구어파서를 위한 생성 인식 언어모델†

정 홍, 황광일
포항공과대학교 전자전기공학과
경북 포항시 남구 효자동 산31번지
우: 790-784
hjeong@postech.ac.kr, lighton@postech.ac.kr

Generation and Recognition Language Model for Spoken Language Parser

Hong Jeong, Kwang Il Hwang
Intelligent Signal Processing Lab.,
Dept. of Electronic and Electrical Engineering, POSTECH.

요약

구어는 프로그래밍 언어와는 달리 주어진 문장 내에서의 해당 어휘의 뜻(semantic information)을 알고 다른 어휘들과의 연관성(grammatical information)을 알아야만 적절한 형태소분석이 가능하다. 또한 구어는 방대한 양의 어휘들로 구성되어 있으며 사용하는 사람마다의 다양한 응용과 공식화되기 어려운 수많은 예외들로 운용되기 때문에 단순히 찾아보기표와 오토마타만으로는 형태소분석에 한계가 있다. 이에 본 논문에서는 주어진 어휘집과 그 어휘들로 만들어진 다양한 문장들로부터 구어운용의 근본기제를 스스로 학습해나가는 강화학습중심의 언어모델을 제안하고 실제로 한국어 형태소분석에 적용하여 그 성능과 특성을 파악해보았다. 구어파서의 입력은 음절단위의 발음이며 인간이 문장을 듣거나 보는 것과 동일하게 시간에 따라 순차적으로 입력된다. 파서의 출력 또한 시간에 따라 변화되면서 나타나며 입력된 연속음절을 형태소단위로 분리(segmentation)하고 분류(labeling)한 결과를 나타낸다. 생성인식 언어모델이 기존의 언어모델과 다른 점은 구어 파싱에 있어서 필수적인 미등록어에 대한 유연성과 앞단의 음성인식기 오류에 적절한 반응(fault tolerance)을 나타내는 것이다.

1. 서론

음성인식의 가장 큰 문제는 잡음 하에서 발생된 구어의 불완전성이다. 완전한 형태의 어휘와 문장구조를 요구하는 파싱기법으로는 이러한 실세계의 문제를 다루는 데에 한계가 있다. 본 논문에서는 실세계의 잡음 하에서 발생된 구어를 음절단위로 표기한 입력데이터로부터 45개의 형태소분류[1]를 인간과 흡사한 방식으로 수행하고 구어문법을 학습할 수 있는 구어파서의 수학적 모델을 연구하였으며, 형태소단위 품사태깅시스템에 적용하여 성능과 특성을 분석하였다. 핵심적인 아이디어는 인간의 음성발생 및 인식의 구조와 흡사한 유한상태기계들의 동적 시스템을 설계하는 것이며, 형태소분류를 계층적으로 처리하기 위하여 음성생성모델은 내부적으로 계층적인 구조로 설계하였으며 생성의 최종결과만을 이용하는 반복적인 생성을 통하여 생성습성과 어휘발생습성을 수정할 수 있도록 강화학습을 적용하였다.

1.1 기존의 언어모델

분야일반 대용량 연속음성인식의 실현을 위해서는 단순히 신호처리수준에서의 음성처리 외에도 상위계층의 언어정보와 의미정보가 하위계층으로 전달되어야만 하며, 이를 위하여 구어의 특수성을 반영하고자하는 여러 가지 계산언어모델의 연구가 시도되어 왔다. 그간의 연구방향을 대별하면 HMM, 신경망, 통합기반문법 접근방식을 들 수 있다. HMM을 기반으로 하는 통계적 언어모델은 주로 트라이그램 언어모델을 사용하는데, 대용량 어휘를 구현하려면 탐색공간이 급격히 증가하여 실시간 계산이 어려워지며 마콥 프로세스를 전제하므로 언어생성과정에서 나타나는 순행동화현상이나 역행동화현상 등의 다양한 변화들을 포용하기 어렵고 어휘확장이나 문법수정을 하기에 유연성이 부족하다는 단점이 있다.[2] 신경망기반의 언어모델은 학습능력과 일반화능력이 있지만 대용량 어휘를 구현하기 어렵고 통계적 빈도수가 적은 문장과 관례적으로 쓰이는 예외적 표현들을 학습하기 어렵다는 단점이 있다.[3] 통합기반문법기반의 언어모델은 문어체의 자연어 파

