지능정보연구 제15권 제3호 2009년 9월(pp.53~65)

영한 기계번역에서의 영어 품사결정 모델

김성동
한성대학교 컴퓨터공학과
(sdkim@hansung.ac.kr)

박성훈
한성대학교 컴퓨터공학과
(garion01@naver.com)

영한 기계번역에서 영어 단어의 품사결정은 번역할 문장에 사용된 어휘의 품사 모호성을 해소하기 위해 필요하다. 어휘의 품사 모호성은 구문 분석을 복잡하게 하고 정확한 번역을 생성하는 것을 어렵게 한다. 본 논문에서는 이러한 문제점을 해결하기 위해 어휘 분석 이후 구문 분석 이전에 품사 모호성을 해소하려 하였으며 품사 모호성을 해소하기 위한 CatAmRes 모델을 제안하고 다른 품사태깅 방법과 성능 비교를 하였다. CatAmRes는 Penn Treebank 말뭉치를 이용하여 Bayesian Network를 학습하여 얻은 확률 분포와 말뭉치에서 나타나는 통계 정보를 이용하여 영어 단어의 품사를 결정을 한다. 본 논문에서 제안한 영어 품사결정 모델 CatAmRes는 결정할 품사의 적정도 값을 계산하는 Calculator와 계산된 적정도 값에 근거하여 품사를 결정하는 POSDeterminer로 구성된다. 실험에서는 CatAmRes의 동작과 성능을 테스트 하기 위해 WSJ, Brown, IBM 영역의 말뭉치에서 추출한 테스트 데이터를 이용하여 품사결정의 정확도를 평가하였다.

논문접수일 : 2009년 07월 14일 논문수정일 : 2009년 09월 03일 게재확정일 : 2009년 09월 10일 교신저자 : 김성동

1. 서론

영한 기계번역 시스템은 영어로 작성된 문서를 한국어로 된 문서로 번역하는 기능을 하는 시스템을 일컫는다. 자연언어처리의 역사와 더불어 기계번역은 상당히 오랜 역사를 가지고 있지만, 국내에서는 1980년대 중반이 되어서야 Prolog로 구현된 영한 기계번역 시스템이 효시가 되어 본격적인 영한 기계번역의 연구가 시작되었다(김영택 외, 2001). 이러한 연구가 있기 전에는 사전을 구성하는 방법(최형석, 1984)이나 국어 구문 구조 분석(한성국, 1981), 기계번역을 위한 한국어 품사의 자동분류(박상규, 1984)와 같이 시스템 구성의 부분적인 문제를 해결하거나 언어 현상 자체에 관한 연구가 진행되었다. 영어와 한국어는 문화적인 차이를 가지는 환경에서 만들어지고 성장한 언어이기 때문에 구조나 표현하는 방식에 있어 언어적으로 차이를 보이며 이와 같은 문제로 인해 영어와 한국어 간의 기계번역은 상당히 어려운 분야로 손꼽힌다(김영택 외, 2001).

최근에는 상용화된 영한 기계번역 시스템이 출시되어 전문 도서나 인터넷 문서의 번역에 사용되기도 하며 특허 같은 특수한 분야에서 사용되는 시스템도 개발되는 등 영한 기계번역 시스템에 대한 요구가 증대되고 있는 상황이다. 그러나 일한(Japanese-Korean)번역이나 한일(Korean-Japanese)번역 시

* 본 연구는 2008년도 한성대학교 교내연구비 지원과제 임.

스템 정도의 자연스러운 번역을 생성하지 못해 시장의 요구에 부합하지 못하고 있다. 이는 기계번역 고유의 모호성 문제(ambiguity problem)에 기인한 것으로 이로 인해 기계번역의 복잡도가 증가하여 자연스럽고 정확한 번역이 매우 어렵기 때문이다. 특히 영한 번역의 경우 원시 언어(source language)인 영어와 목적 언어(target language)인 한국어 간에는 상당히 많은 차이가 존재하며 이로 인한 모호성 문제는 일한, 한일 번역의 경우에 비해 매우 심각하다고 할 수 있다. 본 논문에서는 보다 효과적인 영한 기계번역을 위해 언어가 가지는 여러 가지 모호성 중 영어 단어가 가지는 품사 모호성(part-of-speech ambiguity)을 해결하기 위한 방법을 제안한다. 이를 통해 기계번역의 복잡성을 줄여 좀 더 정확한 번역을 얻고자 함이 제안한 방법의 목적이다.

영어 단어의 품사 모호성 해소를 위한 방법으로 품사태깅(part-of-speech tagging) 방법이 자연언어처리 분야에서 널리 사용되고 있다. 이 방법은 어휘 분석 이전에 입력 문장만을 가지고 문장을 구성하는 각 단어의 품사를 결정한다. 기계번역을 위해서는 단어의 품사뿐만 아니라 다른 정보[1]가 필요하며 이를 위해 어휘 사전을 이용한 어휘 분석 과정이 필요하다. 따라서 본 논문에서는 어휘 분석을 수행한 이후에 어휘 분석 결과를 바탕으로 품사를 결정하기 위한 방법을 제안한다. 즉 영한 기계번역에서 사용하는 어휘 분석 결과를 이용하여 일반적으로 어휘의 분석을 배재한 기존의 품사태깅 방법보다 더 나은 정확도를 얻을 수 있는 방법을 찾기 위해 연구를 시작하였다.

