DOI: 10.3745/KIPSTB.2010.17B.6.459

효율적 영한기계번역을 위한 확률적 품사결정

김 성 동[†] · 김 일 민[††]

요 약

자연언어처리는 여러 가지 모호성 문제를 가지는데, 특히 영한기계번역은 번역 과정의 각 단계마다 해결해야 할 모호성 문제를 가진다. 본 논문에서는 실용적인 영한기계번역 시스템의 개발을 목적으로 영어 분석의 효율성을 높이기 위해 영어 단어의 품사 모호성 해소 문제에 초점을 두었다. 기계번역의 효율성 제고를 위해 영한기계번역 시스템에 통합하기 위한 품사결정 모듈은 빠른 시간에 정확한 품사결정을 하면서도 오류를 최소화 하여야 한다. 본 논문에서는 확률적 품사결정 방법을 제안하고 3가지 품사결정 확률 모델을 제시하였다. Penn Treebank 말뭉치로부터의 통계 정보를 이용하여 확률 모델을 구축하였으며 실험을 통해 제안한 품사결정 방법의 정확성과 품사결정에 의한 기계번역 시스템의 효율 향상 정도를 제시하였다.

키워드 : 기계번역, 품사 모호성 해소, 확률 모델

Probabilistic Part-Of-Speech Determination for Efficient English-Korean Machine Translation

Sung-Dong Kim[†] · Ilmin Kim[††]

ABSTRACT

Natural language processing has several ambiguity problems, and English-Korean machine translation especially includes those problems to be solved in each translation step. This paper focuses on resolving part-of-speech ambiguity of English words in order to improve the efficiency of English analysis, which is in part of efforts for developing practical English-Korean machine translation system. In order to improve the efficiency of the English analysis, the part-of-speech determination must be fast and accurate for being integrated with machine translation system. This paper proposes the probabilistic models for part-of-speech determination. We use Penn Treebank corpus in building the probabilistic models. In experiment, we present the performance of the part-of-speech determination models and the efficiency improvement of the machine translation system by the proposed part-of-speech determination method.

Keywords : Machine Translation, Part-of-Speech Ambiguity Resolution, Probabilistic Models

1. 서 론

자연언어처리 분야 중의 하나인 기계번역은 매우 오랜 역사를 가지고 있으며 다양한 방법론이 연구되어 온 분야이다. 본 논문에서는 규칙 기반(rule-based)의 영한기계번역 시스템을 대상으로 실용적인 번역 시스템을 개발하기 위하여 효율적인 영어 분석을 위한 방법에 대해서 연구하였다. 규칙 기반의 영한기계번역은 어휘 분석(lexical analysis)과 구문 분석(syntactic analysis, parsing)으로 구성되는 영어 원문 분석과 변환(transfer) 및 생성(generation) 과정을 통해 영어 문장을 한국어 문장으로 번역한다. 서로 다른 문화권에 속하는 영어와 한국어는 서로 많은 차이가 있어 두 언어 간의 기계번역은 매우 어려운 문제이다. 이는 기계번역의 모든 과정에 수반된 모호성 문제에 기인한다. 어휘 분석에서는 품사 모호성(part-of-speech ambiguity), 구문 분석에서는 구조적 모호성(structural ambiguity), 변환에서는 의미 모호성(semantic ambiguity) 등의 문제가 적절하게 해결되어야 정확하고 올바른 번역 결과를 얻을 수 있다. 그 중 품사 모호성은 한 단어가 여러 가지 품사를 가질 수 있는 것을 의미하여 이로 인해 구문 분석과정에서 다양한 분석 결과를 생성할 수 있으며 이는 곧 구문 분석의 복잡도를 증가시켜 정확한 분석 결과를 얻는 것을 매우 어렵게 한다. 따라서, 본 논문에서는 품사 모호성 해소를 위한 방법을 제안하고 이를 기존의 영한기계번역 시스템과 통합하여 기계번역의 효율성 향상을 얻고자 하였다.

※ 본 연구는 2009년도 한성대학교 교내연구비 지원과제임.
† 정 회 원 : 한성대학교 컴퓨터공학과 부교수
†† 종신회원 : 한성대학교 컴퓨터공학과 교수
논문접수 : 2010년 9월 10일
수 정 일 : 1차 2010년 11월 4일
심사완료 : 2010년 11월 22일

자연언어처리 분야에서 품사태깅(part-of-speech tagging) 방법은 여러 가지 응용의 전처리(preprocessing) 과정으로 활용될 수 있어 많은 연구가 이루어진 분야이다. 품사태깅이란 문장을 구성하는 단어의 연속(word sequence)에 대해 가장 적절한 어휘 부류(lexical category)의 연속을 결정하는 문제로서 주로 Penn 태그 집합(tag set)을 이용하여 연구가 진행되었다. 본 논문에서 제시하는 품사결정(part-of-speech determination)은 목적 태그 집합이 Penn 태그 집합이 아닌 영한기계번역에서 활용되는 품사를 대상하는 하는 문제를 지칭한다. 또한 영한기계번역 시스템에 통합되기 때문에 어휘 분석 과정 이후에 분석 결과를 이용하여 품사를 결정하는 것을 의미한다. 즉 기존의 품사태거는 입력 문장에 대해서 품사태깅을 수행하지만, 본 논문에서 제안하는 품사결정 모듈은 어휘 분석 이후에 위치하여 품사모호성 문제를 해결하는 방안으로 연구되었다. 단어가 가질 수 있는 품사 중 하나를 구문 분석 이전에 결정한다면 품사결정의 오류로 인해 그 이후의 번역 과정은 무의미한 것이 될 수 있다. 따라서 품사결정 모듈은 어휘 분석에서 제공하는 영어 단어가 가지는 품사들 중 불필요한 품사를 제거하는 역할을 한다. 일반적으로 규칙 기반의 기계번역에서 구문 분석의 복잡도는 문장의 단어 수가 n개일 때 $O(n^3)$의 복잡도를 가진다. 이 때, 각 단어의 가능한 모든 품사에 대해서 모든 가능한 분석을 수행하기 때문에 각 단어의 품사 수를 줄이는 것은 구문 분석의 효율성 향상에 기여할 것이다.