† 이 논문은 1999년도 과학기술부 뇌연구개발사업 지원에 의한 결과임

싱에는 적절하나 구어의 불완전성에 의해 급격한 성능저하가 나타나며 사전구축과 예외처리에 인간의 수작업이 대량 필요하다는 단점이 있다.[4] 이러한 단점들의 근본적인 원인은 기존의 언어모델이 인간의 언어처리기제를 반영하지 못하는데 있다고 보여진다.

1.2 강화학습 이론

강화학습은 단순히 학습기법의 한 종류라기보다 습관화(automatization)가 일어나는 학습유형에 대한 생물학적 기제이다. 일반적인 교사학습과 비교사학습의 경우에는 결국 일종의 함수근사나 데이터군집화라고 할 수 있지만 강화학습의 경우에는 그 토대를 통계학과 동물학습이론에 두고 있다. 이제까지 강화학습은 주로 제어분야와 게임분야에 적용되어왔으나 언어생성과정이 일종의 습관화라는 인지심리학의 연구보고[5]에 따르면 언어모델에도 적용하는 것이 타당하다고 볼 수 있다. 언어모델의 학습기제로 강화학습을 적용할 때 얻게되는 가장 큰 장점은 초기학습 후의 언어모델은 실시간학습을 통하여 태깅이 없는 문장들로도 학습이 가능하다는 것이다. 생성 인식 언어모델은 생성기제에 의해 발생한 의사음성과 관측음성의 비교에 의한 강화신호만으로 실시간학습이 가능하다.

2. 생성 인식 언어모델

구어파서를 위한 생성 인식 언어모델의 전체적인 구조는 [그림 1]과 같다. 이 파서의 구조는 기존의 파서에서 채용하고 있는 상향식 데이터분석구조가 아니라 음성신호의 과거데이터와 파서의 생성습성에 영향을 받는 의사음성의 발생을 이용하여 다음 입력신호를 추정하는 예측필터기반의 하향식 의미생성구조이다. 파서의 예측음성신호는 실제 입력된 음성신호와 비교되어 생성습성을 수정하기 위한 강화신호를 발생시킨다. 형태소분류체계가 다단계임을 반영하기 위하여 생성 언어모델은 계층적 구조로 이뤄져 각 단계마다 분류할 뿐 아니라 상하계층간에도 상호작용을 하게 된다.

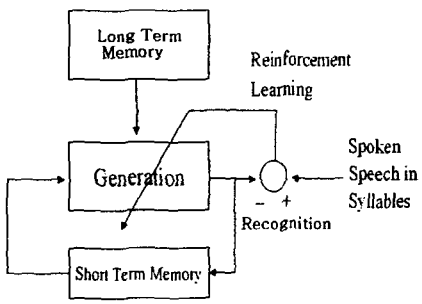

그림 1 생성 인식 언어모델의 전체구조

예측과 학습에 의하여 파서는 하나의 음성신호에 여러 후보를 제시하는 HMM에 쉽게 접목되며, 파서의 예측은 하부의 HMM을 제어하여 탐색공간을 대폭적으로 줄여주게 되어 전체 음성인식시스템의 속도와 성능을 향상시키고 실시간 대용량 어휘처리가 가능해지게 된다.

그러므로 생성 인식 언어모델은 단순한 오토마타 기반의 언어모델과는 달리 능동적으로 생성하여 인지하는 구어파서를 만드는 기초이론이 되며, 보다 인간의 정보처리방식에 가까운 자연스러운 접근방식이라고 할 수 있다.

2.1 생성 인식 계산모형

생성 인식 언어모델은 다음의 동작을 반복하는 유한상태기계들의 동적 시스템이다.

① 주어진 현재의 문맥과 관측데이터를 이용하여 다음 문맥을 예측한다.

② 그러한 예측을 토대로 의사관측데이터를 생성시키고 이것은 실제의 관측데이터와 음향음성학적으로 비교되어서 가장 가까운 후보가 선택된다.

