가 무엇으로 번역되어야 하는가를 번역 사전에 일일이 명시하는 것으로 했는데, 이것은 상당한 시간을 요하는 어려운 작업이다. 이러한 어려움 때문에 본 논문에서 제시하는 전치사 번역 모델에서는 4와 같이 복잡한 번역 사전을 사용하는 대신 다음과 같이 문맥 정보가 전혀 없는 단순한 형태의 전치사 번역 사전을 사용하는 것을 가정한다.

(5) into : -까지, -으로, -에, -을, ···

그런데 전치사의 번역어가 나열된 것에 지나지 않는 5와 같은 단순한 형태의 번역 사전으로는 주어진 문장의 문맥에 맞는 전치사 번역어를 결정하기란 그리 쉬운 일이 아니다. 예를 들어 5와 같은 번역 사전을 이용하면 6의 (가)를 (나)와 같이 번역할 수 있음은 알 수 있지만 이 중에서 어떤 것이 올바른 번역인지를 결정할 수는 없다.

(6) 가. They converted a warship into a liner.

나. 그들은 군함을 정기선까지 전환했다.
그들은 군함을 정기선으로 전환했다.
그들은 군함을 정기선에 전환했다.
그들은 군함을 정기선을 전환했다.
...

개념이란 사람의 머리 속에 있는 추상적인 것으로 이것을 표현하는 수단인 언어와 무관하다고 가정을 한다면, 목표 언어의 말뭉치에서 수집한 통계 정보를 원시 언어의 단어 의미 결정에 이용할 수 있을 것이다. 예를 들어 한국어 말뭉치를 조사했을 때 '전환하다'가 '-까지', '-에', '-을'과 공기하는 비율보다 '-으로'와 공기하는 비율이 훨씬 높다면 이는 '전환하다'라는 개념은 '무엇을 무엇으로 전환하다'라는 식으로 사용되는 것이 보편적임을 의미하는데, 이러한 보편성을 영어에서도 찾을 수 있다면 6의 (가)는 (나)의 두 번째 문장으로 번역되어야 한다는 결정을 내릴 수 있을 것이다.

이와 유사한 개념으로 Dagan 등은 목표 언어의 미가공 말뭉치로부터 수집된 통계 정보를 이용하여 의미 모호성이 있는 단어에 대한 번역어를 결정하는 방법을 제안하였다 (Dagan, Itai and Schwall, 1991; Dagan and Itai, 1994). 이 방법의 기본 개념은 아주 단순한데, 구문 분석기로 주어진 문장을 분석하여 통사적으로 관련이 있는 단어 쌍들을 파악한 다음 이 단어 쌍들에 대한 여러 가지 번역 중에서 가장 그럴

듯한 것을 선택한다는 것이다. 각 단어 쌍에 대한 가능한 번역은 번역 사전(bilingual dictionary)에서 얻을 수 있는데, 이 중에서 가장 그럴듯한 번역을 선택하는 데에는 목표 언어의 말뭉치에서 수집한 통계 정보가 이용된다. 다음 문장을 예로 들어 Dagan 등이 제안한 방법에 따라 영어를 한국어로 번역하는 과정을 살펴보자.

(7) They have plotted a preemptive attack.

이 문장에서 동사 plot과 명사 attack은 동사-목적어 관계에 있으며, 형용사 preemptive와 명사 attack은 수식어-피수식어 관계에 있다. 영한 사전에 보면 동사 plot은 '-을 몰래 계획하다', '-을 지도에 표시하다', '-을 그리다' 등의 의미로 사용될 수 있으며, 명사 attack은 '공격', '발작', '착수' 등의 의미로 사용될 수 있음을 알 수 있다. 따라서 문맥을 고려하지 않고 생각한다면 위 문장에서 동사 plot과 명사 attack은 다음과 같은 9가지 방법으로 번역될 수 있다.

(8) 가. 공격을 몰래 계획하다.

나. 공격을 지도에 표시하다.

다. 공격을 그리다.

라. 발작을 몰래 계획하다.

마. 발작을 지도에 표시하다.

바. 발작을 그리다.

사. 착수를 몰래 계획하다.

아. 착수를 지도에 표시하다.

자. 착수를 그리다.

한국어 말뭉치를 분석하여 동사-목적어 관계에 있는 명사-용언 쌍에 대한 출현 빈도수를 구한 다음, 이것을 기반으로 8의 각 번역에 대한 선호도를 계산하고 그 중에서 가장 우수한 것을 동사 plot과 명사 attack에 대한 번역으로 선택한다는 것이 Dagan 등이 제안한 방법의 기본 개념이다.

2.2 번역어 결정 모델

본 논문에서 제안하는 전치사 번역어 결정 모델은 한국어 미가공 말뭉치에서 수집한 통계 자료를 바탕으로 번역 사전에서 제시된 여러 가지 가능한 번역어 중에서 최적 번역어를 결정하는 통계 모델(statistical model)로 이 모델은 다음과 같은 가정 하에 설립되었다.

가정 1: 번역을 하기 전에 전치사 접속 모호성 문제는 모두 해결되어 있다.
가정 2: 한 문장에 동사나 형용사를 수식하는 전치사구 하나만 존재한다.
가정 3: 주어진 문장에서 전치사를 제외한 다른 어휘들은 모두 번역되어 있다.
가정 4: 전치사는 '조사' 또는 '조사 체언+조사'나 '조사 용언+어미'의 형태로만 번역된다.