영한기계번역 시스템과 통합될 품사결정 모듈은 빠른 시간에 정확하게 품사를 결정해야 하며 기계번역 시스템에 대한 추가의 부담이 적어야 한다. 이를 위해, 본 논문에서는 여러 요인을 고려하는 복잡한 모델보다는 소수의 요인만을 고려하는 간단한 확률 모델들을 제시하고 가장 성능이 좋은 것을 택하여 영한기계번역 시스템에 적용하고자 하였다. 확률적 품사결정 모델을 구축하기 위해 Penn Treebank의 Wall Street Journal 분야의 품사가 태그된 말뭉치를 이용하였다. 확률적인 방법에 의한 품사결정은 확률 모델이 미리 구축되어 기계번역 시스템과 통합되기 때문에, 실행시간 부담이 거의 없다. 결과적으로 영어 구문 분석은 품사결정 모델을 적용하지 않았을 때 보다 적은 개수의 품사를 가지고 분석을 수행하기 때문에 시간/공간적 효율성 향상을 기대할 수 있다. 제안한 품사결정 모델의 성능을 평가하기 위해 Penn Treebank의 Wall Street Journal 분야, Brown 말뭉치, IBM 매뉴얼 등 3가지 영역에서 테스트 데이터를 생성하여 테스트 하고, 가장 성능이 좋은 품사결정 모델을 기존의 영한기계번역 시스템과 통합하여 품사결정 모델의 적용에 의한 효율성 향상 및 번역 품질에의 영향을 평가하였다.

본 논문은 다음과 같이 구성된다. 2장에서는 품사태깅에 관한 기존의 방법을 살펴본다. 3장에서는 다양한 확률적 품사결정 모델 구축 방법에 대해서 설명하고 4장에서는 품사결정 모델의 성능과 영한기계번역 시스템에 대한 기여도를 구문 분석의 효율성 향상과 번역 품질 향상의 측면에서 제시한다. 5장에서 본 논문을 마무리하며 앞으로의 과제를 제시한다.

2. 관련 연구

자연언어처리 분야에서 품사태깅은 매우 유용하게 적용될 수 있는 전처리 과정으로서 다양한 연구가 진행되었다. 품사태거는 크게 규칙 기반 방식의 태거(rule-based taggers)와 확률적 태거(stochastic, probabilistic taggers)로 구분할 수 있다. 규칙 기반 태거는 기본적으로 전문가가 구축한 규칙에 의해 단어의 태그를 결정한다. 이는 매우 많은 전문가의 노력을 필요로 하기 때문에 그 적용률(coverage, recall) 면에서 성능이 떨어진다. 규칙 기반 태거의 예로는 TAGGIT[1], 유한 기계(finite-state machine)를 이용한 태거[2] 등이 있다. Brill의 연구 [3]에서는 태깅 규칙을 자동적으로 획득하는 규칙 기반의 태거를 소개하였는데 이는 규칙 기반 방식의 단점을 보완하면서도 확률적 태거에 근접하는 성능을 보였다.

자연언어처리에 말뭉치를 이용한 통계적인 방법이 널리 활용되면서 품사태깅을 위해서도 확률적인 방법이 적용되기 시작하였다. 확률적인 방법은 대체로 추가적인 분석없이 규칙 기반 태거보다 품사태깅의 정확성(accuracy)이 높다고 알려져 있다. 은닉 마코프 모델(hidden Markov model)이 적용된 태거들이 많이 연구되었는데 대표적으로 [4, 5]가 있으며 [6]에서는 이를 한국어 품사태깅에도 활용하였다. 은닉 마코프 모델을 적용한 태거는 기본적으로 품사 태그를 결정하는데 활용되는 어휘 및 문맥 정보를 표현하는데 마코프 모델을 활용하였으며, 품사가 태그된 말뭉치(tagged corpus)나 태그되지 않은 말뭉치(untagged corpus)에 대해서 forward-backward 알고리즘을[7] 이용하여 마코프 모델의 파라미터를 추정하는 방법으로 태거를 구현한다. 여기서는 최대 유사도 추정(maximum likelihood estimation) 방법으로 파라미터를 추정하였으며 최대 유사도 추정 방법은 확률 추정에 있어 간단하고 널리 사용되는 방법이다. 최대 엔트로피 모델(maximum entropy model)을 적용한 연구도 있다[8]. 여기서는 품사태깅을 위한 지역적 문맥 정보(local contextual information)를 이진 특성(binary-valued feature)으로 표현하여 이들 특성에 의한 품사의 확률값을 계산하는 확률 모델을 최대 엔트로피 원리에 기반하여 생성한다. 여기서 특성이란 품사태그의 결정에 영향을 미치는 문맥적 요인을 의미하는데, 각 요인이 품사태그 결정에 영향을 미치는 정도를 iterative scaling 알고리즘을[9, 10] 이용하여 계산함으로써 확률 모델을 구축하게 된다. 이 방법은 다양한 특성을 이진 값으로 표현하므로 추가적인 특성의 추가나 기존 특성의 제거 등을 쉽게 함으로써 확률 모델의 성능 개선이 용이하다는 장점을 가진다. Support vector machine(SVM)을 이용한 품사태거[11]가 연구되었는데 여기서는 문맥 윈도우(context window) 내에 존재하는 특정 단어와 태그에 대한 이진 특성을 사용하여 품사태그를 결정한다. 또한 전통적인 규칙 기반 방법인 결정트리(decision tree)에 확률적 방법을 혼용한 방식도 연구되었다[12].