1) 예, 동사의 경우 목적어를 가질 수 있다는 정보, 절을 목적어로 취할 수 있다는 정보, 5형식으로 사용될 수 있다는 정보 등 품사별로 여러 가지 추가적인 정보를 필요로 하며 이는 어휘 분석 사전에서 제공된다

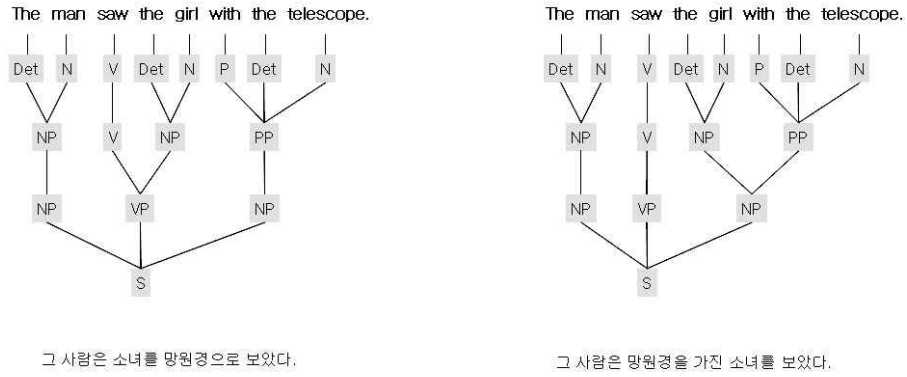
본 논문은 다음과 같이 구성된다. 제 2장에서는 관련 연구 및 기계번역의 문제점을 살펴본다. 제 3장에서는 논문에서 제안하는 품사결정 모델인 CatAmRes에 대해서 자세하게 설명한다. 제 4장에서는 CatAmRes 모델과 다른 품사태깅 방법의 성능을 평가하고 제 5장에서 논문을 마무리한다.

2. 관련 연구 및 기계번역의 문제점

2.1 관련 연구

국내에서 품사태깅에 관한 연구는 많이 이루어졌지만, 영한 기계번역의 분야에서 품사의 모호성을 해소하는 방향에 관련된 연구는 그다지 많지 않다. (이성욱, 1999)에서는 영한 기계번역에서 사용되는 품사와 품사태깅에서 사용하는 품사의 대응관계에 대해서 연구하였다. 이는 품사 태거(part-of-speech tagger)를 영한 기계번역 시스템에 적용하기 위한 방법의 일부로 볼 수 있다. 일반적으로 품사태깅 방법은 어휘의 정보를 고려하지 않고 품사를 결정하는 방법을 이용하기 때문에, 본 논문에서 제안하는 시스템과는 구조적인 차이점을 보인다. 본 논문에서는 어휘 분석이후에 품사결정을 시도하는 방식으로 품사 모호성 해소 문제를 다루고 있다. 국내에서 품사 모호성을 해소하기 위한 모델에 관한 연구로는 최대 엔트로피(maximum entropy) 부스팅 모델(boosting model)을 이용하여 품사의 모호성을 해소하는 방법(박성배, 2003)이 있다. 또한 한국어의 품사태깅의 성능 향상을 위해 어절 주변 형태소의 자질 정보를 이용하여 품사별 분류기(classifier)를 학습시키고 품사태거의 출력을 후처리 하는 방법(최원종 외, 2006)이 제안되기도 하였다. 이는 기존의 품사태거의 에러를 개선하는 효과를 얻었다.

(Nakamura, 1995)에서는 음성 언어(spoken lan-



<그림 1> 구조적으로 모호한 구문 분석 트리의 예.

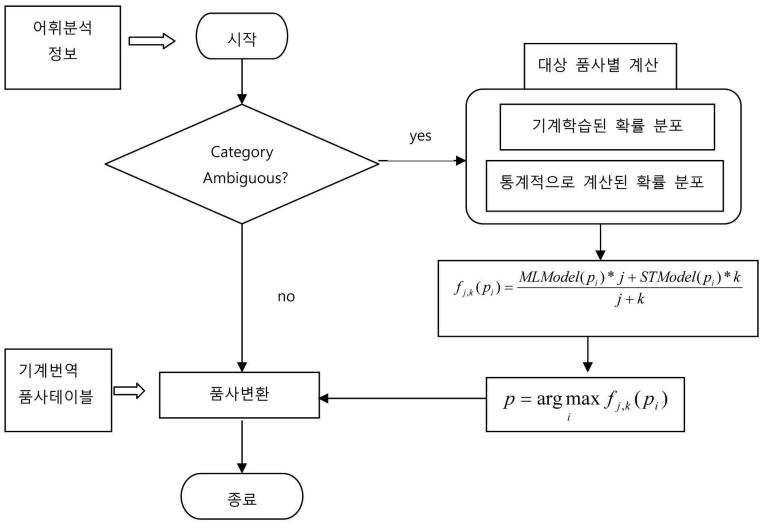
guage) 분석에서 단어의 부류(word category)를 예측하기 위하여 neural network과 n-gram 특성을 사용하였다. 단어의 부류는 본 논문에서 대상으로 하는 품사보다는 태그와 유사한 것으로서 약 85% 정도의 인식 정확률을 보였다. 품사태깅에 관한 많은 연구가 이루어졌으며 이는 품사태깅이 자연언어처리의 다양한 응용분야에서 전처리 단계로 널리 활용되기 때문이다. 규칙을 이용한 태깅방법, 확률적인 방법에 대한 연구가 진행되었으며 최근에는 기계학습 방법인 Support vector machine을 활용한 품사태거(Gimenez et al., 2004), 양방향 퍼셉트론 학습(bidirectional perceptron learning)을 이용한 품사태거(Shen et al., 2007) 등이 개발되기도 하였다.