가정 1에 따라 번역을 하기 전에 전치사 접속 모호성이 해소되어 있어야 한다. 일반적으로 구조적 모호성을 해소하고 난 후에 번역 단계에 들어가므로 이러한 가정을 하는 것은 타당하다. 가정 2에 따라 한 문장에 두 개 이상의 전치사구가 오는 경우는 본 연구의 범위에서 제외된다. 가정 3은 전치사의 수식을 받는 내용어와 전치사 목적어에 해당하는 내용어가 번역이 되어 있어야 두 내용어 사이의 관계를 나타내는 적절한 전치사의 번역어를 선택할 수 있다고 보고 설정한 가정이다. 가정 4에 의하면 전치사가 '-에서', '-의 위로', '-에 대하여', '-을 위하여' 등으로 번역되는 경우만이 고려 대상이 되므로 다음과 같은 몇 가지 특수한 유형은 본 연구의 논의 대상에서 제외하기로 한다.

(9) 가. 전치사가 다른 어휘와 함께 숙어를 이루는 경우
The weather changed on a sudden.

나. 전치사가 특정 유형의 구문을 형성하는 경우
He sleeps with the door closed.

다. 의역이 필요한 경우
He helped the lady out of a car.

9의 (가)에서 'on a sudden'은 '갑자기'로 번역되어야 하는데 이는 'a sudden'에 대한 번역과 'on'에 대한 번역의 조합(composition)으로 번역될 수 있는 부분이 아니기 때문에 5와 같이 주어진 전치사 번역 사전으로는 번역이 불가능하다. 이와 같이 전치사가 다른 어휘와 함께 숙어를 이루는 경우에는 전치사만 따로 번역할 수 있는 것이 아니라 숙어를 하나의 단위로 보고 한꺼번에 번역해야 하므로 본 논문의 논의 대상에서 제외하였다. 이런 경우에는 별도의 숙어 사전을 이용하여 번역해야 할 것이다. 9의 (나)와 같이 전치사가 특수 구문을 형성하는 경우도 본 논문의 논의 대상에서 제외되는데, 이런 특수 구문의 경우에는 전처리 단계에서 번역을 하면 될 것이다. 9의 (다)와 같이 의역이 필요한 경우에는 번역 사전에서 주어지는 번역어만으로는 적절한 번역문을 생성하기가 곤란할 뿐만 아니라 경우에 따라서는 문맥을 이해해야만 올바른 번역을 할 수 있기 때문에 기계적인 번역의 대상으로 삼기에는 곤란하여 본 논문의 논의 대상에서 제외하였다.

5와 같은 단순한 형태의 번역 사전으로는 6의 (가)를 (나)와 같은 여러 가지 방법으로 번역할 수 있다는 것은 알 수 있지만 이 중에서 어느 것이 가장 적절한 번역인지는 결정할 수 없었다. 본 논문에서는 통계적인 방법으로 전치사의 최적 번역어를 결

정하는 모델을 제안한다. 번역어 결정 모델 설명을 위해 원문의 동사 또는 형용사를 v라 하고 이를 수식하는 전치사를 p라 하자. 가정 3에 의해 v는 이미 번역이 완료된 상태인데 이를 $\overline{v}$로 표기하기로 한다. 전치사 p의 가능한 번역어는 5와 같은 형식의 단순 번역 사전으로부터 제공되는데, 이들을 각각 $\overline{p_1}, \overline{p_2}, \cdots, \overline{p_n}$으로 표기한다고 했을 때 전치사 p의 번역어 $\overline{p}$는 다음과 같이 결정된다.

$$\overline{p} = \arg\max_{\overline{p_i}} s(\overline{p_i}) \tag{1}$$

$s(\overline{p_i})$는 $\overline{p_i}$가 전치사 p의 번역어로 채택될 점수를 나타낸다. 이 점수는 한국어 말뭉치에서 수집한 통계 정보로부터 계산된다. 아래에서는 이러한 점수를 계산하는 세 가지 번역 모델을 제시한다.

2.2.1 기능어–용언 공기 정보. 가정 4에 따르면 전치사 p의 번역 후보인 $\overline{p_i}$는 '조사', '조사 체언+조사', '조사 용언+어미'의 형태로 주어지는데, $\overline{p_i}$의 오른쪽 끝에 오는 조사나 어미 등의 기능어와 용언 $\overline{v}$ 사이의 통계적 선호도는 p의 번역어를 결정하는데 이용할 수 있다. 예를 들어 6의 (가)를 (나)와 같이 번역할 수 있는데, 약 5,000만 단어의 한국어 미가공 말뭉치를 조사한 결과 '-까지 전환하다'는 48번, '-으로 전환하다'는 4,023번, '-에 전환하다'는 366번, '-을 전환하다'는 2,085번 나타났다.[1] 이 조사 결과는 6의 (가)에서 into가 '-으로'로 번역될 가능성이 높음을 시사한다. 이 예는 한국어 미가공 말뭉치에서 조사한 기능어와 용언 사이의 공기 (co-occurrence) 정보는 영어 전치사의 번역어를 결정하는 자료로 이용할 수 있음을 보여준다. 그런데 다음 예에서 보듯이 기능어와 용언 사이의 공기 빈도수가 높다고 해서 항상 올바른 번역어를 결정하는 것은 아니다.