③ 위의 ①과 ②를 한 문장이 끝날 때까지 음절이 입력될 때마다 반복한다.

O는 파서가 인식을 위한 결정을 하는데 사용되는 음성관측데이터를 나타낸다. 여기서 인식이란 파서가 관측하는 발음을 토대로 그러한 자극을 발생시킨 외부화자의 마음에 어떤 형태소들이 활성화되었던 것인지를 추정하는 것으로 정의한다.

O는 유한한 크기의 음절집합 Θ의 원소들의 특정한 조합으로서 임의의 음절열을 구성한다. m개의 음절로 구성된 관측데이터는 다음과 같다.

$$O = o_1, o_2, \cdots, o_m \qquad o_i \in \Theta \qquad [1]$$

Q는 관측데이터 O를 발생시킨 화자의 마음속에 발화되었던 형태소열을 나타낸다. Q는 파서와 화자 모두에게 알려져 있는 유한한 크기의 어휘집합 Φ의 원소들의 특정한 조합으로서 임의의 형태소열을 구성한다. n개의 형태소로 구성된 형태소열은 다음과 같다.

$$Q = q_1, q_2, \cdots, q_n \qquad q_i \in \Phi \qquad [2]$$

$P(Q|O)$는 주어진 관측데이터 O가 존재할 때, Q가 화자의 마음속에서 활성화되었을 확률을 의미한다. 당연히 파서는 주어진 관측데이터 O에 대하여 확률값 $P(Q|O)$를 가장 높게 하는 $\hat{Q}$를 선택하여야 한다.

$$\hat{Q} = \arg\max_Q P(Q|O) \qquad [3]$$

일반적으로 $P(Q|O)$를 직접 구하기는 어려우므로 베이지안 규칙을 적용하여 [수식 4]와 같이 변형시킨다.

$$\hat{Q} = \arg\max_Q P(O|Q)P(Q) \qquad [4]$$

하지만 [수식 4]는 언어 인식기제에 있어서 단순히 통계적인 면만을 반영하므로 인간의 인식과정과 비교할 때 자연스럽지 않으며 인간의 적극적인 생성과정으로서의 인식기제는 고려되지 않은 통계처리이다.

그러므로 본 논문에서는 인간의 언어능력에 관계된 인지모델을 기반으로 새로운 계산언어모델을

고안하고 그 모델의 생성 인식 학습기제를 정의하였다.

2.2 어휘집 표현(lexicon representation)

어휘집이란 특정 언어에서 사용되는 형태소들의 전체집합을 말한다. 영어를 사용하는 성인에게는 대략 8만 개의 형태소가 어휘집에 구성되어 있다는 인지심리학의 실험 결과가 있었다[6]. 하나의 형태소는 앞 뒤 형태소의 배열에 의해 발생하는 문맥에 따라 다양한 음소형태로 발생하게 된다. 한국어의 예를 들면 자음접변, 구개음화, 르불규칙 등이 있다. 어휘집의 형태소 분류체계는 '한국어 표준 형태 통사 태그 표준안'[1]에서 제시하는 45개의 품사태그를 따랐다. 세 개의 계층으로 구성된 형태소 수준의 품사체계는 [표 1]과 같다.

A_1	A_2	A_3
n	nc	
	(ncp)	ncpa ncps
	(ncn)	ncn
	nq	nq
	nb	nbu nbn
	np	npp npd
	nn	nnc nno
p	v	pvd pvg
	a	pad paa
	px	px
m	mm	mmd mma
	ma	mad maj mag
i	ii	ii
j	jc	jcs jco jcc jcm jcv jca jcj jct jcr
	jx	jxc jxf
	jcp	jcp
e	ep	ep
	ec	ecc ecs ecx
	et	etn etm
	ef	ef
x	xp	xp
	xs	xsn xsv xsm xsa

표 1. 생성 인식 언어모델의 품사체계

생성 인식 언어모델에서 문법은 생성습성 A의 확률가중치집합으로 표현된다. A는 A_1, A_2, A_3의 계층적인 피라미드구조로 구성되어 있어서 주어진 문장을 여러 수준에서 형태소분석을 수행하게 된다. A_1은 주어진 문장의 형태소분석을 7개의 대분류 품사로만 분류하며 A_1의 동작은 A_2의 생성에 영향을 준다. A_2에서는 20개의 세부품사로 분류하고 A_3의 생성에 영향을 준다. 마지막으로 A_3에서는 45개의 확장세부품사로 분류하게 된다. A_3의 출력은 어휘발생습성 B에 의하여 기저형에서 표층형으로 변환되어 발생하게 된다.