이외에도 품사태깅에 대한 매우 많은 연구가 진행되어 왔

다. 그러나 대부분의 이들 연구들은 특정한 자연언어처리의 응용을 위해 특화되었다고 볼 수 없으며 품사태깅 자체의 성능을 향상시키는 것에 초점을 맞추어 다양한 방식을 시도한 결과라 할 수 있다. 본 논문에서는 영한기계번역에서 효율적인 영어 구문 분석을 위한 확률적 품사결정 모델을 제안한다. 확률적 품사결정 모델은 기존의 확률적 품사태깅 방법과 유사하나, 기계번역 시스템과의 통합을 고려하여 확률 모델을 구축하였다. 즉, 적은 수의 문맥적 특징(contextual features)을 고려한 확률 모델을 구축하였으며 일반적인 품사태거의 주요한 문제인 미지어(unknown) 단어에 대한 품사태깅 문제를 고려하지 않음으로써 기존의 태거에 비해 복잡도를 줄여 보다 간단하게 품사결정 모델을 구축할 수 있다. 미지어는 고유명사로 간주하여 기계번역 시스템의 고유명사 분석모듈에서 처리될 수 있도록 하였다. 일반적으로 영어 품사태거는 Penn 태그 집합을 목표 집합으로 하여 품사태깅을 하는 태거이며 따라서 영한기계번역에 적용하기 위해서는 태그의 변환이 필요하다. [13]에서는 영한기계번역에서 사용되는 품사와 품사태깅에서 사용하는 품사의 대응관계에 대해서 연구하였다. 그러나 본 논문에서는 기존의 품사태거에 [13]에서 제시된 대응 관계를 적용하여 품사를 결정하는 방법 대신에 [13]에서 제시된 대응 관계를 이용하여 말뭉치를 변환하고 이를 이용하여 품사결정 확률 모델을 생성하는 방법을 택하였다. Penn 태그 집합에는 36개의 태그가 존재하며 영한기계번역에서는 8개의 품사를 사용한다. 목표 품사태그의 개수가 적어지므로 보다 작은 크기의 학습 데이터를 사용하여도 신뢰할만한 통계 정보를 획득할 수 있다. [14]에서는 어휘 분석 이후에 어휘 분석 결과를 활용하여 통계적 방법과 기계학습 방법을 혼용하여 품사를 결정하는 방법을 제시하였다. 기계번역을 위한 품사태거로서 트라이그램(trigram)을 이용한 이차 은닉 마코프 모델(second-order hidden Markov model)에 기반한 태거가 제시되었다[15]. 이는 규칙 기반의 기계번역 시스템을 위한 품사태거에 대한 연구로서 의미가 있으나 은닉 마코프 모델의 인자를 추정(estimation)하기 위한 노력과 실행시간에 Viterbi 알고리즘을 적용해야 하는 부담이 있다.

3. 확률적 품사결정 모델 구축

본 절에서는 3가지 확률적 품사결정 모델을 설명한다. 기존의 품사태깅에서 사용되는 Penn 태그 집합 대신에 본 논문에서는 영한기계번역에서 사용되는 8개의 품사를 대상으로 품사결정을 하는 확률 모델을 구축하였다. 8개의 품사는 다음과 같다: NOUN, VERB, ADJ, ADV, DET, PREP, PRON, CONJ.

3.1 모델 I

품사결정 대상 단어의 품사결정에 전-후 단어의 품사가 중요한 영향을 준다는 가정하에 품사결정 문제를 대상 단어와 전-후 단어의 품사로 구성되는 품사 트라이그램 중 가장 적절한 트라이그램을 찾는 문제로 간주하였다. 특정 단어의 품사는 전-후 단어 자체가 아닌 전-후 단어의 품사에 영향을 받는다는 판단에 의해 위와 같은 가정을 하여 품사결정 모델을 설정하였다. 많은 확률적인 품사태거들은 품사태깅 대상 단어의 앞의 하나 또는 두 단어의 정보를 이용하였으나 단어의 품사는 앞 단어 뿐만이 아니라 다음 단어와도 연관성 있으므로 본 논문에서는 이를 고려하여 품사결정 모델을 구축하였다. 대상 단어의 앞-뒤 문맥을 고려하는 연구로는 [16, 17]등이 있다. 그리고 [18]에서는 베이지안 방식(Bayesian approach)이 간단한 최대 유사도 추정(maximum likelihood estimation) 보다 좋은 성능을 보인다는 것을 품사태거를 이용하여 제시하였다. 본 논문에서는 품사결정 문제를 전-후 단어의 품사를 고려하여 현재 단어(w)에 대해서 가장 적절한 품사 트라이그램(T)을 찾는 문제로 간주하여 베이지안 방법에 기반한 확률 모델을 식 (1)과 같이 제시하였다.

$$\arg\max_{T} \Pr(T \mid w) = \frac{\Pr(T)\Pr(w \mid T)}{\Pr(w)} \tag{1}$$

즉, 품사결정이란 단어가 주어졌을 때 그 단어를 중심으로 하는 가장 적절한 품사 트라이그램을 찾는 문제로 간주하였으며 트라이그램 중 두번째 품사를 해당 단어의 품사로 결정하게 된다. 식 (1)에서 분모의 Pr(w)는 모든 트라이그램 T에 대해서 동일하게 계산되기 때문에 argmax T를 찾을 때 계산할 필요가 없으므로 식 (2)와 같이 표현할 수 있다.