(10) 가. He lived in a city.

나. 그는 도시에서 살았다.

다. 그는 도시를 살았다.

같은 말뭉치를 조사했을 때 '-에서 살다'는 3,278번 나온 반면, '-을 살다'는 11,679 번 나왔다. 따라서 공기 빈도수의 대소만으로 번역어를 결정한다면 10의 (가)는 (다)로 번역될 것이다. 이런 문제가 발생하는 이유는 일반적으로 '-을'이 '-에서'보다 자주 사용되기 때문이다. 실제로 말뭉치를 조사해보면 '-을'이 4,307,962번 출현한 데 비하여 '-에서'는 702,400번 출현하였다. 따라서 절대 수치상으로는 '-에서 살다'의 공기 빈도수가 '-을 살다'의 공기 빈도수보다 낮긴 하지만, '-에서'의 출현 빈도수가 '-을'의 출현 빈도수의 1/6 밖에 되지 않음을 고려한 상대적인 공기 빈도수는 '-에서 살다'가 '-을

[1] 여기서 사용된 말뭉치는 1991–1995년 사이의 일간지 4,090만 어절, 한국 과학 기술원에서 개발한 대한민국 국어 정보 베이스 평가판 버전 0.1 660만 어절, 21세기 세종계획을 통해 1999년에 구축된 한국어 말뭉치 버전 5.0 140만 어절, 계몽 백과 사전 110만 어절을 통합하여 만든 것이다. 이 말뭉치의 분석에는 MACH(Shim and Yang, 2002)를 사용하였다.

살다'보다 오히려 높다고 할 수 있다. 물론 이와 반대의 경우도 발생할 수 있지만, 기능어의 출현 빈도수를 고려한 상대적인 공기 빈도수를 이용하여 번역어를 결정한다면 그렇지 않을 때보다 맞을 가능성이 더 높을 것이다.

이러한 점을 고려하여 기능어 $\overline{p_{ij}}$와 용언 $\overline{v}$ 사이의 공기 정보를 이용한 점수 $s_v(\overline{p_{ij}}, \overline{v})$는 상호 정보(mutual information)를 이용하여 다음과 같이 정의하였다.[2] 여기서 $\overline{p_{ij}}$는 전치사 p의 번역어 $\overline{p_i}$의 오른쪽 끝에 오는 조사나 어미 등의 기능어를 나타낸다.

$$\begin{aligned} s(\overline{p_i}) &= s_v(\overline{p_{ij}}, \overline{v}) \\ &= I(\overline{p_{ij}}, \overline{v}) \\ &= \log \frac{P(\overline{p_{ij}}, \overline{v})}{P(\overline{p_{ij}})P(\overline{v})} \\ &\approx \log \frac{N \cdot f(\overline{p_{ij}}, \overline{v})}{f(\overline{p_{ij}})f(\overline{v})} \end{aligned} \tag{2}$$

위 식에서 N은 말뭉치의 크기를 나타내며, $f(\overline{p_{ij}})$와 $f(\overline{v})$는 말뭉치에서 $\overline{p_{ij}}$와 $\overline{v}$가 개별적으로 출현하는 빈도수를, $f(\overline{p_{ij}}, \overline{v})$는 $\overline{p_{ij}}$와 $\overline{v}$가 한 문장 내에서 동시에 출현하는 공기 빈도수를 각각 나타낸다. 위 식을 이용하여 I(에서,살다)를 계산하면 1.88이고, I(를,살다)를 계산하면 1.09였다. 따라서 기능어-용언 공기 정보를 이용하여 번역어를 선택하면 10의 (가)에서 in은 '-를'이 아니라 '-에서'로 번역된다.

2.2.2 체언-조사 공기 정보. 앞에서 본 6이나 10과 같은 경우 기능어-용언 공기 정보를 이용하여 적절한 번역어를 결정할 수 있었다. 그런데 이러한 방법으로 11의 (가)를 번역했더니 (나)와 같이 번역되지 않고 (다)와 같이 번역되었다.

(11) 가. They hung a calendar on the wall.

나. 그들은 벽에 달력을 걸었다.

다. 그들은 벽에게 달력을 걸었다.

이것은 식 (2)에 따라 기능어-용언 사이의 점수를 계산했을 때 '-에게 걸다'의 점수가 '-에 걸다'의 점수보다 높기 때문에 발생하는 문제이다. 이러한 문제는 기능어-용언 공기 정보를 이용하여 점수를 계산할 때 기능어 앞에 오는 단어의 성격을 전혀 반영하지 않기 때문에 발생한다. 가정 4에 따라 전치사 p의 번역 후보 $\overline{p_i}$는 '조사', '조사 체언+조사', '조사 용언+어미'의 형태로 주어지는데, 이들의 제일 왼쪽에는 공통적으로 조사가 나온다. 이 조사와 $\overline{p_i}$의 왼쪽에 오는 체언 사이의 공기 정보는 위의 11에서 본 것과 같은 문제를 해결하는데 이용할 수 있다. 예를 들어 기능어-용언 사이의 공기 정보로 판단했을 때 on의 번역어로 '-에'보다는 '-에게'가 더 적절한 것으로

[2] 상호 정보는 $I(x,y)$는 x와 y 사이의 통계적 상관 관계의 긴밀성을 나타내는 것으로서 $I(x,y)$ 값이 클수록 x와 y 사이의 긴밀성은 지수적으로 증가한다(Church and Hanks, 1990).