품사태깅에 적용된 방법들이 마찬가지로 품사결정에 활용될 수 있으며 본 논문에서는 베이지안 학습 방법 및 확률적인 방법을 결합한 방식으로 품사결정 모델을 제안한다.

2.2 기계번역의 문제점

자연언어처리에 있어 기계번역은 자연언어처리의 많은 기술들을 포괄하는 하나의 집합체라 볼 수 있다. 그렇기 때문에 자연언어처리 기술에서 생기는 기술적인 문제들이 대부분 기계번역에 있어서의 기술적인 과제라 볼 수 있다. 덧붙여 기계번역 기술은 번역의 목표가 되는 다른(번역 대상 언어와 대치되는)언어로의 변환, 생성이라는 추가적인 기능을 수행해야 하고 기존의 기술적 과제보다 더 많은 과제들을 해결해야 한다(김태완, 1997).

기계번역 기술이 해결하여야 하는 기술적인 언어분석의 문제(analysis problem)와 언어 변환의 문제(transfer problem)가 있다(Sergei Nirenburg, 1987). 언어 분석의 문제는 크게 어휘적 모호성(lexical ambiguity)과 구조적 모호성(structural ambiguity)으로 나눌 수 있다. 어휘적 모호성은 어떤 하나의 단어에 해당하는 의미가 둘 이상으로 분석 가능한 것을 나타내며 품사 범주 모호성(part-of-speech category ambiguity), 동형이의어(homograph), 다의(polysemy), 변환 모호성(transfer ambiguity) 등이 있다(김태완, 1997). 품사 범주 모호성은 하나의 단어가 여러 개의 품사를 갖고 있는 경우를 의미한다. 하나의 단어가 여러 품사를 갖게 되는 경우 분석 단계에서 부적절한 품사를 선택하게 되면 다른 의미로 해석된 번역문이 생성될 수 있다. 예를 들면, “I can do it” 이라는 문장에서 “can”이 조동사로 해석될 때 “나는 그것을 할 수 있다”로 해석되고, 명사로 해석



<그림 2> 품사결정 순서도

될 때 “나는 깡통 그것 한다”로 해석된다. 하나의 단어가 여러 품사로 해석될 수 있을 경우 category ambiguity를 해결하는 적절한 품사의 결정은 언어 분석에 있어서 중요한 과제라 할 수 있다. 다음으로 동형이의어는 하나의 단어가 서로 상이한 두 가지 의미를 가질 때에 생기는 문제이다. 이 역시 올바른 의미로 분석하지 못하면 원문에서 의도하는 바를 잘못 표현하는 번역이 된다. 예를 들어, “bank”라는 단어는 “은행”이라는 의미로 해석 될 수 있지만, “강둑” 이라는 의미로도 해석 될 수 있다. 다의는 하나의 단어가 의미적인 유사성은 가지나 사용되는 상황이 전혀 다를 경우를 말한다. 예를 들면, “mouth of river”라는 단어는 “강의 입”이라 해석될 수 있지만 “강의 하구”로 해석 할 수도 있다. 구조적 모호성은 번역 대상 언어의 문장을 구문 분석(parsing)하여 구문 분석 트리(parsing tree)를 만드는 과정에서 여러 개의 구문 분석 트리가 생성되는 경우이다. 구문 분석 트리의 구조는 곧 번역문의 구조와 직접적인 영향이 있기 때문에 서로 다른 구조로 만들어진 구문 분석 트리에 의해 각 번역문의 의미가 큰 차이를 보인다. <그림 1>은 구조적 모호성을 보이는 문장의 구문 분석 트리이다. <그림 1>에서 나타난 바와 같이 동일한 문장의 구문 분석 트리도 구조적으로 모호하면 서로 다른 의미를 가질 수 있다.

위에서 기술한 기계번역의 문제점 중에서 본 논문에서는 언어 분석의 문제 중 어휘적 모호성에 속하는 품사 범주 모호성 문제를 다루며 제 3장에서 품사 범주 모호성 해결을 위한 CatAmRes 모델에 대하여 설명한다.

3. CatAmRes : 품사 모호성 해소를 위한 품사결정 모델

3.1 CatAmRes의 품사결정 방식

CatAmRes는 기계학습의 결과와 통계적으로 계산한 확률 값을 이용하여 품사 모호성을 가지는 단어의 품사를 결정한다. <그림 2>의 품사결정 순서도에 의하면 품사 모호성(category ambiguity)을 가지는 단어에 대해서 기계학습을 통해 구성된 확률 값과

<표 1> 어휘분석 결과의 구조.

분류	속성	설명
문장 분석 자료	word	단어
	position	문장 내 위치
	left_pos	대상어휘 좌측품사
	right_pos	대상어휘 우측품사
결정 대상 품사 목록	lex	단어 원형
	pos	결정된 품사
	case	인칭

통계적으로 계산된 확률 값을 이용한 품사 적정도 값을 이용하여 품사를 결정한다. 품사의 적정도(degree of appropriateness) 값은 식 (1)과 같이 표현된다.