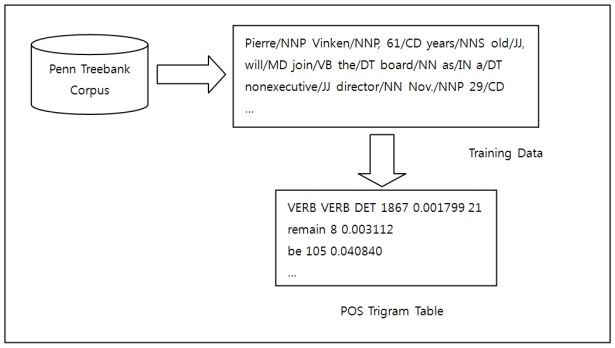
$$\arg\max_{T} \Pr(T \mid w) = \Pr(T)\Pr(w \mid T) \tag{2}$$

식 (2)에서 $Pr(T)$는 품사 트라이그램의 확률이며, $Pr(w|T)$는 품사 트라이그램에 대해서 특정한 단어가 나타날 확률을 의미한다. 품사 트라이그램 T의 문맥에서 여러 단어가 중간에 올 수 있는데, 그 중 단어 w가 중간 단어로 나타날 확률을 의미한다. 따라서 첫번째 확률은 적절한 트라이그램을 고려하는 것이고 두번째 확률은 현재 단어에 대해서 적절한 트라이그램을 고려하는 것으로 이를 결합한 품사결정 모델이라 할 수 있다. 이 두가지 확률 값을 식 (3), (4)를 적용하여 학습 데이터로부터 최대 유사도 추정 방법으로 계산하여 품사결정 모델을 구축하게 된다.

$$\Pr(T_i) = \frac{|T_i|}{\sum_{k=1}^{n} |T_k|} \tag{3}$$

$$\Pr(w_j \mid T_i) = \frac{|w_j \in T_i|}{\sum_{w_k \in T_i}^{m} |w_k|} \tag{4}$$

식 (3), (4)에서 n은 학습 데이터에 나타난 품사 트라이그



(그림 1) 모델 I의 확률을 계산하기 위한 데이터 생성 과정

램의 개수이며 m은 특정 품사 트라이그램(T_i)이 나타났을 때 두번째 품사에 해당하는 단어(w_k)의 개수를 의미한다.

품사결정 모델을 구축하기 위해서 먼저 Penn Treebank의 품사 및 구문 태그된 말뭉치를 품사태그된 데이터로 변환해야 한다. 변환된 학습 데이터로부터 각 품사 트라이그램의 출현 빈도수와 두번째 품사에 대응하는 단어의 목록을 단어의 출현 빈도수와 함께 포함하는 품사 트라이그램 테이블(POS trigram table)을 생성한다. 품사결정 과정에서 품사 트라이그램 테이블은 위의 식 (3), (4)에 해당하는 확률 값을 제공하게 된다.

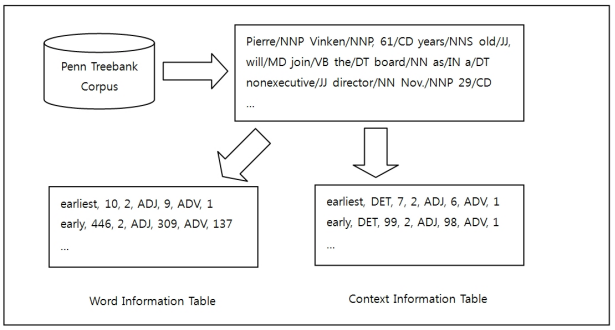
(그림 1)은 품사 트라이그램 테이블의 생성 과정을 예와 함께 제시한다. 구문 분석 된 Penn Treebank 말뭉치로부터 단어/품사태그들을 추출하여 학습데이터를 생성하는데, 이때 품사태그는 Penn 태그 집합에 속하는 것이다. 따라서 이들 품사태그를 기계번역에서 사용하는 8개의 품사로 변환하는 과정을 수행한다. 이러한 변환 과정을 통해 품사 트라이그램에 대한 통계정보를 계산하여 품사 트라이그램 테이블을 생성한다.

3.2 모델 II

앞 단어의 품사가 현재 단어 품사에 영향을 미칠 수 있다는 가정하에 복수의 후보 품사를 가지는 단어의 품사결정을 위해 그 단어의 품사 확률뿐만 아니라 앞 단어의 품사를 함께 고려한다. 모델 II는 이러한 품사결정 방식을 식 (5)와 같이 표현한다.

$$\arg\max_{P_i} \Pr(P_i \mid w_i) \times \Pr(P_i \mid P_{i-1}, w_i) \tag{5}$$

식 (5)에서 w_i는 품사결정 대상이 되는 단어이고 P_{i-1}는 앞 단어의 후보품사이며 P_i는 w_i의 후보 품사이다. 두 확률은 아래 식에 의해 학습데이터로부터 최대 유사도 추정(maximum likelihood estimation)에 의해 계산된다. 식 (6)은 단어 w_i의 품사가 P_i일 확률이며 이는 가장 자주 사용된 품사를 고려하기 위한 식이다. 식 (7)은 가장 적절한 품사의 연속(sequence)을 찾기 위한 식이다.



(그림 2) 모델 II의 확률을 계산하기 위한 데이터 생성 과정

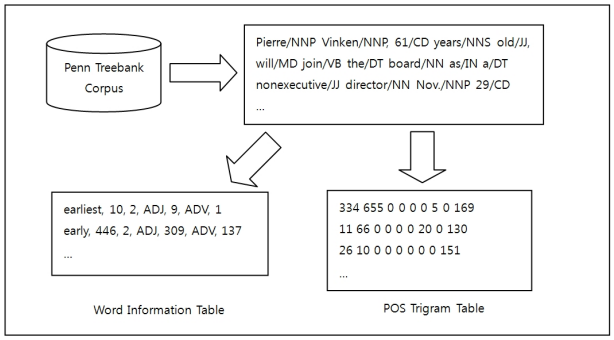
$$\Pr(P_i \mid w_i) = \frac{|w_i \cap P_i|}{|w_i|} \tag{6}$$

$$\Pr(P_i \mid P_{i-1}, w_i) = \frac{|P_i \cap w_i|}{|P_{i-1} \cap w_i|} \tag{7}$$