나왔다 하더라도, 체언-조사 사이의 공기 정보에 의해 '벽에게'보다는 '벽에'가 더 적절한 것으로 나온다면 11의 (가)를 (나)로 번역할 수 있을 것이다.

이러한 점을 고려하여 조사와 조사의 왼쪽에 오는 체언 $\overline{n}$ 사이의 공기 정보를 이용한 점수 $s_n(\overline{n}, \overline{p_{i0}})$는 상호 정보를 이용하여 다음과 같이 정의하였다. 여기서 $\overline{p_{i0}}$는 번역어 $\overline{p_i}$의 왼쪽 끝에 오는 조사를 나타낸다.

$$\begin{aligned} s(\overline{p_i}) &= s_n(\overline{n}, \overline{p_{i0}}) \\ &= I(\overline{n}, \overline{p_{i0}}) \\ &= \log \frac{P(\overline{n}, \overline{p_{i0}})}{P(\overline{n})P(\overline{p_{i0}})} \\ &\approx \log \frac{N \cdot f(\overline{n}, \overline{p_{i0}})}{f(\overline{n})f(\overline{p_{i0}})} \end{aligned} \tag{3}$$

위 식에서 N은 말뭉치의 크기를 나타내며, $f(\overline{n})$과 $f(\overline{p_{i0}})$는 말뭉치에서 체언 $\overline{n}$과 조사 $\overline{p_{i0}}$가 개별적으로 출현하는 빈도수를, $f(\overline{n}, \overline{p_{i0}})$는 $\overline{n}$과 $\overline{p_{i0}}$가 한 어절 내에서 연속해서 나타나는 공기 빈도수를 각각 나타낸다. 위 식을 이용하여 I(벽,에)를 계산하면 3.35이고, I(벽,에서)를 계산하면 0.14였다. 따라서 체언-조사 공기 정보로 번역어를 선택하면 11의 (가)에서 on은 '-에서'가 아니라 '-에'로 번역될 것이다.

다음은 기능어와 용언 사이의 공기 정보만을 이용했을 때에는 전치사의 번역어가 잘못 선택되었으나 체언과 조사 사이의 공기 정보를 함께 이용했을 때에는 올바르게 번역된 예를 보인 것이다. 이 예에서 첫 번째 번역문은 기능어와 용언 사이의 공기 정보만을 이용했을 때 얻은 결과이며, 두 번째 번역문은 체언과 조사 사이의 공기 정보를 함께 이용했을 때 얻은 결과이다.

(12) 가. Such an accident happens through carelessness.
그런 사고는 부주의까지 발생한다.
그런 사고는 부주의로 발생한다.

나. I had a nervous breakdown from overwork.
나는 과로로부터 신경쇠약에 걸렸다.
나는 과로로 신경쇠약에 걸렸다.

다. He came by an expensive car.
그는 비싼 자동차로부터 얻었다.
그는 비싼 자동차를 얻었다.

라. He is famous for his wit.
그는 그의 재능으로서 유명하다.
그는 그의 재능으로 유명하다.

2.2.3 기능어–용언 거리 정보. 임의의 문장에 대하여 구문 분석을 해 줄 수 있는 견고한 구문 분석기가 없기 때문에 말뭉치에서 기능어–용언 공기 정보를 추출할 때 구문 분석 및 구조적 모호성 해소는 하지 않는 것으로 가정하였다. 대신 일정한 크기의 창(window)을 가정하고 이 창 안에 들어오는 어절 사이의 공기 정보를 추출하였다. 이 창을 통해 말뭉치를 관찰했을 때 다음 어절이 창 안에 들어 왔다고 하자.

(13) 숲속에서 지구로 떨어지는 별똥을 발견하였다.

구문 분석과 구조적 모호성 해소를 정확하게 할 수 있다면 '-에서 발견하다', '-로 떨어지다', '-을 발견하다'와 같은 기능어–용언 공기 정보가 추출되겠지만 구조적 모호성 해소를 하지 않기 때문에 추가로 '-에서 떨어지다', '-로 발견하다' 등과 같은 기능어–용언 공기 정보도 추출된다. 따라서 구문 분석과 구조적 모호성 해소를 하지 않은 상태에서 추출된 기능어–용언 공기 정보에는 잡음(noise)이 포함될 수밖에 없다. 기능어와 용언 사이의 거리가 멀면 멀수록 잡음 발생 가능성도 높아진다. 따라서 잡음으로 인한 오류를 줄이려면 기능어와 용언 사이의 거리도 고려해 주어야 한다. 그러나 기능어와 용언 사이의 절대 거리가 멀다고 해서 무조건 잡음 가능성이 높은 것으로 단정을 해서는 안 된다.

한국어가 어순이 자유로운 언어에 속하기는 하지만 그래도 14의 (가)보다는 (나)가 보편적으로 사용되는 경향이 있다.

(14) 가. 별똥을 숲속에서 발견하였다.