$$f_{j,k}(p_i) = \frac{MLModel(p_i)*j + STModel(p_i)*k}{j + k} \quad (1)$$

식 (1)에서 p_i는 단어의 품사 후보 중 하나의 품사 값을 나타내고, $MLModel(p_i)$은 그 품사 값의 기계학습을 통해 구성된 확률 값이며, $STModel(p_i)$은 통계적으로 계산된 품사 확률 값을 나타낸다. MLModel (Machine Learning Model)과 STModel (Statistical Model)은 식 (2)와 식 (3)과 같이 표현된다.[2] 그리고 j는 기계학습을 통해 얻어진 품사 확률을 품사결정에 반영하게 될 비율이고, k 는 통계적으로 계산된 인접 단어와의 연관성을 품사결정에 반영하게 될 비율의 상수 값이다(단, $j + k = 1$, $0 <= j, k <= 1$). 본 논문에서 제안한 품사결정 모델에서는 품사결정의 대상이 되는 단어의 후보 품사들 중에서 품사의 적정도(degree of appropriateness) $f_{j,k}(p_i)$의 값을 최대로 하는 품사를 선택한다.

$$MLModel(p_i) = \Pr(p_{i-1}|P_i) \times \Pr(p_{i+1}|p_i) \quad (2)$$

$$STModel(p_i) = \Pr(p_i|p_{i-1}, p_{i+1}) = \frac{|p_{i-1}, p_i, p_{i+1}|}{|p_{i-1}, p_{i+1}|} \quad (3)$$

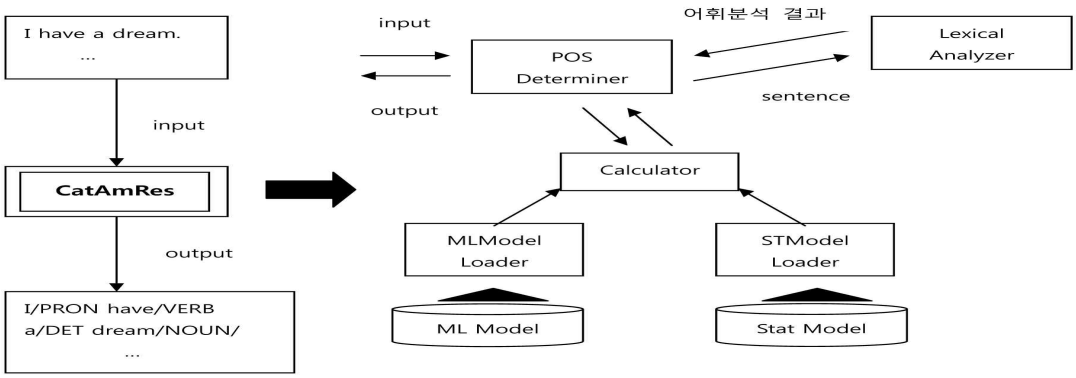
식 $f_{j,k}(p_i)$의 값은 j, k에 따라 달라질 수 있으며, 본 논문에서는 j, k 값에 따른 성능을 비교하였다. 본 논문에서 제안하는 품사결정 모델인 CatAmRes는 입력된 문장의 어휘분석 결과를 이용하여 기계학습을 이용하여 품사별 확률 값과 인접 단어와의 연관성 값을 구한다. <표 1>은 어휘분석 결과의 구조를 보여준다.

<표 1>의 어휘분석 결과를 이용하여 품사를 결정하는 것은 기계학습의 관점에서 볼 때 일종의 분류 문제(classification problem)라 할 수 있다. 본 논문에서는 특정 단어의 품사를 결정하는데 있어서 그 단어의 앞, 뒤 단어의 정보가 중요하다고 판단하여 앞, 뒤 단어들과 대상 단어의 어휘분석 결과를 이용하여 대상 단어의 품사를 결정한다. 앞, 뒤 단어와의 관계에서 현재 단어의 품사의 확률, 즉 조건부 확률(conditional probability) 분포를 학습하고자 하였으며 이러한 관계를 간결하고 이해하기 쉬운 형식으로 표현하는데 적절한 기계학습 방법인 베이지안 네트워크(Bayesian network)를 구성하였다(Ben-Gal, 2007). 즉 앞, 뒤 단어의 품사와 현재 단어의 품사간의 관계를 표현하는 베이지안 네트워크를 구성하고 <표 1>의 어휘분석 결과를 기계학습의 속성값(attribute)으로 하여 베이지안 네트워크를 학습함으로써 앞, 뒤 단어의 품사에 대한 현재 단어 품사의 조건부 확률 분포(conditional probability distribution)를 구성한다. 그리고 통계적인 방법으로 인접 단어와의 연관성 정도를 계산하기 위해 해당 단어 앞, 뒤 단어의 품사, 그리고 해당 단어와의 품사의 조합에 대한 확률 값을 학습 데이터

2) p_{i-1}, p_{i+1}은 각각 앞 단어 품사, 뒤 단어 품사를 의미한다.

<표 2> 태그-기계번역 품사 대응 테이블

	Penn Tags	태그설명	기계번역에서의 대응 품사
1	CC	Coordinating conjunction	CONJ
2	CD	Cardinal number	NUM
3	DT	Determiner	DET
4	EX	Existential there	ADV
5	FW	Foreign word	NOUN
6	IN	Preposition or subordinating conjunction	CONJ or PREP
7	JJ	Adjective	ADJ
8	JJR	Adjective, comparative	ADJ
9	JJS	Adjective, superlative	ADJ
10	LS	List item maker	NOUN
11	MD	Modal	VERB
12	NN	Noun, singular or mass	NOUN
13	NNS	Noun, plural	NOUN
14	NNP	Proper noun, singular	NOUN
15	NNPS	Proper noun, plural	NOUN
16	PDT	Pre-determiner	ADJ
17	POS	Possessive ending	NOUN
18	PRP	Personal pronoun	PRON
19	PRP$	Possessive pronoun	ADJ or PRON
20	RB	Adverb	ADV
21	RBR	Adverb, comparative	ADV
22	RBS	Adverb, superlative	ADV
23	RP	Particle	PREP
24	SYM	Symbol	NOUN
25	TO	to	PREP
26	UH	Interjection	ADV
27	VB	Verb, base form	VERB
28	VBD	Verb, past tense	VERB
29	VBG	Verb, gerund or present participle	VERB
30	VBN	Verb, past participle	VERB
31	VBP	Verb, non-3rd person singular present	VERB
32	VBZ	Verb, 3rd person singular present	VERB
33	WDT	Wh-determiner	ADJ or PRON
34	WP	Wh-pron	PRON
35	WP$	Possessive wh-pronoun	PRON
36	WRB	Wh-adverb	PRON