2.3 언어 생성기제(Generation)

인간은 다른 사람의 말을 들을 때 자신의 장기기억에 있는 어휘집 Φ를 토대로 단기기억에 자신의 생각 $\hat{Q}$을 생성해가면서 상대방의 의미 Q를 인식해간다. 이러한 자신의 생각은 문맥주의집중 D로 작용하여 현재의 인식에 영향을 미친다. 문맥주의집중은 청자에게 가해지는 음성자극 O의 영향을 받아서 자극과의 차이가 적어지는 상태로 변화되어간다. 또한 인간은 단순히 현재의 자극을 수동적으로 분석하는 것이 아니라 다음에 가해질 음성도 미리 기대해가며 기대가 맞을 경우에는 자신의 생각 $\hat{Q}$이 상대방의 생각 Q와 흡사하다는 확신을 가지게 되고 틀릴 경우에는 다른 의미로 해석해보려고 시도하게 된다. 확신도는 파싱가상온도로 나타나며 온도가 낮은 것은 확신도가 높다는 것을 의미한다. 예측은 파서의 생성습성 A과 문맥주의집중 D의 영향을 받으므로 예측과 관측열의 비교로부터 생성되는 강화학습신호를 이용하여 특정문맥에서의 생성습성을 변화시켜주면 반복학습에 의하여 예측오차가 줄어들게 된다. 강화학습신호는 억제신호는 사용하지 않으며 강화신호만을 사용하여 보상분배(credit assignment)문제를 해결한다. 실제로 어린아이가 언어를 배울 때에 긍정반응(positive feedback)만을 사용한다는 아동심리학의 연구보고[7]가 있었다. 이러한 인간의 언어 생성기제를 계산모형으로 고안한 것이 [그림 2]이다.

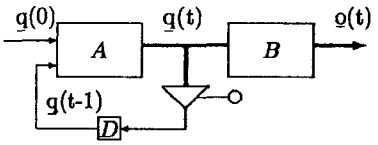

그림 2. 언어 생성기제

[그림 2]에서 굵은 선은 병렬로 여러 개의 후보출력들이 지나가는 경로를 의미하며 가는 선은 인식기제에 의하여 선택된 하나의 출력만이 문맥주의집중 D로 입력되는 것을 나타낸다. 생성기제에 의해 발생하는 의사음성은 관측음성과 음향음성학적 거리가 측정되며 관측음성에 가장 흡사한 의사음성이 선택되면 그러한 의사음성을 발생시킨 형태소열이 함께 선택되는 원리이다. 반복적인 선택은 그러한 발생에 관여한 생성습성이 강화되는 효과를 나타내게 된다. 파서는 말뭉치의 다양한 문장들을 이용하여 생성 인식 학습동작을 수없이 반복하게 되면서 자신만의 고유한 생성습성으로 수렴하게 된다. 언어생성은 파싱가상온도의 영향을 받도록 softmax action selection 방식으로 동작한다. 파서의 t 시점에서의 생성은 다음과 같다.

$$V_t(a|s) = \frac{r_1 + r_2 + \cdots + r_{k_a}}{k_a} \qquad [5]$$

$$\Pr ob(a) = \frac{e^{V_{t-1}(a|s)/\tau}}{\sum_b e^{V_{t-1}(b|s)/\tau}}, \text{ given state s} \quad [6]$$

k_a는 주어진 상태 s에서 특정 행위 a가 일어난 전체회수이다. 각각의 r_i는 0 또는 1의 값을 갖게 되며 1은 보상을 의미한다. b는 주어진 상태 s에서 일어날 수 있는 모든 a들의 집합의 모든 원소를 나타내는 인덱스이다. $\Pr ob(a)$는 주어진 상태 s에서 특정 행위 a가 일어날 확률이므로, 언어생성은 확률과정이 된다.
파싱가상온도는 강화신호의 발생 빈도에 반비례하며 변화되고 생성은 파싱가상온도에 영향을 받으므로 언어생성은 정궤환 특성을 가지게 된다.