나. 숲속에서 별똥을 발견하였다.

5,000만 어절의 말뭉치를 조사한 바에 따르면 조사 '-에서'와 용언 '발견하다' 사이의 평균 거리는 2.71인 반면 조사 '-을'과 용언 '발견하다' 사이의 평균 거리는 1.58이었다.[3] 이와 같이 기능어에 따라서 용언과의 평균 거리가 달라지는 경향이 있는데 이 점을 고려하여 기능어와 용언 사이의 거리 정보를 이용한 점수 $s_d(\overline{p_{ij}}, \overline{v})$를 다음과 같이 정의할 수 있다.

$$s(\overline{p_i}) \quad = \quad s_d(\overline{p_{ij}}, \overline{v}) = \frac{1}{|V|} \sum_{\overline{a} \in V} D(\overline{p_{ij}}, \overline{a}) - D(\overline{p_{ij}}, \overline{v}) \tag{4}$$

여기서 V는 용언 집합을, $D(\overline{a}, \overline{b})$는 $\overline{a}$와 $\overline{b}$ 사이의 거리를 각각 나타낸다. 말뭉치를 조사했을 때 '-을'과 '읽다' 사이의 평균 거리 D(을,읽다)는 1.78이고 '-을'과 '가라앉다' 사이의 평균 거리 D(을,가라앉다)는 3.55였다. 같은 말뭉치에서 조사 '-을'과 모든 용언 사이의 평균 거리는 2.17이다. 따라서 식 (4)에 따라 거리 정보를 이용한 점수를 계산하면 다음과 같다.

$$\begin{aligned} s_d(\text{을,읽다}) &= 2.17 - 1.78 = 0.39 \\ s_d(\text{을,가라앉다}) &= 2.17 - 3.55 = -1.38 \end{aligned}$$

[3] 인접한 두 어절 사이의 거리를 1로 보았다.

이 점수는 '-을 읽다'는 비교적 신뢰도가 높은 공기 정보인 반면 '-을 가라앉다'는 신뢰도가 낮은 공기 정보임을 나타내는 것으로 해석할 수 있다.

다음은 기능어와 용언 사이의 공기 정보만을 이용했을 때에는 전치사의 번역어가 잘못 선택되었으나 체언과 조사 사이의 공기 정보를 함께 사용했을 때에는 올바르게 번역된 예를 보인 것이다. 이 예에서 첫 번째 번역문은 기능어와 용언 사이의 공기 정보만을 이용했을 때 얻어진 결과이며, 두 번째 번역문은 거리 정보를 함께 사용했을 때 얻어진 결과이다.

(15) 가. They caught a bird with birdlime.
그들은 덫에 새를 잡았다. $[s_d(\text{에,잡다}) = -1.00]$
그들은 덫으로 새를 잡았다. $[s_d(\text{으로,잡다}) = 0.54]$

나. He foamed at the mouth.
그는 입을 거품이 일었다. $[s_d(\text{을,일다}) = -1.07]$
그는 입에서 거품이 일었다. $[s_d(\text{에서,일다}) = -0.21]$

3. 실험 및 평가

본 논문에서 제안한 전치사 번역어 결정 모델의 평가를 위하여 1991-1995년 사이의 일간지(4,090만 어절), 한국 과학 기술원에서 개발한 대한민국 국어 정보 베이스 평가판 버전 0.1(660만 어절), 21세기 세종계획을 통해 1999년에 구축된 한국어 말뭉치 버전 5.0(140만 어절), 계몽 백과 사전(110만 어절)을 통합한 총 5,000만 어절의 미가공 말뭉치(raw corpus)로부터 공기 및 거리 정보를 추출하였다. 말뭉치로부터 공기 및 거리 정보를 추출하는 방법은 매우 간단한데, 본 논문에서 제안된 모델에 따라 점수를 계산하기 위해서는 조사, 어미와 같은 기능어와 체언, 용언과 같은 내용어의 출현 빈도수 및 기능어-용언 관계와 체언-조사 관계의 어휘 쌍에 대한 공기 빈도수를 알아야 한다. 따라서 말뭉치에 포함된 문장을 형태소 분석하여 각 어절을 형태소 단위로 분리한 다음 각 형태소의 출현 빈도수를 알아내고, 또 인접한 두 어절이 기능어-용언 관계에 있거나 한 어절이 체언-조사 관계에 있는 경우 이들의 공기 빈도수도 헤아린다. 기능어와 용언 사이의 거리 정보도 유사한 방법으로 구할 수 있는데, 문장을 형태소 분석하여 용언이 발견되는 지점에서 5 어절 이내에 있는 어절에 기능어가 포함된 경우 이 기능어와 용언 사이의 (몇 어절 떨어졌는가를 기준으로 측정한) 거리를 계산한다. 말뭉치에 대한 형태소 분석은 MACH(Shim and Yang, 2002; 심광섭 외, 2004)를 사용하여 하였다.