<그림 3> CatAmRes의 동작 흐름

를 이용하여 계산한다. CatAmRes에서는 어휘분석 결과를 입력으로 하여 기계학습을 통해 구성된 단어의 품사 확률 분포와 해당 단어의 인접 단어와의 연관성 정도를 반영하는 통계적인 확률 값을 각각 식 (1)의 j, k 만큼 반영하여 계산된 품사 적정도 값을 바탕으로 품사를 결정한다. 기계학습과 통계적인 확률 값 계산을 위한 데이터는 Penn Treebank의 Wall Street Journal 영역의 품사가 태그된 말뭉치를 이용한다(Mitchell, 1993). 그런데 Penn Treebank에서 사용하는 품사 태그는 영한 기계번역 시스템에서 사용하는 품사 집합과 다르기 때문에 <표 2>의 태그-기계번역 품사 대응 테이블을 이용하여 변환한다.[3)]

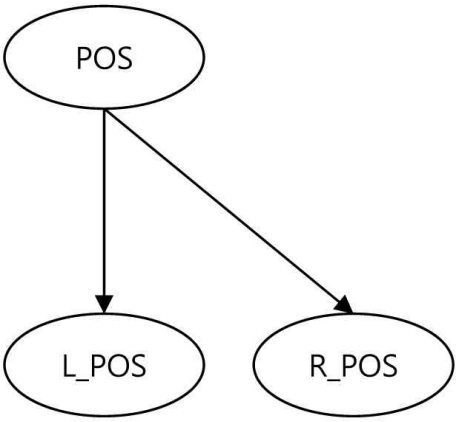
3.2 CatAmRes 구조

CatAmRes(Category Ambiguity Resolution) 모델은 크게 세 부분으로 나뉘어져 있다. 어휘 분석을 하는 LexicalAnalyzer 모듈과 기계 학습된 확률 분포와 통계적으로 계산된 확률 정보를 바탕으로 품사의 적정도 값을 계산하는 Calculator 모듈, 그리고 품사를 결정하는 POSDeterminer 모듈이 있다. <그림 3>은 CatAmRes 모델의 품사결정 과정을 보여준다. 입력 문장이 POSDeterminer 모듈에 전달되면, LexicalAnalyzer 모듈에서 입력 문장에 대한 어휘분석을 수행한다. 어휘분석 결과를 이용하여 품사 모호성을 가지는 단어, 즉 2개 이상의 품사를 가지는 단어에 대해서 주변 단어와의 품사 조합을 생성하고, Calculator 모듈에서 각 품사 조합에 대해 MLModel과 STModel의 확률 값을 이용하여 품사 적정도를 계산한다. 그리고 가장 높은 적정도 값을 가지게 되는 품사 조합을 선택하여 품사 모호성을 가지는 단어의 품사를 결정한다.

3.3 MLModel의 구성

MLModel은 Bayesian Network을 사용하여 학습을 통해 구성된 확률 분포를 말하며 Bayesian Network을 사용한 이유는 각 변수들(주변 어휘의 품사 정보)간의 확률적 의존 관계를 고려하고 이를 확률적으로 표현하기 위해서이다. 탐색 방법에 따라 서로 다른 의존 관계를 보이지만, 본 논문에서는 Naïve한 방법으로 탐색하여 얻은 결과를 바탕으로 MLModel을 구성하였다. MLModel의 구성을 위해 Penn Treebank 말뭉치의 일부를 사용하였다. 일부

3) 영한 기계번역 시스템에서 사용하는 품사 집합 : NOUN, PRON, VERB, ADJ, ADV, CONJ, PREP, DET



<그림 4> 학습 후의 그래프

만 사용한 이유는 전체를 학습시키기에 소요되는 공간복잡도가 지나치게 크기 때문에 특성을 대표할 수 있는 일부만을 학습시켰다. 학습에 사용된 툴은 Weka(Witten et al., 1996)이고, Weka의 Bayesian network 학습을 위한 기본 설정을 사용하여 학습하였으며 <그림 4>는 학습한 결과의 그래프이다. 이 그래프로 좌 품사(앞 단어의 품사)와 우 품사(뒤 단어의 품사)에 대해 각각 결정 대상 품사와 연관을 가짐을 알 수 있다. Bayesian Network 학습을 통해 다음 2가지의 확률 분포를 구성한다: Pr(*Left_POS*| *POS*), Pr(*Right_POS*| *POS*). MLModel은 현재 단어를 중심으로 앞, 뒤 단어의 품사와의 결합 관계를 고려하는 것으로서, 두 가지 확률 분포를 이용하여 품사 모호성을 가지는 단어의 후보 품사들에 대해서 앞, 뒤 단어 품사와의 조건부 확률을 제공한다.