2.3 언어 인식기제(Recognition)

언어인식단계에서는 생성기제에 의하여 발생된 여러 의사음성들 가운데서 관측음성과 음향음성학적으로 가장 가까운 후보가 선택된다. 인식기제의 블록다이어그램은 [그림 3]과 같다.

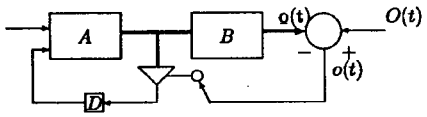

그림 3 언어 인식기제

유사도측정은 DTW를 자음모음 수준에서 동작하도록 수정하여 적용하였다. 의사음성 음절열과 관측음성 음절열의 각각의 음절은 초성 중성 종성으로 분해되어 자모열이 되고 두 개의 자모열이 DTW에 의하여 비교되는 것이다.

2.5 언어 학습기제(Learning)

언어학습단계에서 파서는 강화학습신호 $r(t)$에 의하여 A와 B값을 수정해가면서 점차적으로 최적값에 가까워지게 된다. 강화학습이론의 $TD(\lambda)$ 알고리즘을 생성 인식 언어모델에 적합하도록 수정하여 적용하였다. 학습기제의 블록다이어그램은 [그림 4]와 같다.

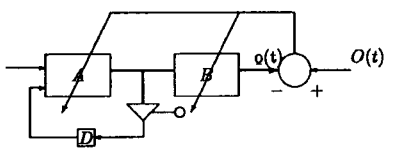

그림 4. 언어 학습기제

학습기제는 파서의 초기에 수행되는 초기학습과 그 이후의 실시간학습의 두 가지 동작모드가 있다. 초기학습에서는 일반적으로 파싱가상온도가 높은 상태이고 학습용 말뭉치로부터 어휘습득과 문법습득이 활발하게 일어나서 파서의 장기기억이 확장된다. 반면에 실시간학습에서는 파싱가상온도가 비교적 낮은 상태가 유지되면서 파서는 자신의 습득한 어휘집과 문법지식을 이용하여 여러 가지 생성들을 반복하면서 생성습성을 정교하게 다듬게 된다. 초기학습에서는 태깅이 되어 있는 말뭉치가 필요하지만 실시간학습에서는 일반 문장들만으로 학습이 가능하다. 실시간학습에서도 파서는 미등록어 자동습득기제에 의하여 어휘와 문법을 확장해 나가지만 초기학습보다는 느리게 확장한다. 실시간학습에서는 $TD(\lambda)$ 알고리즘을 이용하여 $V_t(a|s)$ 값들을 한 문장이 끝나서 강화신호가 발생하기 전에 빠르게 수정해간다. 실시간학습 과정에서 $V_t(a|s)$의 수정은 [수식 7]과 같다.

$$\Delta w_t = \alpha(r(x_t) + V(x_{t+1}, w_t) - V(x_t, w_t)) \sum_{k=1}^{t} \lambda^{t-k} \Delta_W V(x_t, w_t) \quad [7]$$

이 식을 증분형(incremental form)으로 바꾸면 [수식 9]와 같게 된다.

$$\Delta w_t = \alpha(r(x_t) + V(x_{t+1}, w_t) - V(x_t, w_t)) g_t \quad [8]$$

$$g_{t+1} \equiv \sum_{k=1}^{t+1} \lambda^{t+1-k} \nabla_W V(x_t, w_t) \quad [9]$$

$$= \nabla_W V(x_{k+1}, w_t) + \sum_{k=1}^{t} \lambda^{t+1-k} \nabla_W V(x_k, w_t)$$
$$= \nabla_W V(x_{k+1}, w_t) + \lambda g_t$$