(그림 2)는 위의 확률 계산을 위한 데이터 생성과정을 보여준다. (그림 1)에서와 마찬가지로 Penn Treebank 말뭉치에서 품사태그를 기계번역에서 사용되는 품사로 변환한 후 단어 정보 테이블(word information table)과 문맥 정보 테이블(context information table)을 생성한다. 단어 정보 테이블은 단어가 말뭉치에서 특정 품사로 얼마만큼 사용되었는지에 대한 정보를 포함하여 식 (6)에 해당하는 확률 값을 제공한다. 그리고 문맥 정보 테이블은 앞 단어 품사와 현재 단어 품사의 연속의 적절성에 대한 확률 값을 제공한다. 단어 정보 테이블은 쉼표로 분리되는 5개 이상의 필드로 구성된다: 단어, 학습데이터에서의 빈도수, 후보 품사 개수, 그리고 품사와 품사의 빈도수 쌍의 연속. 예를 들어, (earliest, 10, 2, ADJ, 9, ADV, 1)은 'earliest'가 10번 나타났으며 2가지 품사로 사용되었고, 그 중 ADJ로 9번, ADV로 1번 사용되었다는 것을 의미한다. 문맥 정보 테이블은 연속된 두 개의 품사에 대한 정보를 포함하는데, 단어, 이전 단어의 품사, 빈도수, 단어의 가능한 품사의 개수, 그리고 품사와 품사 빈도수 쌍의 연속으로 구성된다. 예를 들어, (earliest, DET, 7, 2, ADJ, 6, ADV, 1)은 'earliest'는 앞 단어의 품사가 DET인 경우가 7번 있었는데 이 경우 2개의 품사로 사용되었으며 ADJ로 6번, ADV로 1번 사용되었음을 의미한다.

3.3 모델 III

품사결정 모델 III은 단어의 품사확률 이외에 앞, 현재, 다음 단어로 구성되는 문맥 정보를 활용한다. 이는 단어의 품사결정에 앞, 뒤 단어가 영향을 준다는 가정에 따른 것이다. 품사결정 확률 모델은 식 (8)과 같다. 식 (8)에서 w_i, P_{i-1}, P_i는 식 (5)에서의 의미와 같다. 그리고 w_{i-1}와 w_{i+1}은 각각 앞 단어, 다음 단어를 나타내고 P_{i-1}와 P_{i+1}은 각 단어들의 후보 품사를 의미한다. 식 (8)에서 두 번째부터 네 번째까지의 단어에 대한 품사 확률 값은 식 (6)으로 계산되며 품사 트라



(그림 3) 모델 III의 확률을 계산하기 위한 데이터 생성 과정

이그램의 확률은 식 (9)를 이용하여 얻을 수 있다.

$$\arg\max_{P_i} \Pr(P_i \mid P_{i-1}, P_{i+1}) \times \Pr(P_i \mid w_i) \times \Pr(P_{i-1} \mid w_{i-1}) \times \Pr(P_{i+1} \mid w_{i+1}) \quad (8)$$

$$\Pr(P_i \mid P_{i-1}, P_{i+1}) = \frac{|P_{i-1}P_iP_{i+1}|}{|P_{i-1}P_{i+1}|} \quad (9)$$

(그림 3)은 모델 III의 확률 계산을 위한 데이터 생성 과정을 보여준다. 단어 정보 테이블은 (그림 2)와 같으며 품사 트라이그램 테이블은 (그림 1)에서의 그것과는 다른 모습을 가진다. 이는 식 (9)의 계산을 쉽게 하기 위한 것이다. 즉 품사 트라이그램 테이블은 다음과 같은 구조의 3차원 배열로 구성된다: POS_Trigram_Table[P_{i-1}][P_i][P_{i+1}]. 배열의 값은 학습데이터에서의 빈도수를 나타낸다. 3차원 배열은 10개[1]의 2차원 배열로 나타낼 수 있는데 (그림 3)에서는 P_{i-1}이 NOUN일 경우에 대한 2차원 배열의 일부를 제시하였다. 2차원 배열에서 행은 현재 단어, 열은 다음 단어의 품사에 대응한다. 예를 들어, (그림 3)의 품사 트라이그램 테이블의 첫 행 (334 655 0 0 0 0 0 5 0 169)에서 334는 앞 단어의 품사가 NOUN이고 다음 단어가 NOUN인 경우 현재 단어가 NOUN인 빈도수가 334번이며 같은 경우에 현재 단어의 품사가 ADJ인 빈도수가 655번인 것을 의미한다. 이를 이용해 품사 트라이그램의 확률을 구하면 Pr(NOUN NOUN NOUN | NOUN NOUN) = 334/1163이고 Pr(NOUN ADJ NOUN | NOUN NOUN) = 655/1163이 된다. 3차원 배열로 품사 트라이그램을 표현함으로써 데이터를 간결하게 유지할 수 있다.

4. 실 험

본 절에서는 학습 및 테스트 데이터를 설명하고 3장에서 기술한 3가지 품사결정 확률 모델의 성능을 제시한다. 그리고 품사결정이 영한기계번역 성능 향상에 기여한 정도를 평가한다.

4.1 데이터

학습 및 테스트 데이터 생성을 위해 Penn Treebank 말뭉치에서 WSJ 영역의 구문 분석된 데이터를 이용하였다. 약 100만 단어로 구성된 49,268 문장으로 이루어진 데이터로부터 학습 데이터를 생성하였다. 학습 데이터를 이용하여 약 14K개의 단어로 구성된 단어 정보 테이블을 생성하였으며, 3장에서 언급된 문맥 정보 테이블과 품사 트라이그램 테이블을 생성하였다. 품사결정 모델들의 성능 평가를 위해, WSJ, Brown, IBM 영역으로부터 테스트 데이터를 추출하였다. <표 1>에서 테스트 데이터에 대한 통계를 제시한다.