전치사 번역의 정확도 측정은 (Lee and Kim, 2002)이 프라임 한영 사전 예문에서 임의 추출하여 만든 945개의 평가용 문장 가운데 동사 또는 형용사를 수식하는 전치사가 포함된 443개의 문장을 가지고 실시하였다. 정확도의 공정한 비교를 위하여 443개의 테스트 문장에 대하여 미리 번역문을 작성해 두었다. 이 번역문에서 전치사 부분은 2.1절의 5에서 본 것과 같은 형태의 전치사 번역 사전에 수록되어 있는 어휘만을 사용하여 작성된 것이다. 참고로 이 전치사 번역 사전의 각 표제어는 평균 15.2개

의 번역어를 가지고 있다. 번역의 정확도는 제안된 모델에 따라 생성된 번역 결과와 미리 준비된 번역문과의 일치 여부에 따라 자동으로 산출되도록 하였다.

본 논문에서 제안한 전치사 번역 모델의 평가에 앞서 전치사 번역 문제의 난이도를 알아보기 위하여 기준 정확도(baseline accuracy)를 측정하였다. (Lee and Kim, 2002)에서는 기준 정확도를 (1) 번역어를 무작위로 선택하는 방법, (2) 사전에서 제일 앞에 나오는 번역어를 무조건 선택하는 방법, (3) 각 단어별로 가장 자주 사용되는 번역어를 선택하는 방법으로 측정하였는데, (1)은 측정할 때마다 정확도가 달리 나오므로 기준 정확도를 측정하기가 곤란하며 (2)는 그럴 듯 해 보이기는 하지만 정확도가 지나치게 낮게 나오는 경향이 있기 때문에 기준 정확도로 삼기에는 부적절하다.[4] 이 때문에 본 논문에서는 (3)과 유사한 방법으로 기준 정확도를 측정하기로 하고 443개의 평가용 문장을 대상으로 각 전치사별로 '가장 자주 사용되는 번역어'를 선정하였는데, 예를 들어 평가용 문장에서 at가 '-에서'로 번역된 경우가 가장 많았다면 at는 모두 '-에서'로 번역하는 것으로 하였다. 이렇게 하여 얻은 기준 정확도는 50.3%였는데, 만약 평가용 문장을 대상으로 하지 않고 일반적으로 '가장 자주 사용되는 번역어'를 선정하여 (3)의 방법을 적용한다면 이보다 낮은 기준 정확도를 얻을 것이다.

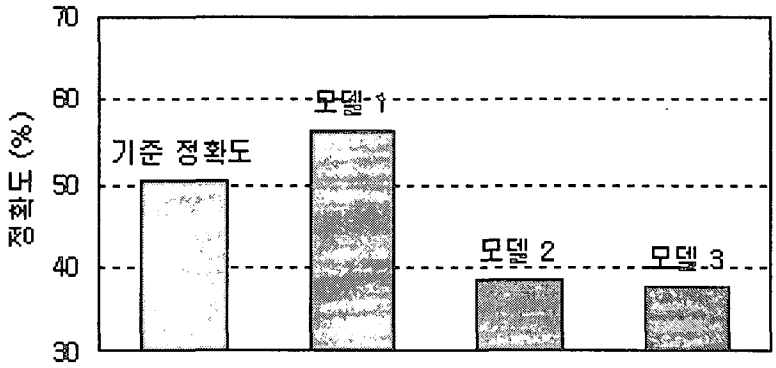
다음은 본 논문에서 제안한 3가지 기본 번역 모델에 대한 개별적인 정확도 평가를 실시해 보았다. 기능어와 용언 사이의 공기 정보를 이용하는 첫 번째 번역 모델(모델 1)의 경우 전체 443 문장 가운데 250 문장이 올바르게 번역되어 번역의 정확도는 56.4%로 나타났다. 체언과 조사 사이의 공기 정보를 이용하는 두 번째 번역 모델(모델 2)의 경우에는 전체 443 문장 중 171 문장이 올바르게 번역되어 번역의 정확도는 38.6%에 지나지 않았다. 마지막으로 기능어와 용언 사이의 거리 정보를 이용하는 세 번째 번역 모델(모델 3)의 경우에는 166 문장만이 올바르게 번역되어 번역의 정확도는 37.5%였다. 각 번역 모델의 정확도를 기준 정확도와 비교하면 그림 1에서 보듯이 모델 1만 기준 정확도보다 약간 우수하고 나머지 두 모델의 정확도는 기준 정확도보다도 훨씬 낮다.

이제 세 가지 기본 번역 모델 중 두 가지를 선택하여 이들을 가중치를 주고 결합했을 때 번역의 정확도가 어떻게 되는지 평가해 보았다. 다음은 서로 다른 두 가지 기본 번역 모델을 결합하는 방법을 나타내는 식이다. 여기서 $s_1(\overline{p_i})$와 $s_2(\overline{p_i})$는 2 장에서 제안한 기본 번역 모델을 나타내며, w_1과 w_2는 가중치를 나타낸다.

$$s(\overline{p_i}) = w_1 \cdot s_1(\overline{p_i}) + w_2 \cdot s_2(\overline{p_i}) \quad (5)$$

먼저 모델 2와 모델 3을 동시에 사용하는 경우에 대한 번역의 정확도를 측정해 보았다. 이 평가에서는 모델 2와 모델 3에 대하여 각각 0.5의 가중치를 부여하였다. 평가 결과 전체 443 문장 중 253 문장이 올바르게 번역되어 정확도는 57.1%로 나타났다. 따라서 체언-조사 공기 정보와 기능어-용언 거리 정보를 복합적으로 사용하는 경우 이들을 따로 사용할 때에 비하여 약 20% 가량 정확도가 개선되었음을 알 수 있다.