2.6 문법, 어휘 자동습득과 미등록어 처리

초기학습단계에서 생성 인식 언어모델은 어린아이가 경험을 통하여 언어를 배우는 것과 유사한 방식으로 태깅이 되어 있는 말뭉치로부터 강화학습을 통해 문법과 어휘를 자동으로 습득하고 습득된 문법과 어휘는 장기기억에 저장된다. 문법습득은 생성습성 A의 가중치 집합의 값들이 적절하게 조정되는 것을 의미하며 어휘습득은 어휘발생습성 B에서 사용하는 어휘집에 새로운 어휘가 추가되는 것을 의미한다.
실시간학습단계에서 언어모델은 태깅이 없는 일반문장만으로 생성습성 A와 어휘발생습성 B를 보다 정교하게 다듬게 되며 어휘집에 있는 기존의 어휘와 같은 용도로 쓰이는 미등록어들을 습득하게 된다. 생성습성 A에 의하여 동일한 문맥으로 파악되는 부분문장에서 구어파서는 미등록어의 품사를 추정할 수 있기 때문이다. 인간은 두 단어가 의미적으로 유사한지를 판단할 때 그 단어들이 사용된 문맥의 유사성을 살핀다는 인지심리학의 연구보고[8]에 따르면 생성 인식 언어모델의 미등록어 처리방식은 보다 인간에 가까운 자연스러운 방식이라고 할 수 있다.

3. 실험 및 평가

생성 인식 언어모델의 특성을 파악하기 위하여 형태소단위 품사태깅시스템을 구현하여 태깅의

정확도를 측정하는 실험을 하였다. 또한 기존의 언어모델과의 특성비교를 위하여 수정된 CYK 알고리즘을 기반으로 구현된 KOMA (Korean Morphological Analyzer)의 형태소분석결과와 비교분석해 보았다. 구현된 품사태깅시스템을 학습시키고 시스템의 품사태깅정확도와 유연성을 측정하기 위하여 태깅된 말뭉치가 사용되었다. 사용된 말뭉치의 통계적 특징은 [표 2]와 같다. 본 말뭉치는 한겨레신문 정치면의 한 페이지내용을 사람이 태깅해 놓은 것이다.

문장 개수	어절 개수	형태소 개수	중의성 없는 어절	어절 평균 중의성
229개	2,525개	5,385개	669개 (26.5%)	3.3개

표 2. 학습에 사용된 말뭉치의 통계적 특성

학습능력실험은 실험에 사용된 말뭉치 전체를 사용하여 태깅시스템을 학습시키고 학습된 태깅시스템을 사용하여 말뭉치 전체를 다시 태깅하였을 때 원래 태그와의 일치도를 측정하는 실험이다. 이때 정확도는 어절단위로 측정하였으며 어절의 형태소분리와 분리된 각 형태소의 품사분류가 모두 정확하게 이루어진 경우에만 정확하게 태깅된 것으로 인정하였다.

3.1 말뭉치 학습능력 실험

[그림 5]는 구어파서가 하나의 문법을 학습할 때의 A1수준에서의 생성 그래프이다. 점으로 표시된 그래프는 구어파서의 생성을 기록한 것이며 각각의 생성에서의 파싱온도를 선으로 표시하여 겹쳐 그렸다. 구어파서가 말뭉치와 강화학습을 통하여 7개의 가능한 행위 중에 주어진 문맥에서의 적절한 행위를 찾아내는 과정을 볼 수 있다. 구어파서의 생성이 수렴하는 것과 파싱온도가 0으로 수렴하는 것이 겹치는 것을 볼 수 있으며 가상파싱온도가 구어파서의 출력에 대한 확신도와 반비례한다는 것을 볼 수 있다.