〈표 1〉 테스트 데이터에 대한 통계

	단어 수	문장 수
Brown	67,086	3,310
IBM	60,890	4,324
WSJ	73,061	3,631

4.2 품사결정 모델의 성능평가

3장에서 제시한 3가지 품사결정 모델의 성능을 기존의 품사결정 방법과 비교하여 논문에서 제시한 방법의 유효성을 제시한다. 기존의 bigram 특성과 Viterbi 탐색을 적용한 HMM 방식의 품사결정 방법을 성능비교의 대상으로 삼았다. 이 방법은 $Pr(P_i|P_{i-1})*Pr(w|P_i)$의 값을 최대로 하는 P_i를 단어 w의 품사로 결정하는 방식이다. 품사결정 모델의 성능평가를 위해 품사결정의 정확성(accuracy)을 측정하여 <표 2>에 결과를 제시하였다. 성능평가 결과는 입력 문장의 모든 어휘를 대상으로 계산하였다.

<표 2>의 결과에 의하면 앞 단어와 현재 단어의 품사 연속을 고려한 모델 II가 모든 영역에서 가장 성능이 좋은 것으로 나타났다. 모델 I, 모델 III에서는 품사 트라이그램을 고려하였는데, 품사 바이그램(bigram)을 사용한 모델의 성능이 낫다는 것은 보다 적은 수의 특성을 고려한 모델이 학습 데이터에서 보다 신뢰할 만한 통계정보를 획득하였기 때문이라고 볼 수 있다. 본 논문은 품사결정의 정확성을 높이기 위한 추가 정보를 결합하는 간단한 확률 모델을 제시하였으며, 간단한 확률 모델은 영한기계번역 시스템에 통합하는데 추가적인 부담을 수반하지 않기에 실용적이라 할 수 있다.

〈표 2〉 품사결정 모델들의 성능평가 결과 (정확성을 %로 나타냄)

	HMM 방식	모델 I	모델 II	모델 III
Brown	96.26	94.58	97.43	96.44
IBM	94.87	94.94	96.10	94.50
WSJ	96.56	94.11	97.69	95.69
Average	95.94	94.52	97.12	95.58

1) 8개의 기계번역 품사 이외에 알파벳이나 숫자가 아닌 단어의 품사를 위한 PUNC, 존재하지 않는 단어(문장 첫 단어의 앞의 가상의 단어와 문장 마지막 단어 다음의 가상의 단어)의 품사를 표현하기 위한 NULL을 포함하여 10개의 품사에 대한 2차원 배열이 존재한다.

4.3 품사결정 모델의 영한기계번역 시스템에 대한 기여

품사결정 모델의 영한기계번역 시스템에 대한 기여를 측정하기 위해 제시한 3개의 모델 중 가장 성능이 좋았던 모델 II를 기계번역 시스템에 통합하였다. 확률 모델을 위해 기계번역 시스템에 추가된 데이터의 종류와 크기는 <표 3>과 같다.

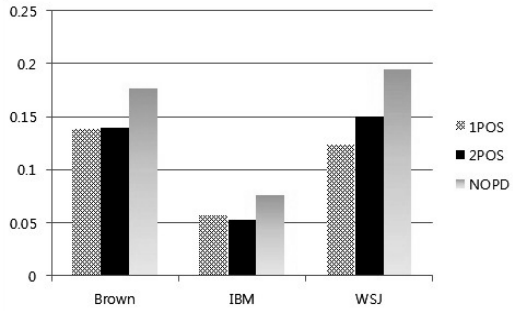
<표 3>에서 제시된 정보 중 단어 정보 테이블은 학습 데이터에서 3번 이상 나온 단어에 대한 품사 정보를 포함하고 있으며 문맥 정보 테이블은 10번 이상 나온 문맥 정보를 포함한 단어를 대상으로 문맥 정보를 보유하고 있다. 전체 데이터 크기는 약 1.1MB로서 영한기계번역 시스템이 보유한 약 40MB 정도의 데이터에 비해 매우 적은 부분에 해당하므로 기계번역 시스템에 추가의 부담은 거의 없다고 할 수 있다.

구문 분석에서 단어의 후보 품사 모두에 대해서 구문 구조를 생성하지 않고 결정된 품사에 대해서만 구문 구조를 생성하므로 우선적으로 구문 분석의 시간, 공간의 효율 향상을 기대할 수 있다. 그러나 품사결정 에러로 인해 구문 분석에서 올바른 구조 생성을 하지 못하는 경우도 가능하다. 품사결정 에러로 인한 구문 분석 실패를 줄이기 위해 하나의 품사가 아닌 두 개의 품사를 활용하는 방법도 적용하여 구문 분석의 효율성 및 번역품질 평가를 수행하였다. (그림 4, 5)에서는 품사결정 모델에 의한 시간, 공간적 효율 향성 정도를 제시하였다. 4.2절의 품사결정 모델 성능 평가를 위해 사용된 Brown, IBM, WSJ 영역에서 각각 100 문장씩을 추출하여 2.13GHz Core2 CPU와 1GB 메모리를 갖춘 컴퓨터를 이용하여 번역을 수행하고 결과를 측정하였다.

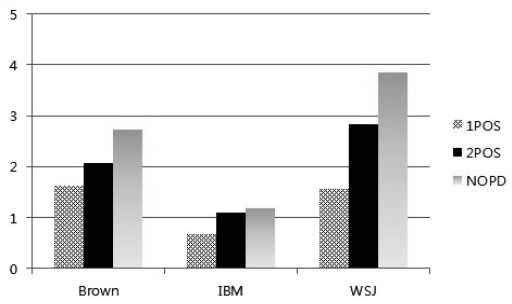
(그림 4, 5)에서 1POS, 2POS는 품사결정 모델에서 단어당 각각 1개, 2개의 품사를 결정하여 구문분석에서 사용하는 경우를 지칭하며 NOPD는 품사결정 모델을 적용하지 않은 경우를 의미한다. 그리고 그래프는 각 영역의 문장을 번역하는데 소요되는 평균 시간 및 메모리 사용량을 나타낸다. 평균적으로 1POS의 경우 약 28%의 속도 향상이 이루어졌으며 약 50% 정도의 메모리 사용량 절감 효과가 있었다. 2POS의 경우에는 약 24%의 속도 향상과 약 22%의 메모리 절감 효과를 나타냈다.