[4] 참고로 Lee and Kim (2002)에서 (1), (2), (3)의 방법에 따라 측정된 기준 정확도는 각각 6.77%, 11.86%, 38.68%였다.



[그림 1] 기본 번역 모델의 정확도 비교

이러한 결과를 얻게 된 것은 체언-조사 공기 정보와 기능어-용언 거리 정보가 상호 보완 관계에 있기 때문인 것으로 해석된다.

다음은 모델 1과 모델 3을 동시에 사용했을 때의 정확도 평가를 실시하였다. 이 평가에서는 모델 1과 모델 3에 대하여 각각 0.7과 0.3의 가중치를 부여하였다. 평가 결과 241 문장이 올바르게 번역되어 정확도는 54.4%로 나타났다. 이것은 모델 1을 단독으로 사용했을 때에 비하여 오히려 낮은 수치이다. 모델 3이 모델 2와 같이 사용되었을 때에는 정확도 향상에 많은 기여를 하였지만 모델 1과 함께 사용되었을 때에는 정확도 향상에 효과가 없는 것은 기능어와 용언 사이의 거리 정보는 기능어와 용언 사이의 공기 정보에 내포되기 때문인 것으로 해석된다.

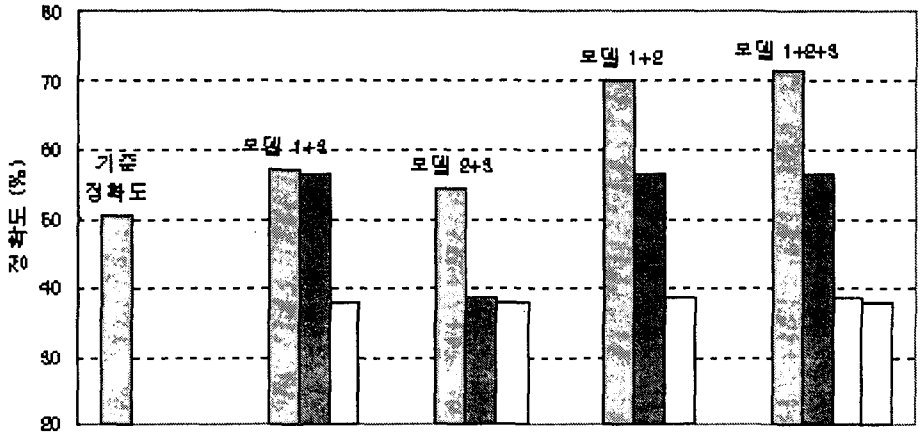
마지막으로 모델 1과 모델 2를 함께 사용하는 경우에 대한 정확도 평가를 실시하였다. 이번 평가에서는 모델 1과 모델 2에 대하여 각각 0.7과 0.3의 가중치를 부여하였다. 평가 결과 총 311 문장이 올바르게 번역되어 정확도는 70.2%로 나타났다. 이는 각 기본 번역 모델을 개별적으로 사용할 때에 비하여 약 14-32% 가량 정확도가 개선된 것이다. 이와 같이 두 모델을 복합적으로 사용했을 때 정확도가 많이 향상된 것은 두 모델에서 사용되는 체언-조사 공기 정보와 기능어-용언 공기 정보가 상호 보완 관계에 있기 때문인 것으로 해석된다.

마지막으로 세 가지 기본 번역 모델을 종합적으로 사용한 경우에 대한 정확도를 평가해 보았다. 세 가지 기본 번역 모델은 다음과 같은 방법으로 결합하였다.

$$s(\overline{p_i}) = w_1 \cdot s_v(\overline{p_{ij}}, \overline{v}) + w_2 \cdot s_d(\overline{p_i}, \overline{v}) + w_3 \cdot s_n(\overline{n}, \overline{p_{i0}}) \quad (6)$$

여기서 w_1, w_2, w_3는 각 모델에 대한 가중치를 나타내는데, 각각에 대하여 0.7, 0.15, 0.15의 값을 부여하고 실험을 한 결과 전체 443 문장 가운데 316 문장이 올바르게 번역되어 정확도가 71.3%로 나타났다. 따라서 세 가지 기본 번역 모델을 동시에 사용하는 경우의 정확도는 기준 정확도에 비하여 최고 21% 가량 성능이 향상된 것을 알 수 있다. 그림 2는 기본 번역 모델을 복합적으로 사용했을 때의 정확도를 기본 번역 모델만 사용했을 때의 정확도와 비교한 그래프이다.

Lee 등은 영영한 사전에서 단어 의미를 설명하는 문장과 예문에 포함된 내용어와 번역할 문장에 나타난 내용어 사이의 일치 정도에 따라 단어 의미(sense)를 결정하고, 이 의미로 사용될 수 있는 여러 가지 가능한 번역어 중에서 하나를 선택하기 위하여 Dagan 등의 방법을 사용하는 방안을 제시하였다 (Lee and Kim, 2002). 영영한 사전에서 단어 의미를 설명하는데 사용된 문장이나 예문은 양적인 측면에서 보았을 때 말뭉치에 비하여 제한적이기 때문에 Lee 등은 WordNet를 이용하여 사전의 문장과 주어진 문장의 각 내용어에 대한 유사도를 측정하였다. Lee 등은 이 방법으로 전치사를 제외한 영어의 명사, 동사, 형용사, 부사를 한국어로 번역하는 실험을 하였는데, 각각 55.23%, 42.22%, 42.86%, 53.45% 정도의 정확도를 달성하였다.