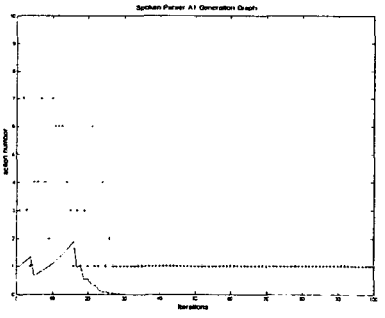

그림 5. 구어파서의 생성 그래프

[그림 6]은 구어파서가 말뭉치 전체를 반복적으로 학습할 때의 성능향상을 나타낸 학습 그래프이다. 하나의 이터레이션에서 구어파서는 말뭉치를 문장 단위로 분리하여 생성 인식 학습을 수행한다. 그래프에서 특이한 점은 60 이터레이션까지는 인식률이 거의 0%에 가깝다가 이후로 급격히 올라가서 80 이터레이션 이후로는 100%에 가까운 인식률을 나타내는 것이다. 이러한 특성은 문법습득과 어휘습득에 강화학습을 적용하였기 때문에 나타나며 60 이터레이션을 기준으로 강화학습이 탐사(exploration)에서 발휘(exploitation)로 전환되었음을 나타낸다.

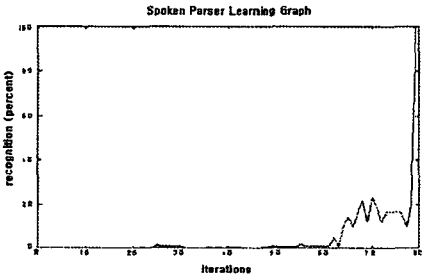

그림 6 구어파서의 학습 그래프

[표 3]은 구어파서가 학습한 이후에 임의의 문장에 대한 출력이다. '나는 학교에 간다'는 문장의 어절을 모두 붙여서 띄어쓰기정보를 제거하고 음절단위로 입력할 때 각 단계에서의 파서의 출력을 나타내었다.

. 나 (observation)
: **나이키**/nq (expectation)
. 나는
: **나가**/px+**는**/etm **관행**/ncn+**에**/jca
. 나는학
: **나가**/px+**는**/etm **한국**/nq
. 나는학교
: **나가**/px+**는**/etm **한국**/nq **파트너**/ncn+**와**/jct
. 나는학교에
: **나가**/px+**는**/etm **한국**/nq+**에**/jca **화승**/nq+**과**/jct
. 나는학교에간
: **나가**/px+**는**/etm **한국**/nq+**에**/jca **가**/pvg+**ㄴ**/etm **기업**/ncn+**도**/jxc
. 나는학교에간다
: **나가**/px+**는**/etm **한국**/nq+**에**/jca **가**/pvg+**ㄴ**/etm+**다**/ef

표 3. 구어파서의 입력과 출력

구어파서가 자신의 생성습성 A와 어휘발생습성 B를 이용하여 관측데이터와 흡사한 부분문장들을 생성해 나가는 것을 볼 수 있다. 최종출력에서 '나는'과 '학교'를 '나가는'과 '한국'으로 잘못 분석한 것은 파서가 단지 2525 어절의 한겨레신문 한 페이지만을 학습한 상태이어서 '나는'과 '학교'라는 어휘가 어휘집에 없기 때문에 어휘집에 있는 어휘 중에서 가장 흡사한 것을 생성한 것이다. 구어파서의 동작에서 중요한 특징은 다음과 같다.

. 구어파서에게 인가하는 문장에서 띄어쓰기 정보를 완전히 제거했는데도 불구하고 사람과 흡사한 반응을 나타내어서 원래의 문장과 비슷한 문

장으로 인식하여서 분할과 분류를 수행한다.
. 구어파서는 주어진 관측데이터를 바탕으로 다음에 입력될 것이라고 기대되는 데이터를 추정한다.
. 구어파서는 어휘의 양에 민감하므로 다양한 말뭉치를 이용하여 어휘의 양을 늘려주어야 한다.
동일한 입력을 KOMA에 넣었을 때의 출력은 [표 4]와 같다.