〈표 3〉 모델 II를 위해 기계번역 시스템에 추가된 데이터

	단어 정보 테이블	문맥 정보 테이블
단어 개수	8,453	8,443
데이터 크기	190KB	910KB



(그림 4) 품사결정 모델에 의한 시간 효율 향상 정도(sec)



(그림 5) 품사결정 모델에 의한 메모리 효율 향상 정도(MB)

<표 4>에서는 번역 품질을 평가하였다. 5명이 품사결정 모델이 적용된 번역과 적용되지 않은 번역 결과를 비교함으로써 번역 품질 평가를 수행하였다. 첫째 행은 하나의 품사만을 사용하였을 때와 품사결정 모델을 적용하지 않았을 때, 번역 품질이 보다 좋은 문장의 수를 대비하였고 둘째 행은 두 개의 품사를 사용하였을 때와의 비교 결과를 5명에 의한 평가 결과를 평균하여 제시하였다.

<표 4>에서 첫째 행 첫째 열의 [8 : 17]의 의미는 하나의 품사만을 이용하여 번역을 하였을 때의 번역이 더 좋았던 문장의 8개이고 품사결정 모델을 적용하지 않았을 때의 번역 결과가 더 좋았던 문장이 17개인 것을 의미한다. 즉 100 문장 중 75문장은 번역 품질이 같으며 25문장만 품질이 다르며 그 결과가 표에 수치로 표현되었다. <표 3>의 결과를 보면 하나의 품사를 이용했을 때는 약 78% 문장의 번역 품질은 같으며 22% 문장에 대해서 다른 결과를 보였으며 두 개의 품사를 이용한 경우에는 약 89% 문장에 대해서는 같은 번역 품질을 보였으며 11% 문장에 대해서만 다른 품질의 번역을 생성하였음을 알 수 있다.

일반적으로 품사결정 모델을 적용한 경우에 번역 품질이 더 좋지 않았는데, 이는 품사결정 오류로 인해 올바른 분석 결과를 생성하지 못하였기 때문이다. 품사결정 모델 적용시 더 좋은 번역이 생성된 경우는, 품사결정으로 인해 올바른 분석 결과의 선택이 가능했기 때문으로 판단할 수 있다. 많은 경우 번역 결과의 품질이 같았는데, 품사결정 모델의 적용이 결과에 영향을 미치지 않으면서도 시간/공간의 효율성을 향상시킬 수 있다는 품사결정 모델의 긍정적인 면이라 할 수 있다.

〈표 4〉 품사결정 모델 적용에 의한 번역 품질 평가 결과

	Brown	IBM	WSJ
1POS vs NOPD	8 : 17	0 : 25	5 : 11
2POS vs NOPD	5 : 11	0 : 15	1 : 2

5. 결 론

본 논문에서는 영한기계번역의 효율성 제고를 위해 어휘분석 결과에서 생성되는 단어들의 후보 품사들 중에서 단어의 품사를 결정하는 품사결정을 도입하였다. 이를 통해 구문분석에서 생성하는 구조의 수를 줄여 시간/공간적인 측면에서의 효율성 향상을 얻고자 하였다. 품사결정은 기존 영한기계번역 시스템에 통합되어야 하므로 빠르고 정확하게 품사를 결정해야 하며 추가적인 부담을 최소화 하여야 한다. 이를 위해 본 논문에서는 간단한 확률적인 품사결정 모델을 3가지 제시하고 성능을 평가하여 가장 좋은 성능을 보인 모델을 선택하여 영한기계번역 시스템에 통합하였다. 그리고 기계번역의 성능 향상을 측정하고 품사결정 모델에 의한 번역 품질에 대한 영향을 관찰하였다.

품사결정을 위해 기본적으로 단어의 품사 확률을 고려하고 단어의 좌-우 한 단어까지의 문맥적 특징을 고려하는 3가지 품사결정 확률 모델을 설정하고 Penn Treebank 말뭉치를 이용하여 확률 모델을 구축하였다. 하나의 복잡한 모델 대신 간단한 3개의 모델을 따로 구축하였는데, 이는 영한기계번역 시스템과 통합할 때 추가적인 부담을 최소화 하며 빠르게 동작해야 한다는 요건을 고려하였기 때문이다. 3가지 모델 중 품사결정 시 앞 단어를 추가적으로 고려한 모델 II의 성능이 가장 좋았으며 이를 영한기계번역 시스템에 통합하였다. 품사결정으로 인한 번역 품질 저하를 줄이기 위해 하나의 품사만을 선택하는 것 이외에 2개의 품사를 선택하는 경우에 대해서도 번역 성능을 측정하였다. 하나의 품사를 선택하는 품사결정 모델을 적용한 경우 시간 면에서 28%, 공간 면에서 약 50%의 효율이 개선되었으며 2개의 품사를 선택하는 경우에는 시간 면에서 24%, 공간면에서 22%의 개선이 있음을 확인하였다. 품사결정에 의한 번역 품질에 대한 영향은 품사결정을 하지 않은 경우에 비해서 품질이 좋지 않았으나, 하나의 품사를 선택할 때 약 80%, 두 개의 품사를 선택할 때 약 90% 정도는 영향을 받지 않음을 확인하였다. 결과적으로 품사결정에 의해 영한기계번역의 효율성의 향상은 많으나 번역 품질에서의 영향은 상대적으로 적음을 확인하였으며 제안한 품사결정 모델이 유효하다고 판단할 수 있다.