3.4 STModel의 구성

본 논문에서의 STModel은 통계적으로 계산한 품사 조합의 확률을 제공한다. STModel 또한 Penn Treebank 말뭉치의 일부를 사용하며 <그림 5>는 품사 조합의 확률 계산을 위해 변환한 데이터 파일의 모습이다. 이 데이터로부터 현재 단어와 앞, 뒤 단어 품사 조합의 확률을 제 3.1절의 식 (3)을 이용하여 계산한다. STModel은 앞, 뒤 단어의 품사와 현재 단어의 품사를 동시에 고려한 품사 조합의 확률을 제공한다.

STModel은 품사 모호성을 가지는 단어의 주변 단어의 품사를 기준으로 현재의 단어가 가지는 품사 후보들의 품사 확률을 제공한다. 이는 품사 모호성을 가지는 단어를 중심으로 하여 앞, 뒤 단어 품사와의 결합 정도를 제공하는 MLModel과 비교된다.

```
PREP DET NOUN PRON VERB VERB DET ADJ NOUN PREP NOUN PREP DET ADJ NOUN PUNC NOUN
DET NOUN PREP PUNC PUNC VERB VERB NOUN PREP DET ADJ NOUN PREP NOUN PRON VERB VEF
DET ADJ NOUN NOUN VERB VERB VERB PREP DET PUNC PUNC PREP NOUN ADV ADJ VERB PREP
DET ADJ NOUN PREP NOUN PREP ADJ NOUN VERB VERB PREP ADJ PUNC PUNC NOUN CONJ PREF
DET PUNC VERB ADV VERB DET NOUN PREP NOUN CONJ NOUN VERB PUNC PRON VERB VERB DET
CONJ PUNC PUNC PUNC DET NOUN PREP PUNC CONJ PUNC PRON VERB DET NOUN PREP VERB DE
PUNC DET NOUN PREP NOUN PREP PUNC VERB ADV ADJ PREP PRON ADJ VERB PREP NOUN PUNC
PUNC PUNC VERB DET NOUN VERB ADJ PREP VERB NOUN VERB NOUN PREP VERB ADJ NOUN PRC
ADJ NOUN VERB ADV VERB NOUN NOUN CONJ VERB VERB VERB PREP NOUN PUNC VERB CONJ AD
PUNC CONJ PUNC PUNC PRON PREP NUM NOUN VERB DET ADJ NOUN NOUN PREP PUNC PUNC VEF
ADJ VERB DET VERB NOUN PUNC DET ADJ NOUN CONJ DET NOUN NOUN PRON VERB VERB NUM N
PUNC CONJ PUNC ADV VERB NUM NOUN VERB PREP NOUN ADJ NOUN PREP ADV ADJ NOUN PUNC
ADV PREP DET NOUN NOUN VERB VERB PREP NOUN NOUN NOUN PUNC
ADV PUNC ADV VERB DET NOUN NOUN PUNC NOUN NOUN VERB PREP NOUN NOUN VERB NOUN PUN
PUNC PUNC PUNC DET VERB NOUN PREP PUNC CONJ PUNC VERB PUNC
PUNC ADJ VERB ADJ PREP DET NOUN PREP VERB NOUN PUNC PUNC
```

<그림 5> 품사 조합의 확률 계산을 위해 변환된 데이터의 모습

4. 실험

실험에서는 CatAmRes 모델의 확률 계산을 위한 학습 데이터와 성능 평가를 위한 테스트 데이터를 제시하고 실험 환경을 설명한다. 또한 품사결정의 정확도 측정을 통해 본 논문에서 제안한 CatAmRes 모델이 올바르게 동작하는지를 검증하고 다른 품사 태깅 시스템과 성능 비교를 한다.

CatAmRes는 Windows XP를 기반으로 Visual Studio 6.0을 사용하여 C++로 구현이 되어있으므로 Windows 환경에서 테스트를 수행하였다. 다른 품사 태깅 시스템의 환경도 Windows 환경을 사용한다. <표 3>은 성능 평가를 위한 하드웨어 환경을 보여준다.

<표 3> 성능 평가 환경.

구성 요소	사양
CPU	Core2Duo 2.13 GHz(6400)
Memory	DDR2 SDRAM 2GB

학습 데이터는 Penn Treebank 말뭉치의 품사가 태그된 데이터 중 WSJ(Wall Street Journal) 분야의 약 620,000 단어로 구성된 25,000 문장을 사용하였다. 테스트 데이터 역시 Penn Treebank 말뭉치를 이용하였는데 학습 데이터와 같은 영역 뿐만 아니라 다른 영역에서도 테스트 데이터를 추출하였다. <표 4>는 테스트 데이터의 구성을 보여준다.

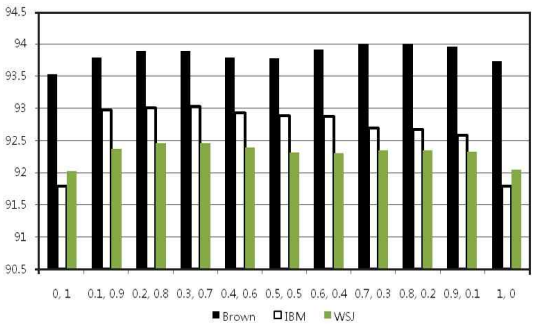
<표 4> 테스트 데이터의 구성

	Brown	IBM	WSJ
문장 수	3,310	4,324	3,631
단어 수	67,170	60,904	73,101

테스트 데이터를 이용하여 CatAmRes 모델의 품사결정 정확도(accuracy)를 평가하고 다른 방법과 정확도를 비교하였다. 정확도 평가 과정에서 제 3.1절의 식 (1)의 품사 적정도 $f(j, k)$에서 인자인 j, k 값을 변화시켜 각각에 대한 정확도를 측정하였다. 인자 j와 k의 합이 1이 되도록 하면서 각각의 값을 0~1까지 변화시키면서 정확도를 측정하였으며 <표 5>에서 결과를 제시하였다. <그림 6>은 이 결과를 그래프로 표현한 것이다. <표 5>와 <그림 6>을 보면 영역별로 정확도의 차이가 나고