. 나 (input)
[1] 나(T 대명사)
[2] 나(MC 보통명사)
[3] 나(DA 보조동사) + 아(m 어말어미)
. 나는
[1] 나(DA 보조동사) + 는(m 어말어미)
[2] 나(MC 보통명사) + 는(j 조사)
[3] 날(DB 본동사) + 는(m 어말어미)
[4] 나(T 대명사) + 는(j 조사)
. 나는학
미등록어
. 나는학교
미등록어
. 나는학교에
미등록어
나는학교에간
미등록어
나는학교에간다
미등록어

표 4. KOMA 형태소 분석기의 입력과 출력

'나'와 '나는'에 대하여 KOMA는 여러 개의 분석결과를 출력하며 '나는학' 이후부터는 띄어쓰기정보가 없으므로 모두 미등록어로 처리되는 것을 볼 수 있다. 인간의 경우에는 띄어쓰기정보에 그다지 영향을 받지 않고 문장을 분석할 수 있으며 주어진 문장에서 최적의 의미 하나만을 생각하므로 CYK 기반의 알고리즘은 인간의 언어처리과정과는 많은 차이가 있음을 알 수 있으며 생성 인식 언어모델이 인간의 방식과 비슷하다는 것을 알 수 있다.
이상의 실험 결과를 종합해 볼 때 제안된 모델이 기존의 형태소단위 품사태깅모델보다 구어 품사태깅에 보다 더 적합한 모델이며 강화학습이 구어파서의 학습기제로서 유용하다는 것을 알 수 있다.

4. 결론 및 향후 연구

본 논문에서는 구어의 특성을 반영하는 언어모델을 제안하고 형태소단위의 품사태깅시스템에 적용해보았다. 제안된 모델은 매우 단순한 계산 구조를 가지고 있기 때문에 고속하드웨어로 구현이 용이하며, 인간의 언어 인지모델을 토대로 하였기 때문에 구어의 잡음에 강한 특성을 가지게 할 수 있다.
제안된 모델의 말뭉치 학습능력을 검증하는 실험을 수행하였으며 동작특성을 기존의 CYK 기반의 태깅시스템과 비교하였다. 실험결과 생성 인식 언어모델이 구어파서에 적합한 언어모델이며 기존의 알고리즘에 비하여 잡음에 강하여서 입력문장의 띄어쓰기정보에 민감하지 않은 특성을 볼 수 있었다.
현재 다양한 잡음내성실험을 수행하여 구어파서가 음성인식기의 출력에서 흔히 나타날 수 있는 제거 삽입 대치 반복 발음효과(coarticulation) 오류에 내성을 갖도록 연구하고 있으며 궁극적으로는 구어파서의 앞단에 HMM 기반의 음소인식기를 결합시켜서 구어파서의 예측을 이용하여 대용량 HMM의 탐색공간을 줄여줌으로서 전체적인 음성인식시스템의 성능을 높이면서도 잡음에 강한 능력을 갖도록 연구할 계획이다.

참고 문헌

[1] 최기선, 남영준, 김진규, 한영균, 박석문, 김진수, 이춘택, 김덕봉, 김재훈, 최경진, "한국어정보베이스를 위한 형태·통사 태그 표준에 관한 연구," *한국인지과학회 논문지*, Vol7,No.4, pp.43-61, 1996

[2] G. Maltese, F. Mancini, "An automatic technique to include grammatical and morphological information in a trigram-based statistical language model," *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, vol. I, pp. 157-60, San Francisco, CA, March 1992.

[3] Jain, A. N., Parsing Complete Sentences with Structured Connectionist Networks, *Neural Computation* 3, pp. 110-1, 1991.

[4] Shiber, S. M., An Introduction to Unification-Based Approaches to Grammar, *CSLI*, 1986

[5] Petitto, L., & Marentette, P. F. (1991). Babbling in the manual mode: Evidence for the ontogeny of language. *Science, 251*(5000). 1493-1499.

[6] Miller, G. A., & Gildea, P. M. (1987). How children learn words. *Scientific American*, 257(3), 94-99

[7] Brown, R., Cazden, C. B., & Bellugi, U. (1969). The child's grammar from 1 to 3. In J. P. Hill (Ed.), *Minnesota Symposium on Child Psychology* (Vol.2). Minneapolis: University of Minnesota Press.

[8] Miller, George A. and Walter G. Charles. Contextual Correlates of Semantics Similarity. *Language and Cognitive Processes*, Vol. 6, No. 1, pp. 1-28, 1991.