본 논문은 기존의 품사태거보다 높은 정확성을 보이는 모델을 제안하기 보다는 영한기계번역 시스템에 통합하기 위한 간단한 품사결정 모델을 제시하고 기계번역 시스템의 성능에 대한 영향을 제시하려는 것이 주요 목적이다. 기존의 품사태거에 비해 보다 적은 데이터를 이용하여 성능이 떨어지지 않은 품사결정 모델을 빠르게 생성하고 이를 기존의 기계번역 시스템에 간단하게 통합하는 일련의 과정을 정립하는 것은 기계번역 시스템의 성능향상을 위해 중요하다는 판단이며 본 논문에서는 이러한 과정을 정립했다는 데 의의가 있다.

본 논문에서 제안한 품사결정 모델은 영한기계번역 시스템에서 어휘분석 이후에 적용되므로 어휘분석 정보를 활용할 수 있다. 따라서 이를 이용하여 품사결정의 정확성을 높일 수 있을 것으로 판단된다. 또한, 품사결정에 의한 품질 저하를 최소화하는 방안에 대한 연구가 필요하다. 이를 위해 품사결정 오류 분석하는 것 뿐만 아니라 번역 품질에 영향을 미치는 품사결정 오류에 대한 분석이 필요하다. 품사결정의 궁극적인 목표는 효율성 향상 및 번역 품질 향상이다. 이를 위해서는 품사결정 오류를 복구 또는 오류에 대응하는 방법에 대한 연구가 수반되어야 할 것이다.

참 고 문 헌

[1] B. B. Greene and G. M. Rubin, “Automatic grammatical tagging of English,” Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, 1971.

[2] K. Koskenniemi, “Finite-state parsing and disambiguation,” *In Proceedings of the 13th International Conference on Computational Linguistics*, pp.229-232, Helsinki 1990.

[3] Eric Brill, “A Simple Rule-Based Part of Speech Tagger,” *Proceedings of the Applied Natural Language Processing*, pp.152-155, 1992.

[4] J. Jupiec, “Robust part-of-speech tagging using a hidden Markov model,” *Computer Speech and Language*, Vol.6, pp.225-242, 1992.

[5] B. Merialdo, “Tagging English text with a probabilistic model,” Computational Linguistics, Vol.20, No.2, pp.155-172, 1994.

[6] Jae-Hoon Kim and Jungyun Seo, “A Hidden Markov Model Imbedding Multiword Units for Part-Of-Speech Tagging,” *Journal of Electrical Engineering and Information Science*, Vol.2, No.6, pp.7-13, 1997.

[7] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic function of Markov chains,” Annals of Mathematical Statistics, Vol.41, No.1, pp.164-171, 1970.

[8] Adwait Ratnaparkhi, “A Maximum Entropy Model for Part-Of-Speech Tagging,” *Proceedings of the Empirical Method in Natural Language Processing*, pp. 133-142, 1996.

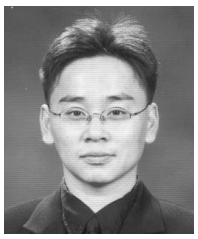
[9] Adam Berger, “The Improved Iterative Scaling Algorithm: A Gentle Introduction,” www.cs.cmu.edu/afs/~aberger/www/ps/scaling.ps

[10] Rong Jin, Rong Yan, Jian Zhang and Alex Hauptmann, “A Faster Iterative Scaling Algorithm for Conditional Exponential Model,” *Proceedings of the 20th International Conference on Machine Learning*, pp.282-289, 2003.

[11] Jesus Gimenez and Lluis Marquez, “SVM-Tool: A General POS tagger generator based on support vector machines,” *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004.

[12] Helmut Schmid, “Probabilistic Part-Of-Speech Tagging

Using Decision Trees," *Proceedings of International Conference on New Methods in Language Processing*, 1994.
[13] 이성욱, 이공주, 서정연, "영한기계번역 품사 집합과 펜트리뱅크 코퍼스 품사 집합간의 품사 대응", *한국정보과학회 1999년도 가을 학술발표논문집*, 제26권 제2호(II), pp.184-186, 1999.
[14] 김성동, 박성훈, "영한기계번역에서의 영어 품사결정 모델", 지능정보연구, 2009.
[15] Zaid Md Abdul Wahab Sheikh, Felips Sanchez-Martinez, "A Trigram Part-of-Speech Tagger for the Apertium Free/Open-Source Machine Translation Platform," *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pp.67-74, 2009.
[16] K. Toutanova, D. Klein, C. D. Manning, Y. Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," Proceedings of HLT-NAACL, pp.252-259, 2003.
[17] Michele Banko and Robert C. Moore, "Part of Speech Tagging in Context," *Proceedings of the 20th international conference on Computational Linguistics*, pp.556-561, August, 23-27, 2004.
[18] S. Goldwater and T. L. Griffiths, "A Tully Bayesian Approach to Unsupervised Part-of-Speech Tagging," Proceedings of the ACL, pp.744-751, 2007.



김 성 동

e-mail : sdkim@hansung.ac.kr
1991년 서울대학교 컴퓨터공학과(학사)
1993년 서울대학교 컴퓨터공학과(석사)
1999년 서울대학교 컴퓨터공학과(박사)
1999년~2001년 ㈜엘앤택 기술이사
2001년~현 재 한성대학교 컴퓨터공학과 부교수
관심분야 : 기계번역, 자연언어처리, 데이터마이닝

김 일 민

e-mail : ikim@hansung.ac.kr
1984년 경북대학교 전자계산학과(학사)
1995년 아리조나 주립대학 컴퓨터공학과 (공학박사)
1997년 3월~현 재 한성대학교 컴퓨터공학과 교수
관심분야 : 자바 활용 분산처리, 가상화, 모바일 응용