<표 5> 인자 j, k 값의 변화에 따른 품사결정 정확도(%)

(j, k)	Brown	IBM	WSJ
(0, 1)	93.53	91.79	92.02
(0.1, 0.9)	93.79	92.97	92.37
(0.2, 0.8)	93.88	93	92.45
(0.3, 0.7)	93.88	93.02	92.45
(0.4, 0.6)	93.79	92.93	92.39
(0.5, 0.5)	93.77	92.88	92.31
(0.6, 0.4)	93.91	92.87	92.3
(0.7, 0.3)	94	92.69	92.34
(0.8, 0.2)	94	92.67	92.34
(0.9, 0.1)	93.95	92.58	92.32
(1, 0)	93.73	91.79	92.04



<그림 6> 인자 j, k 값의 변화에 따른 품사결정 정확도 그래프

<표 6> 품사결정 정확도 비교(%)

	Brown	IBM	WSJ
Baseline method	87.13	86.17	86.41
HMM based tagger	90.25	90.5	90.4
CatAmRes	93.84	92.75	92.30

있지만 인자 j, k값의 변화에 따른 정확도의 차이는 매우 적음을 알 수 있다. 즉 인자 j, k는 품사결정 성능에 크게 영향을 미치지 않으며 논문에서 제시한 품사결정 모델은 영역에 독립적이며 인자 값에 관계없이 일정한 수준의 정확도를 가진다고 할 수 있다.

성능의 비교를 위해 품사확률 정보[4]를 이용하여 가장 높은 확률을 가지는 품사를 품사결정의 결과로 삼는 기본 방식(baseline method)에 의한 정확도와 HMM 기반의 태거를 이용한 정확도를 <표 5>의 CatAmRes 정확도의 평균 정확도를 함께 <표 6>에서 제시하였다. 본 논문에서 제안한 CatAmRes 모델은 영한 기계번역을 위한 품사결정 시스템이기 때문에 <표 2>에 나타나는 영한 기계번역에서 사용하는 품사 집합을 사용한다. 일반적으로 품사 태거는 Penn의 태그 집합(tag set)을 사용하므로 품사 태거의 결과를 <표 2>를 이용하여 영한 기계번역에서 사용하는 품사 집합에 맞추어 변환하였다. <표 6>의 결과를 보면, 제안한 CatAmRes 모델이 3영역 모두에서 가장 우수한 정확도를 나타냈으며 따라서 평균적으로 우수한 품사결정 정확도를 보인다고 할 수 있다.

5. 결론

본 논문에서는 영한 기계번역 시스템의 성능 향상을 위해 단어가 가지는 품사 모호성 해소를 위한 방법을 연구하였으며 그 결과로 CatAmRes 모델을 제안하였다. CatAmRes 모델은 기계학습을 통해 구성된 확률 모델과 통계적으로 계산된 확률 정보를 이용하여 품사 모호성을 가지는 단어의 품사를 결정한다. 기계학습과 확률 계산을 위해 Penn Treebank의 WSJ 분야의 품사가 태그된 말뭉치를 이용하였다. 기계학습 방법으로는 베이지안 네트워크 학습 방법을 적용하였는데, 앞-뒤 단어와 현재 단어 간의 품사 관계를 표현한 베이지안 네트워크를 구성하여 이들 간의 조건부 확률 분포를 생성한다. 그리고 앞-뒤 단어와 현재 단어로 구성되는 품사 트라이그램의 확률을 학습 데이터로부터 계산하며 이 두 가지를 이용하여 품사의 적정도 값을 계산하고 이를 바탕으로 품사 모호성을 가지는 단어의 품사를 결정한다.

실험 결과 CatAmRes 모델은 다른 품사태거에 비해 높은 정확도를 보이는 것으로 나타났는데, 이는 CatAmRes 모델은 어휘분석 결과를 이용하여 품사를 결정하는 것이며 품사태거는 어휘분석 이전에 단어만을 보고 품사태그를 결정하는데 따라서 CatAmRes가 보다 많은 정보를 이용하기 때문인 것으로 판단된다. 본 논문에서 설명한 품사 결정 모델은 영한 기계번역을 위해 제안된 것이며 기계번역을 위한 구문 분석 및 이후의 분석을 위해 어휘분석이 반드시 필요하기 때문에 어휘분석 과정을 수행해야 하며 CatAmRes는 어휘분석 결과를 활용하여 품사 결정을 한다. 본 논문에서 제안한 CatAmRes 모델은 영한 기계번역의 구문 분석의 효율성 및 정확성을 제고하는데 기여할 것이다.

영한 기계번역 시스템에서 어휘분석 이후에 품사 모호성을 해소함으로써 이후의 구문 분석 과정에서 분석의 복잡도를 상당히 줄일 수 있다. 따라서 CatAmRes 모델을 영한 기계번역 시스템에 결

4) 품사확률 정보는 말뭉치에서 각 단어가 나타나는 빈도수와 특정한 품사로 사용된 빈도수와의 비율로 계산되며 어휘 분석을 위한 사전에 품사확률 정보가 포함되어 있다

합하여 기계번역 시간/공간적 성능 향상 및 번역률 개선에 기여하는 정도를 평가하고 이를 통해 보다 많은 성능 향상을 위해 필요한 문제를 발견하는 연구를 수행할 필요가 있다. 여기에는 CatAmRes 모델의 에러를 분석하여 품사결정으로 인한 성능 저하를 방지하는 방법에 대한 고려도 포함된다.