[그림 2] 복합 번역 모델의 정확도 비교

4. 결론 및 향후 과제

본 논문에서는 영한 기계 번역에서 영어 전치사를 한국어로 번역하는 모델을 제안하였다. 제안한 모델에서는 전치사 번역어만 나열된 단순한 형태의 번역 사전을 이용하며, 여러 가지 가능한 번역어 중 최적 번역어를 선정하기 위하여 한국어 미가공 말뭉치에서 추출한 체언-조사 공기 정보, 기능어-용언 공기 정보, 기능어-용언 거리 정보를 이용한다. 본 논문에서 제안한 모델의 타당성 검증을 위하여 (Lee and Kim, 2002)가 프라임 한영 사전의 예문에서 임의 추출하여 만든 945개의 평가용 문장 가운데 동사 또는 형용사를 수식하는 전치사가 포함된 443개의 문장에 대하여 번역 정확도를 평가하였다. 체언-조사 공기 정보, 기능어-용언 공기 정보, 기능어-용언 거리 정보를 개별적으로 사용하여 번역어를 선정했을 때에는 번역 정확도가 그다지 높지 않았으나, 세 가지 정보를 종합적으로 고려하여 번역어를 선정했을 때에는 번역 정확도가 71.3%로 나타났다. 이는 각 전치사에 대하여 가장 대표적인 번역어로 번역했을 때 얻은 기준 정확도(baseline accuracy)에 비하여 약 21% 가량 개선된 수치이다.

실제 문장에서는 한 문장에 두 개 이상의 전치사구가 나오는 경우도 흔한데, 본 연구에서는 한 문장에 하나의 전치사구가 나오는 경우로 제한하였다. 따라서 향후 연구에서는 전치사구의 개수에 제한이 없는 일반적인 문장에 적용할 수 있도록 본 연구의 모델을 확장할 필요가 있다. 또 전치사 번역 사전은 번역어가 '조사', '조사 체언+조사', '조사 용언+어미' 형태로만 주어지는 것으로 제한하였는데 이러한 제한이 없는 일반적인 형태의 전치사 번역 사전을 사용할 수 있도록 개선하는 것도 필요하다.

이러한 제약에도 불구하고 수작업으로 만든 복잡한 형태의 번역 사전을 이용하지 않고 자동화된 방법으로 미가공 말뭉치로부터 추출한 통계 정보만을 이용하여 전치사를 번역할 수 있다는 점에서 본 연구의 가치가 있다고 보며, 이 연구를 더욱 발전시킨다면 향후 기계 번역에 실질적인 도움을 줄 수 있을 것으로 기대한다.

<참고문헌>

심광섭, 김영택. 1992. 변환 사전 기술 언어. *한국정보과학회 논문지* 19.1, 1-11.

심광섭, 양재형. 2004. 인접조건 검사에 의한 초고속 한국어 형태소 분석. *한국정보과학회 논문지 (B)* 31.1, 89-99.

Brown, Peter F., et al. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* 16.2, 79–85.

Brown, Peter F., et al. 1993. The Mathematics of Statistical Machine Translation : Parameter Estimation. *Computational Linguistics* 19.2, 263–311.

Church, Kenneth W. and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16.1, 22–29.

Collins, Michael and James Brooks. 1995. Prepositional Phrase Attachment through a Backed-Off Model. *Proceedings of the 3rd Workshop on Very Large Corpora*, 27–38.

Copestake, Ann, et al. 1994. Acquisition of Lexical Translation Relations from MRDs. *Machine Translation* 9.3-4, 183–219.

Dagan, Ido, Alon Itai and Ulrike Schwall. 1991. Two languages are more informative than one. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 130–137.

Dagan, Ido and Alon Itai. 1994. Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics* 20.4, 563–596.

Koehn, Philipp and Kevin Knight. 2002. Learning a Translation Lexicon from Monolingual Corpora. *Proceedings of the Workshop of ACL SIGLEX*, 9–16.

Lee, Hyun Ah and Gil Chang Kim. 2002. Translation Selection through Source Word Sense Disambiguation and Target Word Selection. *Proceedings of the 19th International Conference on Computational Linguistics*, 530–536.

Pantel, Patrick and Dekang Lin. 2000. An Unsupervised Approach to Prepositional Phrase Attachment using Contextually Similar Words. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 101–108.

Shim, Kwangseob and Jaehyung Yang. 2002. MACH: A Supersonic Korean Morphological Analyzer. *Proceedings of the 19th International Conference on Computational Linguistics*, 939–945.

Volk, Martin. 2002. Combining Unsupervised and Supervised Methods for PP Attachment Disambiguation. *Proceedings of the 19th International Conference on Computational Linguistics*, 1065–1071.

Yamada, Kenji and Kevin Knight. 2001. A Syntax-based Statistical Translation Model. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 523–530.

접수 일자: 2004년 4월 19일
게재 결정: 2004년 6월 12일