참고문헌

김영택 외 25인, "자연 언어 처리", 제1판, *생능출판사*, 2001.

김태완, "기계 번역 기술의 개요 및 동향", *대한전자공학회 전자공학회지*, 24권 9호(1997), 1095~1102.

박상규, "기계 번역을 위한 한국어 품사의 자동 분류 방법", *한국과학기술원 석사학위 논문*, 1984.

박성배, 장성탁, "최대 엔트로피 부스팅 모델을 이용한 품사 모호성 해소", *한국정보과학회 2003년도 봄 학술발표논문집*, 30권 1호(B) (2003), 522~524.

심광섭, 김영택, "기계 번역 시스템", *한국정보과학회 정보과학회지*, 12권 8호(1994), 17~23.

이성욱, 이공주, 서정연, "영한 기계번역 품사 집합과 펜트리뱅크 코퍼스 품사 집합간의 품사 대응", *한국정보과학회 1999년도 가을 학술 발표논문집*, 26권 2호(1990), 184~186.

최원종, 이도길, 임해창, "어휘별 분류기를 이용한 한국어 품사 부착의 성능 향상", *제18회 한글 및 한국어 정보처리 학술대회 논문집*, (2006), 133~139.

최형석, "국어의 처리를 위한 기계사전에 관한 연구", *인하대학교 석사학위논문*, 1984

한성국, "한국어의 Machine Translation을 위한 구문 구조 분석", *인하대학교 석사학위논문*, 1981.

Ben-Gal I., "Bayesian Networks", Encyclopedia of Statistics in Quality and Reliability, Wiley and Sons, 2007.

Brill, E., "Transformation-Based Error-Driven Learning and Natural Language Processing : A Case Study in Part-Of-Speech Tagging", *Computational Linguistics*, Vol.21, No.4(1995), 543~565.

Gimenez J., and Marquez L., "SVMTool : A General POS Tagger Generator based on Support Vector Machines", *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004.

Kupiec, J., "Robust part-of-speech tagging using a hidden Markov model", *Computer Speech and Language*, Vol.6(1992), 225~242.

Mitchell, P. M., B. Santorini, and M. A. Marcinkiewicz, "Building a Large Annotated Corpus of English : The Penn Treebank", *Computational Linguistics*, Vol.19, No.2(1993), 313~330.

Nakamura, M., Tsuda K., and J.-I. Aoe, "Word category prediction based on neural network", *International Journal of Computer Mathematics*, Vol.57, No.3(1995), 169~181.

Sergei Nirenburg, *Machine Translation-Theoretical and methodological issues*, Cambridge University Press, 1987.

Shen L., Satta G., and Joshi A., "Guided Learning for Bidirectional Sequence Classification", *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (2007), 760~767.

Witten, I.H., E. Franck, L. Trigg, M. Hall, G. Holmes, and S.J. Cunningham, "Weka: Practical machine learning tools and techniques with Java implementations", *Proceedings of ANNES'99 International Workshop on emerging Engineering and Connectionist-based Information Systems,* (1999), 192~196.

Abstract

A Model of English Part-Of-Speech Determination for English-Korean Machine Translation

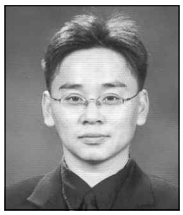
Sung-Dong Kim · Sung-Hoon Park

The part-of-speech determination is necessary for resolving the part-of-speech ambiguity in English-Korean machine translation. The part-of-speech ambiguity causes high parsing complexity and makes the accurate translation difficult. In order to solve the problem, the resolution of the part-of-speech ambiguity must be performed after the lexical analysis and before the parsing. This paper proposes the *CatAmRes* model, which resolves the part-of-speech ambiguity, and compares the performance with that of other part-of-speech tagging methods. *CatAmRes* model determines the part-of-speech using the probability distribution from Bayesian network training and the statistical information, which are based on the Penn Treebank corpus. The proposed *CatAmRes* model consists of *Calculator* and *POSDeterminer*. *Calculator* calculates the degree of appropriateness of the part-of-speech, and *POSDeterminer* determines the part-of-speech of the word based on the calculated values. In the experiment, we measure the performance using sentences from WSJ, Brown, IBM corpus.

Key Words : Machine Translation, Part-Of-Speech Determination, Part-Of-Speech Tagging, Machine Learning, Statistical Methods

* Department of Computer Engineering, Hansung University

저 자 소 개



김성동

서울대학교 컴퓨터공학과를 졸업(1991)하고, 동 대학원에서 컴퓨터공학 석사(1993) 및 박사(1999)학위를 취득하였다. 서울대학교 컴퓨터신기술공동연구소에서 특별연구원으로, ㈜엘앤텍에서 기술이사로 근무하였으며 2001년부터 한성대학교 컴퓨터공학과 부교수로 재직 중이다. 주요 관심분야는 자연언어처리, 기계번역, 데이터마이닝 등이다.

박성훈

한성대학교 컴퓨터공학과를 졸업(2006)하고, 동 대학원에서 컴퓨터공학 석사(2008) 학위를 취득하였다. 현재 ㈜모비젠 R&D 연구소에서 주임 연구원으로 재직 중이다. 관심분야는 기계번역, 기계학습, 데이터마이닝 등이다.